



# Contribution à l'identification automatique du locuteur sur des critères acoustiques et phonétiques

Odile Mella

## ► To cite this version:

Odile Mella. Contribution à l'identification automatique du locuteur sur des critères acoustiques et phonétiques . Informatique et langage [cs.CL]. Université de Nancy I, 1993. Français. NNT : 1993NAN10411 . tel-01739696

**HAL Id: tel-01739696**

**<https://inria.hal.science/tel-01739696>**

Submitted on 21 Mar 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : [ddoc-theses-contact@univ-lorraine.fr](mailto:ddoc-theses-contact@univ-lorraine.fr)

## LIENS

Code de la Propriété Intellectuelle. articles L 122. 4

Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

[http://www.cfcopies.com/V2/leg/leg\\_droi.php](http://www.cfcopies.com/V2/leg/leg_droi.php)

<http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm>



Contribution à l'identification automatique  
du locuteur sur des critères acoustiques et phonétiques

THÈSE

présentée et soutenue publiquement le **22 février 1993**

pour l'obtention du

Doctorat de l'Université de Nancy I  
(Spécialité Informatique)

par

Odile MELLA

Composition du jury :

*Président :* M. Jean-Paul HATON

*Rapporteurs :* M. René CARRÉ  
M. Pierre MARCHAND  
M. Guy PÉRENNOU

*Examineurs :* Mme Marie-Christine HATON  
M. François LONCHAMP

Au terme de ce travail, je voudrais plus particulièrement remercier ceux qui me font l'honneur de participer à mon jury :

- Monsieur René Carré, Professeur à l'Ecole Nationale Supérieure des Télécommunications, qui a accepté d'être rapporteur de ce travail malgré ses nombreuses charges ;
- Madame Marie-Christine Haton, Professeur à l'Université de Nancy I, qui m'a encadrée au cours de ce travail et dont j'ai eu la chance et le plaisir de partager le bureau pendant plusieurs années. Comme beaucoup l'ont déjà exprimé, c'est une merveilleuse directrice de thèse, toujours souriante et positive. Mais elle est pour moi beaucoup plus que cela ;
- Monsieur Jean-Paul Haton, Professeur à l'Université de Nancy I et responsable de l'équipe R.F.I.A., qui me fait l'honneur de présider ce jury. Il est à l'origine de ce travail de recherche mais également de mon métier d'enseignant-chercheur. Grâce à lui, il y a quelques années déjà, lors de discussions au restaurant universitaire, j'ai découvert le monde de la recherche en informatique et notamment celui de la reconnaissance automatique de la parole. Cette thèse est pour moi l'occasion de lui exprimer mon admiration tant pour ses qualités professionnelles que pour ses qualités humaines ;
- Monsieur François Lonchamp, Professeur à l'Université de Nancy II, enthousiasmant conteur.... de phonétique qui m'a formée dans ce domaine et qui est également à l'origine de ce travail ;
- Monsieur Pierre Marchand, Professeur de l'Université de Nancy I en détachement, qui a accepté d'être rapporteur d'un mémoire de plus de deux cents pages dans lequel figurent très peu d'égalités mathématiques. Cette thèse est pour moi l'occasion d'exprimer mes remerciements et mon affectueuse admiration à mon ancien Chef de Département Informatique ;
- Monsieur Guy Pérennou, Professeur à l'Université Paul Sabatier de Toulouse, qui a accepté d'être rapporteur de ce travail, ajoutant ainsi une tâche supplémentaire aux nombreuses charges qu'il assume.

Je voudrais également associer dans une même pensée amicale et remercier :

- tous les locuteurs et locutrices bénévoles qui m'ont fait don de leur voix et de leur temps ;
- l'ensemble des membres du Centre de Recherche en Informatique de Nancy et du Département Informatique de la Faculté des Sciences, grâce auxquels, j'ai pu trouver au fil des années une ambiance et un environnement de travail des plus favorables ;
- Martine Kuhlmann, secrétaire de l'équipe R.F.I.A., pour sa compétence et sa gentillesse ;
- Fabien Collin sans qui la typographie des tableaux situés en annexe n'aurait pas été possible ;
- Anne Boyer pour son soutien moral et logistique dans les dernières heures qui ont précédées la soutenance ;
- tous mes proches, qu'ils soient parents ou amis, qui m'ont aidée et encouragée tout au long de ce travail ;
- tous ceux que je n'ai pas pu citer et qui ont contribué directement ou indirectement et jusqu'à la dernière minute à la bonne réalisation de cette thèse.

Enfin, je tiens à exprimer une pensée particulière à mes parents à qui ce travail appartient.



# INTRODUCTION GENERALE

Notre travail se situe dans le cadre de la reconnaissance automatique du locuteur. Nous considérons que les recherches dans ce domaine peuvent se répartir en deux catégories. La première regroupe celles qui mettent en œuvre des techniques mises au point ou validées dans le cadre de la reconnaissance automatique de la parole comme la programmation dynamique, les modèles de Markov cachés ou les réseaux neuronaux. Ces méthodes utilisent de manière implicite la variabilité interlocuteur. La seconde catégorie regroupe les études qui cherchent à extraire du signal de parole des paramètres acoustiques et phonétiques qui caractérisent au mieux le locuteur. Ces études cherchent à exploiter explicitement la variabilité interlocuteur et la variabilité intralocuteur. Notre travail se situe dans cette dernière catégorie.

Nous avons établi une liste de paramètres acoustiques, phonétiques et phonologiques susceptibles de caractériser au mieux le locuteur. A partir de ces paramètres, nous avons élaboré un corpus qui a été enregistré par un certain nombre de locuteurs et de locutrices. Afin d'obtenir des paramètres fiables, nous avons étiqueté manuellement une partie de ce corpus. Puis, nous avons commencé l'étude de la pertinence des paramètres par celle des trois premiers formants de certaines voyelles orales dans un contexte phonologique bien défini. Pour cela, nous avons élaboré une méthodologie de détermination de formants robustes. Puis, nous avons étudié quelles étaient les voyelles orales les mieux adaptées à l'identification du locuteur, en déterminant pour chacune d'elles quelles étaient les combinaisons de formants et d'écarts entre formants les plus discriminantes.

Nous avons détaillé ces différentes étapes de notre travail ainsi que leurs résultats dans la troisième partie de ce mémoire. Nous avons également présenté les limites de notre méthodologie d'étude et indiqué quels prolongements il serait intéressant d'apporter à cette étude.

Auparavant, nous avons souhaité faire apparaître dans la première partie l'ensemble des connaissances nécessaires à la compréhension et à la justification de notre travail. Les premiers chapitres concernent aussi bien le domaine de la linguistique que celui de l'anatomie de la production de la parole ou celui de la paramétrisation du signal de parole. Le dernier chapitre de cette partie A est entièrement consacré à la description des manifestations et des origines de deux caractéristiques primordiales de la parole qui sont sa variabilité intralocuteur et sa variabilité interlocuteur. Cette partie est assez longue pour plusieurs raisons. Tout d'abord, nous avons cherché à être exhaustive en ce qui concerne la variabilité de la parole car elle constitue la pierre d'achoppement de bien des recherches entreprises dans le domaine de la parole, tant en reconnaissance automatique de la parole qu'en reconnaissance automatique du locuteur. Par ailleurs, nous avons voulu regrouper des connaissances qui étaient dispersées dans de nombreux ouvrages et publications. En outre, lors de la rédaction de cette partie, nous pensions avoir le temps d'étudier la pertinence des voyelles nasales. Aussi avons-nous détaillé les différents paragraphes les concernant.

Dans la deuxième partie, nous avons voulu présenter le domaine dans lequel se situe notre travail ainsi que certaines études qui ont été réalisées dans ce domaine. Pour ce dernier point, nous avons respecté la dichotomie que nous avons définie au début de cette introduction. En

ce qui concerne les études de la première catégorie, nous avons surtout décrit les travaux mettant en œuvre des techniques utilisées récemment pour la reconnaissance du locuteur comme les réseaux de Markov. Pour la seconde catégorie, dans laquelle se situe notre travail, nous avons essayé d'être plus exhaustive et d'exposer la plupart des recherches entreprises. Malheureusement, la plupart des travaux sur la caractérisation du locuteur ont été effectués à partir de corpus en anglais américain ou britannique. Or, les paramètres acoustico-phonétiques qui sont susceptibles de caractériser les locuteurs sont dépendants de la langue.

Nous avons terminé ce mémoire en résumant notre démarche et en indiquant dans quelle perspective générale se situait notre travail.

L'annexe rassemble tous les tableaux de formants "intermédiaires" et de formants "finaux" que nous avons établis lors de notre étude. Ces tableaux comprennent également d'autres résultats comme ceux de la vérification des formants sur les spectrogrammes de parole.



S.C.D. - U.H.P. NANCY 1  
BIBLIOTHÈQUE DES SCIENCES  
RUE du Jardin Botanique  
54600 VILLERS-LES-NANCY

# **PARTIE A**

## **LA PAROLE**



## TABLE DES MATIÈRES DE LA PARTIE A

INTRODUCTION	1
<b>I QUELQUES ELEMENTS DE LINGUISTIQUE</b>	<b>3</b>
I.1 Introduction	3
I.2 Phonétique et phonologie	3
I.3 Sons, phonèmes et archiphonèmes	3
I.3.1 Le phonème	4
I.3.2 Les allophones	5
I.3.3 Les systèmes phonologiques	5
I.3.4 La neutralisation et l'archiphonème	6
I.4 Notations	7
I.5 Quelques autres définitions	7
I.6 Conclusion	8
<b>II LA PRODUCTION DE LA PAROLE</b>	<b>9</b>
II.1 Introduction	9
II.2 L'appareil respiratoire inférieur et le souffle	9
II.3 Le larynx et la phonation	11
II.3.1 Les cartilages	11
II.3.2 Les muscles	12
II.3.3 Les cordes vocales	14
II.3.4 Le mécanisme de la phonation	15
II.3.5 Les caractéristiques de l'onde glottale	16
II.3.5.1 Les caractéristiques statiques	16
II.3.5.2 Les caractéristiques dynamiques	17
II.3.5.3 La "voix craquée" (creaky voice) et la "friture vocale" (vocal fry)	17
II.3.6 Les autres modes de fonctionnement du larynx	18
II.4 Les cavités supraglottiques	19
II.4.1 Description générale	19
II.4.2 Le pharynx	20
II.4.3 La cavité buccale	21
II.4.3.1 Articulateurs fixes et mobiles	21
II.4.3.2 Le palais	21
II.4.3.3 La langue	21
II.4.3.4 Les lèvres	21
II.4.4 Les cavités nasales et les sinus paranasaux	22
II.4.4.1 Introduction	22
II.4.4.2 Les cavités nasales	22
II.4.4.3 Les sinus paranasaux	27



II.5	Conclusion . . . . .	29
III	LES SONS DU FRANCAIS . . . . .	31
III.1	Introduction . . . . .	31
III.2	Les voyelles . . . . .	31
III.2.1	Les voyelles orales . . . . .	32
III.2.2	Les voyelles nasales . . . . .	34
III.2.3	Le phonème / & / . . . . .	38
III.3	Les consonnes . . . . .	39
III.3.1	Les occlusives . . . . .	40
III.3.2	Les nasales . . . . .	40
III.3.3	Les fricatives . . . . .	40
III.3.4	La consonne latérale / l / . . . . .	41
III.3.5	Les différents allophones du phonème / r / . . . . .	41
III.3.6	Les semi-consonnes . . . . .	42
III.3.7	Remarques . . . . .	42
III.4	Conclusion . . . . .	43
IV	PARAMETRISATION DU SIGNAL DE PAROLE . . . . .	45
IV.1	Introduction . . . . .	45
IV.2	Echantillonnage du signal de parole . . . . .	45
IV.3	L'analyse spectrale de la parole . . . . .	46
IV.3.1	Introduction . . . . .	46
IV.3.2	Analyse spectrale à court terme . . . . .	47
IV.3.3	Les spectrogrammes analogiques et numériques . . . . .	49
IV.3.4	Les formants . . . . .	49
IV.4	L'analyse par prédiction linéaire . . . . .	52
IV.4.1	Introduction . . . . .	52
IV.4.2	Le modèle linéaire de production de la parole . . . . .	53
IV.4.3	Le modèle d'analyse par prédiction linéaire . . . . .	54
IV.4.4	Détermination des coefficients de prédiction $a_i$ . . . . .	55
IV.4.5	La méthode d'autocorrélation . . . . .	56
IV.4.6	La méthode de covariance . . . . .	56
IV.4.7	Analyse spectrale issue de l'analyse LPC . . . . .	56
IV.4.8	Stabilité du modèle . . . . .	57
IV.4.9	Ordre de l'analyse LPC . . . . .	58
IV.4.10	La fenêtre d'analyse . . . . .	58
IV.4.11	Estimation des formants à partir de l'analyse LPC . . . . .	59
IV.4.12	Les limites de l'analyse par prédiction linéaire . . . . .	59
IV.5	Conclusion . . . . .	60



<b>V LA VARIABILITE DE LA PAROLE</b>	61
<b>V.1 Introduction</b>	61
<b>V.2 Variabilité due à la coarticulation</b>	62
V.2.1 Généralités	62
V.2.2 Les assimilations du français	64
V.2.2.1 Quelques définitions	64
V.2.2.2 Les assimilations consonantiques	65
V.2.2.3 L'assimilation vocalique	65
V.2.3 Autres influences entre consonnes	66
V.2.4 Influence des consonnes sur les voyelles	66
V.2.4.1 Nasalisation et assourdissement	66
V.2.4.2 Variation des paramètres suprasegmentaux	66
V.2.4.3 Variabilité des fréquences formantiques	67
V.2.5 Autres influences entre voyelles	70
V.2.6 Influence des voyelles sur les consonnes	70
V.2.7 Conclusion	71
<b>V.3 Variabilité d'origine linguistique</b>	71
V.3.1 La prosodie	71
V.3.1.1 Introduction	71
V.3.1.2 Variations prosodiques de la fréquence fondamentale	73
V.3.1.3 Variations prosodiques de la durée	74
V.3.1.4 Variations prosodiques de l'intensité acoustique	76
V.3.1.5 Variations prosodiques des paramètres segmentaux	76
V.3.1.6 Conclusion	76
V.3.2 Le style	77
V.3.3 Le débit d'élocution	78
<b>V.4 Variabilité liée au locuteur</b>	79
V.4.1 Introduction	79
V.4.2 Variabilité interlocuteur	80
V.4.2.1 Différences physiologiques	80
V.4.2.2 Idiomes et habitudes linguistiques	85
V.4.3 Variabilité intralocuteur	91
V.4.3.1 Introduction	91
V.4.3.2 Variabilité minimale	91
V.4.3.3 Influence de l'âge	92
V.4.3.4 Variabilité due à l'état émotionnel du locuteur	92
V.4.4 Conclusion sur la variabilité liée au locuteur	94
<b>V.5 Conclusion</b>	94
<b>BIBLIOGRAPHIE</b>	95

## Liste des figures

Figure A.1	La décomposition du langage selon Ferdinand de Saussure, d'après . . . . .	4
Figure A.2	L'appareil phonatoire. . . . .	10
Figure A.3	Schéma simplifié du larynx, d'après Lullies dans Encyclopedia Universalis. .	11
Figure A.4	Les muscles extrinsèques du larynx, d'après Lumby dans . . . . .	12
Figure A.5	Le rôle des muscles crico-aryténoïdiens, d'après . . . . .	13
Figure A.6	Coupe frontale du larynx d'après Lamby dans . . . . .	14
Figure A.7	Configuration des cordes vocales et onde glottale lors de la phonation, d'après Hirano dans . . . . .	15
Figure A.8	Ondes glottales d'une voix intense (A) et d'une voix faible (B), d'après . .	16
Figure A.9	Différentes formes d'onde glottale dans le cas de la "friture vocale", d'après .	18
Figure A.10	Position de la glotte pendant, la respiration normale (A), la respiration forte (B), la voix chuchotée (C) et la phonation (D), d'après . . . . .	19
Figure A.11	Les cavités supraglottiques et les principales divisions anatomiques utilisées dans la description articulatoire des sons, d'après Encyclopedia Universalis. .	20
Figure A.12	Os et cartilages du nez, d'après . . . . .	22
Figure A.13	Paroi latérale des fosses nasales (cornets supprimés), représentation schématique d'après . . . . .	23
Figure A.14	La cloison nasale, d'après . . . . .	24
Figure A.15	Coupe schématique frontale des fosses nasales, d'après l'Encyclopédie Pratique de Médecine et d'Hygiène. . . . .	24
Figure A.16	Paroi latérale des fosses nasales : les cornets et les structures sous-jacentes, d'après . . . . .	25
Figure A.17	Les différents types de rhinopharynx, d'après Pailoux dans . . . . .	26
Figure A.18	Les différents types de sinus frontaux, d'après . . . . .	27
Figure A.19	Les différents types de sinus sphénoïdaux, d'après . . . . .	28
Figure A.20	Les différentes tailles de sinus maxillaires, d'après . . . . .	29
Figure A.21	Le trapèze articulatoire des voyelles orales du français, les voyelles soulignées sont arrondies. Chacun des numéros permet de repérer l'articulation de la voyelle correspondante sur la figure A.22. . . . .	32
Figure A.22	Articulation de quelques voyelles orales, d'après . . . . .	33
Figure A.23	Constriction pharyngale lors de l'articulation d'un [ i ] et d'un [ O ], d'après .	33
Figure A.24	Positions articulatoires des voyelles [ E ], [ * ] et [ E ] nasalisé et des voyelles [ O ], [ ) ] et [ O ] nasalisé d'après Zerling dans . . . . .	35
Figure A.25	Positions articulatoires des voyelles [ A ], [ @ ] et [ A ] nasalisé et des voyelles [ O ], [ % ] et [ O ] nasalisé d'après J.P. Zerling dans . . . . .	36
Figure A.26	Comparaisons des positions articulatoires des voyelles [ * ] — [ ) ] et [ @ ] — [ % ], d'après J.P. Zerling dans . . . . .	37
Figure A.27	Les positions articulatoires de [ 9 ] et [ 3 ] par rapport aux voyelles françaises, d'après . . . . .	38
Figure A.28	Articulations schématiques d'un [ s ] et d'un [ S ], d'après . . . . .	41
Figure A.29	Articulations schématiques de [ r ] et [ R ], d'après . . . . .	42
Figure A.30	Décomposition spectrale de la production des sons voisés d'après . . . . .	46
Figure A.31	Analyse par FFT de 512 échantillons d'un [ a ] ; (a) : signal temporel ; (b) : signal multiplié par une fenêtre de Hamming (1), Hanning (2), trapézoïdale (3) ; (c) : le spectre d'amplitude pour chaque fenêtre ; d'après . . . . .	48



Figure A.32	Les différentes étapes de calcul d'un spectre instantané, d'après . . . . .	48
Figure A.33	Spectrogrammes numériques : (a) signal temporel, (b) spectrogramme à bande étroite, (c) spectrogramme à large bande, d'après . . . . .	50
Figure A.34	Voyelles orales du français dans le plan ( $F_1$ , $F_2$ ), d'après . . . . .	51
Figure A.35	Spectre d'un [ a ] synthétique présentant un formant glottal dont la fréquence est 125 Hz, d'après . . . . .	52
Figure A.36	Modèle linéaire de production de la parole. . . . .	53
Figure A.37	Un exemple de spectres LPC obtenus à partir de 160 échantillons de la voyelle anglaise [ A ] échantillonnée à 8 kHz ; (a) le signal temporel, (b) le spectre FFT, (c) les spectres LPC pour différentes valeurs de l'ordre de prédiction ; d'après . . . . .	57
Figure A.38	Articulation réduite de [ u ] dans un contexte dental (/ dut /), d'après . . . . .	62
Figure A.39	Allophones dorso-vélaire ([ ku ]) et dorso-palatal ([ ky ]) du phonème / k /, d'après . . . . .	63
Figure A.40	Prépositionnement de la langue lors de l'articulation d'un / b / prévocorique, d'après . . . . .	64
Figure A.41	Variabilité des deux premiers formants du [ a ] dans l'expression "à Madagascar", d'après . . . . .	68
Figure A.42	Influence du [ i ] sur les deux premiers formants du [ a ] dans le mot "carabine", d'après . . . . .	68
Figure A.43	Transitions formantiques de [ \$ ] dans un contexte vélaire et uvulaire, d'après . . . . .	69
Figure A.44	L'intonation modale en français, d'après Encyclopedia Universalis. . . . .	73
Figure A.45	Exemples de contours mélodiques en français : "le riche fermier breton" et "le petit homme que Denis connaissait" ont la même structure prosodique composée de deux mots prosodiques, d'après . . . . .	74
Figure A.46	Variations de durée (en pourcentage) d'une syllabe à l'autre dans le début de la phrase : "Nous pouvons vous proposer du café noir, du café au lait ...", d'après . . . . .	75
Figure A.47	Le mot "legat" prononcé par trois locuteurs suédois avec trois accentuations différentes du / g /, d'après . . . . .	77
Figure A.48	Les effets de l'effort vocal sur le spectre des voyelles. (A) : Spectrogramme du mot anglais "bib", (1) dans le cas d'une voix intense, (2) dans le cas d'une voix faible ; (B) : spectre instantané du [ i ] pour les deux voix ; d'après . . . . .	81
Figure A.49	Exemples de pente spectrale moyenne estimée par J. Mártony pour des locuteurs suédois, après une préaccentuation de + 12 dB par octave, dans . . . . .	82
Figure A.50	Zones de dispersion des voyelles orales du français dans le plan $F_1$ - $F_2$ (échelle Bark), en traits pleins pour les locuteurs, en traits pointillés pour les locutrices. Les hachures délimitent les zones de recouvrement entre voyelles pour un même sexe, d'après . . . . .	83
Figure A.51	Les facteurs d'échelle entre les fréquences formantiques féminines et masculines des voyelles orales ; en traits pointillés pour l'étude de F. Lonchamp, avec des losanges pour les données de G. Fant et en traits pleins pour la simulation de H. Traunmüller ; d'après . . . . .	84
Figure A.52	Comparaison des spectres LPC de deux [ n ], l'un est prononcé dans un état normal, l'autre avec un léger rhume, d'après . . . . .	85

Figure A.53	Temps de pause accumulé par rapport au temps de parole lue pour cinq locuteurs, d'après . . . . .	86
Figure A.54	Contours mélodiques sur une phrase pour un locuteur donné suivant quatre états émotifs, l'état normal, la colère, la peur et le chagrin ; d'après . . . . .	93

Liste des tables

Table A.1 Les phonèmes du français avec leur représentation dans l'Alphabet  
Phonétique International et dans celui utilisé par le logiciel SNORRI. 6

Table A.2 Classification articulatoire des consonnes. . . . . 39

## INTRODUCTION

Cette première partie constitue à la fois les fondements et la justification des deux autres parties de ce mémoire. Nous avons souhaité y introduire toutes les connaissances nécessaires à la compréhension de la problématique de la reconnaissance automatique du locuteur, en général, et à celle de la caractérisation du locuteur à partir de paramètres acoustiques et phonétiques, en particulier. Nous avons donc voulu décrire les principaux niveaux du processus de communication que constitue la parole, tout en mettant en évidence le caractère variable de celle-ci, aussi bien en ce qui concerne les différences entre les locuteurs qu'en ce qui concerne les différences entre les messages d'un même locuteur.

Le premier chapitre, très court, introduit le concept de parole du point de vue des sciences du langage et définit quelques notions de linguistique qui seront utilisées par la suite.

Le deuxième chapitre a pour objet la description détaillée du mécanisme de production de la parole. Dans celle-ci, il nous a semblé important de mettre l'accent à la fois sur la complexité et la souplesse des organes mis en œuvre et sur les différences anatomiques des locuteurs. Simultanément, nous avons introduit les premiers éléments d'acoustique du signal de parole. Lorsque nous avons rédigé cette partie, nous pensions avoir le temps, dans le cadre de ce travail, d'étudier la pertinence des voyelles nasales dans la caractérisation du locuteur. Ceci explique la précision de la description anatomique des cavités nasales et paranasales.

Le troisième chapitre présente l'application du processus général de production de la parole à la prononciation des phonèmes du français. Il décrit le mode et le lieu d'articulation des réalisations physiques normalisées des phonèmes considérés isolément. Toutefois, dans la description de l'articulation des voyelles nasales, est abordée une première approche de la liberté que possède le locuteur pour articuler un phonème.

Le traitement automatique de la parole nécessite une analyse acoustique et une paramétrisation du signal de parole. Ces thèmes font l'objet du quatrième chapitre. Le domaine du traitement du signal de parole est à la fois vaste et pointu. Aussi nous sommes-nous limitée à la présentation des notions et des outils indispensables à la compréhension de notre travail, comme les formants ou l'analyse par prédiction linéaire.

Le dernier chapitre est entièrement consacré à la variabilité de la parole sous toutes ses formes. Du point de vue général, le contenu de ce chapitre montre comment les paramètres acoustiques liés à la production d'un phonème isolé sont modifiés par son environnement phonémique, par les niveaux supérieurs de l'acte de communication, par la spécificité du locuteur et par le contexte dans lequel celui-ci s'exprime. Du point de vue de la reconnaissance du locuteur, ce chapitre envisage les sources de différences entre les locuteurs qui peuvent se répercuter au niveau du signal de parole (variabilité interlocuteur). Mais il établit aussi que leurs conséquences acoustiques sont sensibles à toutes les autres formes de variabilité. Par ailleurs, les paragraphes consacrés aux phénomènes de coarticulation permettent également d'expliquer les difficultés rencontrées dans l'étiquetage des corpus de parole et qui seront développées dans la troisième partie de ce mémoire.





## CHAPITRE I QUELQUES ELEMENTS DE LINGUISTIQUE

### 1. Introduction

Le but de notre étude est une approche analytique de la reconnaissance du locuteur se concrétisant par le choix, l'extraction et l'analyse de paramètres phonétiques et phonologiques. Il nous semble donc nécessaire de commencer la rédaction de ce mémoire en essayant de définir de façon simple — assurément trop pour les spécialistes du domaine — ces deux qualificatifs, ainsi que quelques autres objets linguistiques comme le phonème, l'allophone ou l'archiphonème.

### 2. Phonétique et phonologie

La parole est une manifestation du langage humain qui est, lui-même, un mécanisme de communication d'information entre les êtres humains. Pour cette raison, l'étude de la parole, que ce soit au niveau de sa production, de sa transmission ou de sa réception, est l'objet des différentes branches de la linguistique que A. Martinet [Martinet 80] définit comme l'étude scientifique du langage humain.

La phonétique et la phonologie sont deux branches importantes de la linguistique que nous essaierons de qualifier en nous fondant sur la pensée saussurienne dont une représentation très schématisée est donnée par la figure A.1. Selon Ferdinand de Saussure, tout message linguistique est décomposable en un contenu (ou signifié) et en une expression (ou signifiant), chacune de ces deux entités étant elle-même décomposable en une forme (ou structure) et une substance (ou réalisation) [Malmberg 74]. Ainsi, si nous considérons l'énoncé "*Nancy est une belle ville*", celui-ci a un contenu, ce dont il fait part, et une expression qui peut être soit une suite de caractères dans le cas d'un énoncé écrit, soit une suite de sons dans le cas d'un énoncé oral.

La phonétique et la phonologie sont les deux sciences du langage qui s'intéressent à l'expression orale d'un message linguistique. Bien que la répartition exacte de leurs domaines d'étude respectifs ait évolué au cours des siècles et varie encore de nos jours selon les diverses écoles de pensée linguistique, nous pouvons dire en simplifiant que la phonétique s'occupe de la substance d'un énoncé oral alors que la phonologie s'efforce d'établir et d'étudier les règles structurant la forme de cet énoncé.

### 3. Sons, phonèmes et archiphonèmes

L'acte de parole est avant tout un acte de communication qui, pour réaliser au mieux cette fonction, doit être régi par un code connu du locuteur et de l'auditeur. Ce code doit posséder, entre autres facteurs, un nombre fini d'éléments unitaires distinctifs permettant à un être humain



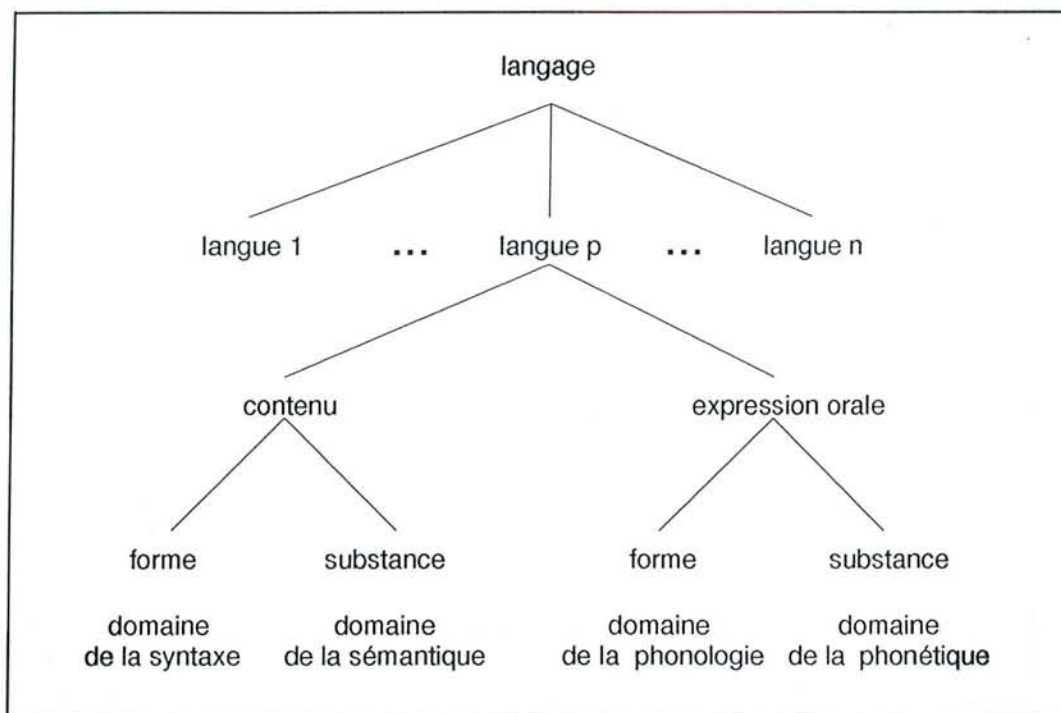


Figure A.1. La décomposition du langage selon Ferdinand de Saussure, d'après [Malmberg 74].

d'engendrer ou de décoder n'importe quel message linguistique. En général, ces codes diffèrent d'une langue à une autre mais plusieurs d'entre eux peuvent comporter des éléments communs. Dans le cas du langage parlé, certains de ces éléments s'appellent des phonèmes. L'une des tâches du phonologue est d'établir l'inventaire des phonèmes d'une langue tout en associant à chacun d'eux un ensemble de traits distinctifs qui le caractérise.

### 3.1. Le phonème

Le phonème est une unité phonologique abstraite et distinctive d'une langue, c'est-à-dire un élément minimal de la langue parlée qui permet de distinguer un mot d'un autre lorsque tous les autres éléments de ces deux mots sont identiques. On dit alors que ces deux mots constituent une paire minimale.

Prenons comme exemple la paire minimale formée des mots "*peau*" et "*beau*" dont les transcriptions phonétiques sont [ po ] et [ bo ]. Les prononciations de ces deux mots ne diffèrent que par leur premier élément. Les sons [ p ] et [ b ] sont dits en opposition pour cette paire minimale (notation : [ p ] ~ [ b ]) et le trait distinctif qui les oppose est le voisement. En effet, [ p ] est une occlusive bilabiale sourde alors que [ b ] est une occlusive bilabiale sonore<sup>1</sup>, les cordes vocales vibrant pendant sa réalisation. Lorsque le phonologue a déterminé un nombre significatif de paires minimales pour lesquelles ces deux sons s'opposent, il leur attribue la qualité de phonèmes (notation : / p / et / b /), puis cherche à leur opposer d'autres phonèmes dans le même contexte.

<sup>1</sup> La description articulatoire des sons du français est détaillée dans le chapitre III. Le lecteur peut se reporter à la figure A.11 page 20 pour situer ces articulations.

Par analogie mathématique, on peut également définir le phonème comme une classe d'équivalence de sons dont la relation d'équivalence serait à quelque chose près : " les sons  $s_1$  et  $s_2$  sont en relation pour la langue L si et seulement s'il n'existe aucune paire de mots de L discriminés par le couple  $(s_1, s_2)$  et si les deux sons possèdent le même ensemble de traits distinctifs".

### 3.2. Les allophones

Lors de l'établissement de l'inventaire des phonèmes d'une langue, le phonologue est confronté au problème d'identifier comme un même phonème des sons différents. Car, selon son contexte, un phonème peut se réaliser en deux segments phonétiques différents et, réciproquement, il est possible qu'un même son puisse être la réalisation de deux phonèmes distincts.

Considérons par exemple les trois mots "cou", "cas" et "qui" dont les transcriptions phonologiques sont / **ku** /, / **ka** / et / **ki** /. Ils commencent tous par la même consonne / **k** /. Pourtant, une analyse acoustique mettrait en évidence trois sons différents, [  $k_1$  ], [  $k_2$  ] et [  $k_3$  ]<sup>1</sup>, le son [  $k_1$  ] correspondant à une réalisation dorso-vélaire de / **k** /, le son [  $k_3$  ] à une réalisation dorso-palatale et [  $k_2$  ] à une articulation moyenne située entre les deux précédentes<sup>1</sup>. Les sons [  $k_1$  ], [  $k_2$  ] et [  $k_3$  ] sont les variantes allophoniques (ou allophones) d'une entité fonctionnelle distinctive unique : le phonème / **k** /. En d'autres termes, les sons sont les réalisations phonétiques concrètes des unités phonologiques abstraites que sont les phonèmes.

Cet exemple illustre le cas des variantes combinatoires d'un phonème. Il existe une autre catégorie de variantes, les variantes individuelles ou idiolectales dont un exemple typique est donné par les différentes variantes françaises du phonème / **r** / qui sont décrites au paragraphe III.3.5.

### 3.3. Les systèmes phonologiques

A chaque langue correspond un système maximal de phonèmes qui lui est propre. Si, par exemple, / **l** / et / **r** / sont, pour tous les locuteurs français, deux phonèmes qui s'opposent, notamment dans "rire" et "lire", ils ne représentent pour les locuteurs japonais que deux variantes du même phonème [Duchet 86]. Le système phonologique d'une langue est, en quelque sorte, le sur-ensemble des systèmes phonologiques des locuteurs parlant cette langue. En effet, chaque locuteur possède un système phonologique spécifique dépendant à la fois de sa localisation géographique, de son appartenance à un groupe socioculturel et de ses habitudes linguistiques.

La langue française possède potentiellement trente-huit phonèmes mais l'un d'eux, / **ŋ** /, n'apparaît que dans les cas d'assimilation consonantique de nasalité<sup>2</sup> (par exemple dans l'expression "Banque de France"), et un autre, / **ɛː** /, dernière réminiscence des oppositions vocaliques de longueur du XVII<sup>e</sup> siècle ("faites" ~ "faîte"), a pratiquement disparu [Walter 76].

La table A.1 présente les phonèmes du français transcrits simultanément dans l'Alphabet Phonétique International (API) et dans l'alphabet phonétique utilisé par le logiciel SNORRI [Laprie 88] qui a servi à visualiser et à étiqueter le corpus de données que nous avons élaboré pour notre étude.

<sup>1</sup>  $k_1$ ,  $k_2$  et  $k_3$  sont des symboles arbitrairement choisis afin de ne pas utiliser le symbolisme très technique des diacritiques de l'Alphabet Phonétique International (API).

<sup>2</sup> Les différents cas d'assimilation consonantique de nasalité sont traités dans le paragraphe V.2.2.2.



VOYELLES			CONSONNES		
Snorri	API		Snorri	API	
/ i /	/ i̥ /	pie	/ p /	/ p /	pipe
/ e /	/ e /	pétale	/ t /	/ t /	titre
/ ai /	/ ɛ /	père	/ k /	/ k /	kilo
/ a /	/ a /	patte	/ b /	/ b /	bible
	/ ɑ /	pâte	/ d /	/ d /	dix
/ ) /	/ ɔ /	port	/ g /	/ g /	gui
/ o /	/ o /	peau	/ f /	/ f /	film
/ u /	/ u /	poule	/ s /	/ s /	six
/ y /	/ y /	pur	/ ch /	/ ʃ /	chichi
/ eu /	/ ø /	peu	/ v /	/ v /	vivre
/ œ /	/ œ /	peur	/ z /	/ z /	zizanie
/ & /	/ ə /	petit	/ gh /	/ ʒ /	girafe
/ in /	/ ẽ /	pain	/ l /	/ l /	lilas
/ an /	/ ɑ̃ /	pente	/ R /	/ r /	rire
/ on /	/ ɔ̃ /	pont	/ m /	/ m /	mime
/ un /	/ œ̃ /	un	/ n /	/ n /	nid
			/ nj /	/ ɲ /	dignité
SEMI-CONSONNES				/ ŋ /	camping
Snorri	API				
/ j /	/ j̥ /	pied			
/ w /	/ w /	pois			
/ ui /	/ ɥ /	puir			

Table A.1. Les phonèmes du français avec leur représentation dans l'Alphabet Phonétique International et dans celui utilisé par le logiciel SNORRI.

3.4. La neutralisation et l’archiphonème

Nous avons considéré jusqu’à maintenant le cas idéal où l’opposition entre deux phonèmes a pu être établie dans toutes les distributions phonologiques possibles. Mais, pour quelques paires de phonèmes, il peut exister des distributions pour lesquelles l’opposition n’est pas valable. On dit alors que l’opposition est neutralisée pour ces distributions et certains phonologues regroupent le couple de phonèmes sous le terme d’archiphonème.

La neutralisation d’une opposition dans une distribution se manifeste soit par la disparition de l’un des phonèmes au profit de l’autre, soit par le fait que l’utilisation par le locuteur de l’un ou l’autre des phonèmes ne change pas la signification du mot prononcé.

Ainsi, en français, l'opposition de timbre [ e ] ~ [ ε ] se réalise en syllabe finale ouverte<sup>1</sup> ([ **maRe** ] ~ [ **maRε** ]) et engendre les phonèmes / e / et / ε /. Mais elle se neutralise en syllabe fermée au profit de [ ε ] ([ **pεRdy** ]), ainsi qu'en syllabe ouverte non finale où le locuteur peut prononcer indifféremment [ e ] ou [ ε ] ([ **mezõ** ] ou [ **mεzõ** ]). Dans ce dernier cas, il est plus simple d'utiliser dans la transcription phonologique de " *maison* " l'archiphonème / E / (/ mEzõ /).

## 4. Notations

Dans les exemples utilisés dans les paragraphes précédents, nous avons employé trois notations différentes selon qu'il était question de transcription orthographique, phonologique ou phonétique. Dans la mesure du possible, nous essaierons de conserver ce système de notation dans la suite de l'ouvrage :

- "Nîmes"            <—        transcription orthographique,
- / **nîm** /            <—        transcription phonologique,
- [ **nĩm** ]            <—        transcription phonétique<sup>2</sup>.

## 5. Quelques autres définitions

Il faut noter enfin que chacune de ces grandes disciplines linguistiques que sont la phonétique et la phonologie, se divise en plusieurs branches selon les différents aspects sous lesquels elle étudie la parole :

- en phonétique :
  - la phonétique articulatoire ou physiologique décrit le fonctionnement de l'appareil phonatoire ;
  - la phonétique acoustique analyse les sons du langage d'un point de vue physique en tant que vibrations des molécules d'air ;
  - la phonétique perceptive s'intéresse à la transformation par l'oreille de ces vibrations en influx nerveux et à leur interprétation par le système nerveux central ;
  - la phonétique combinatoire, quant à elle, étudie tous les phénomènes de coarticulation entre les sons ;

<sup>1</sup> Une syllabe ouverte est une syllabe qui ne se termine pas par une ou plusieurs consonnes : dans / **sistematik** / les deuxième et troisième syllabes sont des syllabes ouvertes, les autres des syllabes fermées.

<sup>2</sup> Le symbole **ĩ** indique une nasalisation du [ **i** ].

- en phonologie :
  - la phonologie synchronique a pour objet l'établissement du système phonologique d'une langue, d'un dialecte ou d'un groupe d'individus à un instant donné ;
  - la phonologie diachronique ou évolutive étudie l'évolution chronologique des systèmes phonologiques en essayant d'établir des théories structurales des changements ;
  - la phonologie générative se situe au carrefour des domaines précédents. Son but est d'exprimer, sous forme de règles, le passage d'une transcription orthographique d'un énoncé écrit aux transcriptions phonétiques de tous les énoncés oraux potentiellement réalisables, tout en transitant par la transcription phonologique la plus générale possible. Prenons un exemple simple extrait de [Duchet 86] : pour obtenir les transcriptions phonétiques des énoncés écrits "*petit ami*" et "*petit garçon*", les phonologues utilisent la transcription phonologique et la règle suivantes :
    - "*petit*" → / pətɪt /
    - "une consonne ne se prononce pas lorsqu'elle est suivie d'une pause ou d'une autre consonne"

d'où :

- "*petit ami*" → / pətɪt / + / ami / → [ pətɪtami ]
- "*petit garçon*" → / pətɪt / + / ɡarsɔ̃ / → [ pətɪtɑ̃ɡarsɔ̃ ]

Parmi ces règles de réécriture se trouvent toutes les règles de coarticulation donnant les allophones possibles d'un phonème selon son contexte. Certaines d'entre elles permettent d'expliquer et de coder l'évolution des systèmes phonologiques.

## 6. Conclusion

Ces quelques pages nous ont permis de définir un certain nombre de termes linguistiques qui seront souvent employés dans la suite de ce mémoire. Ces définitions ne sont peut-être pas assez précises du point de vue du spécialiste en linguistique mais elles sont à notre avis suffisantes pour permettre à un informaticien d'appréhender le domaine de la parole. La phonétique articulatoire est l'objet des deux prochains chapitres puisque nous y décrivons successivement le processus de production de la parole et l'articulation des phonèmes du français. Quelques éléments de phonétique acoustique sont introduits dans le chapitre IV. Enfin, de nombreuses connaissances de phonétique combinatoire et quelques autres de phonologie synchronique et évolutive sont développées dans le dernier chapitre de cette partie qui est consacré à la variabilité de la parole. En ce qui concerne les autres disciplines, nous conseillons au lecteur de se reporter aux ouvrages cités en bibliographie ; notamment [Walter 76] pour la phonologie évolutive du français, [Duchet 86] et [Martinet 80] pour la phonologie générale et [Calliope 89] et [Zwicker 81] pour la perception.



## CHAPITRE II LA PRODUCTION DE LA PAROLE

### 1. Introduction

Les sons de la parole, qui, comme nous l'avons souligné dans le chapitre précédent, constituent les réalisations physiques des phonèmes, sont des ondes de pression. Plus précisément, ce sont des variations de la pression d'air produites par l'appareil phonatoire d'un locuteur qui se propagent, par l'intermédiaire des molécules d'air, jusqu'au tympan d'un auditeur. Hormis le larynx, l'homme ne se sert d'aucun organe spécifique pour produire ces sons mais utilise d'une façon particulière des organes originellement réservés à des fonctions biologiques primaires comme la respiration ou la déglutition.

Le processus de production d'un son par l'appareil phonatoire peut se décomposer en trois phases : la création d'un écoulement d'air en provenance des poumons, la transformation de ce courant d'air en une énergie sonore par la vibration des cordes vocales (sons voisés) et/ou par la création de turbulences dues à un rétrécissement ou à une obstruction du conduit vocal, et enfin le filtrage de cette énergie sonore par les cavités supraglottiques que sont le pharynx, le conduit buccal et les cavités nasales.

Comme nous pouvons l'observer sur la figure A.2, cette décomposition du processus de production de la parole en trois phases correspond approximativement au découpage de l'appareil phonatoire en trois zones : l'appareil respiratoire inférieur, le larynx et les cavités supraglottiques situées au-dessus du larynx. Les trois étapes de ce processus sont donc développées dans les paragraphes suivants par l'intermédiaire de la description de ces trois zones.

Cette description est volontairement détaillée pour deux raisons. La première est de souligner à la fois la complexité, la souplesse d'utilisation et la diversité des appareils phonatoires humains, et de permettre ainsi une première approche de la variabilité de la parole, que ce soit pour un même locuteur ou pour des locuteurs différents. La deuxième est de mettre en évidence les limites actuelles des connaissances concernant certains organes, tant au point de vue anatomique qu'au point de vue de leur rôle exact dans la production de la parole.

### 2. L'appareil respiratoire inférieur et le souffle

L'appareil respiratoire inférieur est constitué d'un réservoir d'air, les poumons, dont le volume dépend de la morphologie du locuteur mais aussi de ses capacités respiratoires acquises, et, d'un ensemble de muscles chargés de remplir (inspiration) et de vider (expiration) ce réservoir. L'inspiration est réalisée grâce à la contraction du diaphragme et des muscles intercostaux externes. L'expiration est en général une phase passive : le relâchement des muscles inspireurs, sauf lorsqu'il est nécessaire d'avoir une expiration contrôlée, ce qui est le

cas dans la production de la parole. Elle met alors également en œuvre les muscles intercostaux internes ainsi que la plupart des abdominaux [Marchal 80].

Bien qu'il soit possible de réaliser des sons lors de l'inspiration (parler sangloté des enfants, par exemple), la production de la parole s'effectue au cours d'une expiration prolongée où tous les muscles interagissent, afin d'avoir un débit d'air à peu près constant de 0,2 litre par seconde [O'Shaughnessy 87]. Selon la longueur du groupe de phonation, qui est la partie d'un énoncé située entre deux pauses respiratoires, on assiste tout d'abord à un contrôle des muscles inspireurs afin d'éviter l'affaissement trop rapide de la cage thoracique puis à une contraction des muscles expirateurs pour allonger l'expiration. Par ailleurs, la phase d'inspiration est réduite par rapport à celle de la respiration normale : 5 à 10% au lieu de 25%.

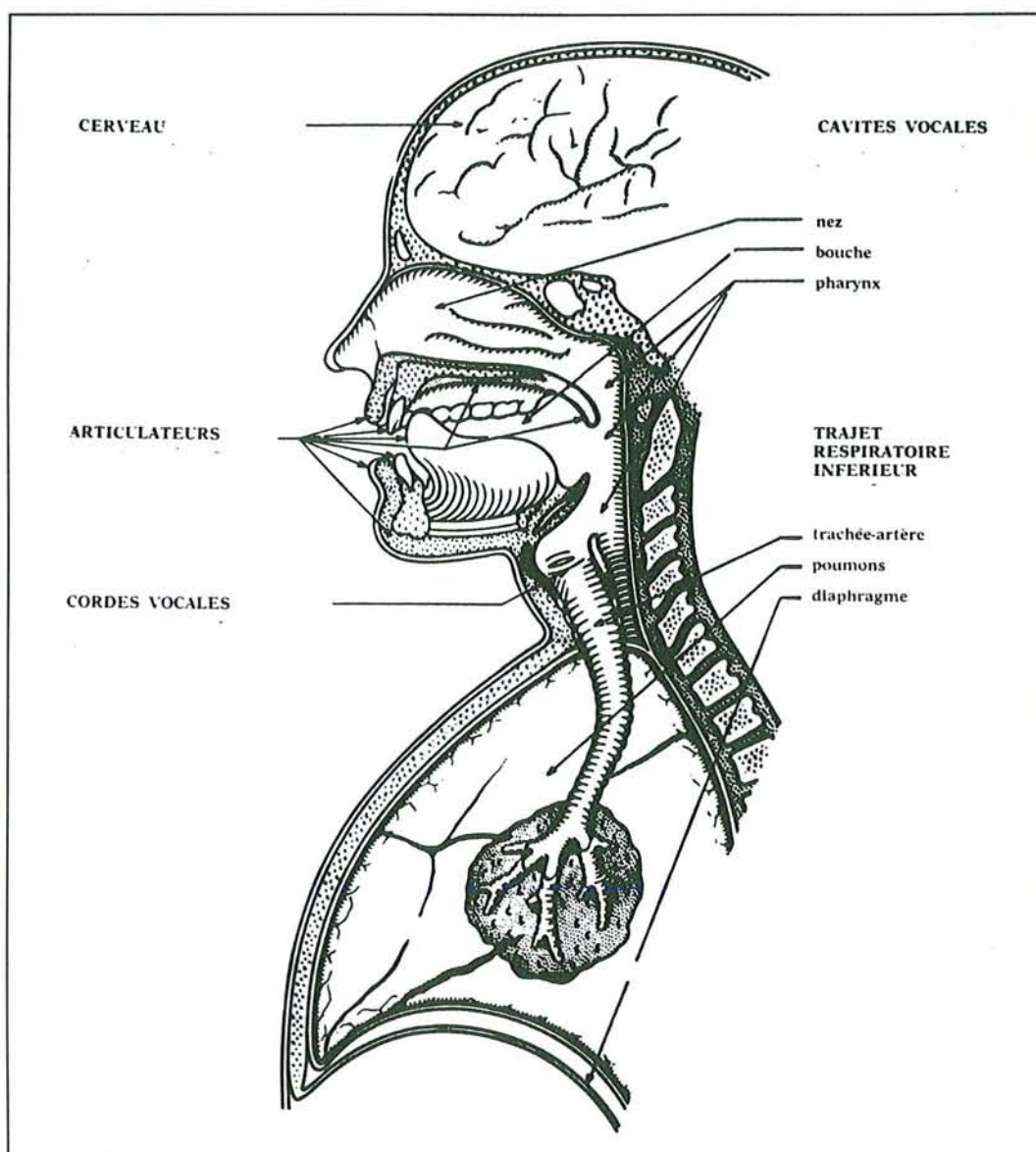


Figure A.2. L'appareil phonatoire.

### 3. Le larynx et la phonation

L'air chassé des poumons arrive au niveau du larynx par l'intermédiaire de la trachée. Le larynx, grâce notamment à deux de ses éléments appelés cordes vocales, joue un rôle primordial dans le mécanisme de la production de la parole. Il permet de moduler le flot d'air en provenance des poumons en une onde de débit d'air possédant un certain nombre de propriétés.

Dans un premier temps, nous allons décrire les principaux constituants du larynx, ensuite nous détaillerons son mode de fonctionnement puis nous terminerons par une description des caractéristiques de l'onde glottale.

Comme le montre la figure A.3, le larynx est composé de cinq cartilages principaux, qui tendent à s'ossifier chez l'adulte et le vieillard, reliés par un ensemble complexe de ligaments et de muscles [Bouchet 83].

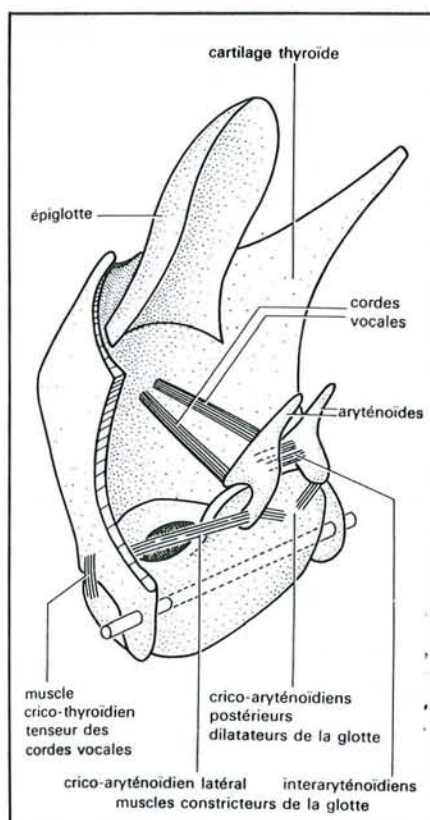


Figure A.3. Schéma simplifié du larynx, d'après Lullies dans Encyclopedia Universalis.

#### 3.1. Les cartilages

- **le cartilage épiglottique** ou épiglotte est une lame fibreuse qui permet d'obstruer le larynx lors de la déglutition,
- **le cartilage thyroïde** est un dièdre dont l'arête, beaucoup plus saillante chez les hommes, est plus connue sous le nom de pomme d'Adam. Il s'ossifie à l'âge adulte et fait office de bouclier du larynx,



- **le cartilage cricoïde** en forme de chevalière, situé à la base du larynx, constitue la fondation de celui-ci,
- **les cartilages aryténoïdes** sont deux petites pyramides placées sur le chaton du cricoïde. Mobiles à la fois en translation et en rotation, ils permettent l'étirement ainsi que l'écartement ou le rapprochement des cordes vocales auxquelles ils sont reliés au niveau des apophyses vocales.

### 3.2. Les muscles

Ceux-ci peuvent se décomposer en deux groupes :

- **les muscles extrinsèques** dont les principaux sont représentés sur la figure A.4 et qui relient le larynx au reste du squelette. Ces muscles servent essentiellement à abaisser ou à élever le larynx mais ils ont une double influence sur la production de la parole, premièrement en modifiant la forme et la taille des cavités résonantes situées au-dessus du larynx, deuxièmement en exerçant une action indirecte sur la fréquence du ton laryngien. Ainsi, la contraction du muscle sterno-hyoïdien entraîne l'abaissement du larynx mais provoque également un raccourcissement, un épaississement et un relâchement des cordes vocales conduisant à une fréquence fondamentale plus basse [Atkinson 78]. Ceci indiquerait qu'un imitateur ne peut pas modifier de manière indépendante son timbre, qui dépend de la taille et de la forme de son conduit vocal, et sa fréquence fondamentale. D'après D. O'Shaughnessy [O'Shaughnessy 87], le même phénomène expliquerait pourquoi, lors de la production des voyelles fermées comme / i / et / u /, la fréquence de vibration des cordes vocales est plus élevée que pendant celle des voyelles ouvertes comme / a / ;

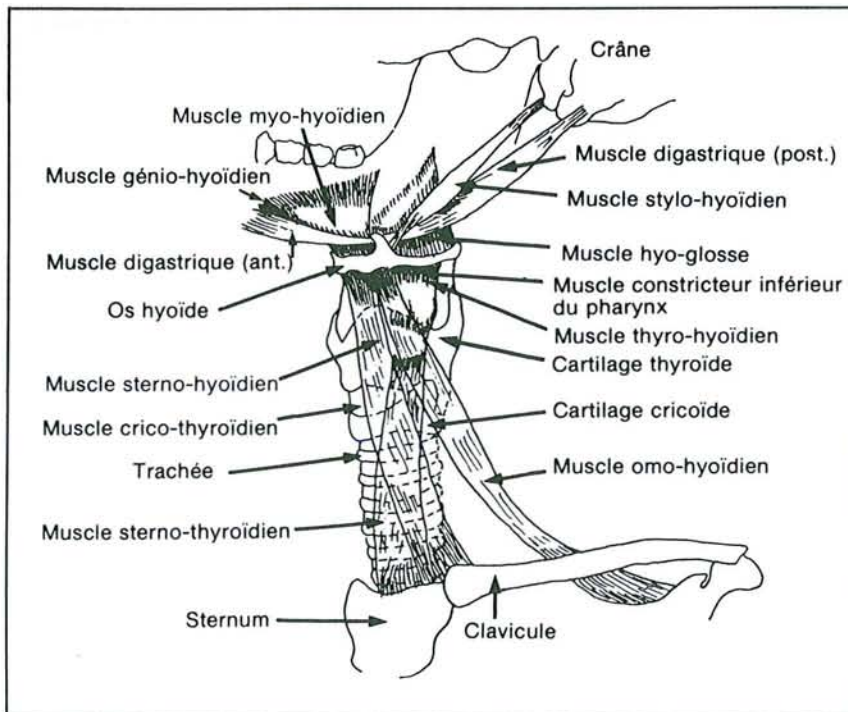


Figure A.4. Les muscles extrinsèques du larynx, d'après Lumby dans [Marchal 80].

- les **muscles intrinsèques** qui interconnectent les différents cartilages du larynx :
  - les *muscles thyro-aryténoïdiens* ou muscles vocaux. Ces deux muscles horizontaux sont attachés ensemble au cartilage thyroïde et séparément à chaque cartilage aryténoïde. Ils constituent la structure interne des cordes vocales et obstruent complètement le larynx sauf en un espace central, circonscrit par les cordes vocales et les faces internes des aryténoïdes, appelé la glotte. Les muscles thyro-aryténoïdiens sont responsables du volume, de la consistance et de la tension des cordes vocales,
  - les *muscles crico-thyroïdiens*. Leur contraction entraîne soit un basculement vers l'avant du cartilage thyroïde soit un recul du cartilage cricoïde (cf. figure A.3). Dans les deux cas, elle provoque un allongement des cordes vocales d'où leur appellation de muscles tenseurs des cordes vocales,
  - les *muscles crico-aryténoïdiens*. Ils gèrent le degré d'ouverture de la glotte. Les muscles postérieurs sont les seuls muscles dilatateurs de la glotte [Bouchet 83]. Grâce à leur contraction, les aryténoïdes glissent sur l'arête du cricoïde tout en pivotant vers l'extérieur, ce qui a pour effet d'ouvrir la glotte. Quant aux muscles latéraux, ils provoquent le pivotement des aryténoïdes sur eux-mêmes, ce qui a pour effet de fermer la glotte. La figure A.5 présente les mouvements des aryténoïdes lors de la contraction des différents muscles crico-aryténoïdiens,
  - le *muscle inter-aryténoïdien*, tendu entre les deux aryténoïdes. Sa contraction entraîne le rapprochement de ceux-ci et la fermeture de la glotte.

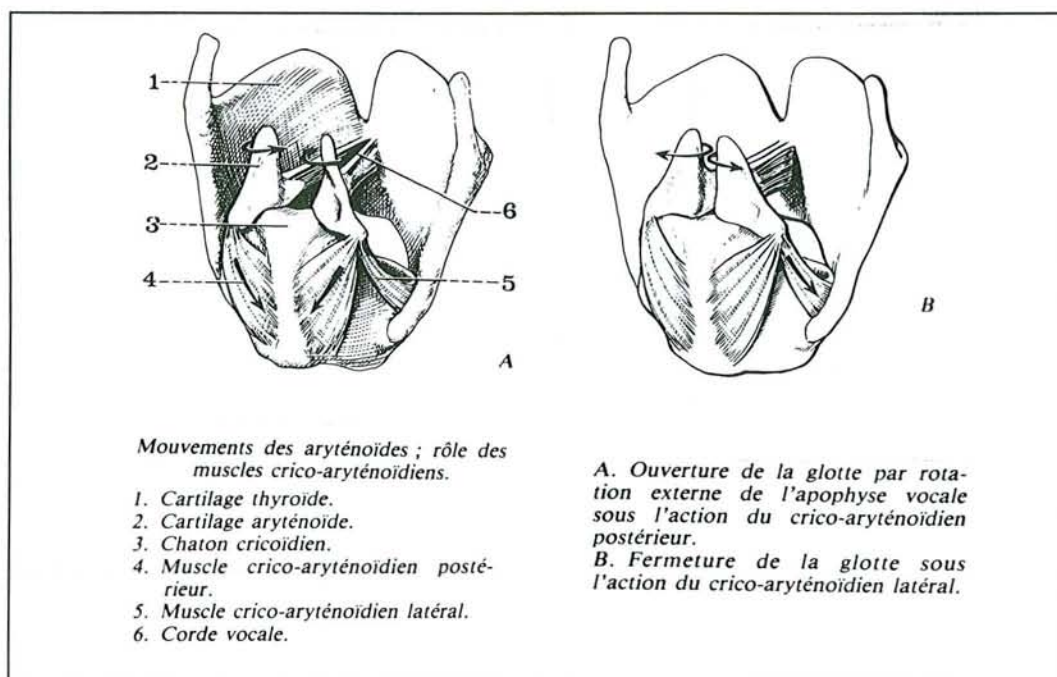


Figure A.5. Le rôle des muscles crico-aryténoïdiens, d'après [Bouchet 83].

Les rôles assignés aux différents muscles du larynx dans la description précédente le sont d'un point de vue général. Le processus de phonation étant très complexe, certains muscles peuvent avoir une toute autre fonction lors de la réalisation de sons ou groupes de sons particuliers [Lofqvist 84].

### 3.3. Les cordes vocales

Comme le montre la coupe frontale du larynx présentée sur la figure A.6, les deux cordes vocales ressemblent plus à des lèvres qu'à des cordes. Chacune d'elle est constituée d'un muscle, le muscle vocal, et d'un ligament, le ligament vocal, recouverts d'une muqueuse.

Juste au-dessus des cordes vocales se trouve une autre paire de lèvres appelées bandes ventriculaires (ou fausses cordes vocales). Bien qu'elles n'aient aucun rôle direct dans la phonation, elles créent deux petites cavités résonantes supplémentaires, les ventricules de Morgani (ou ventricules laryngés) dont l'effet ne semble pas avoir été complètement étudié [Malmberg 74] mais que nous pouvons supposer minime par rapport à celui produit par les cavités supraglottiques. Il est possible que ces fausses cordes vocales exercent un rôle de protection, notamment microbienne, vis-à-vis des cordes vocales proprement dites.

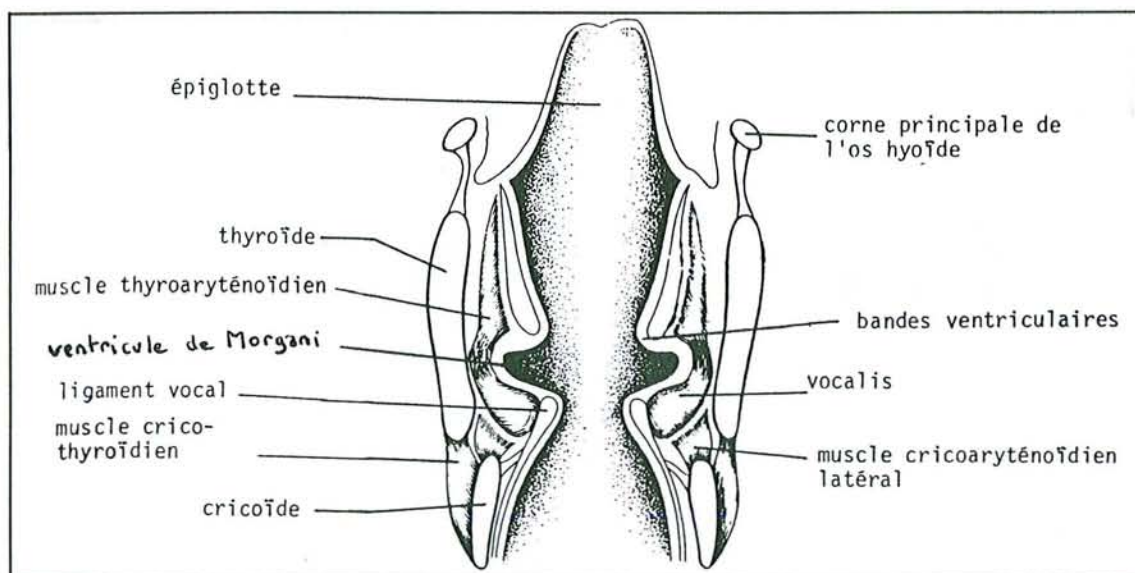


Figure A.6. Coupe frontale du larynx d'après Lamby dans [Calliope 89].



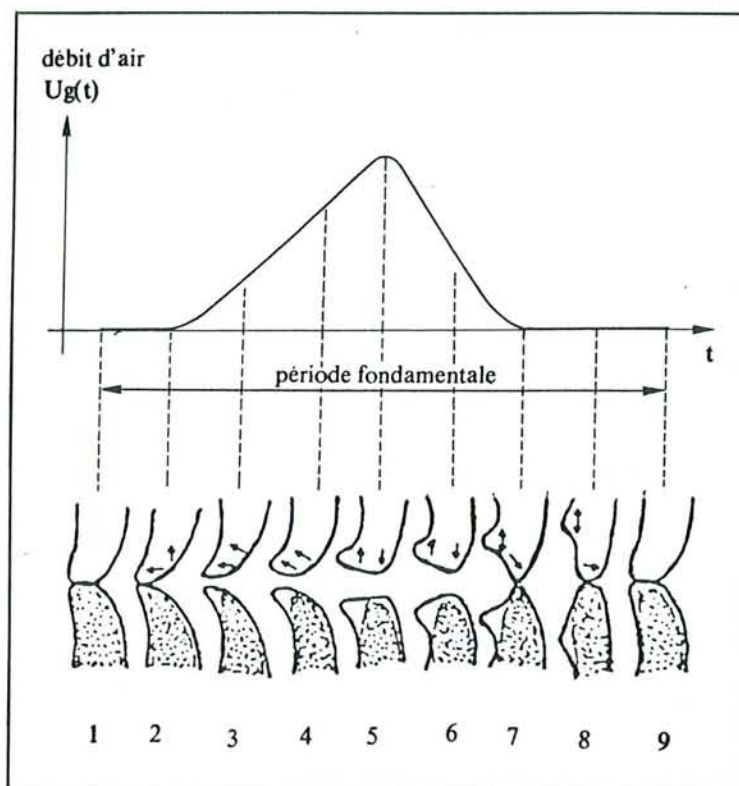


Figure A.7. Configuration des cordes vocales et onde glottale lors de la phonation, d'après Hirano dans [Calliope 89].

### 3.4. Le mécanisme de la phonation

Au cours des siècles, plusieurs théories ont essayé de rendre compte du fonctionnement du larynx [Marchal 80] [Malmberg 74]. Finalement, les techniques modernes d'exploration du larynx ont permis de valider la théorie myoélastique et aérodynamique de Van den Berg (1957) [Marchal 80] [Calliope 89].

Pendant la respiration, les cordes vocales sont tenues écartées pour permettre le passage de l'air. Le processus de la phonation débute donc par un accolement musculaire des cordes vocales. Au fur et à mesure de l'expiration, l'air s'accumule sous elles avec pour conséquence une augmentation de la pression sous-glottique qui finit par écarter les cordes vocales, laissant un passage libre à l'air. Sous l'action de leur masse, de leur tension et de leur élasticité, les cordes vocales reviennent vers le centre du larynx pour s'accoler à nouveau grâce à l'effet aérodynamique de Bernoulli : suite au rapprochement des cordes vocales, l'air passe entre elles plus rapidement, produisant une baisse de pression qui aspire les cordes vocales l'une vers l'autre. Puis, le cycle, schématisé sur la figure A.7 recommence.

Ces fermetures et ouvertures successives de la glotte — respectivement par adduction et abduction des cordes vocales — engendrent une onde de débit d'air, appelée onde glottale. Cette onde périodique ou quasi périodique, de forme approximativement triangulaire est à l'origine des sons voisés. Sa fréquence (quasi-fréquence), appelée fréquence fondamentale, est souvent notée  $F_0$ .

### 3.5. Les caractéristiques de l'onde glottale

Le fonctionnement de l'appareil phonatoire décrit ci-dessus montre que, si en première approximation les caractéristiques de l'onde glottale (forme, amplitude et fréquence) sont fonction des pressions supraglottique et sous-glottique, elles dépendent également de l'état des cordes vocales au moment de la phonation (taille, masse, tension et consistance).

Ces caractéristiques varient donc d'un locuteur à l'autre mais aussi, pour un locuteur donné, en fonction du registre phonatoire qu'il utilise. Dans le seul but de simplifier leur formulation, nous qualifierons de statiques les caractéristiques de l'onde glottale directement liées à l'anatomie du locuteur et de dynamiques celles qui sont soumises à son contrôle qu'il soit conscient ou inconscient.

#### 3.5.1. Les caractéristiques statiques

Chaque individu possède un registre phonatoire moyen appelé registre de poitrine. Dans ce mode, les cordes vocales sont épaisses (3 à 4 mm) et vibrent sur toute leur longueur (15 mm en moyenne pour les hommes et 13 mm pour les femmes). L'amplitude de vibration des cordes vocales est de l'ordre de 3 mm et l'aire d'ouverture maximale de la glotte est de 20 mm<sup>2</sup> pour les hommes et de 14 mm<sup>2</sup> pour les femmes [O'Shaughnessy 87]. A ces différences morphologiques correspond un premier domaine de variation de  $F_0$  : de 80 à 160 Hz pour les hommes et de 150 à 300 Hz pour les femmes.

De même, la forme et l'amplitude de l'onde glottale dépendent de la capacité respiratoire et d'une certaine énergie musculaire, appelée effort vocal. Cette énergie musculaire concerne aussi bien les muscles expirateurs, qui provoquent une plus grande force expiratoire, que les muscles du larynx et les cordes vocales qui sont plus tendus. Ces caractéristiques de l'onde glottale varient donc d'un locuteur à l'autre. En particulier, les voix féminines sont plus faibles que les voix masculines. Comme nous pouvons l'observer sur la figure A.8, les formes d'onde associées à une voix intense et à une voix faible sont très différentes.

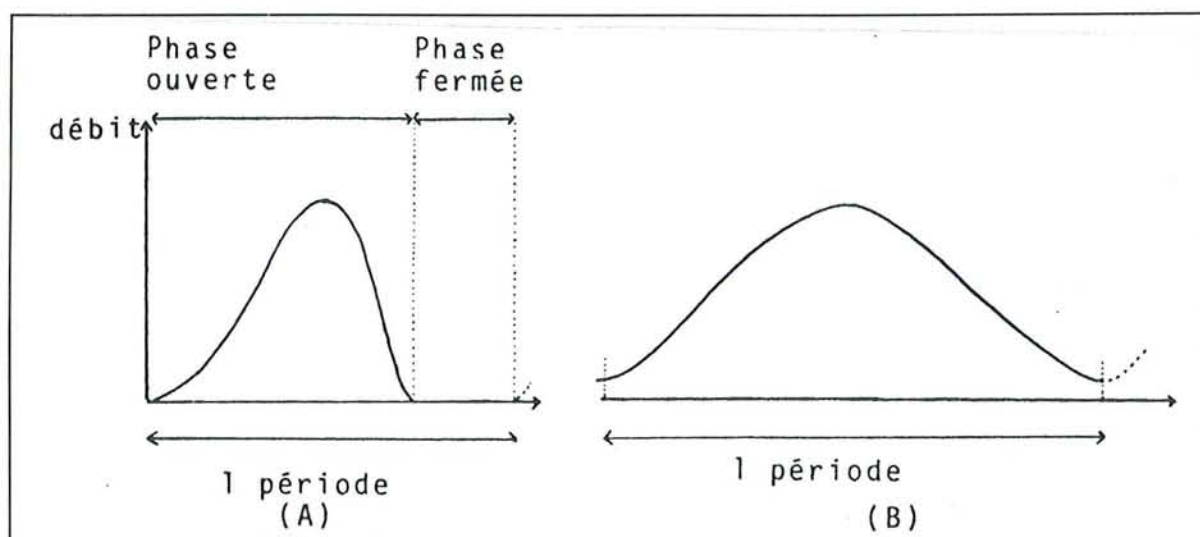


Figure A.8. Ondes glottales d'une voix intense (A) et d'une voix faible (B), d'après [Lonchamp 87b].



Dans le cas d'une voix intense, la glotte reste fermée pendant un intervalle de temps non négligeable qui peut atteindre la moitié de la période [Lonchamp 87b]. En outre, l'onde de débit d'air correspondant à la glotte ouverte est dissymétrique, la phase de fermeture de la glotte étant plus rapide que la phase d'ouverture.

Au contraire, l'onde glottale d'une voix faible est beaucoup plus symétrique et dans la plupart des cas ne s'annule pas. Lors de la phase d'adduction des cordes vocales, celles-ci sont relâchées, ne réalisent pas un accollement complet et laissent échapper un filet d'air non vibrant, ce qui engendre une voix soufflée (*breathy voice*) [Malmberg 74].

### 3.5.2. Les caractéristiques dynamiques

Un locuteur peut faire varier par divers moyens les caractéristiques de l'onde glottale qu'il produit. Ainsi, il peut accroître la fréquence fondamentale de vibration de ses cordes vocales soit en augmentant leur longueur ou leur tension, soit en diminuant leur épaisseur, soit encore en augmentant la pression sous-glottique et ceci jusqu'à un autre registre phonatoire appelé registre de tête [Marchal 80]. Plus communément, dans le langage parlé courant (par opposition à la parole chantée ou déclamée), pour un énoncé donné, le locuteur limite les variations de sa fréquence fondamentale à celles nécessaires à la production de la mélodie de cet énoncé<sup>1</sup>. Ce qui correspond à un domaine de variation de  $F_0$  dont l'étendue est inférieure à une octave : la valeur maximale de  $F_0$  est inférieure au double de la valeur minimale de  $F_0$  [O'Shaughnessy 87].

Dans une étude portant sur un locuteur masculin, J.E. Atkinson [Atkinson 78] remarque que dans le domaine des fréquences basses (de 80 à 100 Hz), les variations de  $F_0$  sont plutôt gérées par les variations de la pression sous-glottique, dont l'action est prépondérante lorsque les cordes vocales sont épaisses et relâchées. Alors que dans le domaine des fréquences plus élevées (de 120 à 160 Hz), les variations de  $F_0$  sont plutôt régies par l'activité des muscles crico-aryténoïdiens latéraux et crico-thyroïdien, qui agissent sur la structure des cordes vocales. Quant au domaine des fréquences basses ou élevées, l'auteur pense qu'il serait atteint par le positionnement du larynx grâce à ses muscles extrinsèques.

L'utilisation de l'allongement des cordes vocales — jusqu'à 4 mm d'après D. O'Shaughnessy [O'Shaughnessy 87] — pour augmenter leur fréquence de vibration semble contredire les observations du paragraphe précédent sur les caractéristiques statiques de l'onde glottale. En fait, l'effet dû à l'allongement des cordes vocales est annulé puis inversé par celui produit par leur amincissement et leur tension plus importante.

Conjointement, l'amplitude de l'onde glottale, et, par là l'intensité acoustique, peut être volontairement abaissée en réduisant la pression sous-glottique ou en diminuant le degré de fermeture de la glotte (cf. paragraphe précédent).

### 3.5.3. La "voix craquée" (*creaky voice*) et la "friture vocale" (*vocal fry*)

Suivant les auteurs, ces deux appellations correspondent à un seul mode de phonation ou à deux modes opposés. Tous sont unanimes sur l'observation d'un même phénomène dont un exemple est présenté sur la figure A.9 : une fréquence fondamentale faible et irrégulière (de 3 à 50 Hz ou de 18 à 65 Hz selon les auteurs), une phase de fermeture des cordes vocales longue (53 à 87% du cycle d'après R.L. Whitehead et al. [Whitehead 84]), précédée de une ou deux voire trois oscillations. Si la plupart des auteurs attribuent ce phénomène à

<sup>1</sup> La mélodie, évolution temporelle de  $F_0$  sur un énoncé, est l'une des manifestations acoustiques de la prosodie qui seront étudiées au paragraphe V.3.1.

un fort accolement des cordes vocales empêchant une grande partie de la corde de vibrer et l'appellent indifféremment "voix craquée" ou "friture vocale" [Marchal 80] [Lonchamp 87b], D. O'Shaughnessy [O'Shaughnessy 87] l'attribue à un relâchement des cordes vocales en fin de groupe de phonation qui sont alors courtes et épaisses. Il le nomme "friture vocale" et l'oppose au fort accolement des cordes vocales qui produit selon lui, une voix cassante, de faible intensité, ayant une fréquence fondamentale irrégulière et qu'il définit comme étant la "voix craquée".

Dans le corpus de données construit pour notre étude et qui sera présenté dans la partie C, ce phénomène semble se produire essentiellement en fin de phrase ; ce qui étayerait plutôt la thèse de D. O'Shaughnessy.

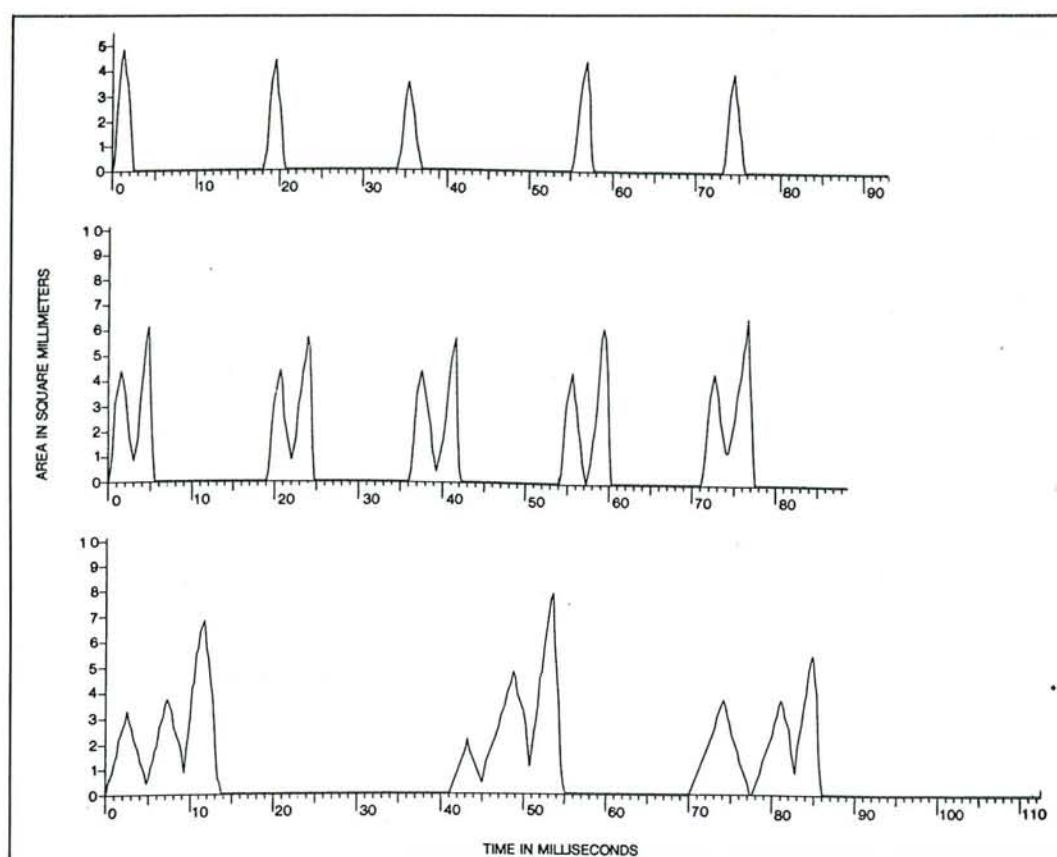


Figure A.9. Différentes formes d'onde glottale dans le cas de la "friture vocale", d'après [Whitehead 84].

### 3.6. Les autres modes de fonctionnement du larynx

Les modes de phonation décrits jusqu'à présent utilisent le larynx pour la production des sons voisés mais celui-ci possède d'autres modes de fonctionnement lui permettant d'élaborer d'autres types de sons :

- **la sourdité** : l'air passe entre les cordes vocales écartées engendrant des sons sourds ;
- **l'occlusion glottale** : les cordes vocales, fortement accolées, empêchent le passage de l'air. Celui-ci s'accumule derrière le barrage qu'elles forment avant d'être brusquement libéré



(explosion) lors de leur relâchement. Dans certaines langues, ce son correspond à un phonème à part entière. En français, il n'a qu'une fonction stylistique et est appelé coup de glotte ;

- **l'aspiration glottale** : lors de la phonation donc de l'expiration, l'air passe entre les cordes vocales plus ou moins rapprochées. C'est par exemple le cas lors de la production du "h aspiré" de l'anglais ("hat") ou de celle des occlusives sourdes aspirées des langues anglo-germaniques ;
- **le chuchotement** (*whispering*) : la glotte est fermée sauf au niveau des cartilages aryénoïdes, le passage de l'air dans cette petite ouverture triangulaire engendre un bruit de friction qui remplace l'onde périodique de la phonation normale.
- **le murmure** : c'est une combinaison du chuchotement et du voisement, la glotte reste ouverte au niveau des aryénoïdes et seule la partie antérieure des cordes vocales vibre.

La figure A.10 présente les degrés d'ouverture de la glotte suivant les principaux modes de fonctionnement du larynx.

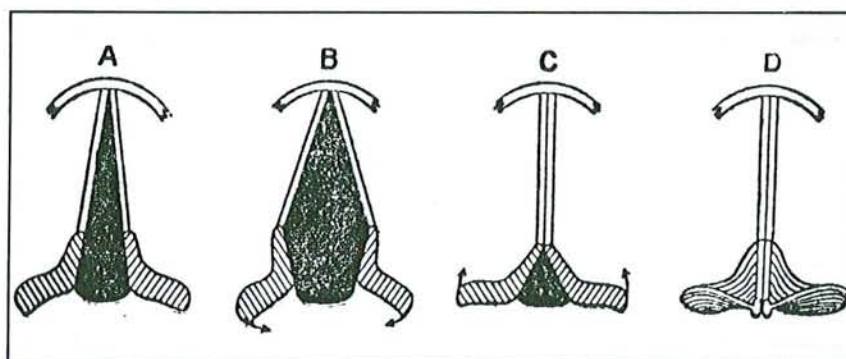


Figure A.10. Position de la glotte pendant, la respiration normale (A), la respiration forte (B), la voix chuchotée (C) et la phonation (D), d'après [Malmberg 79].

## 4. Les cavités supraglottiques

### 4.1. Description générale

La figure A.11 présente les cavités supraglottiques qui se scindent en deux grands conduits<sup>1</sup> : le conduit oral et le conduit nasal. Le conduit oral, dont la longueur varie de 15 cm chez les locutrices à 17 cm chez les locuteurs, s'étend de la glotte aux lèvres et comprend au minimum le pharynx et la cavité buccale, auxquels vient s'ajouter une cavité supplémentaire lorsque les lèvres sont projetées en avant. Lorsque le voile du palais est abaissé, le conduit nasal vient se brancher en parallèle sur le conduit oral. Il comprend les cavités nasales et les sinus de la face.

Les cavités supraglottiques remplissent deux rôles primordiaux dans la production de la parole :

- constituer un tube acoustique résonant de l'onde glottale, lorsqu'elle existe. Le conduit vocal, de forme très variable grâce aux articulateurs mobiles que sont la mandibule, la langue, les lèvres

<sup>1</sup> Ici, "conduit" est un terme général qui ne tient pas compte de la forme effective des cavités le constituant.



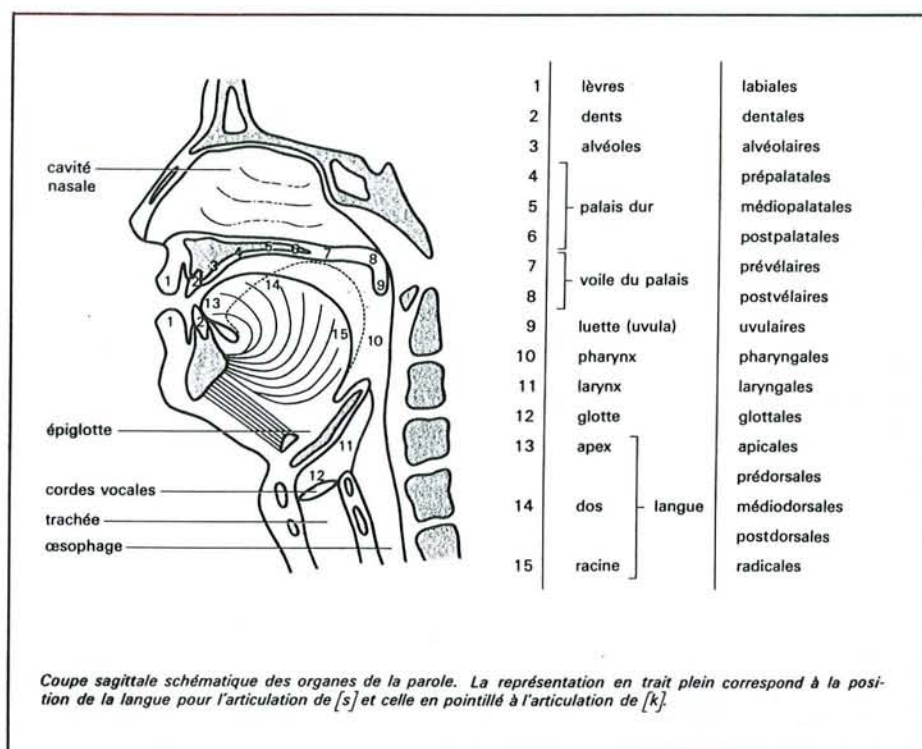


Figure A.11. Les cavités supraglottiques et les principales divisions anatomiques utilisées dans la description articulatoire des sons, d'après Encyclopedia Universalis.

et le voile du palais, possède, à un instant donné, des fréquences de résonance responsables du timbre des voyelles et des sonantes ;

- permettre la production de bruits en perturbant le passage de l'air par une obstruction (consonnes occlusives) ou par un rétrécissement (consonnes constrictives).

## 4.2. Le pharynx

Ce conduit vertical, musculo-membraneux, de taille variable selon le locuteur, s'étend de la glotte à la luette (cf. figure A.11) et peut varier en diamètre et longueur sous l'action de neuf muscles [Marchal 80].

Ces variations permettent de modifier la longueur et la forme du conduit vocal et par là ses fréquences de résonance. L'élargissement du pharynx peut également être utilisé pour prolonger la vibration des cordes vocales pendant la tenue des occlusives voisées<sup>1</sup> (/b, d, g/). Lors de la tenue des occlusives voisées, les cordes vocales vibrent jusqu'à l'égalité des pressions supraglottique et sous-glottique. Or, une augmentation du volume du conduit vocal abaisse la pression supraglottique et retarde l'égalisation des deux pressions.

<sup>1</sup> L'articulation des occlusives voisées est détaillée au paragraphe III.3.1.

### 4.3. La cavité buccale

#### 4.3.1. Articulateurs fixes et mobiles

Lors de la production d'un son, la configuration de la cavité buccale est obtenue par le positionnement des articulateurs mobiles que sont la mâchoire inférieure, la luette, les lèvres et la langue par rapport aux articulateurs fixes que sont la mâchoire supérieure, le pharynx, les dents et le palais. Comme pour le larynx, chaque articulateur est mû par un ensemble de muscles extrinsèques et intrinsèques. En général, le déplacement d'un articulateur s'effectue en deux phases grâce à l'action d'un couple de muscles antagonistes : la contraction de l'un des muscles et la relaxation de l'autre projettent l'articulateur vers la cible puis la contraction du deuxième muscle permet d'ajuster la trajectoire lors de l'approche de la cible [O'Shaughnessy 87].

Nous allons décrire sommairement quelques-uns de ces articulateurs.

#### 4.3.2. Le palais

Le palais se divise en trois parties (cf. figure A.11) : les alvéoles, situées juste derrière les dents, le palais dur qui correspond à l'os palatal et le palais mou ou voile du palais terminé par la luette ou uvule qui, lorsqu'elle est relevée, empêche l'air de passer dans les fosses nasales. A l'exception de cette dernière partie, le palais est un articulateur fixe mais dont la forme varie d'un individu à l'autre.

#### 4.3.3. La langue

Deuxième organe essentiel dans la production de la parole après le larynx, la langue est le plus mobile et le plus complexe des articulateurs. Elle est constituée d'un ensemble de dix-sept muscles recouverts d'une muqueuse. Les muscles intrinsèques sont responsables de la forme de la langue : aplatissement, relèvement des bords ou de la pointe, etc., alors que les muscles extrinsèques provoquent son déplacement dans la cavité buccale [Marchal 80] [O'Shaughnessy 87].

Du point de vue articulatoire, la langue est découpée en trois parties pouvant se mouvoir presque séparément (cf. figure A.11) : la pointe (*apex*) qui est la partie la plus agile, le dos (*dorsum*) et la racine (*radix*)<sup>1</sup>.

#### 4.3.4. Les lèvres

Les lèvres sont deux replis musculo-membraneux. Egalement très mobiles, elles ont comme le reste de la cavité deux fonctions : leur arrondissement et leur projection en avant modifient le timbre des sons voisés par l'ajout d'un quatrième résonateur alors que leur positionnement l'une par rapport à l'autre ou par rapport à un articulateur fixe permet la réalisation de consonnes obstruantes comme / p / et / f /.

<sup>1</sup> Les noms latins précisés entre parenthèses interviennent dans le classement articulatoire des sons (cf. chapitre III).

## 4.4. Les cavités nasales et les sinus paranasaux

### 4.4.1. Introduction

Ces résonateurs supplémentaires sont mis en œuvre par l'abaissement de la luette lors de la production des voyelles et des consonnes nasales. Bien qu'ils soient fixes, leur forme et leur volume sont très variables selon le locuteur et surtout selon son état pathologique (coryza, sinusite, déplacement de la cloison nasale, ...).

Les informations concernant l'anatomie de ces cavités, regroupées dans les paragraphes suivants, sont en grande partie issues de la première partie de la thèse d'Etat de F. Lonchamp consacrée aux indices acoustiques de la nasalité vocalique [Lonchamp 88] et du traité d'anatomie de A. Bouchet et J. Guilleret. [Bouchet 83].

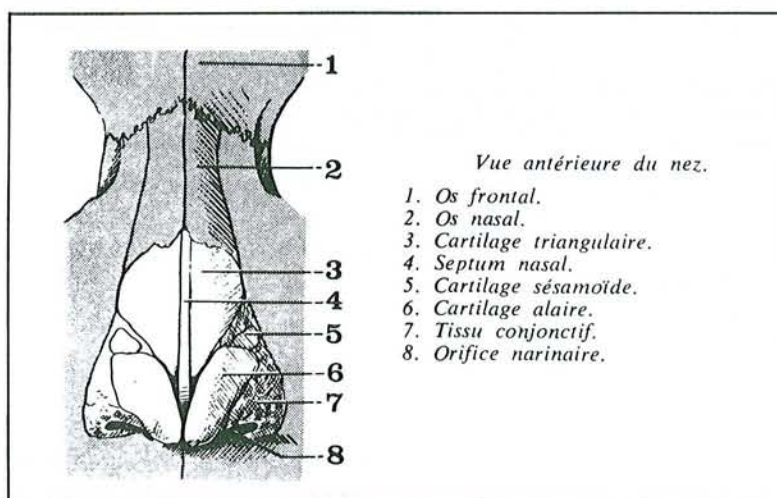


Figure A.12. Os et cartilages du nez, d'après [Bouchet 83].

### 4.4.2. Les cavités nasales

Elles comprennent, d'avant en arrière :

- **le nez** qui, comme le montre la figure A.12, est un ensemble d'os et de cartilages dont l'un d'eux, le septum nasal, le divise en deux parties pas toujours égales ;
- **les fosses nasales**. Ces deux cavités anfractueuses ont globalement la forme d'une pyramide tronquée, longue de 7 à 8 cm, haute de 5 cm et large de 4 cm à sa base et de 1 cm à son sommet.

Les parois de ce volume trapézoïdal sont formées d'un ensemble complexe d'os, de cartilages et de muqueuses, qu'il n'est pas facile de décrire simplement. Schématiquement, elles sont constituées en bas par la voûte palatine, en haut – respectivement d'avant en arrière – par l'os du nez, l'os frontal, la lame criblée de l'os éthmoïde et l'os sphénoïde, et latéralement par l'ensemble des parties marquées d'un astérisque sur la figure A.13.



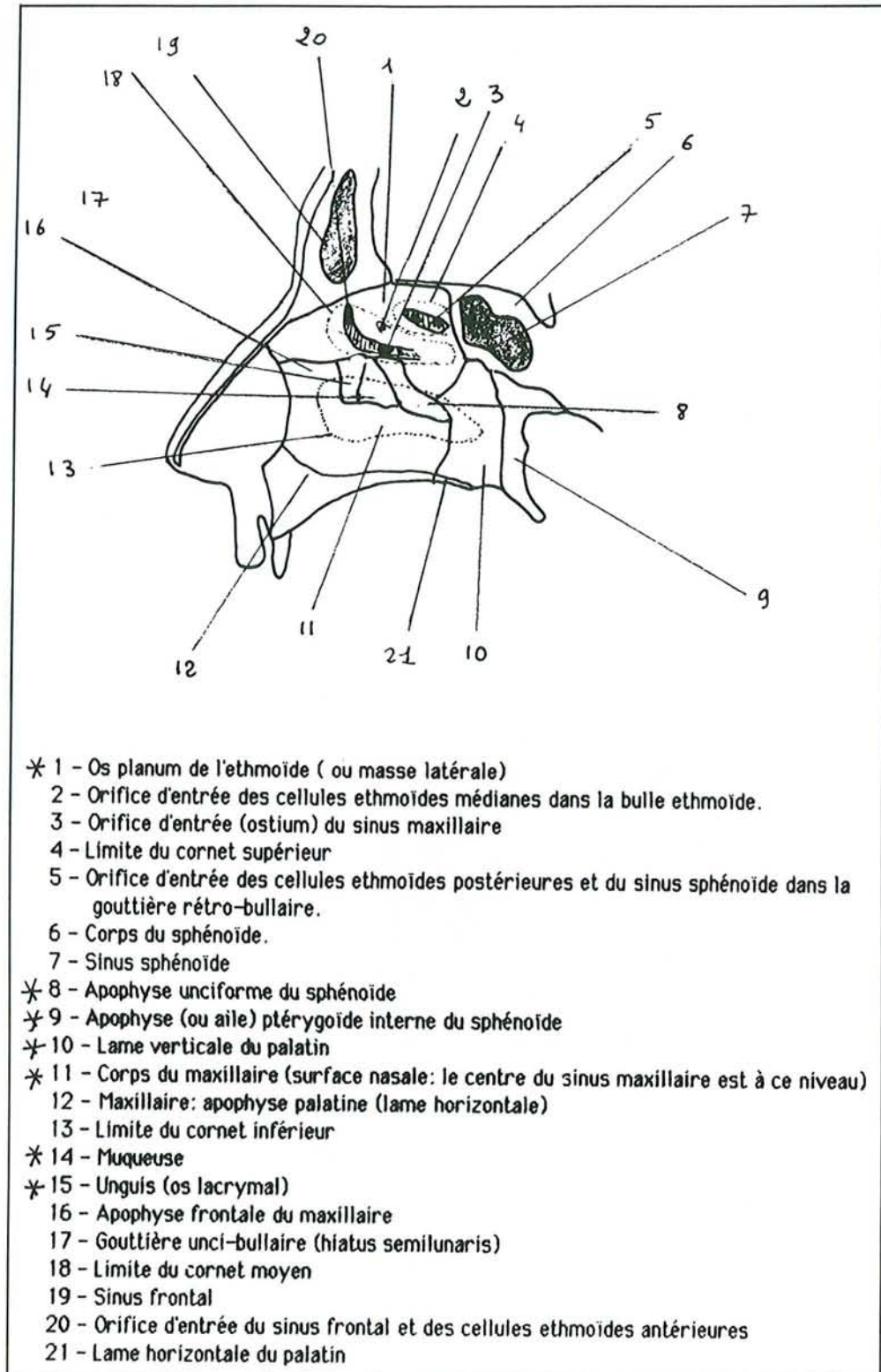


Figure A.13. Paroi latérale des fosses nasales (cornets supprimés), représentation schématique d'après [Lonchamp 88].

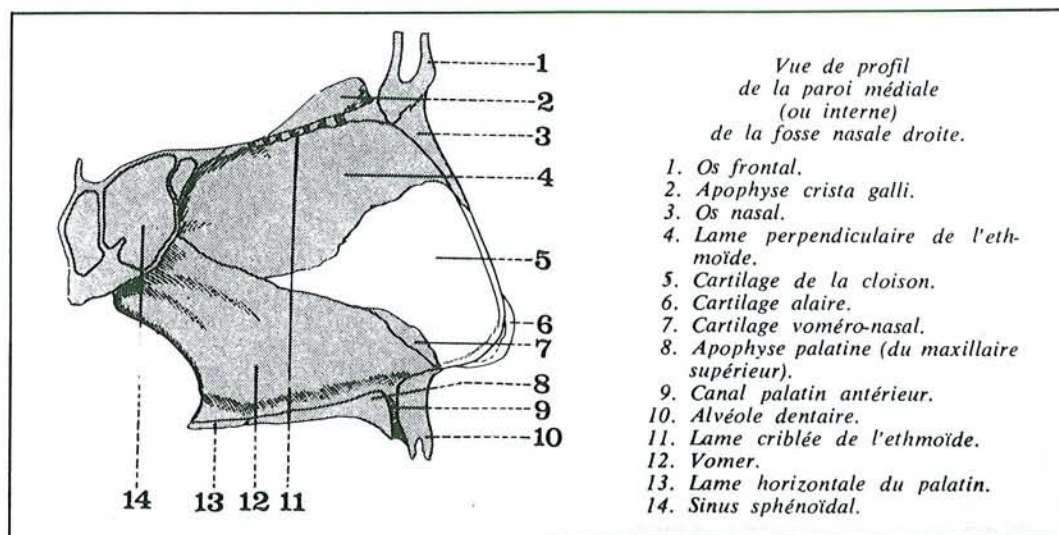


Figure A.14. La cloison nasale, d'après [Bouchet 83].

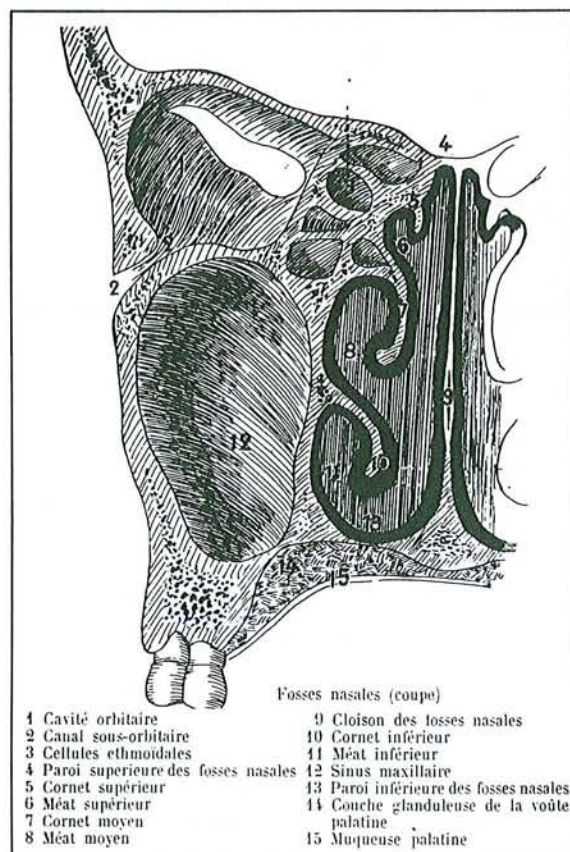


Figure A.15. Coupe schématique frontale des fosses nasales, d'après l'Encyclopédie Pratique de Médecine et d'Hygiène.

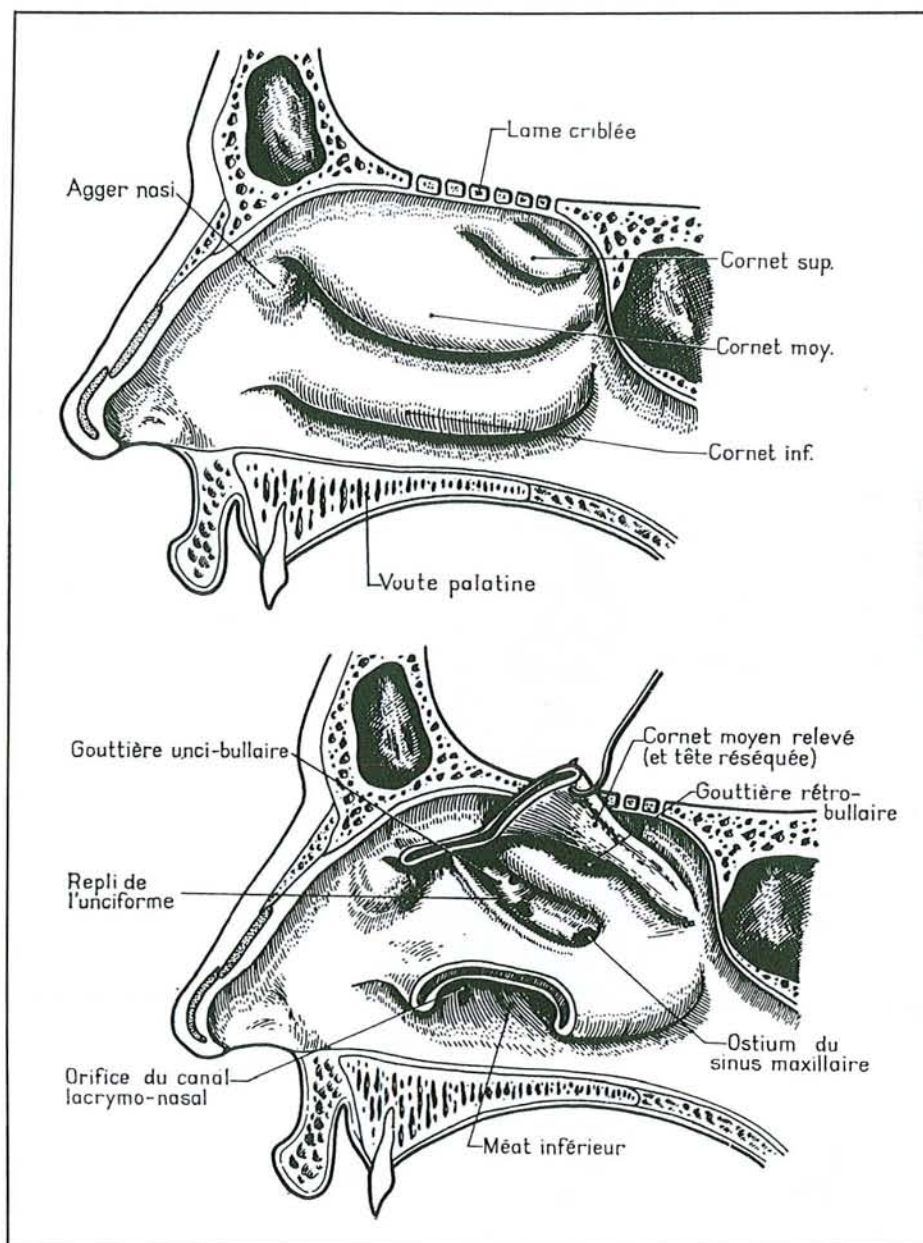


Figure A.16. Paroi latérale des fosses nasales : les cornets et les structures sous-jacentes, d'après [Lonchamp 88].

La cloison nasale, présentée sur la figure A.14, sépare complètement ce volume en deux cavités, le plus souvent dissymétriques.

Les figures A.16 et A.15 révèlent la complexité de la structure interne de ces deux cavités. En effet, la paroi latérale externe de chaque fosse est partiellement masquée par trois replis cartilagino-osseux, recourbés selon leur grand axe, appelés respectivement cornets supérieur, moyen et inférieur. Chaque cornet délimite avec la paroi latérale un espace de quelques millimètres appelé méat. Ce sont dans ces méats que débouchent les sinus de la face ;



- **le rhinopharynx** ou cavum qui correspond à la partie supérieure du pharynx dans laquelle s'ouvrent les fosses nasales. Ce conduit, de forme parallélépipédique, est délimité par la paroi arrière du pharynx, l'os sphénoïde et le voile du palais. Celui-ci, lorsqu'il se contracte, forme une cloison horizontale qui isole complètement le rhinopharynx de l'oropharynx. Les dimensions moyennes du rhinopharynx sont 4 cm verticalement et transversalement et 2 cm dans le sens antéro-postérieur. Toutefois, la figure A.17 montre qu'il peut présenter des sections différentes selon les locuteurs.

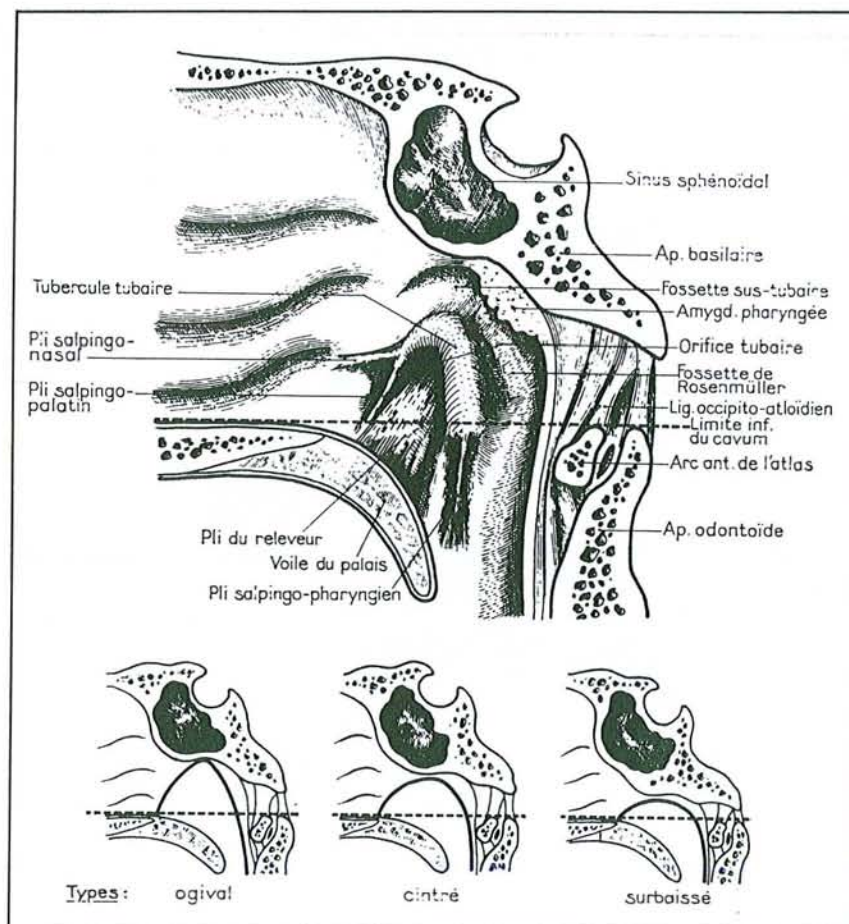


Figure A.17. Les différents types de rhinopharynx, d'après Pailoux dans [Lonchamp 88].

#### 4.4.3. Les sinus paranasaux

Aux cavités nasales déjà décrites s'ajoutent d'autres cavités creusées dans les os de la face, les sinus. Ce sont des diverticules des fosses nasales dans lesquelles ils débouchent par des orifices étroits. Ils sont pairs, quatre de chaque côté, mais peu symétriques.

Les dimensions et le rôle dans la production de la nasalité de ces résonateurs supplémentaires sont encore assez mal connus, seuls les travaux de l'équipe Lindquist-Sundberg de Stockholm et ceux de F. Lonchamp de Nancy [Lonchamp 88] ont essayé de mettre en évidence l'influence des résonances sinuales dans le spectre des voyelles et des consonnes nasales.

Une grande partie de cette méconnaissance provient d'une part de l'importante variabilité anatomique et pathologique de leurs dimensions et de leur état physiologique (polypes, sinusites, coryzas, ...) et d'autre part du fait que cette grande variabilité n'affecte pas le processus humain de compréhension de la parole.

On distingue :

- **les sinus ethmoïdaux** ou cellules éthmoïdales qui sont deux ensembles de petites cavités, plus ou moins indépendantes les unes des autres, creusées dans l'os ethmoïde et qui communiquent avec les fosses nasales au niveau des méats supérieurs et moyens (cf. figures A.13 et A.15). Leur taille et leur nombre varient selon le sujet et le côté de la face. S. Takeuchi, cité dans [Lonchamp 88], évalue leur volume total à  $8,3 \text{ cm}^3$  ;
- **les sinus frontaux**. Ces deux cavités sont situées au-dessus des orbites et incluses dans l'os frontal. Comme le montre la figure A.18, leur taille et leur forme sont aussi très variables

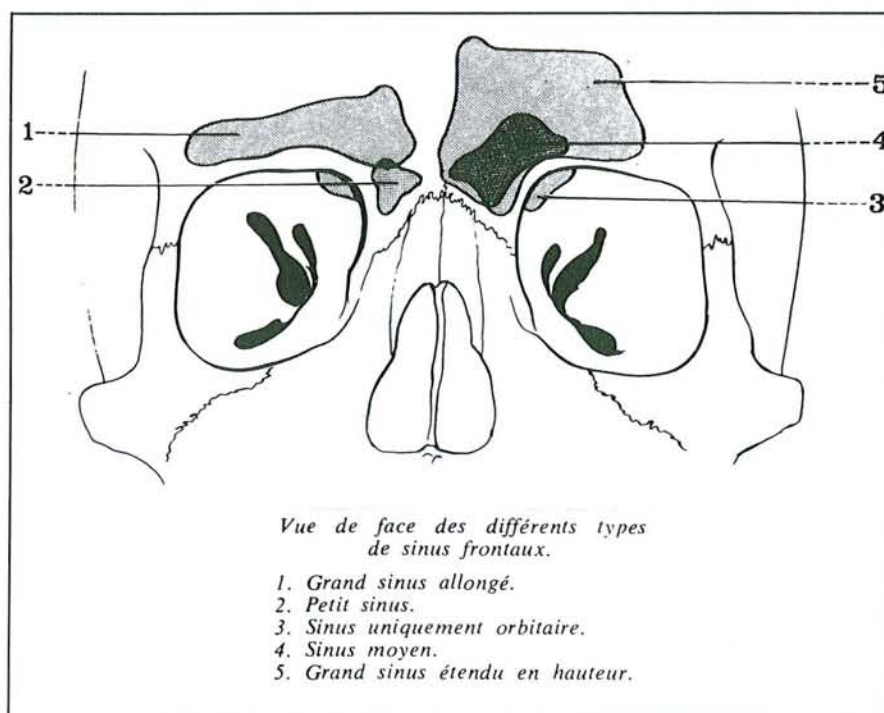


Figure A.18. Les différents types de sinus frontaux, d'après [Bouchet 83].

selon le sujet et le côté de la face. Les petits sinus sont surtout l'apanage des femmes et les grands celui des hommes. F. Lonchamp leur attribue à chacun un volume théorique maximal de  $15 \text{ cm}^3$  tout en citant S. Takeuchi qui se limite à  $5,7 \text{ cm}^3$ . A. Bouchet [Bouchet 83] propose une capacité de  $5 \text{ cm}^3$  pour un volume externe de  $10 \text{ cm}^3$ .

Chaque sinus débouche dans la gouttière unci-bullaire du méat moyen (cf. figures A.13 à A.15) ;

- **les sinus sphénoïdaux** qui sont deux cavités plus ou moins contiguës, creusées dans le corps de l'os sphénoïde et qui débouchent dans les fosses nasales au niveau des méats supérieurs, dans la gouttière rétro-bullaire. Comme les sinus frontaux, ils sont très asymétriques et de taille variable selon les individus. Différentes formes de sinus sphénoïdaux sont présentées sur la figure A.19. Leur capacité varie de  $1 \text{ cm}^3$  à plus de  $10 \text{ cm}^3$  [Bouchet 83] ;

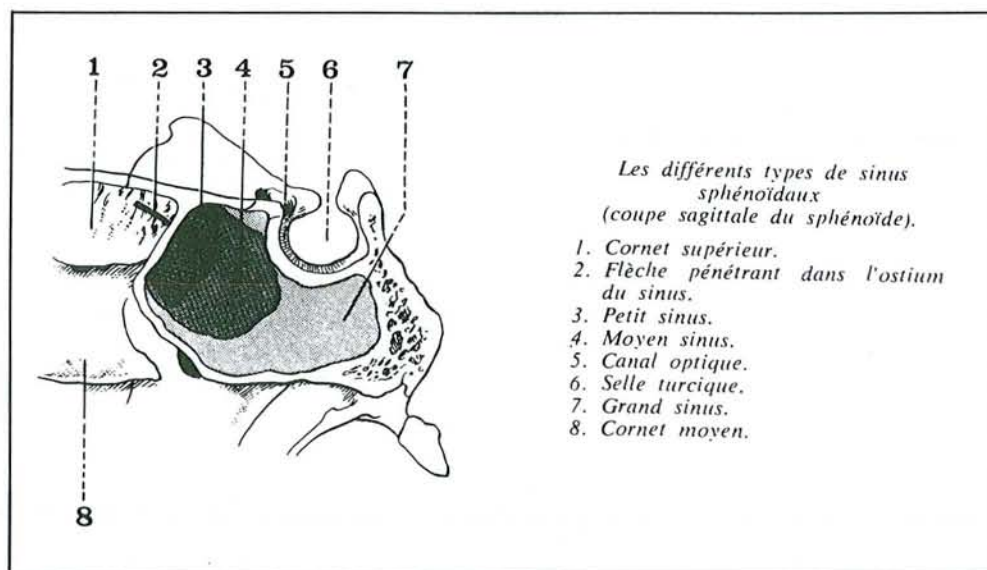


Figure A.19. Les différents types de sinus sphénoïdaux, d'après [Bouchet 83].

- **les sinus maxillaires.** Situées sous les orbites de part et d'autre des fosses nasales, ce sont les deux plus grandes cavités paranasales de la face. Comme le montre la figure A.20, leur volume diffère d'un locuteur à l'autre mais, contrairement aux autres sinus, ils sont symétriques. F. Lonchamp cite une étude de Aust et Drettner dans laquelle ils répertorient des volumes allant de  $7,8 \text{ cm}^3$  pour un sujet féminin à  $27,5 \text{ cm}^3$  pour un sujet masculin. A. Bouchet propose une capacité moyenne de  $12 \text{ cm}^3$  mais il cite des cas de petits sinus de  $2 \text{ cm}^3$  et de grands sinus de  $25 \text{ cm}^3$ , surtout chez les hommes.

Chacun des sinus débouche dans la gouttière unci-bullaire du méat moyen par l'intermédiaire d'un petit canal, l'ostium maxillaire (cf. figure A.16). Une occlusion de l'ostium par un gonflement de la muqueuse nasale due à un léger coryza entraîne une mise hors service du sinus correspondant qui n'intervient plus dans la résonance nasale [Lonchamp 88].



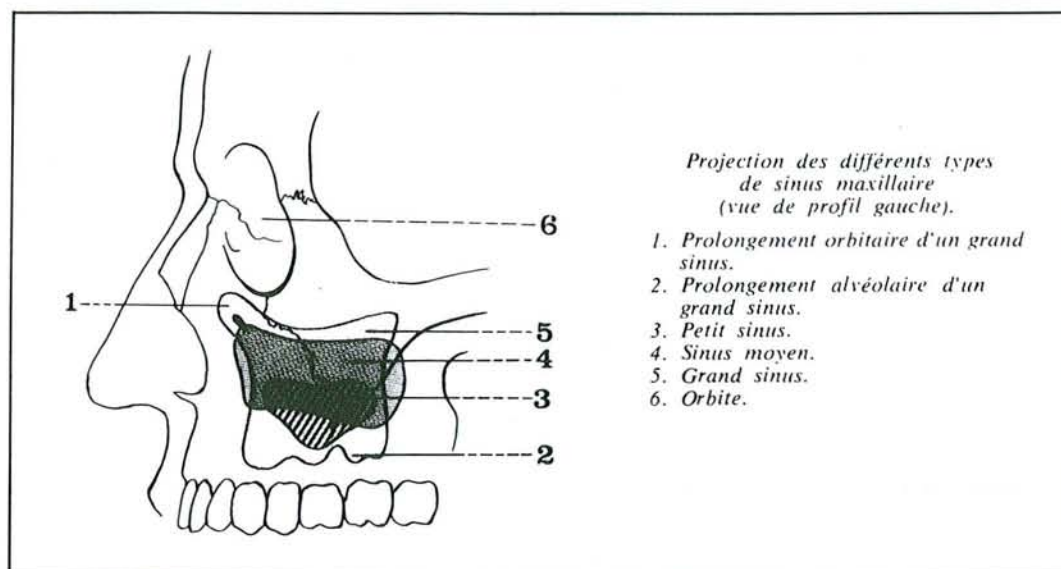


Figure A.20. Les différentes tailles de sinus maxillaires, d'après [Bouchet 83].

## 5. Conclusion

Dans ce chapitre, nous avons décrit le processus de production de la parole en mettant en évidence la complexité de fonctionnement de certains organes comme le larynx. Ceci nous a permis d'introduire les premières caractéristiques de l'onde glottale et de la fréquence fondamentale aussi bien entre locuteurs différents que dans des modes distincts d'utilisation du larynx par un seul locuteur.

Lors de la description des cavités supraglottiques qui interviennent dans ce processus, nous avons particulièrement détaillé les paragraphes se rapportant aux cavités nasales. En effet, souvent dans les ouvrages, la production des voyelles et consonnes nasales est résumée par la phrase sybilline : "l'abaissement du voile du palais permet le passage de l'air dans le conduit nasal". Mais très peu d'informations sont données sur la complexité et la variabilité anatomique de ce résonateur supplémentaire couplé au conduit oral. En fait, l'influence de ces cavités dans la production des voyelles et consonnes nasales est peu connue et sa modélisation demeure un problème difficile.

Tout ce que nous avons développé dans ce chapitre concerne le mécanisme général de production des sons voisés et non voisés et s'applique à toutes les langues. Dans le chapitre suivant, nous allons expliciter la mise en œuvre de ce mécanisme dans la réalisation des phonèmes du français.



## CHAPITRE III LES SONS DU FRANÇAIS

### 1. Introduction

L'objet de ce chapitre est la description articulatoire des réalisations physiques "normalisées" des phonèmes du français.

Comme nous l'avons vu dans le chapitre I, la table A.1 regroupe les principaux phonèmes du français transcrits simultanément dans l'Alphabet Phonétique International (API) et dans l'alphabet phonétique utilisé par le logiciel SNORRI qui a servi à visualiser et à étiqueter notre corpus de données.

Il est possible de classer les phonèmes d'une langue en plusieurs catégories en fonction d'un ensemble de traits distinctifs. Les trois principaux traits distinctifs des sons du français sont :

- **le voisement** : la vibration ou l'absence de vibration des cordes vocales réalise un premier découpage en sons sonores et sons sourds ;
- **la nasalité** : selon que le voile du palais est relevé ou abaissé, les sons se répartissent en sons oraux et sons nasaux ;
- **la présence ou l'absence d'un obstacle** au passage de l'air dans le conduit vocal créent les deux grandes classes que sont les consonnes et les voyelles.

Conformément à ce dernier trait, nous effectuerons un premier découpage de ce chapitre en deux paragraphes, l'un étant consacré aux voyelles, l'autre aux consonnes. Après une présentation rapide de l'articulation des voyelles orales, nous insisterons un peu plus sur celle des voyelles nasales en nous fondant sur les résultats de quelques travaux récents. Puis, nous terminerons ce premier paragraphe par quelques mots sur le cas particulier du "*e muet*". Le deuxième paragraphe décrira l'articulation des consonnes selon un classement qui prendra en compte, dans un premier temps, leur mode d'articulation, puis, leur lieu d'articulation.

### 2. Les voyelles

Les voyelles sont des sons voisés et stationnaires dont les différents timbres résultent du filtrage de l'onde glottale par le conduit vocal.

Le français possède 16 voyelles qui se répartissent en :

- 11 voyelles orales : / i, e, ɛ, a, ɑ, ɔ, o, u, y, ø, œ /,
- 4 voyelles nasales : / ã, õ, ẽ, œ̃ /,
- un phonème au statut particulier / ə /.



## 2.1. Les voyelles orales

La figure A.21, que les phonéticiens nomment triangle ou trapèze articulatoire selon son degré de précision, classe les voyelles orales du français selon les trois critères articulatoires suivants :

- le degré d'élévation de la langue par rapport au palais qui classe les voyelles selon quatre degrés, appelés degrés d'aperture : *ouvert* ([ a ] ), *mi-ouvert* ([ ɛ ] ), *mi-fermé* ([ e ] ) et *fermé* ([ i ] ) ;
- la position longitudinale du point le plus élevé de la langue qui détermine une articulation plus ou moins *antérieure* ([ i ] ) ou *postérieure* ([ u ] ) ;
- le degré de labialité (ou d'arrondissement des lèvres) : lors de la prononciation d'une voyelle (ou d'une consonne), les lèvres peuvent être *écartées* ([ e ] ) ou *arrondies* ([ ø ] ). En français, l'arrondissement des lèvres s'accompagne toujours de leur projection en avant — même si celle-ci peut être plus ou moins importante — ce qui crée une cavité résonante supplémentaire.

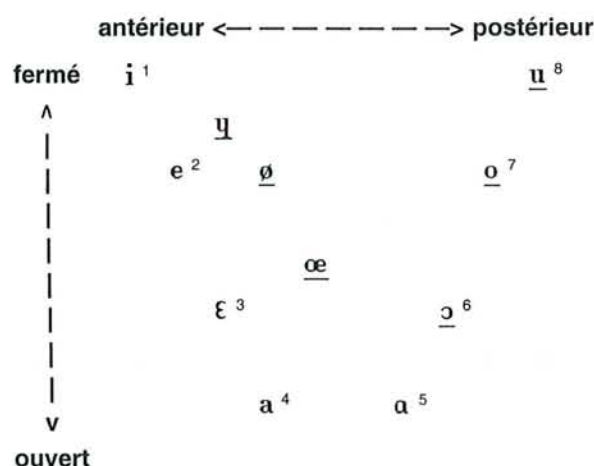


Figure A.21. Le trapèze articulatoire des voyelles orales du français, les voyelles soulignées sont arrondies. Chacun des numéros permet de repérer l'articulation de la voyelle correspondante sur la figure A.22.

La figure A.22 illustre ces trois critères pour les voyelles situées sur les bords du trapèze articulatoire.

Bien qu'il soit toujours très usité, ce trapèze articulatoire date d'une époque, la fin du XVIII<sup>e</sup> siècle, où seule la partie visible du conduit vocal était prise en considération dans la description articulatoire des sons. Par conséquent, il ne met pas en relief certaines composantes articulatoires comme le degré d'ouverture du pharynx. La figure A.23 montre, par exemple, qu'une élévation de la langue ([ i, u ]) provoque un élargissement du pharynx alors que son abaissement ([ ɔ, ɑ ]) rapproche la racine de la langue de la paroi pharyngale.

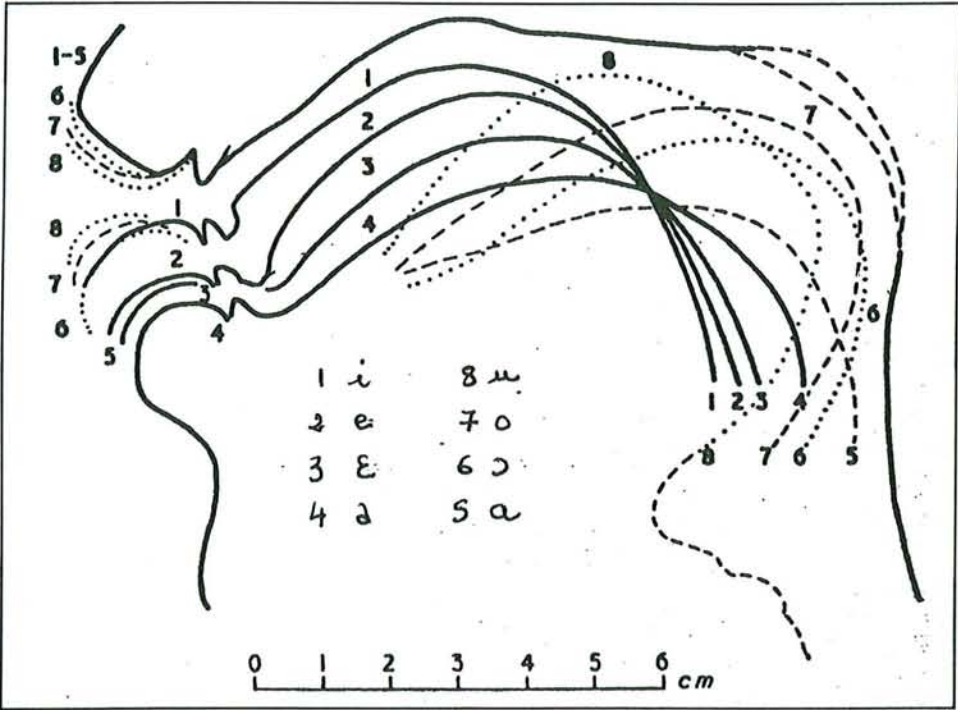


Figure A.22. Articulation de quelques voyelles orales, d'après [Lonchamp 87b].

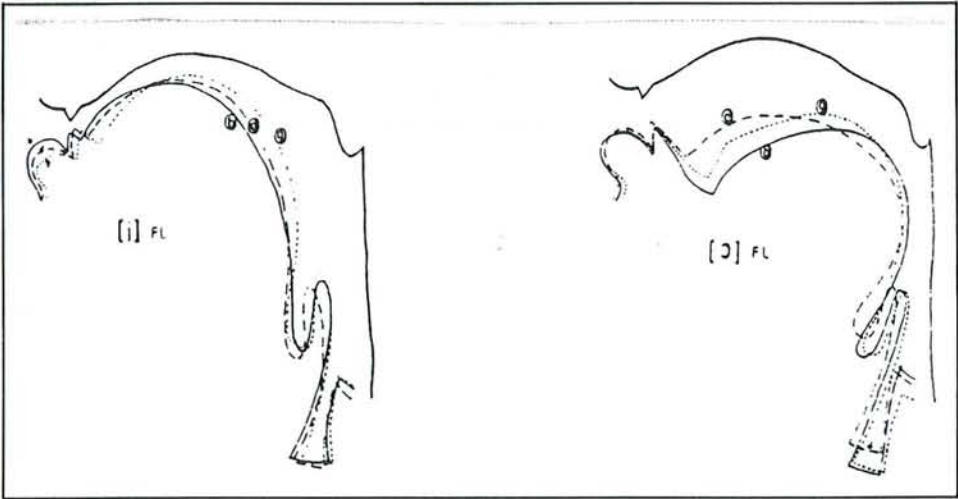


Figure A.23. Constriction pharyngale lors de l'articulation d'un [ i ] et d'un [ ɔ ], d'après [Lonchamp 87b].

## 2.2. Les voyelles nasales

L'abaissement du voile du palais caractérise l'articulation des voyelles nasales françaises /  $\tilde{\epsilon}$ ,  $\tilde{œ}$ ,  $\tilde{ɑ}$ ,  $\tilde{ɔ}$  /. Malgré le symbolisme utilisé, les articulations orales de ces quatre voyelles ne correspondent pas à celles des symboles situés sous le signe " ~ " : [  $\tilde{\epsilon}$  ] n'est pas un [  $\epsilon$  ] nasalisé.

Définir formellement cette correspondance se révèle difficile. Dédaisant ses conclusions d'une étude spectrale de six échantillons par voyelle nasale prononcés par deux locuteurs, F. Lonchamp propose dans [Lonchamp 79] de nouveaux symboles phonémiques traduisant mieux l'articulation linguale des voyelles nasales : respectivement /  $\tilde{æ}$ ,  $\tilde{ɜ}$ ,  $\tilde{ɔ}$ ,  $\tilde{ɔ}$  / à la place de /  $\tilde{\epsilon}$ ,  $\tilde{œ}$ ,  $\tilde{ɑ}$ ,  $\tilde{ɔ}$  /. Mais quelques années plus tard, à la lumière d'une étude réalisée par Zerling sur des cinéradiographies de voyelles nasales prononcées par les mêmes locuteurs, F. Lonchamp met en évidence des articulations plus individuelles [Lonchamp 88].

Nous reprenons ici l'essentiel de sa description comparative des figures A.24, A.25 et A.26, en la complétant en *italique* par nos propres remarques :

- [  $\tilde{\epsilon}$  ] : articulation linguale plus reculée que pour [  $\epsilon$  ] pour les deux sujets mais avec un recul plus marqué, accompagné d'une légère labialisation pour le locuteur FL (cf. figure A.24).

*Ce recul semble incompatible avec la première proposition de Lonchamp puisque, comme le montre la figure A.27, [  $\tilde{æ}$  ] est une voyelle antérieure plus ouverte que [  $\epsilon$  ] ;*

- [  $\tilde{œ}$  ] : curieusement, le sujet JPZ utilise la même position linguale pour [  $\tilde{œ}$  ] et [  $\tilde{\epsilon}$  ]. En revanche, [  $\tilde{œ}$  ] est surlabialisé par rapport à [  $\tilde{\epsilon}$  ] (cf. figure A.26). Cet individu réalise donc l'opposition /  $\tilde{\epsilon}$  / ~ /  $\tilde{œ}$  / par l'opposition (- labial) ~ (+ labial). La figure A.26 montre que pour le locuteur FL aussi l'articulation orale de [  $\tilde{œ}$  ] est très voisine de celle de [  $\tilde{\epsilon}$  ].

*Pour le sujet JPZ, la remarque sur les positions linguales très proches peut aussi s'appliquer à [  $\tilde{œ}$  ] et [  $\tilde{œ}$  ]. De plus, comme les deux sons présentent le même écartement labial (cf. figure A.24), on peut supposer que pour ce locuteur, [  $\tilde{œ}$  ] est bien un [  $\tilde{œ}$  ] nasalisé. Pour le locuteur FL, la comparaison de l'articulation de [  $\tilde{œ}$  ] par rapport à celle de [  $\tilde{œ}$  ], révèle un léger recul de la langue et une délabialisation, ce qui correspond à une réalisation proche d'un [  $\tilde{ɜ}$  ] : voyelle postérieure, mi-ouverte, non arrondie ;*

- [  $\tilde{ɑ}$  ] : par rapport à celle de [  $\tilde{ɑ}$  ], l'articulation de [  $\tilde{ɑ}$  ] est surlabialisée pour les deux sujets (cf. figure A.25). Le locuteur JPZ y ajoute un recul de la langue.

*Ce recul s'accompagne d'une légère fermeture du conduit vocal. Pour ce locuteur, ces trois modifications caractérisent une articulation plus arrondie, plus postérieure et plus fermée, et par conséquent voisine de [  $\tilde{ɔ}$  ] ;*

- [  $\tilde{ɔ}$  ] : par rapport à l'articulation d'un [  $\tilde{ɔ}$  ], les deux sujets augmentent l'arrondissement des lèvres. En outre, l'un recule la langue vers le haut, l'autre la recule horizontalement mais abaisse sa partie antérieure (cf. figure A.25).

*Les deux locuteurs semblent viser le même but, une articulation voisine de [  $\tilde{ɔ}$  ], même s'ils le réalisent différemment.*



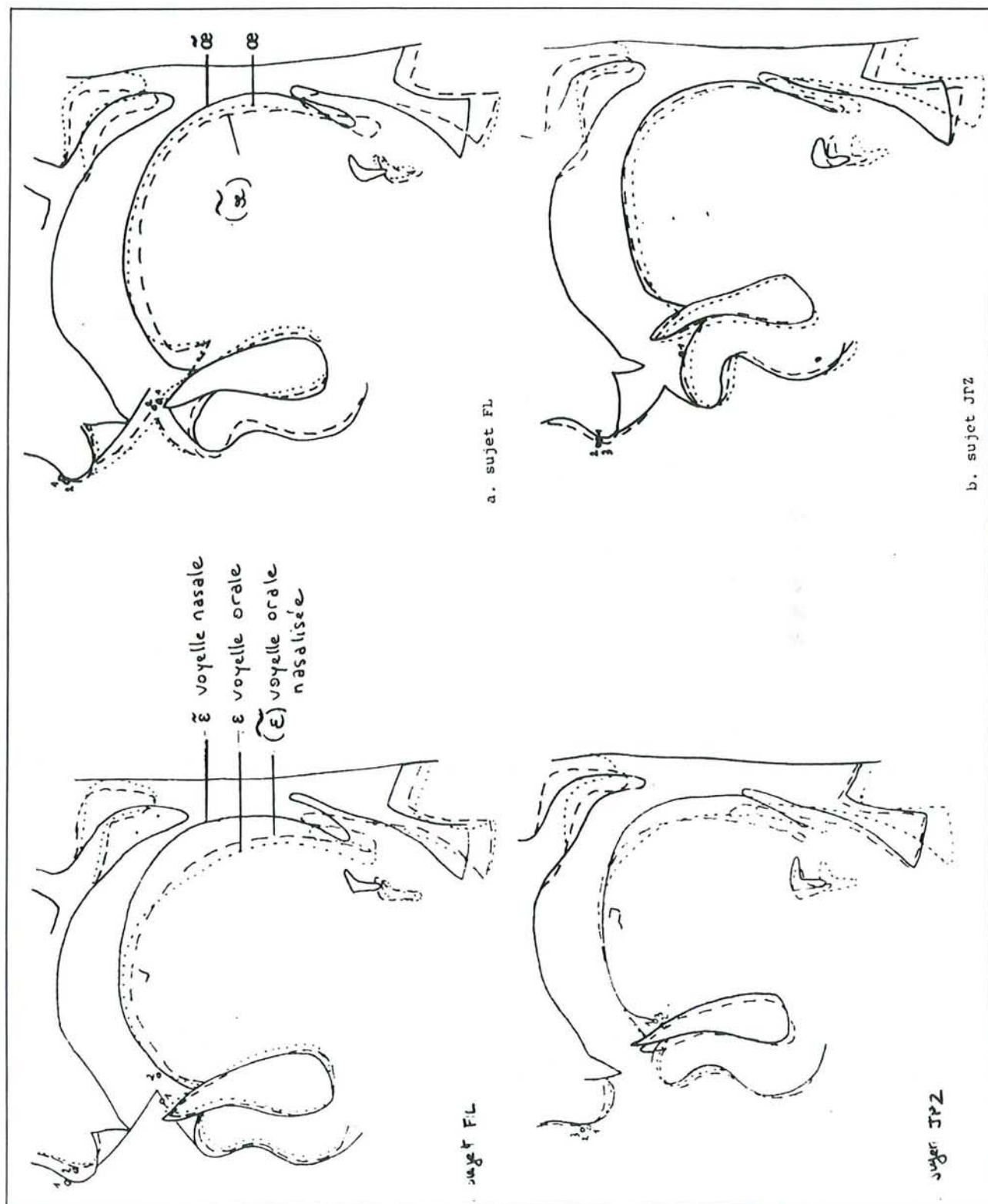


Figure A.24. Positions articutoires des voyelles [ε], [ẽ] et [ε] nasalisé et des voyelles [œ], [œ̃] et [œ] nasalisé d'après Zerling dans [Lonchamp 88].

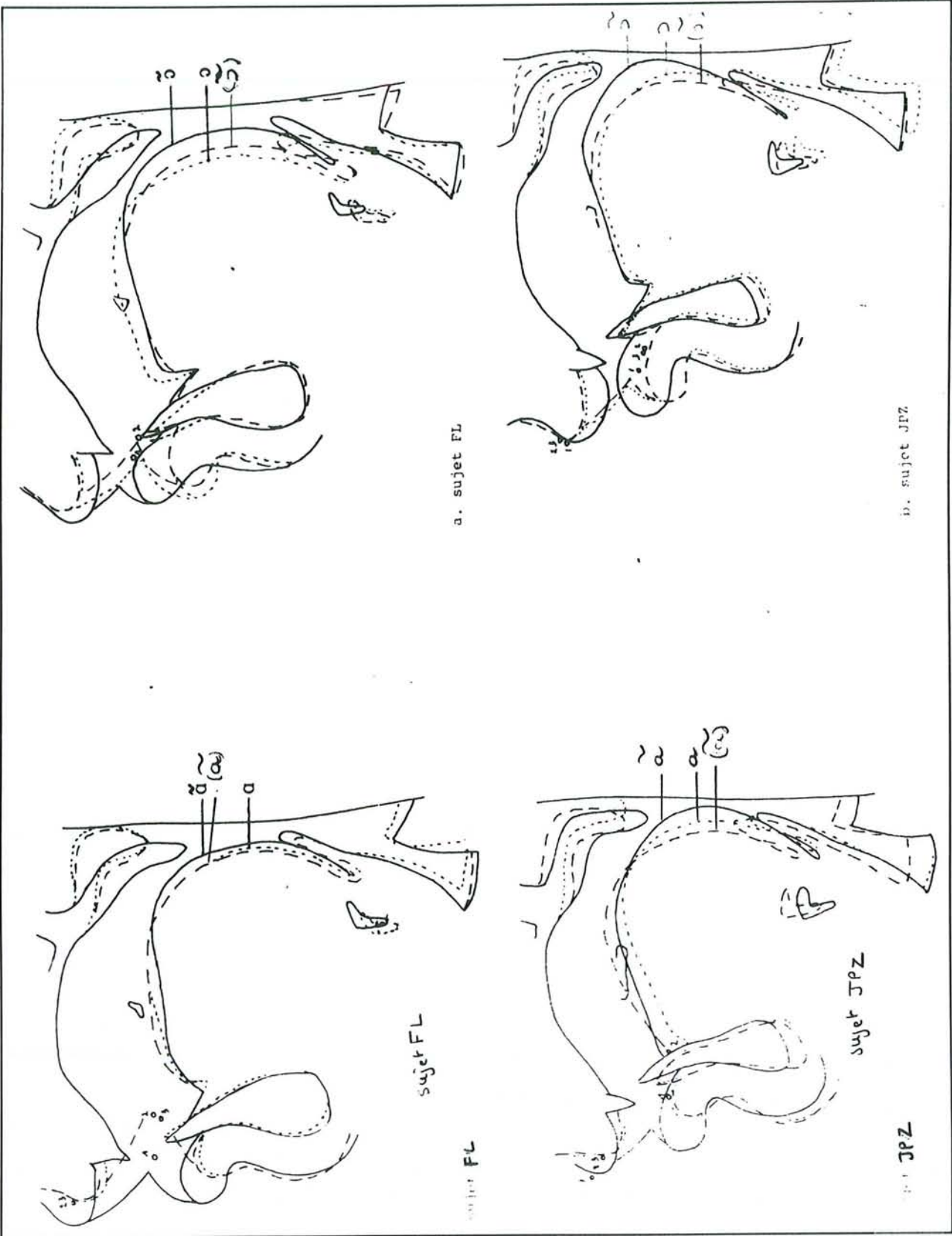


Figure A.25. Positions articutoires des voyelles [ a ], [ ã ] et [ a ] nasalisé et des voyelles [ ɔ ], [ õ ] et [ ɔ ] nasalisé d'après J.P. Zerling dans [Lonchamp 88].

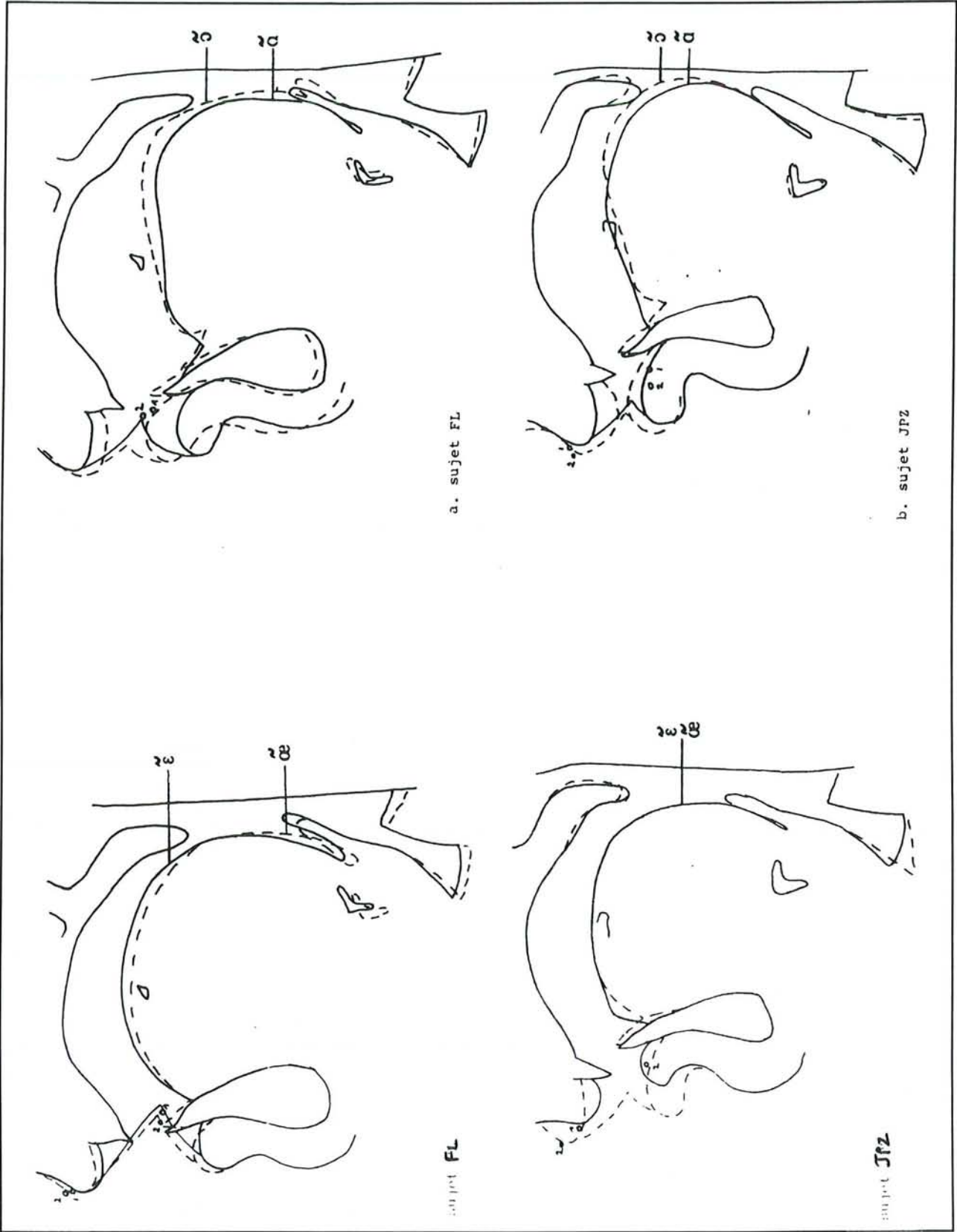


Figure A.26. Comparaisons des positions articutoires des voyelles [ ɛ̃ ] — [ œ ] et [ ɑ̃ ] — [ õ ], d'après J.P. Zerling dans [Lonchamp 88].



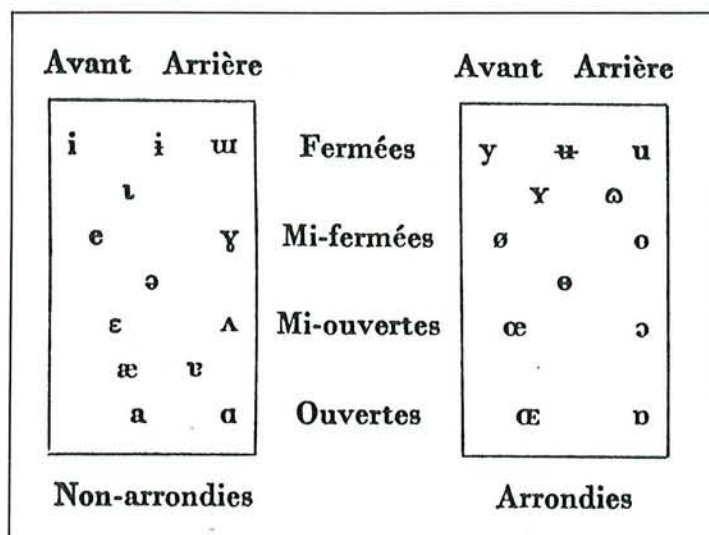


Figure A.27. Les positions articulatoires de [æ] et [ʌ] par rapport aux voyelles françaises, d'après [Duchet 86].

A part pour [ɛ̃], cette nouvelle étude de F. Lonchamp [Lonchamp 88] ne contredit pas ses premières conclusions [Lonchamp 79] : ([ɛ̃, œ̃, ɑ̃, ɔ̃]) = ([æ̃, ʌ̃, ɔ̃, ɔ̃]). Mais elle démontre que ces articulations linguales types ne sont que des tendances et qu'il semblerait que chaque locuteur adopte la stratégie du moindre effort pour adapter son conduit oral à son conduit nasal afin d'obtenir des voyelles nasales perceptivement correctes mais tout juste suffisamment opposées phonologiquement entre elles.

L'enquête réalisée par A. Martinet et H. Walter, entre 1968 et 1973 pour l'élaboration du "Dictionnaire de la prononciation française dans son usage réel" [Martinet 73] met en évidence le même phénomène. Sur dix-sept locuteurs parisiens étudiés, huit ont complètement perdu l'opposition /œ̃/ ~ /ɛ̃/ et trois, parmi les plus jeunes, ont des réalisations perceptivement très proches pour /ɑ̃/ et /ɔ̃/.

### 2.3. Le phonème /ə/

Le phonème /ə/, encore appelé "e muet", "e caduc" ou "schwa", est une des pierres d'achoppement de la phonologie française. Si, jusqu'au VII<sup>e</sup> siècle, ce phonème était toujours prononcé et correspondait à une voyelle centrale non arrondie [ə] au timbre distinct de [ɐ] et de [œ], son statut phonologique a considérablement évolué au cours des siècles suivants pour devenir, en français moderne, beaucoup plus discutable et discuté [Walter 76]. Peu de locuteurs l'opposent encore aux phonèmes /ɐ/ et /œ/ et, à de rares exceptions près comme "pelage" ~ "plage" ou "le hêtre" ~ "l'être", il peut être omis dans la prononciation d'un mot sans en changer l'identité.

Nous définirons /ə/ comme le symbole phonologique d'une voyelle associée aux graphèmes "e" ("petite"), "ai" ("faisan") ou "on" ("monsieur") qui peut-être élidée et dont la réalisation physique, lorsqu'elle est prononcée, a pour timbre [ɐ], [œ] ou un timbre intermédiaire.

### 3. Les consonnes

Les consonnes sont produites par la présence dans le conduit vocal d'un obstacle au libre passage de l'air en provenance des poumons, que cet air soit ou non modulé par la glotte.

Les consonnes se regroupent en grandes classes selon leur mode d'articulation c'est-à-dire selon la nature de l'obstacle : occlusives, nasales, fricatives, latérales, vibrantes et semi-consonnes. A l'intérieur de chaque classe, elles sont répertoriées suivant la position de l'obstacle dans le conduit vocal, cette position étant définie par deux composantes : l'élément articulant et le lieu vers lequel il articule.

La table A.2 présente cette classification. Le lecteur pourra se reporter à la figure A.11 pour retrouver les différents éléments de la cavité buccale correspondant aux termes employés.

Consonnes \ Classe articuloire		bilabiales	labio-dentale	apico-dentale	apico-alvéolaire	prédorso-alvéolaire	dorso-prépalatale	dorso-palatale	dorso-vélaire	dorso-uvulaire	glottale
occlusives	non voisées	p		t					k		ʔ
	voisées	b		d					g		
nasales		m		n				ɲ	ŋ		
fricatives	non voisées		f			s	ʃ				
	voisées		v		ʝ	z	ʒ			ʁ	
liquides	latérale			l							
	vibrante				r					R	
semi-consonnes	non arrondies							j			
	arrondies							ɥ	w		

Table A.2. Classification articuloire des consonnes.

### 3.1. Les occlusives

L'articulation des occlusives comporte quatre phases :

- la mise en place d'une occlusion complète : entre les deux lèvres pour [ p ] et [ b ], entre la pointe de la langue et la face interne des incisives supérieures pour [ t ] et [ d ], entre le dos de la langue et l'arrière du palais pour [ k ] et [ g ] ;
- la tenue de l'occlusion durant plusieurs dizaines de millisecondes pendant lesquelles l'air s'accumule derrière le barrage. Perceptivement, cette tenue correspond à un silence pour les occlusives sourdes [ p, t, k ] et à un léger murmure pour les occlusives voisées [ b, d, g ] ;
- le brusque relâchement de l'occlusion provoquant une intense perturbation acoustique de quelques millisecondes appelée communément explosion ;
- l'écartement plus ou moins rapide des articulateurs qui, lorsqu'il est lent, entraîne l'apparition d'un bruit de friction dû aux turbulences de l'air dans le passage rétréci.

Pour les occlusives françaises, le terme anglais "*burst*" englobe les bruits d'explosion et de friction.

### 3.2. Les nasales

Les quatre consonnes nasales / m, n, ɲ, ŋ / peuvent être considérées comme des occlusives voisées en ce qui concerne leur articulation orale. Cependant, lors de la tenue du barrage buccal, le voile du palais est abaissé, laissant l'air s'échapper par le nez. Ce passage continu de l'air par les fosses nasales diminue la pression derrière l'occlusion, entraîne la disparition du bruit d'explosion lors du relâchement des articulateurs et confère aux consonnes nasales une structure acoustique très différente de celle des occlusives.

Pour [ m ], [ n ], [ ɲ ], les occlusions sont respectivement identiques à celles de [ b ], [ d ], [ g ] ; celle de [ ŋ ] est réalisée entre le dos de la langue et le milieu du palais.

La consonne / ŋ / n'est pas à proprement parler un phonème du français : elle apparaît uniquement à la fin des mots empruntés à l'anglais : "*parking*" ou comme résultat d'une assimilation consonantique de nasalité<sup>1</sup>.

### 3.3. Les fricatives

Dans le cas des consonnes fricatives, encore appelées consonnes constrictives, l'obstacle au passage de l'air n'est pas total. Le rapprochement d'un articulateur mobile et d'un articulateur fixe forme un rétrécissement au niveau du lieu d'articulation qui perturbe le passage de l'air, provoquant des turbulences qui engendrent un bruit de friction. / f, s, ʃ / sont les fricatives sourdes du français, / v, z, ʒ / leurs homologues voisées.

Pour [ f ], les incisives supérieures appuient partiellement sur la lèvre inférieure.

Comme nous pouvons l'observer sur la figure A.28, [ s ] et [ ʃ ] ont des lieux d'articulation très voisins et se différencient principalement par la forme de la langue et des lèvres. Lors de l'articulation de [ s ], la constriction se situe entre la partie antérieure du dos de la langue et les alvéoles ; la pointe de la langue est abaissée et crée un canal étroit et court. Pour [ ʃ ], la constriction se place un peu plus en arrière dans la zone prépalatale ; la pointe de la langue est relevée et les lèvres projetées en avant, ce qui crée un chenal plus long que pour [ s ] [Malmberg 74].

<sup>1</sup> Les différents cas d'assimilation consonantique de nasalité sont traités dans le paragraphe V.2.2.2.



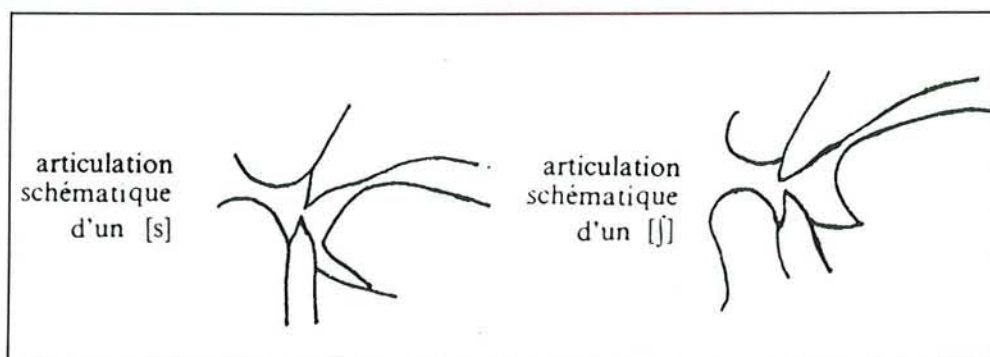


Figure A.28. Articulations schématiques d'un [ s ] et d'un [ ʃ ], d'après [Calliope 89].

### 3.4. La consonne latérale / l /

Comme les consonnes nasales, les latérales ont une certaine parenté avec les occlusives voisées puisqu'il y a contact entre deux articulateurs mais dans le cas des consonnes latérales, celui-ci ne crée pas une fermeture complète du conduit et l'air peut passer, latéralement, des deux côtés du barrage.

Le français ne possède qu'une seule consonne latérale : / l /. La pointe de la langue fait contact avec la surface intérieure des incisives supérieures ou avec les alvéoles. Selon les sujets, l'air passe de chaque côté de la langue ou, ce qui est le cas le plus fréquent, d'un seul côté sans que cela provoque une différence de perception.

### 3.5. Les différents allophones du phonème / r /

Selon les langues et les idiomes,<sup>1</sup> le graphème<sup>2</sup> " r " s'articule soit comme une consonne vibrante, [ r ] ou [ R ], soit comme une consonne constrictive, [ ʀ ] ou [ ʁ ].

[ r ] et [ R ] sont, selon les auteurs, des consonnes vibrantes, roulées ou battues. Elles proviennent des battements d'un organe élastique contre une paroi : au début de l'articulation, l'organe est en contact avec la paroi, le passage de l'air le repousse puis il revient au contact de la paroi grâce à son élasticité et le cycle recommence ...

La figure A.29 présente les deux lieux d'articulation des consonnes vibrantes. Dans le cas du [ r ], communément appelé " r roulé ", l'organe élastique est formé par la pointe de la langue qui bat contre les alvéoles du palais. Pour [ R ], encore appelé " r grasseyé ", c'est la luvette qui vibre contre le dos de la langue.

Lorsqu'au lieu de vibrer, l'organe élastique forme avec la paroi un passage étroit, ces deux consonnes se transforment en fricatives : respectivement [ ʀ ] et [ ʁ ].

Tous ces types de / r / sont intrinsèquement sonores mais peuvent de dévoiser au contact d'autres consonnes sourdes.

Dans certaines langues, ces différents sons correspondent à des phonèmes distincts. En français, ce ne sont que des variantes régionales ou individuelles d'un seul phonème / r /. L'allophone le plus répandu en français moderne, qui est celui de la prononciation

<sup>1</sup> Les idiomes sont les parlers locaux d'une langue que la localisation soit socioculturelle ou géographique.

<sup>2</sup> Un graphème est une unité distinctive de la langue écrite : " cours " ~ " court ".

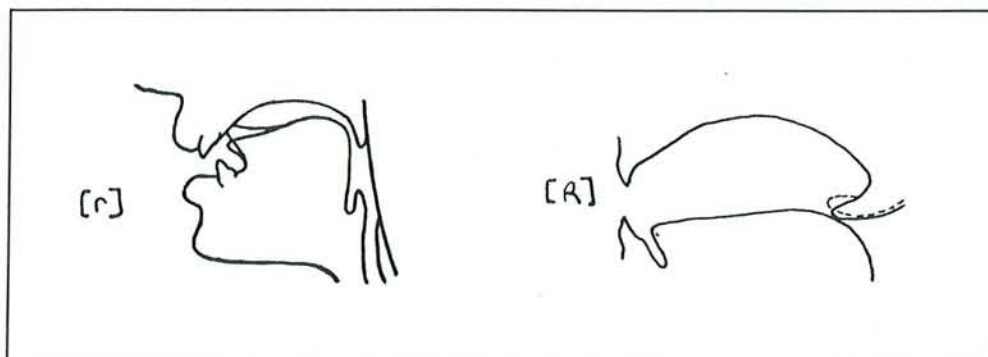


Figure A.29. Articulations schématiques de [r] et [R], d'après [Malmberg 74].

parisienne standard, est [ʀ] : consonne fricative dorso-uvulaire mais, pour des raisons de simplicité typographique, il est presque toujours noté [R], notation que nous adopterons dans ce mémoire.

### 3.6. Les semi-consonnes

Encore appelées semi-voyelles, les semi-consonnes /j, y, w/ doivent cette double appellation à leurs caractéristiques articulatoires et acoustiques qui les situent à la frontière entre les consonnes et les voyelles. En effet [j], [y] et [w] ont des articulations très voisines de [i], [y] et [u] mais légèrement plus fermées et surtout plus brèves, ce qui leur confère, acoustiquement, une structure vocalique très instable — d'où leur troisième appellation de glissantes (*glides*) — mélangée à du bruit.

Du point de vue fonctionnel, leur classement dans le groupe des consonnes se justifie par le découpage syllabique des mots. En français, à une syllabe correspond une et une seule voyelle. En conséquence, un mot comme “*pied*”, /pje/ qui est constitué intuitivement d'une seule syllabe, comporte deux consonnes /p/ et /j/. Il en est de même pour “*lui*”, /lɥi/ et “*loi*”, /lwa/.

### 3.7. Remarques

- Les consonnes /l, R, m, n, p, ɲ, j, y, w/ sont souvent réunies sous les termes de sonantes ou de consonnes à formants. Ces dénominations expriment la faculté qu'à le conduit vocal à résonner lors de leur production comme lors de celle des voyelles.
- Les sonantes ont en commun le fait d'être intrinsèquement sonores mais de pouvoir se dévoiser au contact des consonnes sourdes.
- Les latérales et les vibrantes sont quelquefois regroupées sous le terme de liquides ;
- La consonne [ʔ], ou coup de glotte, n'est pas un phonème du français. Toutefois sa présence dans la table A.2 s'explique par sa présence occasionnelle en français (cf. paragraphe II.3.6).

## 4. Conclusion

Ce chapitre nous a permis de détailler les réalisations physiques “normalisées” des phonèmes du français. Toutefois, cette norme, qui correspond à l'articulation d'un phonème prononcé isolément ne reflète pas exactement ce qui se passe dans la parole réelle. Nous avons déjà constaté que l'articulation de certains phonèmes, comme les voyelles nasales ou le / r /, n'est pas la même pour tous les locuteurs. De même, nous avons souligné le fait qu'une propriété articulatoire d'un phonème, comme la vibration des cordes vocales peut être modifiée au contact d'un autre phonème. Ces deux faits sont des manifestations d'un phénomène plus général et plus complexe, qui sera développé dans le chapitre V, la variabilité de la parole. Dans ce chapitre, nous essaierons d'exposer les différentes sources de la variabilité de la parole ainsi que leurs conséquences sur les paramètres acoustiques communément utilisés dans les études sur la parole. Nous allons donc dans le chapitre suivant introduire ces paramètres ainsi que quelques méthodes de traitement du signal permettant de les obtenir.





## CHAPITRE IV PARAMETRISATION DU SIGNAL DE PAROLE

### 1. Introduction

Le signal de parole est un signal continu, d'énergie finie, non stationnaire, qui renferme des composantes complexes et variables selon les sons émis : quasi périodiques pour les sons voisés, aléatoires pour les sons fricatifs et "impulsionnelles" pour les sons occlusifs. Les informations contenues dans ce signal le sont de manière redondante et codent aussi bien le message utile que des caractéristiques du locuteur qui a émis ce message.

Le traitement automatique de la parole nécessite une paramétrisation de ce signal vocal afin, d'une part, d'en obtenir une représentation plus concise et moins redondante, et, d'autre part, d'en extraire des paramètres significatifs spécifiques de l'application choisie : reconnaissance de la parole, du locuteur ou de la langue, transmission ou synthèse de la parole.

Pour cela, les méthodes d'analyse du signal de parole se fondent sur des techniques traditionnelles de traitement du signal, comme la transformée de Fourier ou la transformée en  $z$ , autour desquelles viennent se greffer des méthodes plus élaborées prenant en compte des modèles de production ou de perception de la parole.

Nous nous proposons d'introduire dans ce chapitre quelques-unes de ces méthodes d'analyse conduisant notamment à l'extraction des paramètres que nous avons utilisés dans notre étude, les formants.

### 2. Echantillonnage du signal de parole

L'échantillonnage et la quantification, que l'on regroupe habituellement sous le terme de numérisation, sont les deux premières opérations subies par le signal de parole en vue de son traitement automatique.

L'échantillonnage consiste à transformer un signal continu  $s(t)$  en un signal discret  $s^* = (\dots, s((n-1)T), s(nT), s((n+1)T), \dots)$ .

$T$  est la période à laquelle est échantillonné le signal analogique. La fréquence d'échantillonnage doit respecter le théorème de Shannon et être supérieure ou égale au double de la plus grande fréquence contenue dans le signal analogique, ce qui suppose un préfiltrage dans le cas de la parole [Rabiner 78].

Afin d'alléger les équations, on a l'habitude d'adopter la notation :  $s(n) = s(nT)$

La quantification consiste à coder chacune des valeurs échantillonnées sur un certain nombre de digits binaires, en général douze ou seize.

### 3. L'analyse spectrale de la parole

#### 3.1. Introduction

Si l'on considère les diverses répétitions d'un son prononcées par un même locuteur, celles-ci se ressemblent plus dans le domaine fréquentiel (énergie, fréquence) que dans le domaine temporel (amplitude, temps). Plus généralement, les caractéristiques du signal vocal qui sont le plus pertinentes au sens de la production et de la perception de la parole sont d'ordre fréquentiel. De telles constatations ont privilégié depuis longtemps l'analyse spectrale de la parole.

De plus, l'analyse spectrale s'applique particulièrement bien à l'hypothèse fondamentale de la modélisation du processus de production de la parole. Celle-ci suppose que le signal de parole est la réponse d'un résonateur, le conduit oral éventuellement couplé au conduit nasal, à une source d'excitation, et que la source et le résonateur sont indépendants. Le spectre du signal de parole s'écrit alors comme le produit du spectre de la source par la transformée de Fourier de la réponse impulsionnelle du résonateur.

Le schéma de la figure A.30 montre l'application de cette décomposition spectrale à la production des sons voisés.

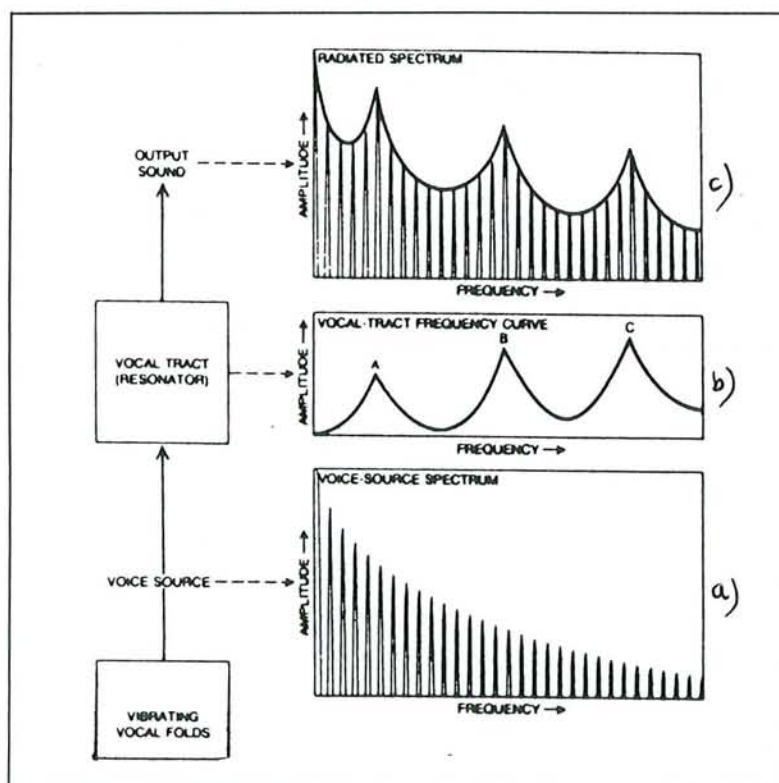


Figure A.30. Décomposition spectrale de la production des sons voisés d'après [Fant 60].



### 3.2. Analyse spectrale à court terme

Considérons l'onde de débit d'air produite par la vibration des cordes vocales lors de la production des sons voisés. C'est une onde quasi périodique, de quasi-période  $1/F_0$ , ayant une forme triangulaire asymétrique dont la phase croissante, correspondant à l'écartement des cordes vocales, est plus longue que la phase décroissante, correspondant au rapprochement des cordes vocales (cf. figures A.7 et A.8).

D'après le théorème de Fourier [O'Shaughnessy 87], cette onde se décompose en une somme de sinusoïdes dont les fréquences sont des multiples entiers de la fréquence de l'onde initiale. Le premier terme de la somme est appelé fondamental ( $F_0$ ) ou parfois premier harmonique et a pour fréquence  $F_0$ . Les termes suivants constituent les harmoniques et ont des fréquences qui sont des multiples de  $F_0$ .

Cette décomposition de l'onde glottale en fonctions sinusoïdales, peu exploitable dans le domaine temporel, a la particularité de se représenter sous une forme simple dans le domaine fréquentiel qui est, comme le montre la figure A.30-a, un spectre de fréquence constitué de raies. Ce spectre de raies harmoniques a une pente moyenne de  $-12$  dB par octave<sup>1</sup> à partir de 200–300 Hz (pente réelle de  $-10$  dB à  $-18$  dB). Cette pente spectrale varie avec l'effort vocal, l'atténuation des fréquences élevées étant plus importante pour les voix faibles que pour les voix fortes [Lonchamp 87b] [O'Shaughnessy 87].

La propagation de l'onde glottale dans le conduit vocal engendre une onde complexe qui reste quasi périodique mais dont les harmoniques ont été plus ou moins amplifiés selon le voisinage de leurs fréquences par rapport aux fréquences de résonance du conduit vocal. La figure A.30-b schématise la fonction de transfert du conduit vocal qui visualise notamment les fréquences de résonance de ce conduit.

Le spectre de la figure A.30-c est celui du signal de parole à la sortie du conduit vocal et correspond au produit des deux précédents. Il possède une pente spectrale moyenne de  $-6$  dB par octave qui provient du rehaussement de 6 dB par octave de la pente du spectre de l'onde glottale sous l'effet du rayonnement aux lèvres. Ce spectre instantané, qui est en fait le seul des trois spectres dont on peut disposer directement, est soit obtenu à partir d'un spectrographe analogique soit calculé à partir du signal numérisé.

Les spectres instantanés numériques comme celui présenté sur la figure A.31 sont issus d'un calcul de transformée de Fourier discrète rapide (FFT) sur  $N$  échantillons du signal de parole :

$$S(k) = \sum_{n=0}^{N-1} s(n) \cdot e^{-j2\pi k \frac{n}{N}} \quad \text{où } k \text{ varie de } 0 \text{ à } N-1.$$

Auparavant, le signal est préaccentué afin de compenser la pente négative du spectre des sons voisés et multiplié par une fenêtre de taille fixe (cf. figure A.32), le plus souvent une fenêtre de Hamming, afin d'atténuer les discontinuités aux frontières [Rabiner 78] [Calliope 89].

Le passage du signal temporel au spectre de fréquence entraîne la perte de l'information de phase mais cette dernière n'est généralement pas considérée dans les études sur la parole, en particulier, parce qu'elle n'est pas perceptivement pertinente.

<sup>1</sup> Une octave correspond à un doublement de fréquence.

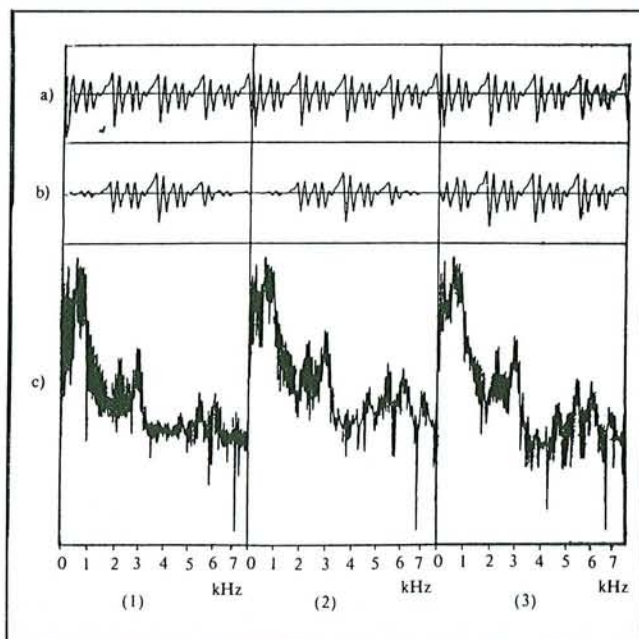


Figure A.31. Analyse par FFT de 512 échantillons d'un [ a ] ;  
 (a) : signal temporel ; (b) : signal multiplié par une fenêtre de Hamming (1), Hanning (2),  
 trapézoïdale (3) ; (c) : le spectre d'amplitude pour chaque fenêtre ; d'après [Calliope 89].

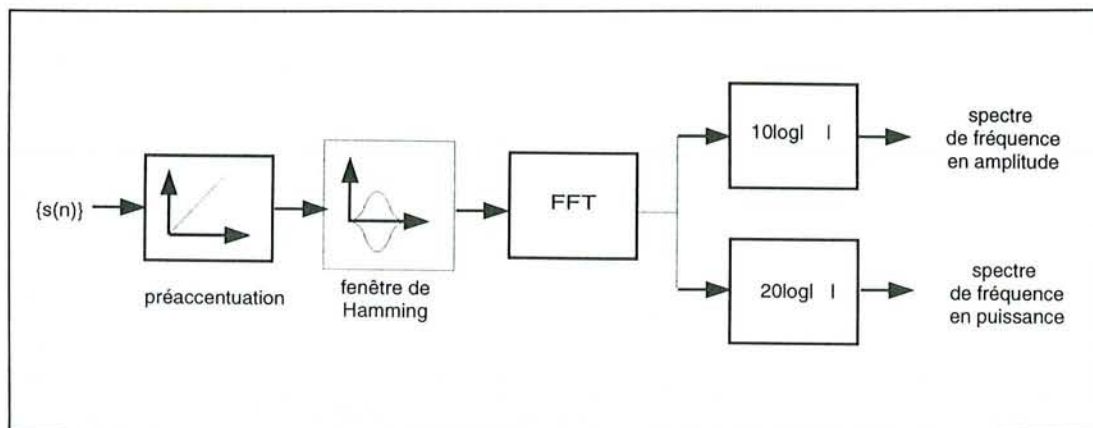


Figure A.32. Les différentes étapes de calcul d'un spectre instantané, d'après [Calliope 89].

L'évolution temporelle des spectres instantanés engendre une forme tridimensionnelle (énergie, fréquence, temps) appelée spectrogramme. La troisième dimension, l'énergie, est le plus souvent représentée par la noirceur du tracé.



### 3.3. Les spectrogrammes analogiques et numériques

En 1946, quelques années après la réalisation manuelle du premier spectrogramme de parole, les premiers spectrographes analogiques furent annoncés par les "Bell Telephone Laboratories" [Bolt 70]. Balayant la gamme de fréquences de 50 à 16 000 Hz, les spectrographes analogiques emploient soit un filtre de 300 Hz de largeur de bande ("spectrogramme à large bande") pour mettre en évidence la structure formantique des sons et les changements de forme du conduit vocal, soit un filtre de 45 Hz de largeur de bande ("spectrogrammes à bande étroite") afin de suivre au mieux l'évolution des harmoniques des sons voisés.

Les spectrogrammes numériques sont obtenus à partir des spectres instantanés numériques en déplaçant la fenêtre de calcul sur le signal de parole. Leur largeur de bande est une fonction de la taille de cette fenêtre (cf. figure A.33). Par rapport à leurs prédécesseurs, les spectrogrammes numériques :

- possèdent une meilleure dynamique, environ 60 dB au lieu de 30 dB,
- sont complètement paramétrables, aussi bien en ce qui concerne l'échelle des fréquences qui peut être linéaire ou de type psychoacoustique, en Mels ou en Barks [Zwicker 80] [Zwicker 81], qu'au niveau de la largeur de bande, de la dynamique ou du type de fenêtre,
- peuvent être visualisés en couleur sur les stations de travail les plus récentes.

Malgré cela, les connaissances accumulées par les phonéticiens sur les spectrogrammes analogiques font que les spectrogrammes numériques essaient de recopier les caractéristiques de leurs prédécesseurs notamment lorsqu'ils sont utilisés pour l'étiquetage manuel du signal de parole ou dans les expériences de lecture de spectrogramme.

Les spectrogrammes à large bande ont une meilleure résolution temporelle que les spectrogrammes à bande étroite et laissent apparaître sur les tracés une alternance de raies sombres et claires mettant en évidence la période de vibration des cordes vocales ainsi que la répartition de l'énergie pendant cette période.

La figure A.33 fournit un exemple de chacun des grands types de spectrogrammes.

### 3.4. Les formants

Les formants des sons voisés sont définis comme les différentes zones du spectre du signal de parole où les harmoniques sont les plus intenses. Ils correspondent donc aux maxima de l'enveloppe du spectre instantané (cf. figure A.31) ou aux bandes noires plus ou moins horizontales présentes sur les spectrogrammes à large bande (cf. figure A.33). Les fréquences centrales des formants sont très voisines des fréquences de résonance du conduit vocal lorsque les largeurs de bande de celles-ci sont faibles, ce qui est le cas en parole [Lonchamp 88].

En première approximation, les emplacements fréquentiels de ces maxima ne dépendent que de la forme et de l'état du conduit vocal et aucunement de l'onde glottale mais la détermination de ces emplacements peut être rendue difficile voire impossible par un espacement trop important des harmoniques dû à une fréquence fondamentale trop élevée.

Les formants des sons voisés strictement oraux sont numérotés de F1 à Fn suivant l'ordre croissant de leurs fréquences ; dans le cas des sons nasaux, la numérotation est différente puisqu'elle tient compte des formants dus au couplage nasal. Si on se place dans le domaine de validité du modèle acoustique sous-jacent (propagation unidimensionnelle de l'onde glottale dans le conduit vocal), c'est-à-dire pour des fréquences inférieures à 5000 Hz, une voyelle orale



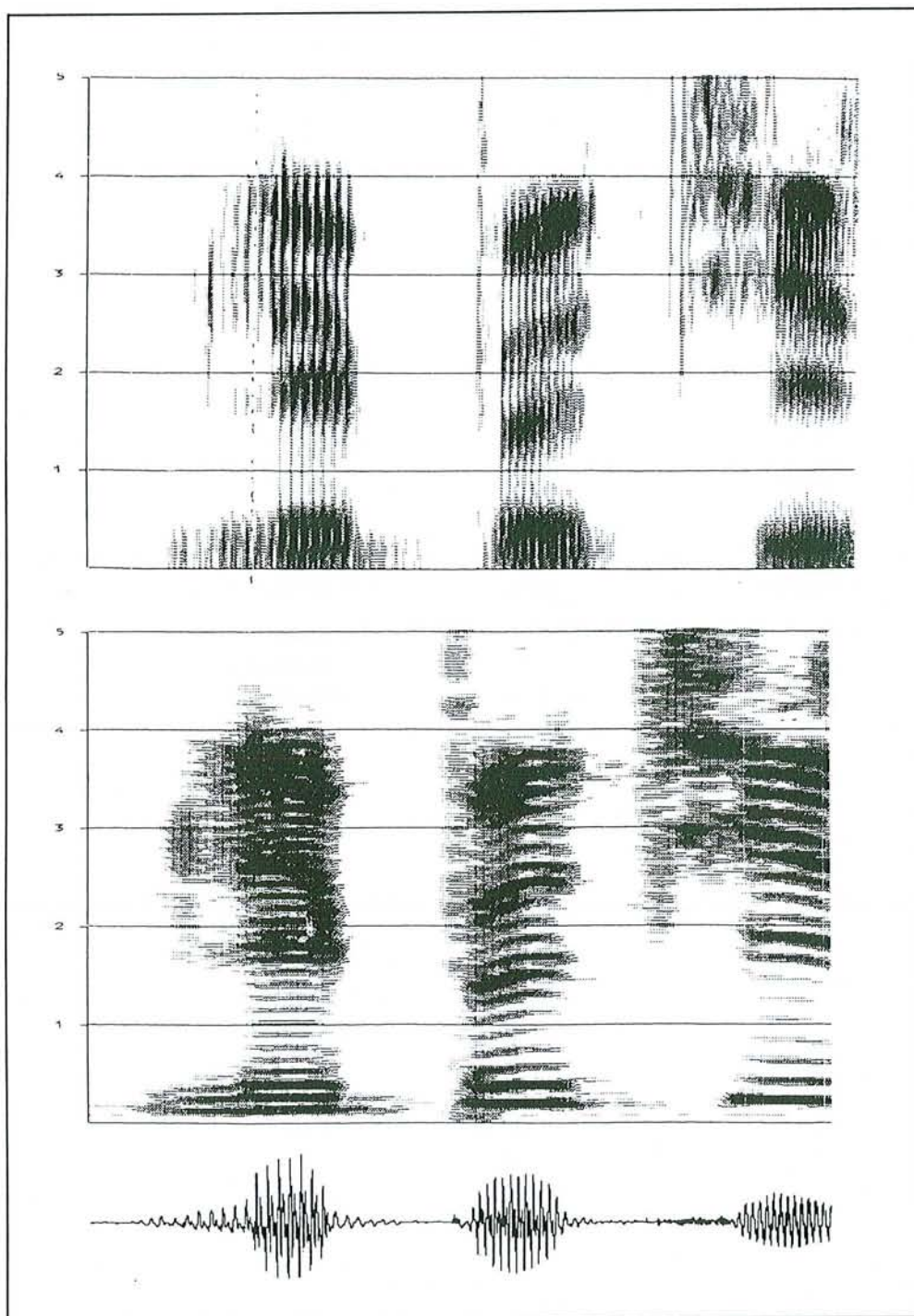


Figure A.33. Spectrogrammes numériques : (a) signal temporel, (b) spectrogramme à bande étroite, (c) spectrogramme à large bande, d'après [Calliope 89].

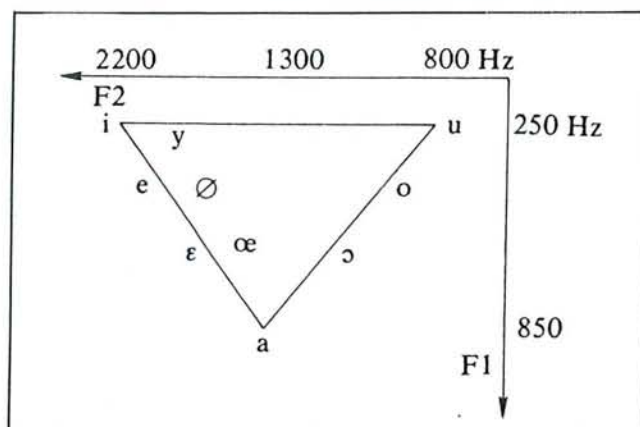


Figure A.34. Voyelles orales du français dans le plan ( $F_1$ ,  $F_2$ ), d'après [Calliope 89].

d'un locuteur masculin possède cinq formants. Cependant, des expériences de synthèse de la parole ont montré que les trois premiers formants suffisaient à une bonne perception du timbre de chaque voyelle [Calliope 89]. Par abus de langage, les fréquences formantiques sont souvent identifiées aux formants eux-mêmes. Nous avons adopté comme convention de représenter les formants par  $F_n$  et les fréquences formantiques par  $F_n$ , de la même façon que nous avons représenté le fondamental par  $F_0$  et la fréquence fondamentale par  $F_0$ .

On a coutume de représenter les voyelles orales dans le plan ( $F_1$ ,  $F_2$ ). Les voyelles [i], [a], [u] se situent aux sommets d'un triangle acoustique (cf. figure A.34) qui peut se mettre en rapport avec le trapèze articuloire présenté au chapitre III.

La similarité entre ces deux représentations des voyelles orales permet d'interpréter une augmentation de  $F_1$  comme une ouverture de la cavité buccale et une augmentation de  $F_2$  comme une antériorisation de l'articulation sans, toutefois, autoriser une affectation simple et plus directe d'un formant à une partie du conduit oral [Lonchamp 87b].

Un pic supplémentaire peut apparaître vers 200–300 Hz sur certains spectres de voyelles ouvertes ([ε], [a], [ɑ], [ɔ]) ou de voyelles nasales, produites avec une voix faible. Malgré l'appellation de formant glottal, ce pic ne provient pas de la fonction de transfert du conduit vocal mais directement de la source glottale. Des simulations réalisées à partir du modèle théorique d'onde glottale de G. Fant ont montré que la fréquence du formant glottal est approximativement égale à  $1/2 tr$ ,  $tr$  étant la durée de la phase strictement croissante d'une période de l'onde glottale [Lonchamp 88].

La figure A.35 présente le spectre d'un [a] synthétisé de manière à ce qu'il présente un formant glottal.

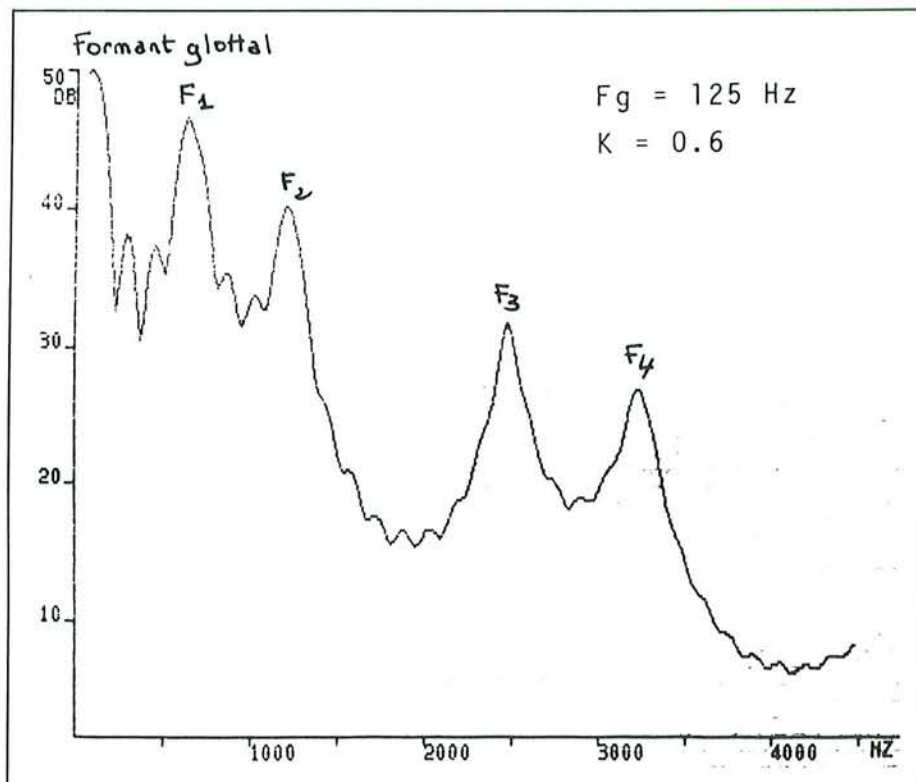


Figure A.35. Spectre d'un [a] synthétique présentant un formant glottal dont la fréquence est 125 Hz, d'après [Lonchamp 88].

## 4. L'analyse par prédiction linéaire

### 4.1. Introduction

L'analyse par prédiction linéaire (ou analyse LPC comme "Linear Predictive Coding") est un outil souvent utilisé dans le traitement automatique de la parole. Sa popularité est due essentiellement à la relative simplicité des calculs qu'elle nécessite ainsi qu'à la paramétrisation compacte du signal de parole qu'elle fournit.

Sous réserve de quelques hypothèses simplificatrices, l'analyse par prédiction linéaire permet de calculer les paramètres d'un filtre linéaire modélisant le processus de production d'un segment de parole. Elle fournit simultanément dix à vingt paramètres codant efficacement ce segment de parole en vue de sa transmission, son stockage ou sa synthèse. En outre, l'analyse LPC permet d'approcher d'autres paramètres de la parole, comme la fonction d'aire du conduit vocal, les formants et la fréquence fondamentale [Markel 76].



## 4.2. Le modèle linéaire de production de la parole

De nombreux modèles mathématiques ont été proposés pour tenter de décrire le processus de production de la parole. Malheureusement, il est impossible d'établir un modèle idéal tenant compte de toutes les caractéristiques de ce processus comme :

- les différentes sources d'excitation du conduit vocal,
- le rayonnement aux lèvres,
- la variabilité temporelle de la forme du conduit vocal,
- le couplage entre le conduit vocal et le conduit nasal,
- le couplage entre les cavités subglottiques et supraglottiques,
- les pertes dues aux échanges thermiques et à la viscosité des parois du conduit, etc.

Toutefois, en se limitant à de courts intervalles de temps, une dizaine de millisecondes, et en faisant un certain nombre d'hypothèses raisonnables sur les caractéristiques physiques à modéliser, il est possible de développer des modèles de production de la parole linéaires et invariants dans le temps, comme celui de G. Fant [Fant 60].

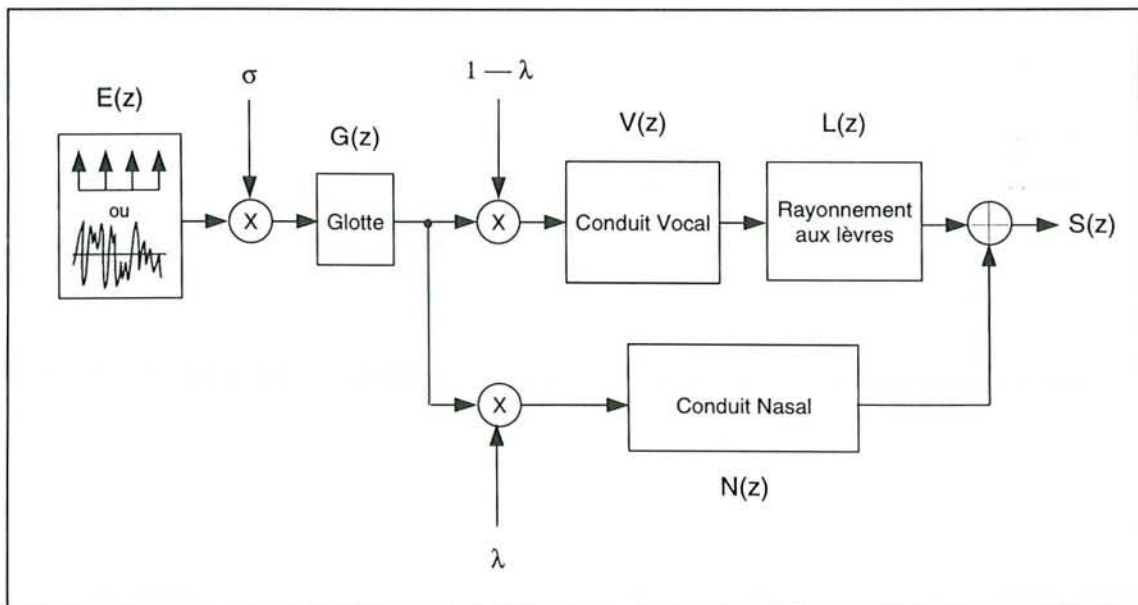


Figure A.36. Modèle linéaire de production de la parole.

Ce modèle, présenté sur la figure A.36, suppose que l'onde glottale est modélisée par un filtre passe-bas  $G(z)$  à deux pôles et de fréquence de coupure 100 Hz, excité par une source qui est soit une suite d'impulsions espacées de  $1/F_0$  pour les sons voisés, soit un bruit blanc pour les sons non voisés. La sortie de ce filtre passe-bas est connectée à deux filtres linéaires  $V(z)$  et  $N(z)$ . Le premier est un filtre tout-pôle modélisant le conduit oral et qui peut se décomposer en une cascade d'un petit nombre de résonateurs à deux pôles complexes conjugués. Le second est un modèle du conduit nasal comportant des pôles et des zéros. Pour terminer, un filtre tout-zéro  $L(z)$  modélise le rayonnement aux lèvres de l'onde acoustique.

Les fonctions de transfert des différents filtres sont les suivantes, au gain près par souci de simplification d'écriture :

- $G(z) = \frac{1}{(1 - e^{-2\pi f_g T} z^{-1})^2}$

où  $T$  est la période d'échantillonnage et  $f_g$  la fréquence de coupure du filtre glottal.

- $V(z) = \frac{1}{\prod_{i=1}^K [1 - 2e^{-\pi B_i T} \cos(2\pi F_i T) z^{-1} + e^{-2\pi B_i T} z^{-2}]}$

où  $T$  est la période d'échantillonnage,  $K$  est le nombre de résonateurs à deux pôles conjugués mis en cascade,  $F_i$  la fréquence de résonance du  $i^{\text{e}}$  résonateur et  $B_i$  la largeur de bande associée.

- $N(z) = \frac{\prod_{i=1}^R [1 - 2e^{-\pi B'_{N_i} T} \cos(2\pi F'_{N_i} T) z^{-1} + e^{-2\pi B'_{N_i} T} z^{-2}]}{\prod_{i=1}^R [1 - 2e^{-\pi B_{N_i} T} \cos(2\pi F_{N_i} T) z^{-1} + e^{-2\pi B_{N_i} T} z^{-2}]}$

où  $F_{N_i}$  et  $B_{N_i}$  sont respectivement la fréquence et la largeur de bande de la résonance du  $i^{\text{e}}$  résonateur nasal, alors que  $F'_{N_i}$  et  $B'_{N_i}$  sont respectivement la fréquence et la largeur de bande de l'antirésonance du  $i^{\text{e}}$  résonateur nasal.

- $L(z) = 1 - z^{-1}$

La transformée en  $z$  du signal de parole résultant est donnée par :

$$S(z) = G(z) \left[ (1 - \lambda) V(z) L(z) + \lambda N(z) \right] E(z)$$

où  $E(z)$  est la transformée en  $z$  de l'excitation du modèle.

Si l'on ne considère que les sons oraux ( $\lambda = 0$ ), l'équation précédente devient :

$$S(z) = G(z) V(z) L(z) E(z)$$

Si, de plus, on suppose que  $1 - z^{-1}$  est approximativement égal à  $1 - e^{-2\pi f_g T} z^{-1}$  ce qui est vrai si  $2\pi f_g T$  est grand devant 1, alors  $S(z)$  peut s'écrire :

$$S(z) = \frac{1}{A(z)} E(z) \quad \text{avec} \quad A(z) = \sum_{i=0}^p a_i z^{-i} \quad p = 2K + 1 \quad \text{et} \quad a_0 = 1$$

### 4.3. Le modèle d'analyse par prédiction linéaire

Le modèle qui vient d'être présenté au paragraphe précédent est celui proposé par G. Fant dès les années 60. L'inverse de la fonction de transfert du modèle,  $A(z)$ , s'exprime simplement sous forme d'un polynôme de degré  $p$  en  $z^{-1}$ .

Nous reprenons cette expression en donnant un caractère de généralité à  $p$  pour exprimer la fonction habituellement retenue en analyse par prédiction linéaire.

En effet, exprimée dans le domaine temporel, l'écriture :

$$A(z) = \sum_{i=0}^p a_i z^{-i} \quad \text{équivalent à} \quad e(n) = \sum_{i=0}^p a_i s(n-i)$$

où  $e(n)$  et  $s(n)$  désignent respectivement l'entrée et la sortie du modèle à l'instant noté  $n$ .

Cette dernière expression s'écrit :

$$e(n) = s(n) - \sum_{i=1}^p -a_i s(n-i) = s(n) - \hat{s}(n)$$

L'entrée  $e(n)$  du modèle linéaire peut donc être interprétée comme l'écart entre l'échantillon  $s(n)$  du signal de sortie et son estimation  $\hat{s}(n)$  obtenue à partir des  $p$  échantillons qui le précèdent. La détermination des  $p$  coefficients  $a_i$  permet de définir le modèle d'analyse  $1/A(z)$  d'un court segment de parole. Ces mêmes paramètres permettent de reconstituer le segment de parole en utilisant le modèle de synthèse défini par  $A(z)$ .

#### 4.4. Détermination des coefficients de prédiction $a_i$

On détermine les coefficients  $a_i$  en minimisant selon le critère des moindres carrés les erreurs de prédiction calculées pour tous les échantillons d'un segment de parole, c'est-à-dire en minimisant la quantité :

$$\varepsilon = \sum_{n=n_1}^{n_2} e^2(n) = \sum_{i=0}^p \sum_{j=0}^p a_i \left[ \sum_{n=n_1}^{n_2} s(n-i)s(n-j) \right] a_j$$

Posons :

$$C_{ij} = \sum_{n=n_1}^{n_2} s(n-i)s(n-j)$$

Minimiser  $\varepsilon$  revient à annuler les dérivées partielles  $\frac{\partial \varepsilon}{\partial a_k}$  pour  $k = 1, \dots, p$ , ce qui conduit à chercher les solutions du système linéaire :

$$\sum_{i=1}^p a_i C_{ij} = -C_{0j}$$

dans lequel les coefficients  $a_i$ ,  $i$  variant de 1 à  $p$ , sont les inconnues.

Ce système peut être résolu par un algorithme classique de résolution d'un système de  $p$  équations linéaires à  $p$  inconnues mais de nombreuses études ont conduit à l'élaboration d'algorithmes plus performants prenant en compte les caractéristiques de la matrice des coefficients  $C_{ij}$ . Ces algorithmes se répartissent en grandes classes de résolution dont les deux principales diffèrent par les valeurs choisies pour l'intervalle  $[n_1, n_2]$ , la méthode de covariance et la méthode d'autocorrélation [Markel 76].



#### 4.5. La méthode d'autocorrélation

Dans ce cas, le calcul des coefficients  $C_{ij}$  est effectué sur l'intervalle théorique  $] -\infty, +\infty[$  :

$$C_{ij} = \sum_{n=-\infty}^{+\infty} s(n-i)s(n-j) = \sum_{n=-\infty}^{+\infty} s(n)s(n+|i-j|) = r(|i-j|)$$

Les coefficients d'autocorrélation  $C_{ij}$  ne dépendent plus que de la différence des indices  $i$  et  $j$ .

Etant donné que seuls sont connus les  $N$  échantillons  $(s(0), s(1), \dots, s(N-1))$  de la fenêtre de parole étudiée, on a :

$$s(n) = 0 \quad \text{pour} \quad n < 0 \quad \text{et} \quad n \geq N$$

et le système à résoudre devient :

$$\sum_{i=1}^p a_i r(|i-j|) = -r(j) \quad \text{avec} \quad r(l) = \sum_{n=0}^{N-1-l} s(n)s(n+l) \quad l \geq 0$$

De nombreux algorithmes récursifs et performants comme ceux de Levinson, Robinson et Durbin ont été conçus pour résoudre ce dernier système [Rabiner 78] [Markel 76].

#### 4.6. La méthode de covariance

Dans ce cas, le calcul des coefficients  $C_{ij}$  est effectué sur l'intervalle  $[p, N-1]$ . Le système linéaire se réécrit donc simplement :

$$\sum_{i=1}^p a_i C_{ij} = -C_{0j} \quad \text{avec} \quad C_{ij} = \sum_{n=p}^{N-1} s(n-i)s(n-j)$$

Il existe également pour cette méthode des algorithmes de résolution performants, comme la décomposition de Cholesky [Rabiner 78], même si la matrice des  $C_{ij}$  possède moins de propriétés que celle de la méthode d'autocorrélation.

#### 4.7. Analyse spectrale issue de l'analyse LPC

L'analyse LPC par autocorrélation est largement utilisée dans l'analyse spectrale de la parole car le spectre réel FFT du signal de parole,  $|S(f)|$ , peut être approché par :

$$\left| \frac{\sigma}{A(e^{j2\pi f})} \right| \quad \text{où } f \text{ varie de } 0 \text{ à } F_e/2 \text{ et où } \sigma^2 = \varepsilon.$$

Comme nous pouvons le constater sur la figure A.37, le spectre obtenu est un spectre lissé qui ne possède plus la structure harmonique due à l'onde glottale du spectre réel.

L'ordre de prédiction  $p$  contrôle le degré de lissage. Néanmoins, le spectre LPC suit mieux le spectre FFT au niveau des "pics" qu'au niveau des "vallées" et il modélise d'autant mieux ces pics que leur amplitude est plus grande [Rabiner 78]. Par conséquent, lorsque la méthode d'autocorrélation est utilisée pour calculer le spectre LPC, le signal de parole doit être

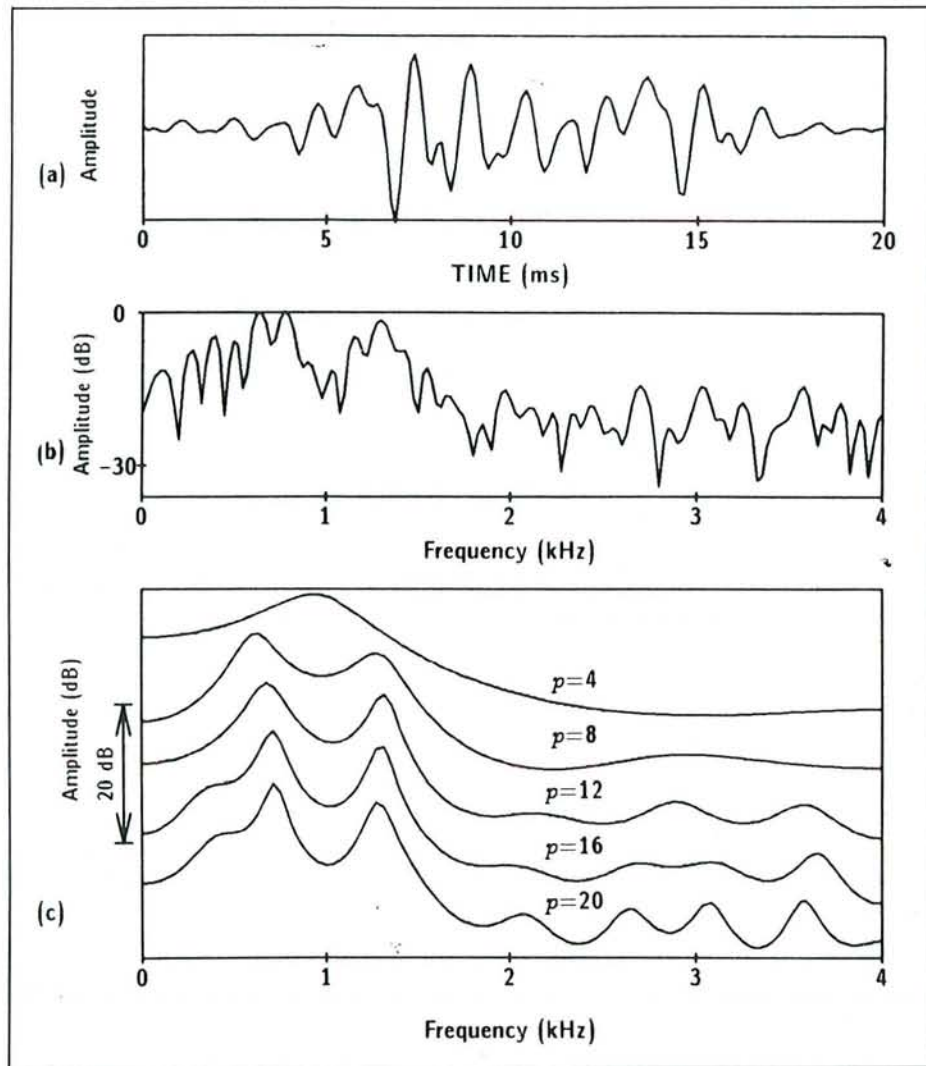


Figure A.37. Un exemple de spectres LPC obtenus à partir de 160 échantillons de la voyelle anglaise [ **ɑ** ] échantillonnée à 8 kHz ; (a) le signal temporel, (b) le spectre FFT, (c) les spectres LPC pour différentes valeurs de l'ordre de prédiction ; d'après [O'Shaughnessy 87].

préaccentué afin de modéliser correctement les formants d'ordre supérieur dont l'amplitude est moins élevée en raison de la pente du spectre de l'onde glottale.

Lorsque l'ordre de prédiction est bien choisi, le spectre LPC estime donc correctement la structure formantique des voyelles orales.

#### 4.8. Stabilité du modèle

La méthode d'autocorrélation est la méthode la plus utilisée pour déterminer les coefficients LPC car le filtre  $1/A(z)$  obtenu est théoriquement stable.

En revanche, la méthode de covariance, qui est plus précise que la méthode d'autocorrélation, notamment lorsque  $N$  est petit, peut engendrer un filtre instable et le test de cette instabilité n'est pas simple.

#### 4.9. Ordre de l'analyse LPC

Le choix de l'ordre du modèle de prédiction linéaire doit résulter d'un compromis entre précision, espace mémoire et temps de calcul. Toutefois sa valeur minimale dépend de la fréquence d'échantillonnage du signal.

La théorie montre que pour représenter correctement le conduit vocal (fréquences de résonance), la mémoire du modèle  $A(z)$  doit être égale à deux fois le temps mis par l'onde acoustique pour traverser le conduit de la glotte aux lèvres [Markel 76], ce qui conduit à :

$$p = \frac{2L}{c}$$
 où  $L$  est la longueur du conduit vocal et  $c$  la vitesse de propagation des sons dans l'air.

Pour un conduit vocal de 17 cm de longueur et une vitesse de 34 cm/ms, cette mémoire est donc de 1 ms ce qui correspond exactement à  $n$  échantillons de parole échantillonnée à une fréquence de  $n$  kHz. Il en découle une relation directe entre la fréquence d'échantillonnage et l'ordre de l'analyse LPC :

$$p = n \quad \text{avec} \quad F_e = n \text{ kHz}.$$

Mais le modèle doit pouvoir aussi modéliser la forme générale du spectre due à l'onde glottale et au rayonnement aux lèvres ainsi que de possibles zéros lors de la production de sons non voisés ou de voyelles nasalisées. Il est donc nécessaire d'augmenter le nombre de coefficients LPC de 2 à 4 unités selon le résultat souhaité [O'Shaughnessy 87].

#### 4.10. La fenêtre d'analyse

Les trois facteurs qui interviennent dans le choix de la fenêtre d'analyse sont sa taille, c'est-à-dire le nombre d'échantillons pris en compte, sa position par rapport à la période de vibration des cordes vocales et l'utilisation d'une multiplication par une fonction de type Hamming. Ces paramètres de l'analyse LPC varient avec la méthode utilisée.

La méthode d'autocorrélation requiert pour l'analyse des sons voisés à la fois un intervalle d'analyse recouvrant quelques périodes de vibrations des cordes vocales (100 à 400 échantillons) et l'emploi d'une fenêtre comme celle de Hamming. Ce fenêtrage est nécessaire pour éliminer les distorsions dues aux discontinuités aux frontières mais il impose une fenêtre d'analyse plus grande pour réduire l'imprécision du modèle provoquée par la pondération des échantillons [Rabiner 78]. Dans ce cas, la position de la fenêtre par rapport au début de la période de l'onde glottale peut-être quelconque ("analyse pitch-asynchrone").

Comme nous l'avons vu au paragraphe 4.7, il faut faire précéder l'utilisation d'une fenêtre de Hamming par une préaccentuation du signal de parole lorsqu'on désire diminuer l'influence de l'onde glottale et du rayonnement aux lèvres.

La méthode de covariance est une méthode qui peut être utilisée localement puisque la taille minimale requise pour l'intervalle d'analyse est de  $2p$  échantillons, ce qui est donc bien inférieur à la période de l'onde glottale. En contrepartie, cette méthode ne fournit des résultats satisfaisants que si les coefficients de prédiction sont estimés pendant la fermeture de la glotte. Le positionnement de la fenêtre d'analyse par rapport au début de la période est donc un élément crucial de l'analyse et nécessite l'utilisation d'un bon détecteur de fondamental ("analyse pitch-synchrone"). Ces difficultés limitent l'emploi de la méthode de covariance localement à la



prédiction de segments de parole très variables. Dans ce cas, la convolution par une fenêtre de Hamming est déconseillée [Markel 76].

Lorsque la méthode de covariance est utilisée sur plusieurs périodes, elle donne des résultats similaires à la méthode d'autocorrélation mais coûte plus cher en temps de calcul.

#### 4.11. Estimation des formants à partir de l'analyse LPC

L'analyse par prédiction linéaire permet d'estimer les formants des sons voisés soit directement par factorisation du polynôme  $A(z)$  soit en recherchant les pics du spectre LPC. Quelle que soit la solution choisie, le résultat obtenu est un ensemble de pôles ou de pics qu'il faut ensuite affecter aux formants, ce qui constitue l'opération la plus délicate.

La première solution garantit de trouver toutes les racines candidates avec une valeur précise de leur fréquence mais elle consomme plus de temps de calcul. Pour un ordre de prédiction  $p$ , on obtient par cette méthode au maximum  $p/2$  couples de pôles conjugués. Pour chacun de ces couples, on élimine le pôle comportant une partie imaginaire négative. Parmi les pôles restants, il faut ensuite éliminer ceux qui ne correspondent pas à des formants et qui modélisent la forme générale du spectre. En général, ces derniers ont des largeurs de bande très grandes comparées à celles des formants. En revanche, il a été noté une sensibilité de la largeur de bande des pôles à la taille et à la position de la fenêtre d'analyse ainsi qu'à la méthode d'analyse [Rabiner 78].

Le principal inconvénient de la seconde méthode est de ne pas pouvoir détecter deux formants qui se cachent sous un seul pic. Par ailleurs, les largeurs de bande estimées sont plus grandes car l'allure de la courbe dépend d'influences mutuelles entre pics.

La préaccentuation du signal, quand elle est utilisée, engendre un léger décalage des fréquences formantiques vers les valeurs élevées par rapport aux mêmes fréquences calculées sans préaccentuation [Markel 76].

Si l'estimation des premiers formants des voyelles orales prononcées par des locuteurs masculins fournit de bons résultats, ceux-ci sont moins satisfaisants lorsque les voyelles sont nasales ou simplement nasalisées ou lorsqu'on est en présence d'une locutrice. Dans le premier cas, le modèle tout-pôle ne suffit plus. Dans le deuxième, la fréquence fondamentale est plus élevée, donc les harmoniques plus espacés. Or, lorsqu'un harmonique est voisin d'un formant, l'analyse par prédiction linéaire a tendance à placer la fréquence du pôle sur cet harmonique.

Les autres méthodes d'estimation de formants se fondent toutes sur la recherche de pics dans un spectre, soit le spectre FFT lissé soit le cepstre. Par rapport aux autres spectres, le spectre LPC a l'avantage de faire ressortir les formants du conduit vocal et de présenter moins de pics supplémentaires comme le formant glottal ou d'autres irrégularités spectrales dues à l'onde glottale.

#### 4.12. Les limites de l'analyse par prédiction linéaire

Bien qu'étant de manipulation facile et agréable, le modèle d'analyse par prédiction linéaire comporte certaines limites qu'il est bon de rappeler. Il suppose que le conduit vocal est un tube résonateur idéal sans pertes. Il n'est pas conçu pour modéliser les occlusives ou les fricatives voisées même s'il ne fournit pas de mauvais résultats lorsqu'il est appliqué à ces sons. En effet, le modèle n'est pas prévu pour mixer les sons voisés et les bruits blancs. Le modèle tout-pôle n'est pas non plus conçu pour modéliser les voyelles et les consonnes nasales. Si en augmentant

l'ordre de prédiction, il est possible d'approcher les zéros d'origine nasale et d'obtenir un bon spectre LPC, l'interprétation physique des racines du polynôme est pratiquement impossible. Enfin, ce modèle suppose que le conduit vocal est stable pendant la fenêtre d'analyse.

## 5. Conclusion

Nous avons présenté dans ce chapitre quelques outils du traitement du signal de parole en détaillant plus particulièrement l'analyse par prédiction linéaire et ce qui se rapporte aux formants, en laissant de côté d'autres méthodes comme l'analyse cepstrale ou les vocodeurs. Mais le lecteur curieux de connaître ces autres méthodes pourra se reporter aux ouvrages cités dans ce chapitre.

Notre choix se justifie par le fait que, comme nous le verrons dans la partie C de ce document, une grande partie de notre travail a consisté en l'estimation des formants de voyelles orales prononcées par des locuteurs masculins. Pour toutes les raisons développées dans ce chapitre, nous avons fondé cette estimation sur l'analyse par prédiction linéaire.



## CHAPITRE V LA VARIABILITE DE LA PAROLE

### 1. Introduction

Les difficultés et les quelques désillusions historiques rencontrées dans les années 70 par les chercheurs en reconnaissance automatique de la parole, tant au niveau de la parole continue qu'au niveau de l'indépendance vis-à-vis du locuteur [Calliope 89] [Vaissiere 84], ont mis en lumière le caractère variable de la parole. La variabilité de la parole peut se définir comme le fait qu'un segment de parole varie considérablement en fonction de son contexte, au sens le plus large du terme. En effet, comme nous le verrons tout au long de ce chapitre, ce contexte peut être temporel, linguistique ou paralinguistique. En d'autres termes, ce qu'on a coutume d'appeler "*variabilité*" est en réalité l'expression de deux propriétés de la parole, sa continuité et sa complexité.

La parole n'est pas une succession d'entités indépendantes et invariables égrenées au fil du temps mais un processus continu dans lequel les phonèmes sont liés les uns aux autres et s'influencent les uns les autres par le biais d'un phénomène naturel appelé coarticulation. Chaque phonème se réalise sous la forme d'allophones différents selon ses contextes antérieur et postérieur. Ces variations allophoniques peuvent aller de l'infime et imperceptible modification acoustique à l'assimilation totale à un autre phonème.

Par ailleurs, la parole est un moyen de communication complexe qui véhicule plusieurs niveaux d'information linguistique auxquels s'ajoutent des informations paralinguistiques. En effet, outre un message phonémique, le signal de parole transmet des données sur la structure lexicale, syntaxique et sémantique du message, ainsi que des renseignements sur l'identité du locuteur, son origine sociogéographique, ses habitudes linguistiques, son état de santé et son humeur.

Toutes ces données se combinent de façon complexe en un seul flux acoustique, à partir duquel on n'est pas capable à l'heure actuelle de retrouver chaque source d'information. Car, si on dispose au niveau segmental de relations quantitatives entre le signal de parole et les unités linguistiques abstraites que sont les phonèmes<sup>1</sup>, de telles relations n'existent pas pour les autres sources d'information linguistique. Les quelques relations établies jusqu'à maintenant dans ces domaines ont un caractère qualitatif et demeurent très fragmentaires [Fant 90a]. En ce qui concerne les informations paralinguistiques, la méconnaissance est encore plus importante puisqu'il n'existe aucun code les décrivant ni les structurant [Lienard 89].

Dans ces conditions, ce que nous allons analyser dans ce chapitre, et que nous nommons variabilité, est l'influence d'une source d'information sur les principaux paramètres acoustiques d'un son élémentaire, toutes les autres sources d'information étant maintenues constantes.

Après avoir détaillé les effets de la coarticulation sur les réalisations physiques des phonèmes, nous examinerons l'influence sur ceux-ci des informations de nature lexicale, syn-

---

<sup>1</sup> Les autres niveaux d'information étant plus ou moins maintenus constants.



taxique et sémantique, en nous fondant essentiellement sur les résultats de recherches sur la prosodie de la parole lue. Enfin, nous terminerons ce chapitre par l'étude de la variabilité qui concerne directement notre sujet de recherche, la variabilité liée au locuteur.

Bien que notre étude se situe dans le domaine du traitement automatique de la parole, nous ne traiterons pas dans ce chapitre de la variabilité de la parole qui résulte des moyens de transmission ou d'acquisition du signal de parole (microphone, filtre, convertisseur A/N, ...).

## 2. Variabilité due à la coarticulation

### 2.1. Généralités

La coarticulation peut se définir comme l'influence articulatoire qu'exerce un son sur un son contigu ou peu éloigné. Deux phénomènes non symétriques interviennent dans cette modification contextuelle : l'inertie mécanique des articulateurs qui provoque une influence progressive du son passé sur le son présent et la réorganisation du geste articulatoire du son présent en fonction des sons futurs (influence régressive) en vertu du principe du "minimum d'effort pour le maximum d'effet" [Lonchamp 87b] [O'Shaughnessy 87].

La coarticulation n'affecte pas uniformément tous les phonèmes : les voyelles y sont plus sensibles que les consonnes et les voyelles courtes plus que les voyelles longues [Vaissiere 84]. Par ailleurs, les phonèmes qui appartiennent à des syllabes inaccentuées subissent plus les phénomènes de coarticulation que ceux appartenant à des syllabes accentuées<sup>1</sup> [Cooper 83]. Enfin, l'ampleur des faits de coarticulation dépend considérablement du débit d'élocution, du locuteur et du contraste articulatoire existant entre les phonèmes adjacents [Cooper 83].

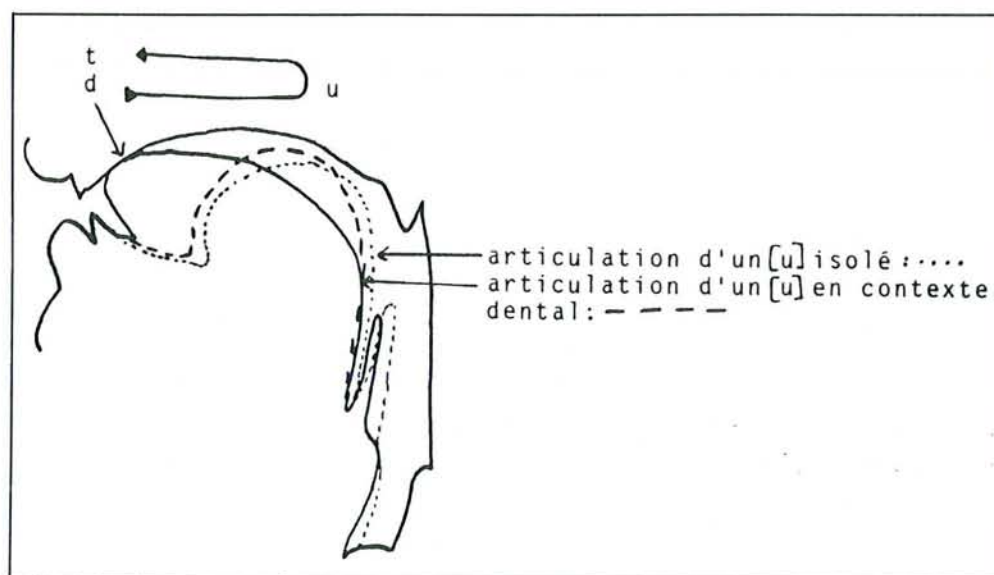


Figure A.38. Articulation réduite de [u] dans un contexte dental (/dut/), d'après [Lonchamp 87b].

<sup>1</sup> Les syllabes accentuées sont les syllabes mises en relief par un accent lexical ou par un accent d'insistance. Les différents types d'accent seront définis plus précisément dans le paragraphe 3.1.

Outre le déplacement des organes articulatoires d'une cible à une autre qui engendre des phénomènes transitoires aux frontières des phonèmes, la coarticulation peut prendre deux autres aspects non exclusifs qui concernent le phonème dans sa totalité : la réduction du geste articulatoire global (*undershooting*) et la superposition dans le temps de mouvements articulatoires qui sont normalement indépendants et locaux à des phonèmes distincts (*overlapping*).

La figure A.38 illustre le cas de la réduction du geste articulatoire : lors de la prononciation d'un phonème en contexte, le conduit vocal tend vers la position articulatoire du phonème prononcé isolément sans l'atteindre. Lors de la prononciation de /*du*t/, la langue doit passer d'une position avancée (dentale) pour l'articulation du /*d*/ à une position reculée pour l'articulation du /*u*/, avant de revenir à la position dentale. Si le débit n'est pas assez lent, la langue n'a pas le temps d'effectuer cet aller et retour sinon au prix d'une dépense considérable d'énergie. Le déplacement de l'articulateur est donc réduit et la voyelle réellement prononcée est un [ *u* ] antériorisé qui sera toutefois perçu par l'auditeur comme le phonème /*u*/. La figure A.39 présente le même processus de réduction pour une consonne.

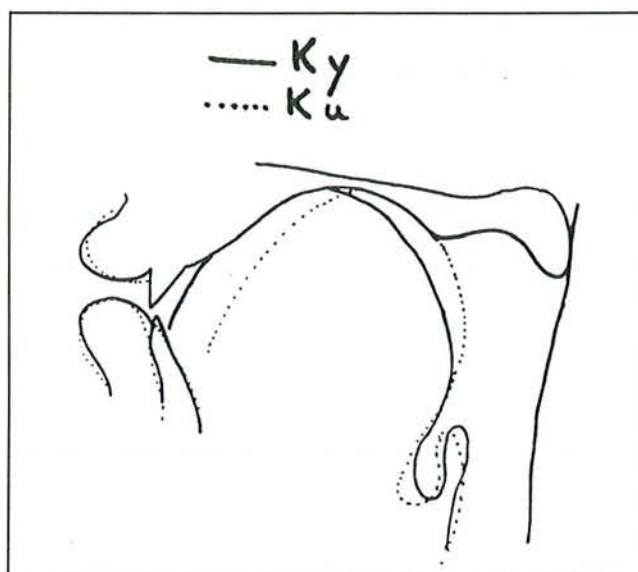


Figure A.39. Allophones dorso-vélaire ([ *ku* ]) et dorso-palatal ([ *ky* ]) du phonème /*k*/, d'après [Lonchamp 87b].

Durant la production d'un son, comme nous l'avons vu au chapitre II, la configuration du conduit vocal est obtenue par le positionnement de quatre articulateurs mobiles, la mandibule, les lèvres, la langue et le voile du palais, auxquels s'ajoute "l'articulateur glottal"<sup>1</sup> lors de l'établissement ou de l'arrêt du voisement. Chaque fois que cela est possible, le déplacement de l'un des articulateurs nécessaires à la production d'un phonème donné est reporté avant ou après la réalisation proprement dite du phonème. Il en résulte un recouvrement temporel de gestes articulatoires appartenant à des phonèmes distincts. Ce recouvrement est rendu possible grâce, d'une part, à la quasi-indépendance de ces cinq articulateurs, d'autre part, au fait qu'ils ne soient pas toujours tous impliqués dans la réalisation d'un phonème ni, surtout, dans sa

<sup>1</sup> Le terme articulateur glottal englobe tous les articulateurs nécessaires à la fermeture ou à l'ouverture de la glotte.

caractérisation en traits distinctifs. L'un des exemples de superposition de gestes articulatoires les plus communément cités en français est la prononciation du mot "structure" qui débute par l'arrondissement des lèvres indispensable à l'articulation du / y / ; la labialisation des trois premières consonnes de la syllabe ne gêne en rien leur perception. Un autre exemple est présenté sur la figure A.40 : la langue, articulateur libre dans la production de l'occlusive voisée / b /, adopte déjà la position nécessaire à la prononciation de la voyelle qui suivra la consonne.

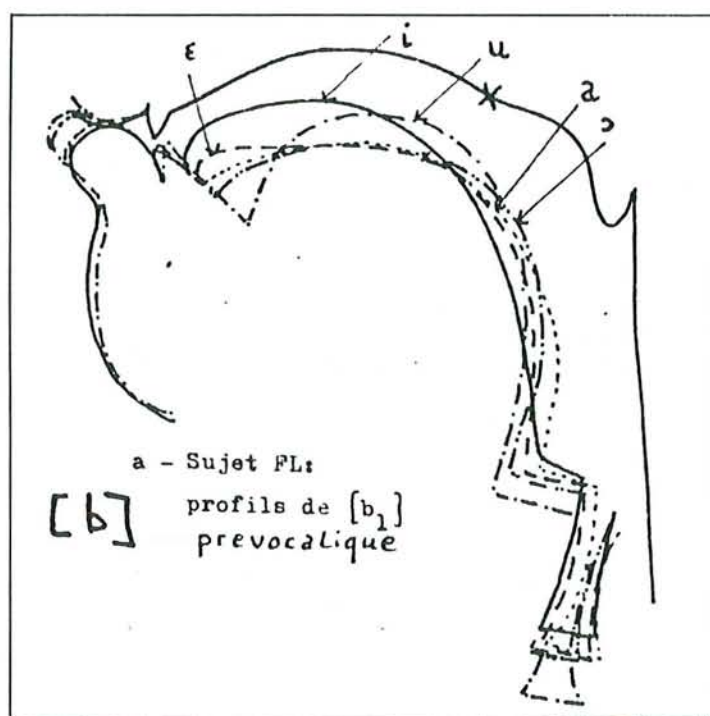


Figure A.40. Prépositionnement de la langue lors de l'articulation d'un / b / prévocalique, d'après [Lonchamp 87b].

Comme le montrent les exemples précédents, la coarticulation peut être mise en évidence directement à l'aide de radiographies et d'électromyographies du conduit vocal mais, le plus souvent, elle s'observe de manière indirecte par ses conséquences acoustiques. En effet, les caractéristiques acoustiques associées à la production d'un phonème isolé sont modifiées lorsque ce phonème est soumis à la coarticulation. Lorsque ces modifications acoustiques sont telles que le son modifié se rapproche d'un autre phonème (par commutation de traits distinctifs) selon une règle commune à une classe de phonèmes, on regroupe ces modifications sous le terme d'assimilation. Après quelques définitions, le paragraphe suivant présente les trois principaux cas d'assimilation du français, les autres conséquences acoustiques de la coarticulation étant traitées dans les paragraphes subséquents.

## 2.2. Les assimilations du français

### 2.2.1. Quelques définitions

Le sens de l'assimilation, sa portée, la classe de phonèmes concernée et la nature du trait modifié définissent le type de l'assimilation. L'assimilation de contact concerne deux



phonèmes adjacents, la dilatation ou assimilation de distance concerne deux phonèmes séparés par d'autres éléments de la chaîne parlée. L'assimilation est dite régressive lorsque le son programmé influence le son réalisé et progressive dans l'autre sens. Enfin, l'assimilation peut être partielle ou complète.

### 2.2.2. Les assimilations consonantiques

Elles sont issues de la coarticulation entre deux consonnes adjacentes, ou qui le deviennent à la suite de l'élision d'un schwa, et portent essentiellement sur deux traits phonétiques, le voisement et la nasalité.

#### • L'assimilation de voisement ou d'assourdissement

Une consonne occlusive ou fricative sourde (respectivement sonore) suivie d'une occlusive ou d'une fricative sonore (respectivement sourde) se sonorise (respectivement s'assourdit) :  
/ absā / > <sup>1</sup>[ apsā ] et / pasdRwa / > [ pazdRwa ].

La consonne assimilée ne possède pas toutes les caractéristiques acoustiques de la consonne opposée mais les différences sont minimales [Lonchamp 87a].

Cette assimilation régressive possède quelques exceptions comme :  
/ defəvo / > [ deffo ].

#### • L'assimilation de nasalité

Une occlusive située entre une voyelle nasale et une consonne autre que / l, R, j, ɥ, w / se nasalise : / p,b / > [ m ], / t,d / > [ n ] et / k,g / > [ ŋ ].

Toutefois, cette assimilation progressive peut n'être que partielle lorsque la deuxième consonne est sourde [Lonchamp 87a] :

/ pātəkot / > [ pānkot ] ou [ pāntkot ].

Si la deuxième consonne est nasale, la règle énoncée comporte quelques exceptions. Notamment, lorsque la consonne assimilée est identique à cette deuxième consonne, la règle ne s'applique pas sauf dans le cas de "maintenant" :

/ lɔ̃gəmā / > [ lɔ̃ŋmā ] et / mɛ̃tənā / > [ mɛ̃nnā ] mais / kārəmā / ne devient pas [ kārmmā ].

De même, la règle ne s'applique pas lorsque la consonne nasale est suivie d'une voyelle nasale identique à celle qui précède la consonne assimilée :

/ lātəmā / ne devient jamais [ lānmā ].

### 2.2.3. L'assimilation vocalique

Encore appelée harmonisation vocalique, cette assimilation de distance résulte d'un effet de coarticulation entre deux voyelles entraînant la modification du timbre de l'une d'elles. En règle générale, ce sont les voyelles accentuées qui influencent le timbre des voyelles inaccentuées. Plus ou moins importante selon les langues, la dilatation vocalique n'est pas systématique en français. La plupart des phonéticiens la décrivent comme une tendance assez nette à faire varier le degré d'aperture de la voyelle d'une syllabe ouverte<sup>2</sup> inaccentuée d'un mot, en fonction du degré d'aperture de la voyelle accentuée du mot : / tɛty / > [ tety ]. Toutefois, A. Lacheret-Dujour, dans une étude sur les variantes phonologiques du français parisien [Dujour 90], note

<sup>1</sup> Le symbole > indique une transformation régulière.

<sup>2</sup> Une syllabe ouverte est une syllabe qui ne se termine pas par une ou plusieurs consonnes : dans / sistematik / les deuxième et troisième syllabes sont des syllabes ouvertes, les autres des syllabes fermées.

qu'en parole continue le sens de l'harmonisation vocalique ne respecte pas cette tendance mais varie selon les locuteurs.

### 2.3. Autres influences entre consonnes

Nous ne citerons que les deux plus remarquables :

- les sonantes /l, R, m, n, p, ɲ, j, ɥ, w/, intrinsèquement sonores, se dévoient complètement ou partiellement au contact d'une consonne sourde :  
/ tuRtəRɛl / > [ tuRtəRɛl ]<sup>1</sup> ou encore / plyvjedɔRe / > [ plyvjedɔRe ].

Bien que cette règle ressemble à une assimilation rétro-progressive, elle n'en est pas une car, en français, les sonantes ne perdent pas leur identité en s'assourdisant.

Les mots se terminant par le suffixe /ism/ présentent une exception à cette règle : selon la nature du mot et l'origine sociolinguistique du locuteur, le suffixe se transforme en [ism̥]<sup>2</sup> ou [ism̩].

Notons enfin que, selon le même principe, les sonantes ont tendance à se dévoiser en finale de groupe ;

- lorsque deux consonnes adjacentes sont identiques ou s'articulent au même endroit, certains des gestes articulatoires qui leur sont associés disparaissent ou se transforment.

Si deux occlusives se succèdent, la première "n'explose pas" et la consonne résultante possède une tenue de durée double.

Lorsqu'une occlusive est suivie de la consonne nasale qui s'articule au même lieu (/bm/), l'explosion buccale ne se produit pas car l'air s'échappe par le nez.

De même, quand une occlusive dentale (/t/, /d/) est suivie d'un /l/, consonne latérale, la langue adopte la position pour l'articulation du /l/ avant l'explosion et celle-ci devient latérale, l'air s'échappant de chaque côté de la langue [Malmberg 74].

### 2.4. Influence des consonnes sur les voyelles

#### 2.4.1. Nasalisation et assourdissement

Les voyelles orales précédées ou suivies d'une consonne nasale ont tendance à se nasaliser. En français, cette assimilation n'est que partielle en raison de la présence des voyelles nasales dans le système phonologique. De la même façon, le début des voyelles s'assourdit au contact des consonnes sourdes.

#### 2.4.2. Variation des paramètres suprasegmentaux

La durée, la fréquence fondamentale (hormis la présence ou l'absence de voisement) et l'intensité acoustique d'un son interviennent peu dans l'établissement des traits distinctifs caractérisant un phonème<sup>3</sup>, par rapport aux paramètres décrivant les répartitions spectrales et temporelles de l'énergie. En revanche, leurs variations au cours d'un énoncé sont des indices importants de sa structuration en éléments linguistiques d'un niveau supérieur (mot, syntagme, ...)

<sup>1</sup> ɔ̥ est le diacritique API signalant un assourdissement.

<sup>2</sup> ɔ̩ est le diacritique API signalant une sonorisation.

<sup>3</sup> Si l'on fait abstraction de certaines langues, comme les langues à tons. Les langues à tons, comme le chinois, sont des langues où il existe des unités distinctives élémentaires, appelées tonèmes, qui ne s'opposent que par la valeur ou le sens de variation de la fréquence de vibrations des cordes vocales.



ainsi que de certaines informations que souhaite transmettre le locuteur (sentiments, attitude sociale, ...). Aussi a-t-on coutume de regrouper ces trois paramètres sous l'appellation de paramètres suprasegmentaux.

La durée des voyelles est principalement influencée par la classe de la consonne suivante. Les voyelles précédant un contexte sonore sont plus longues que celles précédant un contexte sourd ; l'effet est plus important lorsque les deux phonèmes appartiennent à la même syllabe [Vaissiere 88]. En anglais, l'allongement peut atteindre 50 % [Lehiste 75]. De plus, une voyelle est plus courte devant une consonne nasale que devant une occlusive et plus longue devant une fricative que devant une occlusive [Malmberg 79].

En ce qui concerne la fréquence fondamentale, c'est la consonne antécédente qui a une action prépondérante : les voyelles précédées d'une consonne non voisée ont une fréquence fondamentale plus élevée que celles précédées d'une consonne voisée [Vaissiere 88] [Lehiste 75].

L'intensité dépend également du trait de voisement du contexte consonantique, les voyelles en contexte sourd étant plus intenses que celles en contexte sonore.

#### 2.4.3. Variabilité des fréquences formantiques

Les effets de la coarticulation sur les formants d'une voyelle entourée d'un contexte consonantique bilatéral (CVC) sont considérables et leur étude est primordiale en reconnaissance automatique de la parole. Elle l'est également en reconnaissance du locuteur car ces effets varient avec le locuteur. Ils se manifestent d'une part, par une déviation des fréquences formantiques au centre de la voyelle par rapport à celles de la voyelle cible (prononcée isolément) ; d'autre part, par la présence de transitions formantiques (évolution temporelle des fréquences formantiques) au début et à la fin de la voyelle. La figure A.41 présente la variabilité consonantique de  $F_1$  et  $F_2$  de [a] dans "Madagascar" [Vaissiere 87b]. Pour un locuteur donné, la valeur des fréquences formantiques au centre de la voyelle et l'allure des transitions formantiques suffisent dans un bon nombre de cas à déterminer la constitution du groupe CVC [Shoup 75].

Les fréquences formantiques au centre de la voyelle dépendent à la fois de la cible vocalique et de l'ampleur de la réduction du geste articulatoire provoqué par la présence des deux consonnes (cf. paragraphe 2.1), c'est-à-dire, plus précisément, de la durée de la voyelle et du contraste articulatoire entre la voyelle et les consonnes [Vaissiere 87b]. Les plus grands écarts entre les formants réels et les formants de la voyelle cible sont obtenus lors de la prononciation d'une voyelle antérieure dans un contexte postérieur : [kɛk] ou dans la situation duale : [dut] (cf. figure A.38).

Les transitions formantiques traduisent les positions intermédiaires prises par les organes articulateurs entre les lieux d'articulation des consonnes et celui de la voyelle, qu'elle soit ou non réduite. Leurs pentes sont donc plus ou moins fortes selon le débit d'élocution et l'écart entre les lieux d'articulation.



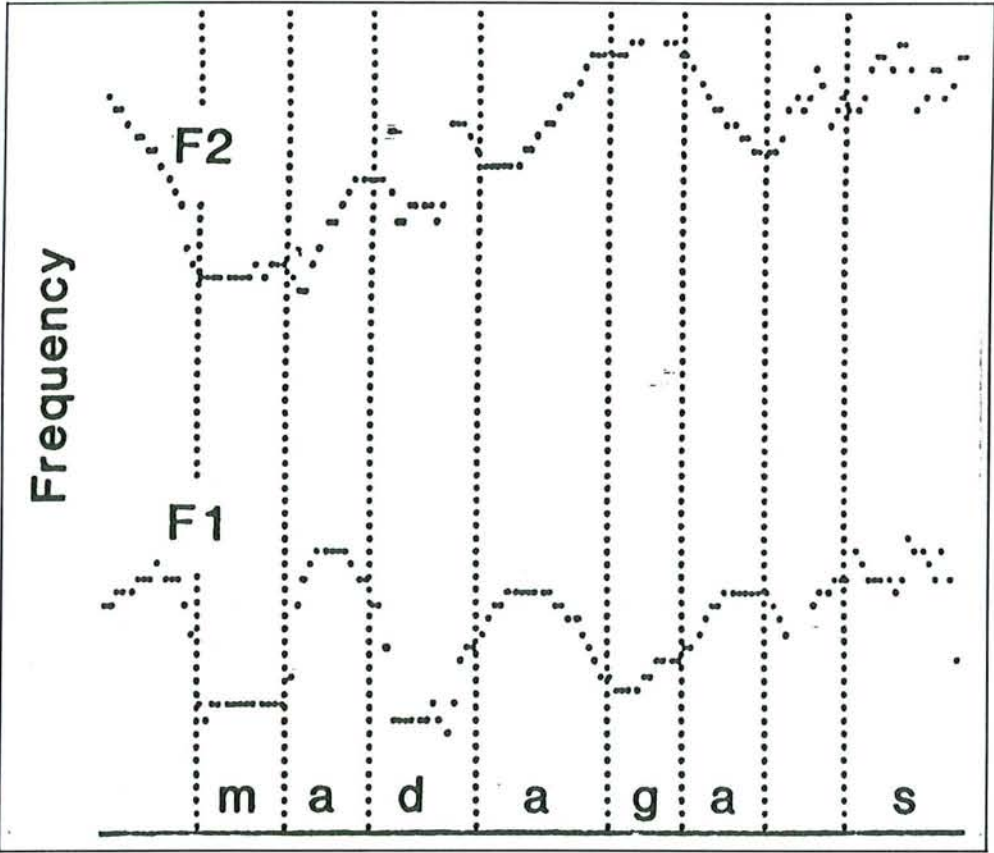


Figure A.41. Variabilité des deux premiers formants du [ a ] dans l'expression "à Madagascar", d'après [Vaissiere 87b].

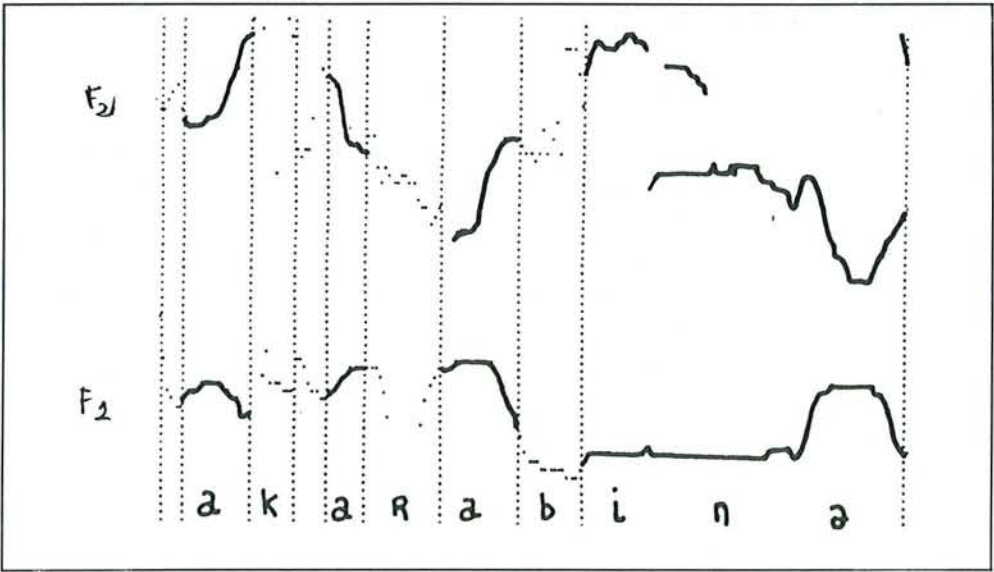


Figure A.42. Influence du [ i ] sur les deux premiers formants du [ a ] dans le mot "carabine", d'après [Vaissiere 87b].

Pour décrire ces transitions, les phonéticiens ont coutume de regrouper les consonnes en quatre classes suivant leur lieu d'articulation [Lonchamp 87b] [Vaissiere 87a]. Les transitions sont considérées dans le sens consonne vers voyelle :

- **contexte labial** : [ p, b, m, f, v ]. Un resserrement aux lèvres a tendance à faire baisser toutes les fréquences formantiques ; les transitions sont donc en général montantes, notamment pour  $F_2$  et  $F_3$  (cf. figure A.41) ;
- **contexte dental** : [ t, d, n, s, z, l ]. Les transitions sont peu marquées pour les voyelles antérieures fermées qui ont un lieu d'articulation proche de celui de la consonne. Les transitions de  $F_2$  et  $F_3$  sont descendantes pour toutes les autres voyelles, celle de  $F_1$  est plate ou montante dans le cas des voyelles ouvertes (cf. figure A.41) ;
- **contexte vélaire** : / k, g /. L'articulation des consonnes dorso-vélaires étant voisine de celle de / u /, les transitions de  $F_2$  et  $F_3$  sont plus importantes lorsque la voyelle est antérieure et ouverte : les transitions montante pour  $F_2$  et descendante pour  $F_3$  forment une "pince vélaire". La transition de  $F_1$  est montante dans tous les cas (cf. figure A.43) ;
- **contexte uvulaire** : [ R ]. Les transitions sont également très caractéristiques. Elles forment pour toutes les voyelles deux pinces entre  $F_1$  et  $F_2$  d'une part et  $F_3$  et  $F_4$  d'autre part ; les transitions sont montantes pour  $F_1$  et  $F_3$  et descendantes pour  $F_2$  et  $F_4$  (cf. figure A.43).

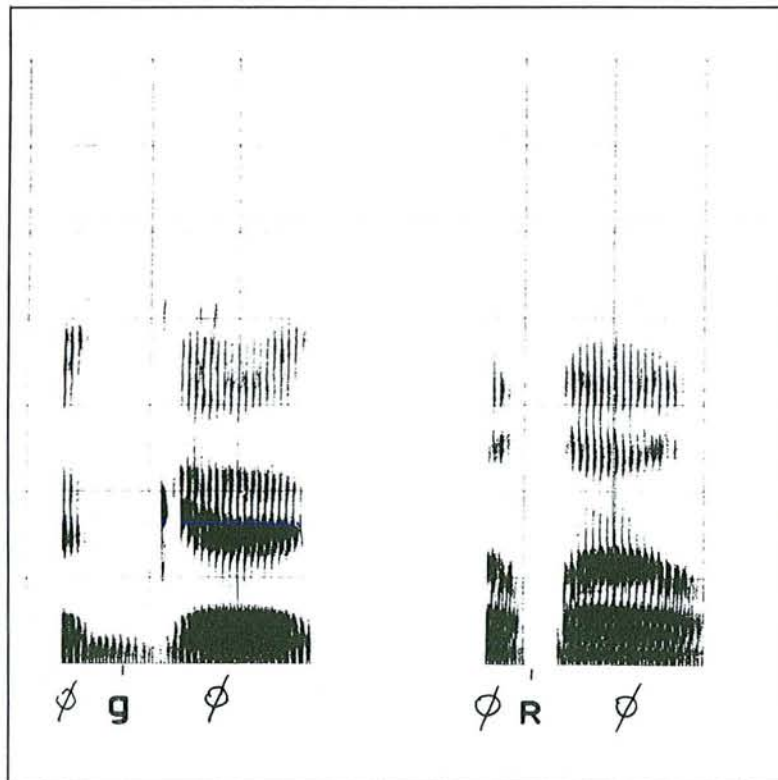


Figure A.43. Transitions formantiques de [ ø ] dans un contexte vélaire et uvulaire, d'après [Eskenazi 88].

Précisons pour clore ce paragraphe que les transitions  $C_1 - V$  et  $V - C_1$  d'une séquence  $C_1VC_1$  ne sont pas symétriques. Cette asymétrie est due à celle qui existe entre les effets conservateurs et les effets anticipateurs de la coarticulation consonantique mais aussi à l'influence régressive de la voyelle qui suit la dernière occurrence de la consonne, influence qui sera étudiée dans le paragraphe suivant.

## 2.5. Autres influences entre voyelles

La comparaison des figures A.41 et A.42 montre que, malgré le même contexte labial, les trajectoires de  $F_2$  à la fin du deuxième [a] de "carabine" et au début du premier [a] de "Madagascar" sont contradictoires : la première est descendante (dans le sens consonne vers voyelle) alors que la deuxième est montante, respectant la tendance indiquée au paragraphe précédent. Cette contradiction s'explique par l'influence anticipatrice du [i] de [abi] sur les formants du [a]. Celle-ci est maximale dans le contexte labial — l'articulation linguale n'étant pas pertinente lors de la production d'un [b] — mais elle demeure dans les autres contextes [Vaissière 87b]. Lorsque le débit est rapide, cette influence affecte également les fréquences formantiques au centre de la voyelle.

En d'autres termes, il existe une coarticulation régressive entre deux voyelles séparées par une consonne. Les cas de coarticulation progressive sont très rares et sembleraient associés à des contours mélodiques particuliers [Vaissière 87b].

En conclusion, les formants d'une voyelle dépendent de la cible vocalique mais aussi des consonnes qui l'entourent et de la voyelle de la syllabe suivante.

## 2.6. Influence des voyelles sur les consonnes

Si deux voyelles interagissent au travers d'une consonne, il est évident que cette consonne est, elle aussi, affectée par les voyelles. Comme nous l'avons souligné dans l'introduction, l'influence des voyelles sur les consonnes peut-être due à une réduction du geste articulatoire ou bien à l'anticipation d'un geste articulatoire concernant un articulateur qui n'est pas mis en œuvre dans la production de la consonne.

L'anticipation du geste articulatoire concerne principalement :

- la langue dans le cas des consonnes bilabiales /p/, /b/ (cf. figure A.40) et /m/ (cf. chapitre III de la partie B),
- les lèvres pour toutes les consonnes sauf celles qui font intervenir les lèvres dans leur articulation, c'est-à-dire /p/, /b/, /m/, /f/ et /v/,
- la luette essentiellement pour /b/ et /d/.

L'influence due à la réduction du geste articulatoire est prépondérante en français [Malmberg 74]. Le cas le plus typique est celui de la consonne [k] dont l'occlusion est dorso-vélaire devant [o] et [u], dorso-palatale devant [i] et [e], en passant par une articulation intermédiaire pour [ɛ] et [a] (cf. figure A.39).

Certaines des modifications articulatoires des consonnes ont reçu une dénomination :

- **la labialisation** ou arrondissement des lèvres qui accompagne les consonnes qui sont au voisinage des voyelles labiales. En français, une consonne labialisée n'est pas perceptivement différente d'une consonne non labialisée sauf dans le cas de la paire (/j/, /ɥ/) où /ɥ/ est l'opposé labial de /j/ [Malmberg 79] ;



- **la vélarisation** qui désigne le recul de l'articulation d'une consonne au contact d'une voyelle postérieure et en particulier d'un [ u ]. En français, la vélarisation s'accompagne toujours d'une labialisation (cf. chapitre III) ;
- **la palatalisation**, qui correspond à la poussée de la masse de la langue vers le palais dur, accompagne les articulations consonantiques en présence d'un [ i ] et dans une moindre mesure d'un [ e ] ou d'un [ y ]. Dans le cas des occlusives / t, d /, on parle d'affrication car l'explosion de la consonne est suivie d'un bruit de friction (/ tydi / > [ t<sup>s</sup>yd<sup>z</sup>i ]) causé par l'étroit chenal existant entre le corps de la langue et le palais. Le français est caractérisé par une très forte tendance à la palatalisation des consonnes par les voyelles notamment dans le dialecte parisien populaire [Malmberg 74].

## 2.7. Conclusion

Du point de vue phonologique, il est important de noter que bien que certaines variations contextuelles soient plus marquées lorsque les phonèmes appartiennent à la même syllabe [Vaissière 88], la coarticulation ne permet pas de définir le concept de syllabe [Lonchamp 87a].

Les nombreux cas traités dans ce paragraphe attestent sans équivoque l'intérêt capital des faits de coarticulation en reconnaissance automatique de la parole. Mais ils sont aussi primordiaux dans le domaine de la reconnaissance du locuteur car, comme nous le verrons plus loin, ils varient en fonction du locuteur et du débit d'élocution.

# 3. Variabilité d'origine linguistique

## 3.1. La prosodie

### 3.1.1. Introduction

La prosodie est un concept linguistique difficile à cerner. Elle possède de multiples définitions selon que leurs auteurs s'intéressent aux entités linguistiques ou paralinguistiques qu'elle recouvre (accent, syntaxe, emphase, ...) [Fant 90a], à ses manifestations acoustiques (fréquence fondamentale, durée, intensité acoustique) [Vaissière 88] [Calliope 89] ou encore à la perception de ces manifestations (hauteur, quantité, intensité sonore) [Lehiste 75].

L'une des définitions les plus communément proposées, notamment en français, définit la prosodie comme l'évolution temporelle dans une phrase des trois paramètres suprasegmentaux, que sont la fréquence fondamentale, la durée et l'intensité acoustique, en fonction de tout facteur linguistique ou paralinguistique qui ne soit pas d'ordre phonématique (segmental).

Dans le cadre de ce paragraphe, nous limiterons le domaine d'étude de la prosodie à des facteurs purement linguistiques comme les emplacements respectifs du phonème dans la syllabe, de la syllabe dans le mot ou du mot dans la phrase, comme le sens de la phrase ou comme l'information que le locuteur souhaite ajouter à son énoncé (insistance, contraste, doute, ...). En revanche, nous étendrons l'examen des manifestations acoustiques de la prosodie aux paramètres liés à l'articulation des phonèmes.

Il est difficile à l'heure actuelle de définir avec précision l'évolution temporelle au cours d'un énoncé des paramètres suprasegmentaux et ceci pour plusieurs raisons :

- l'étude suprasegmentale d'un paramètre est toujours délicate puisqu'il faut d'abord s'affranchir de sa partie segmentale, c'est-à-dire propre au son étudié et à son environnement immédiat ;
- les fonctions linguistiques de la prosodie ne se traduisent pas au niveau du signal de parole par des valeurs absolues de paramètres acoustiques mais par leurs valeurs relatives entre plusieurs phonèmes. De plus, ces variations ont généralement un caractère purement qualitatif (montée ou descente de  $F_0$ , allongement, intensité plus ou moins forte, ...)
- la fréquence fondamentale, l'intensité acoustique et la durée ne sont pas des paramètres prosodiquement indépendants. En effet, la même fonction linguistique peut être obtenue par une modification de l'un ou l'autre d'entre eux ou de plusieurs à la fois. L'accent, qui consiste à mettre en relief une syllabe aux dépens des autres, en est un bon exemple. En effet, il peut résulter d'une variation de la fréquence fondamentale, d'une augmentation de l'intensité ou d'un allongement de la syllabe ou bien d'une combinaison de ces trois phénomènes. De plus, selon les langues, l'une ou l'autre de ces variations est prépondérante pour la perception de l'accent [Lehiste 75] [O'Shaughnessy 87]. Cette interdépendance est essentiellement due au fait qu'il n'existe pas de correspondance biunivoque entre ces trois paramètres acoustiques et leurs équivalents perceptifs qui sont la hauteur, l'intensité sonore et la quantité [Calliope 89] [Zwicker 81] ;
- ces trois paramètres ne sont pas non plus physiologiquement indépendants. Ainsi, une intensité acoustique plus élevée, provoquée par un accroissement de la pression sous-glottique, a pour conséquence une augmentation de la fréquence de vibration des cordes vocales à moins d'un ajustement compensatoire de leur tension ;
- hormis quelques principes généraux, la prosodie revêt des formes très diverses selon le style de parole (spontané, lu, ...) et la langue considérés. Prenons par exemple le cas de l'accent lexical. Les linguistes ont coutume de distinguer les langues à accent fixe<sup>1</sup> comme le français ou le polonais, des langues à accent libre<sup>2</sup> comme l'anglais ou l'italien.

Dans ces dernières, la place de l'accent dans le mot dépend du mot prononcé et joue parfois un rôle phonologique au même titre que les phonèmes (en anglais, / 'import / signifie "*importation*" et / im'port /, "*importer*")<sup>3</sup>. Au sein d'une phrase, cet accent lexical est supplanté par l'accent de phrase qui consiste à ne marquer, parmi les syllabes potentiellement accentuables par l'accent lexical, que celles des mots primordiaux de la phrase [O'Shaughnessy 87] [Lehiste 75].

Dans les langues à accent fixe, comme leur nom l'indique, l'accent tombe toujours sur la même syllabe du mot quel que soit celui-ci. En français, chaque mot présente un faible accent sur la dernière syllabe qui disparaît souvent, comme nous le verrons dans les paragraphes suivants, au profit d'un accent prosodique.

L'accent d'insistance ou de contraste intervient, quant à lui, dans toutes les langues. Placé intentionnellement par le locuteur sur l'un des éléments de la chaîne parlée, il a tendance à marquer des syllabes normalement inaccentuées.

<sup>1</sup> Qui sont encore appelées langues sans accent.

<sup>2</sup> Qui sont encore appelées langues à accent d'intensité, à accent de mot ou à accent tonique (*stress*).

<sup>3</sup> Le symbole ' précise que la syllabe suivante porte l'accent lexical principal.



Dans les paragraphes suivants sont regroupés de manière non exhaustive des résultats sur la variabilité prosodique de la parole lue.

### 3.1.2. Variations prosodiques de la fréquence fondamentale

Deux phénomènes s'observent dans beaucoup de langues :

- l'association d'un type de contour mélodique global (évolution temporelle de  $F_0$  sur une phrase) avec une catégorie modale de phrase. En français, comme l'illustre la figure A.44, le contour est descendant pour les phrases déclaratives et montant pour les phrases interrogatives ;
- l'adéquation entre certains contours mélodiques (montée et descente de  $F_0$ ) et des entités lexicales ou sémantiques comme les mots, les groupes nominaux ou les syntagmes [Lehiste 75]. Selon les auteurs, ces entités sont regroupées sous le terme générique de "mots prosodiques" [Vaissiere 80] ou celui de "groupes phonétiques" [Malmberg 74]. Ces contours mélodiques varient plus ou moins suivant les langues, les locuteurs et le débit d'élocution [Vaissiere 83]. En français, les variations significatives de  $F_0$  se situent aux frontières

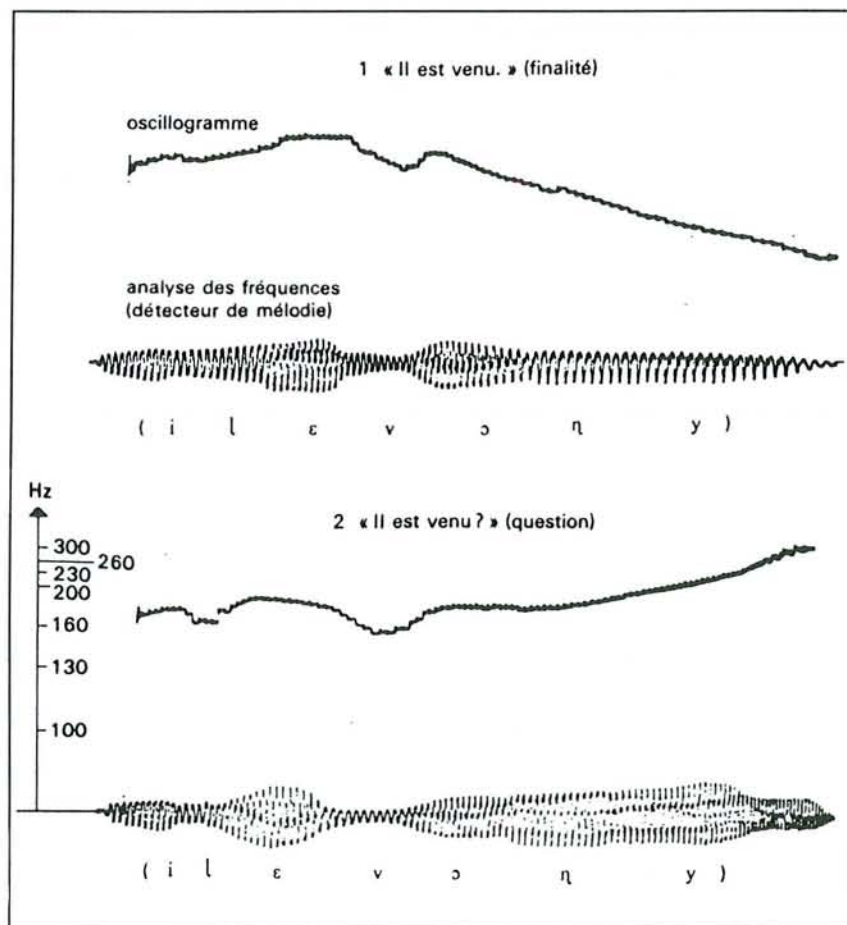


Figure A.44. L'intonation modale en français, d'après Encyclopedia Universalis.



de mots, en anglais au niveau des syllabes accentuées [Vaissiere 88]. La figure A.45 fournit un exemple de contours mélodiques associés à des mots prosodiques.

Par ailleurs, dans les langues à accent tonique (*stress*), un noyau vocalique a une fréquence fondamentale plus élevée lorsqu'il appartient à une syllabe accentuée [Shoup 75] [O'Shaughnessy 87]. En anglais, par exemple l'augmentation de  $F_0$  est le meilleur indice perceptif d'accentuation [Lehiste 75].

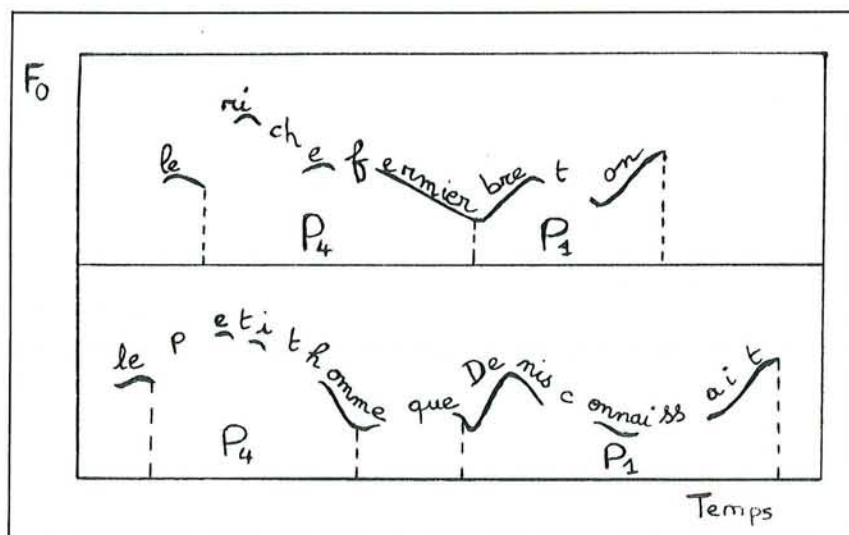


Figure A.45. Exemples de contours mélodiques en français : "le riche fermier breton" et "le petit homme que Denis connaissait" ont la même structure prosodique composée de deux mots prosodiques, d'après [Vaissiere 80].

### 3.1.3. Variations prosodiques de la durée

Du point de vue suprasegmental, la durée d'un son et plus particulièrement d'une voyelle est fixée par plusieurs facteurs dont les effets sont susceptibles de s'ajouter ou de se compenser :

- le débit d'élocution qui est lui-même une fonction du locuteur (identité et état) et du style de parole et qui, pour ces raisons, fera l'objet d'un paragraphe spécial ;
- le nombre de phonèmes dans la syllabe (isosyllabité), de manière à conserver une durée de syllabe globalement invariante, toutes choses étant égales par ailleurs [Vaissiere 88] ;
- le nombre de syllabes dans le mot prosodique. Ainsi plus le suffixe d'un mot construit à partir d'un radical est long, plus la durée du noyau vocalique du radical diminue [Lehiste 75]. Cette tendance, plus importante dans les langues à accent libre, s'explique par le phénomène d'isochronie qui consiste à égaliser les intervalles de temps séparant les syllabes accentuées [Vaissiere 88]. Pourtant, d'après G. Fant [Fant 90b], l'intervalle d'isochronie moyen est de l'ordre de 550 ms et comprend 6 ou 7 phonèmes, que ce soit en suédois, en anglais ou en français, à condition que cet intervalle ne recouvre pas une frontière syntaxique ;
- la structure syntaxico-sémantique de la phrase. Dans de nombreuses langues, on note un allongement de la dernière syllabe d'un groupe de phonation, c'est-à-dire avant une pause. L'allongement est inversement proportionnel à la distance à la pause [Fant 90a]. En anglais, cet

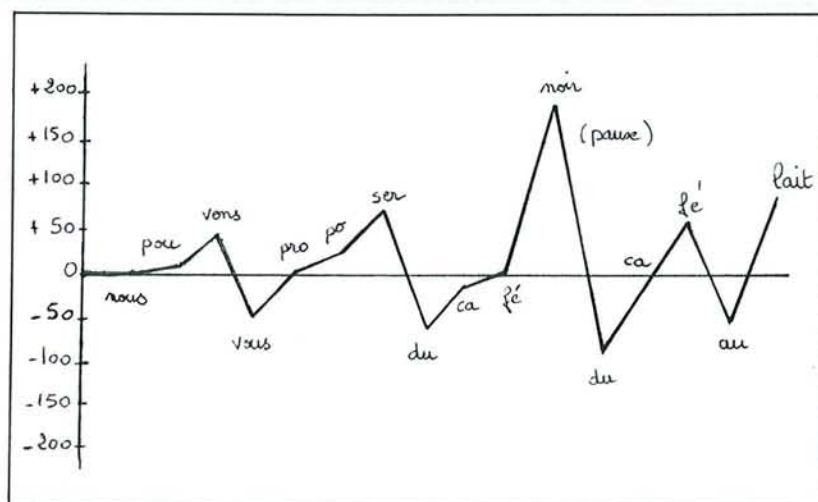


Figure A.46. Variations de durée (en pourcentage) d'une syllabe à l'autre dans le début de la phrase : "Nous pouvons vous proposer du café noir, du café au lait ...", d'après [Vaissière 80].

allongement peut atteindre 200 ms [O'Shaughnessy 87]. Si la pause disparaît, cet allongement demeure comme marqueur de frontière syntaxique, tout en étant moins conséquent [Vaissière 83].

En français, l'allongement en fin de groupe de phonation porte essentiellement sur la dernière voyelle et varie selon celle-ci et son contexte [Lonchamp 87a] [Malmberg 74]. L'allongement est plus important lorsque :

- la voyelle est suivie de / R /, / v /, / z / ou / ʒ / ;
- la voyelle est soit une voyelle nasale ou soit une des voyelles { / o /, / ø /, / a / } et est suivie d'une consonne unique ou d'un doublet / **occlusive-l** / ou / **occlusive-R** /.

Dans une étude multilingue sur la parole lue, G. Fant [Fant 90b] remarque que la durée cumulée de l'allongement de la dernière syllabe d'un groupe de phonation et de la pause qui la suit possède le même ordre de grandeur que l'intervalle d'isochronie. Par ailleurs, les pauses situées entre les phrases ont des durées qui sont des multiples entiers de cet intervalle, le coefficient multiplicatif étant une fonction du locuteur.

L'allongement est également utilisé dans les phénomènes d'accentuation mais avec un degré variable selon la langue. Dans les langues comme l'anglais ou le suédois, il accompagne souvent l'augmentation de  $F_0$  marquant l'accent d'intensité [Shoup 75] [Lehiste 75]. Ainsi, en anglais, les voyelles accentuées sont en moyenne 10 à 20% plus longues que les mêmes voyelles en position inaccentuée [O'Shaughnessy 87]. En français, la dernière syllabe d'un mot est aussi la plus longue mais, comme l'accent est moins marqué, l'allongement est moins important qu'en anglais ou qu'en suédois [Vaissière 83] [Fant 90a]. Malgré tout, la durée des syllabes suit les contours mélodiques, mettant ainsi en relief un découpage prosodique de la phrase : la consonne à l'initiale de mot est plus longue et, surtout, la dernière syllabe d'un mot prosodique est allongée [Vaissière 80]. Par exemple, nous pouvons observer sur la figure A.46 les allongements des dernières syllabes des deux premiers mots prosodiques, qui correspondent ici à des mots lexicaux, et l'allongement en fin de groupe de phonation, qui est nettement plus important.



L'allongement est communément utilisé pour indiquer un accent d'insistance [Vaissiere 83].

#### 3.1.4. Variations prosodiques de l'intensité acoustique

Bien qu'elle soit le paramètre prosodique le moins étudié [Vaissiere 83] et le moins pertinent du point de vue perceptif [Lehiste 75] [O'Shaughnessy 87], l'intensité acoustique intervient dans les phénomènes d'accentuation.

L'importante corrélation, aussi bien physiologique que perceptive, qui existe entre l'intensité et les autres paramètres suprasegmentaux rend difficile la détermination exacte de sa contribution dans la réalisation de l'accent d'intensité. Toutefois des études électromyographiques ont révélé des pics d'activité des muscles intercostaux internes ainsi qu'un accroissement de la pression sous-glottique lors de la réalisation de syllabes portant un accent d'insistance ou un accent de phrase [Lehiste 75].

En français, langue sans accent d'intensité, l'intensité acoustique ne semble pas être utilisée comme marqueur prosodique, le locuteur français ayant tendance à la répartir uniformément sur toutes les syllabes [Vaissiere 83].

#### 3.1.5. Variations prosodiques des paramètres segmentaux

Les faits prosodiques ne se limitent pas aux trois paramètres suprasegmentaux, ils influent indirectement sur d'autres paramètres acoustiques et notamment sur les formants.

Plus une voyelle est longue, mieux elle est articulée, moins elle est assujettie à son contexte et, par conséquent, plus ses formants se rapprochent de ceux de la voyelle cible.

Dans une langue à accent comme l'anglais, l'influence de la prosodie est encore plus considérable. Une syllabe accentuée est prononcée avec plus d'énergie articulatoire et par conséquent avec des gestes articulatoires plus précis. Elle est donc moins sensible à la coarticulation [Cooper 83]. Au contraire, les phonèmes inaccentués sont produits avec une articulation relâchée. Il en résulte, pour les voyelles, une centralisation des formants vers ceux du schwa [Shoup 75] [O'Shaughnessy 87], et pour les consonnes une structure proche de celle des sonantes. La figure A.47 présente le cas du / g / suédois prononcé avec trois niveaux d'accentuation différents.

#### 3.1.6. Conclusion

Rappelons tout d'abord que les informations présentées dans ce paragraphe ne concernent que la prosodie de la parole lue. La prosodie de la parole naturelle est d'une complexité bien plus importante et son étude demande des moyens considérables notamment en ce qui concerne l'élaboration des bases de données. Quelques résultats très partiels la concernant seront mentionnés dans le paragraphe suivant.

Par ailleurs, ces résultats peuvent paraître assez vagues et imprécis mais ils sont d'ordre général et sont quasi indépendants du locuteur. En réalité, comme nous le verrons dans le paragraphe 4.2.2, les manifestations acoustiques de la prosodie sont variables d'un locuteur à l'autre.



### 3.2. Le style

De par sa définition, la prosodie d'un énoncé varie avec son style. Dans une phrase lue, les mots prosodiques, délimités par les contours mélodiques, s'identifient aux mots lexicaux alors qu'en parole spontanée, plus rapide, ils ont tendance à recouvrir plusieurs mots lexicaux voire une subordonnée [Vaissière 80]. Les pauses représentent 50% des énoncés informels et seulement 20% des énoncés formels. En revanche, l'allongement, qu'il soit en fin de groupe de phonation ou lié à l'accent, l'isochronie et les effets consonantiques de durée sur les voyelles précédentes sont beaucoup plus fréquents dans les phrases lues que dans les dialogues réels [O'Shaughnessy 87]. Par ailleurs, une étude réalisée en suédois révèle qu'un changement de style de parole (de la parole continue au mot isolé) modifie les durées absolues des phonèmes mais pas leurs durées relatives (phonèmes longs versus phonèmes courts) [Fant 90a]. L'influence du débit d'élocution intervient à tous les niveaux et fera donc l'objet d'un paragraphe spécial.

D'après F. Nolan [Nolan 83], W. Labov a mis en évidence, dans une étude sur l'américain, le remplacement de certains phonèmes ou allophones par d'autres au fur et à mesure que le style de parole devient plus formel. Les différents degrés de formalisme des énoncés s'échelonnent de "informel" pour la conversation usuelle à "très formel" pour la prononciation de listes de paires minimales de mots. Les exemples fournis par F. Nolan sont le remplacement du / t / de "thing" par / θ / et la prononciation du / r / postvocalique qui est souvent éliminé dans le style informel. W. Labov a également montré que ce phénomène n'atteint pas uniformément tous les mots

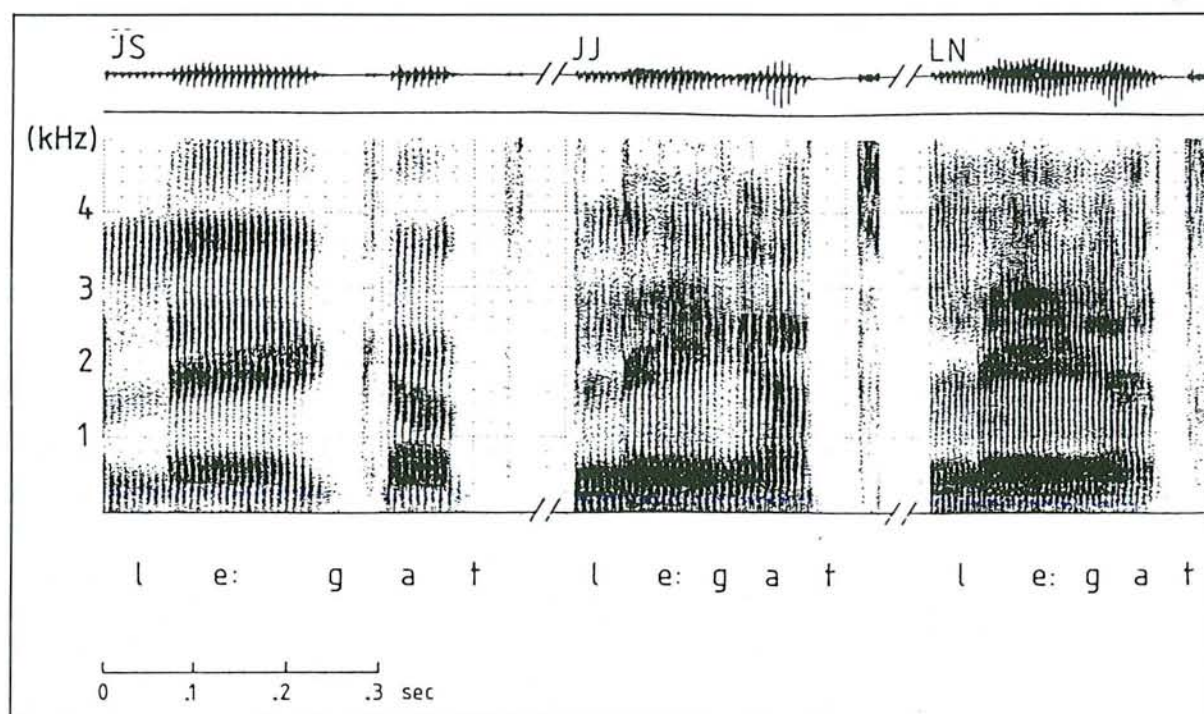


Figure A.47. Le mot "legat" prononcé par trois locuteurs suédois avec trois accentuations différentes du / g /, d'après [Fant 90b].

concernés et qu'il n'est donc pas possible de prédire, pour un style et un locuteur prédéfinis, la prononciation de tous les mots concernés à partir de quelques exemples. Par ailleurs, les résultats de W. Labov ont souligné également la forte imbrication entre la variabilité stylistique et la variabilité sociolinguistique puisqu'elles influencent les mêmes paramètres. Ainsi, il est difficile à partir de ces paramètres de différencier un locuteur de classe moyenne, employant un style formel, d'un locuteur de classe plus élevée, employant un style informel.

En ce qui concerne les formants des voyelles, les conclusions des scientifiques sont contradictoires. W. Labov, cité par P. Ladefoged, relève, dans une étude sur les locuteurs new-yorkais, des décalages systématiques entre le timbre des voyelles d'un discours formel et celui des voyelles prononcées dans un cadre informel. P. Ladefoged [Ladefoged 76] ne constate pas de telles modifications dans une étude portant sur six mots monosyllabiques : *"bee, bow, boy, bed, bad, et bud"* prononcés par neuf locuteurs californiens issus de la même classe socioculturelle, dans six styles de parole différents : dialogue réel, rimes, phrases spontanées, phrases répétées après un interlocuteur, phrases lues et liste de mots. Son étude révèle une plus grande différence de timbre entre les occurrences d'une même voyelle en milieu et en fin de phrase, pour un même style, qu'entre les occurrences de cette voyelle situées au même endroit mais dans des styles distincts. Enfin, les résultats numériques de cette étude montrent que sur les deux tiers des mesures, l'écart-type des fréquences formantiques au centre des voyelles<sup>1</sup> est inférieur à 40 Hz pour F<sub>1</sub> et 90 Hz pour F<sub>2</sub>. Or ces valeurs sont du même ordre de grandeur que les erreurs estimées dans la mesure ou le calcul des fréquences formantiques [Monsen 83].

### 3.3. Le débit d'élocution

Il est difficile de classer la variabilité du signal de parole résultant du débit d'élocution. En effet, celui-ci dépend à la fois du style de parole, de l'identité du locuteur et de l'état dans lequel il se trouve. Aussi avons-nous décidé de lui consacrer un paragraphe à part entière situé à la frontière des paragraphes dédiés à la variabilité d'origine linguistique et à la variabilité liée au locuteur.

La variabilité du débit d'élocution se répercute davantage sur les pauses que sur le temps d'articulation des phonèmes (temps de parole diminué de la durée cumulée des pauses) [Dujour 90].

Lorsque le débit d'élocution est plus lent que la normale, 80% des accroissements de durée concernent les pauses : 25% d'augmentations des pauses existantes et 55% de nouvelles pauses. Dans l'autre sens, c'est-à-dire si le débit est plus rapide que la normale, toutes les durées diminuent de 30% [O'Shaughnessy 87]. Par ailleurs, dans une étude comparant cinq locutrices à débit rapide à cinq locutrices à débit lent [Cooper 83], W. Cooper constate qu'une "locutrice lente" fait trois fois plus de pauses qu'une "locutrice rapide" et que ses pauses durent deux fois plus longtemps alors que l'accroissement de la durée totale d'une phrase ne dépasse pas 25 %. En outre, que la locutrice soit "lente" ou "rapide", lorsqu'elle ralentit son débit, elle double le nombre de pauses et leur durée triple. En revanche, lorsqu'elle accélère son débit, il y a très peu de différences. Dans cette étude, lorsqu'une locutrice "rapide" ralentit son débit naturel, celui-ci demeure plus rapide que le débit naturel d'une locutrice "lente". La réciproque n'est pas vraie. Ceci laisse supposer qu'un locuteur rapide conservera cette spécificité même dans le cadre d'un dialogue homme-machine.

<sup>1</sup> Le centre de la voyelle signifie ici le milieu de la partie la plus stable.



En ce qui concerne le temps d'articulation des phonèmes, les variations du débit d'élocution affectent bien sûr la durée des phonèmes proprement dite mais aussi tous les phénomènes de coarticulation décrits au paragraphe 2 ainsi que certains faits phonologiques comme l'élision du schwa (/ bɛtəmā / > [ bɛtmā ]), l'élision de la liquide de la paire finale /obstruante-liquide/ (/ pœpləfRāsɛ / > [ pœpfRāsɛ ] ou la fusion vocalique (/ tɔbaavɛk / > [ tɔba:vɛk ])<sup>1</sup>.

Les changements de durée des phonèmes concernent essentiellement les syllabes accentuées [Cooper 83] [Fant 90a]. Du point de vue de la coarticulation, les cibles articulaires, qu'elles soient vocaliques ou consonantiques, sont moins souvent atteintes en parole rapide qu'en parole normale ou lente [Vaissière 84], et les phénomènes d'assimilation, d'élision et de fusion sont beaucoup plus fréquents. Toutefois, un débit naturellement rapide n'est pas synonyme d'une articulation négligée. Dans une étude menée par A. Lacheret-Dujour [Dujour 90], le locuteur qui parle le plus rapidement est aussi celui qui prononce tous les phonèmes. De plus, elle constate que le nombre d'élisions de phonèmes en débit accéléré est à peu près le même qu'en débit normal, alors que dans le cas du débit ralenti, les locuteurs ont tendance à prononcer tous les phonèmes et en particulier tous les schwas. Par ailleurs, il semblerait que les faits de coarticulation soient plus sensibles au débit naturel du locuteur qu'au débit forcé (ralenti, normal ou accéléré) : quel que soit leur débit forcé, le phénomène de palatalisation en américain (cf. paragraphe 2.6) est plus important pour les locutrices au débit naturellement rapide que pour les locutrices "lentes" [Cooper 83].

Le débit d'élocution influe aussi sur les autres paramètres prosodiques. Comme nous l'avons mentionné au paragraphe précédent, lorsque le débit augmente, les marques lexicales représentées par les frontières prosodiques disparaissent au profit des marques syntaxiques [Vaissière 84].

Dans les langues à accent d'intensité, les pics de la fréquence fondamentale associés aux syllabes accentuées sont plus élevés pour les locuteurs au débit naturellement rapide que pour ceux au débit naturellement lent mais les pentes des contours mélodiques restent conservées. En revanche, lorsqu'un locuteur accélère son débit, les pics de F<sub>0</sub> sont aussi plus élevés mais les pentes du contour mélodique sont plus raides [Cooper 83].

## 4. Variabilité liée au locuteur

### 4.1. Introduction

Les répétitions d'un même énoncé ne sont pas acoustiquement identiques, qu'elles aient été prononcées par plusieurs locuteurs (variabilité interlocuteur) ou par le même locuteur (variabilité intralocuteur).

Une première approche, plutôt intuitive, consiste à séparer la variabilité interlocuteur issue des différences anatomiques existant entre les appareils phonatoires des locuteurs de celle qui résulte des habitudes articulatoires et linguistiques de chacun d'eux. En fait, ces deux sources de variabilité ne sont pas indépendantes puisque, si les habitudes linguistiques d'un locuteur dépendent de sa situation géographique et de son milieu socioculturel, certaines de ses habitudes articulatoires découlent aussi des particularités anatomiques de son appareil

<sup>1</sup> Le symbole : indique un allongement.



phonatoire. Néanmoins, pour des raisons de simplicité, nous avons conservé ce découpage traditionnel dans notre présentation des différentes sources de variabilité interlocuteur.

La variabilité intralocuteur est aussi une association de variabilités d'origines diverses. Mais le nombre et le type des sources de variabilité intralocuteur dépendent du style de parole considéré : parole lue, dialogue homme-machine ou dialogue naturel.

Nous allons examiner dans les paragraphes suivants de nombreuses sources de variabilité interlocuteur et intralocuteur.

## **4.2. Variabilité interlocuteur**

### **4.2.1. Différences physiologiques**

#### *a) Introduction*

La diversité des appareils phonatoires des locuteurs ne se répercute pas sur le signal de parole directement sous la forme de valeurs de paramètres acoustiques, mais plutôt sous la forme de plages de variation dans lesquelles les paramètres acoustiques sont contraints à rester. Les différences physiologiques se scindent en deux catégories qui sont d'une part les différences statiques ou anatomiques et d'autre part les différences dynamiques.

Ces dernières se rapportent à la commande neuromusculaire des articulateurs. Même si ces différences dynamiques sont très peu connues, l'inégalité des aptitudes des individus dans des domaines comme les travaux manuels ou la pratique d'un instrument de musique permet de supposer que les locuteurs sont plus ou moins agiles dans le domaine de la production de la parole. Le débit d'élocution, que nous déjà traité dans un paragraphe précédent, est une des manifestations de ces différences dynamiques.

Les différences statiques concernent l'appareil respiratoire inférieur, les cordes vocales, la taille et la forme des conduits oral et nasal, l'état de leurs parois et la position relative des articulateurs. Dans le chapitre II, nous avons mis en évidence une partie des différences anatomiques des appareils phonatoires des locuteurs ainsi que quelques-unes de leurs conséquences acoustiques. Nous allons compléter cette description en séparant les variations acoustiques relatives à l'anatomie du larynx et de l'appareil respiratoire inférieur de celles qui se rapportent à l'anatomie des cavités supraglottiques.

#### *b) Le larynx et l'appareil respiratoire inférieur*

La pression sous-glottique, les caractéristiques des cordes vocales et l'effort vocal influent sur la fréquence et la forme de l'onde glottale. Certaines influences ont été mises en évidence lors de la description du processus de phonation et des caractéristiques de l'onde glottale dans le chapitre II. D'autres sont introduites seulement maintenant pour deux raisons, pour ne pas surcharger cette première description et parce qu'elles utilisent des notions définies après le chapitre II.

Le volume des poumons et leur élasticité influent directement sur la pression sous-glottique et sur sa vitesse de variation, par conséquent sur la valeur moyenne de  $F_0$  et sur son domaine d'évolution. De plus, l'impédance présentée par l'appareil respiratoire inférieur est responsable des fluctuations de la pression au cours d'une période et donc de la forme de l'onde glottale [Stevens 77].

En ce qui concerne les cordes vocales, des modifications de leur structure comme une grosseur, une paralysie partielle ou un changement d'état de la muqueuse provoquent des irrégularités dans la forme de l'onde glottale au cours d'une période et entre périodes [Williams 72].

Les effets de l'effort vocal sur la forme de l'onde glottale s'observent par les modifications subies par la pente de son spectre : l'atténuation des fréquences élevées étant plus importante pour les voix faibles que pour les voix fortes [Lonchamp 87b]. La figure A.48 illustre cette propriété. Dans le cas de la voix intense, les amplitudes des formants F2 et F3 sont supérieures à celle de F1 alors que le contraire s'observe pour la voix faible.

Par ailleurs, plusieurs études ont montré que la pente du spectre de l'onde glottale n'est pas la même sur tout le spectre mais qu'elle varie avec la fréquence et que l'ampleur des variations (de  $\pm 10$  dB à  $\pm 15$  dB) est une fonction du locuteur. En particulier, chez les locutrices, le spectre a tendance à chuter brusquement aux environs de 1000 Hz alors que cette rupture de pente se situe vers 2000 Hz pour les locuteurs [Fant 90b]. La figure A.49 montre le résultat d'une estimation de la pente du spectre de l'onde glottale par J. Mártony pour des locuteurs suédois.

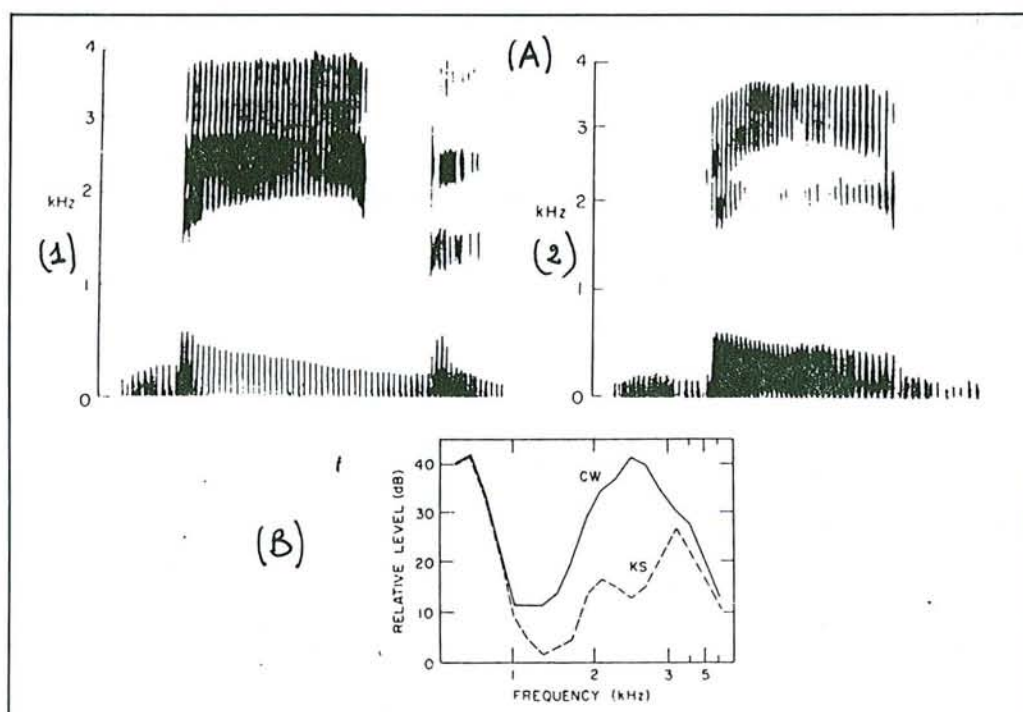


Figure A.48. Les effets de l'effort vocal sur le spectre des voyelles.  
(A) : Spectrogramme du mot anglais "bib", (1) dans le cas d'une voix intense, (2) dans le cas d'une voix faible ; (B) : spectre instantané du [i] pour les deux voix ; d'après [Stevens 77].



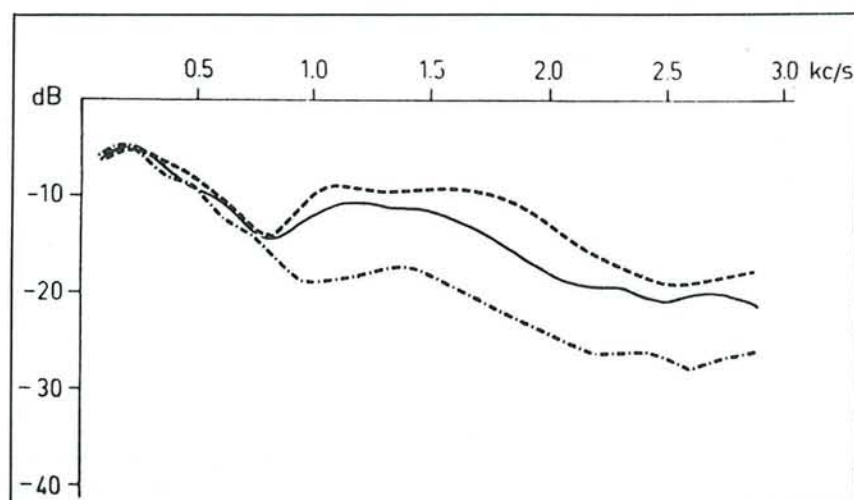


Figure A.49. Exemples de pente spectrale moyenne estimée par J. Mártony pour des locuteurs suédois, après une préaccentuation de + 12 dB par octave, dans [Stevens 77].

### c) Les cavités supraglottiques

La répartition spectrale de l'énergie du signal de parole et plus particulièrement les fréquences formantiques des sons voisés dépendent de la taille et de la forme des cavités orales et nasales. Quelle que soit la langue, la variabilité interlocuteur des formants des voyelles orales a fait l'objet de nombreuses études. Nous allons décrire l'une d'entre elles menée par F. Lonchamp à l'Institut de Phonétique de Nancy [Lonchamp 87b].

La figure A.50 présente les résultats de cette étude réalisée sur 10 locuteurs et 9 locutrices à partir d'un corpus que nous avons enregistré au laboratoire et qui est constitué de deux répétitions des voyelles [ i, ε, a, ɔ, œ ] dans le contexte [ p—R ]<sup>2</sup> et de [ e, o, u, y, ø ] dans le contexte [ p— ]. Si, à contexte identique, les points représentant les voyelles orales d'un locuteur dans le plan (F<sub>1</sub>, F<sub>2</sub>) sont tous clairement espacés, il n'en est pas de même des nuages de points représentant les voyelles d'un ensemble de locuteurs du même sexe. La confusion entre les domaines de dispersion des différentes voyelles est encore plus importante lorsqu'on considère les productions des locuteurs des deux sexes. Cette superposition des polygones relatifs à chaque locuteur montre l'importance des études sur la normalisation des fréquences formantiques dans le cadre de la reconnaissance multilocuteur [Fant 90b] [Bonneau 89].

Ces études cherchent en particulier à établir des facteurs d'échelle simples entre les fréquences formantiques des locuteurs et celles des locutrices mais ceux-ci diffèrent selon la voyelle et le formant considérés. A un conduit vocal plus court correspondent, en première approximation, des fréquences de résonance plus élevées. Les conduits vocaux féminins sont plus courts que les masculins. Mais le rapport des longueurs, dont la valeur moyenne est 15%, n'est pas uniforme sur tout le conduit : il est plus important au niveau du pharynx. Toutefois, ceci ne suffit pas à expliquer la disparité des facteurs d'échelle entre les voyelles. La figure A.51 fournit, pour chaque voyelle, les rapports  $k_1$ ,  $k_2$  et  $k_3$  des fréquences formantiques F<sub>1</sub>, F<sub>2</sub> et F<sub>3</sub> des locutrices à celles des locuteurs pour trois études : celle que nous venons de décrire, une

<sup>2</sup> [ p—R ] = [ p voyelle R ].



étude multilingue de G. Fant et une simulation de H. Traunmüller [Lonchamp 87b]. Outre les rapports des longueurs des conduits vocaux, H. Traunmüller fait intervenir l'ouverture linguale dans ses calculs des facteurs d'échelle.

Les largeurs de bande des formants, qui sont proportionnelles aux pertes d'énergie dans le conduit vocal aux fréquences formantiques (pertes à la glotte, aux parois et aux lèvres), sont plus grandes (environ 50%) chez les locutrices [Fant 90b].

En dehors de leurs dimensions proprement dites et de leurs formes générales, les conduits vocaux des locuteurs se distinguent par les détails morphologiques des organes qui les constituent. Ces différences influent aussi sur la répartition spectrale de l'énergie de certains sons. Ainsi, d'après K. Stevens, les différences observées entre les spectres des constrictives et des voyelles antérieures fermées de plusieurs locuteurs proviendraient à la fois de la variabilité interlocuteur de la zone prépalatale et de la forme de la langue derrière la constriction [Stevens 77].

Par ailleurs, nous avons souligné, dans le paragraphe II.4.4, combien les dimensions et les configurations internes des cavités nasales et des sinus paranasaux sont variables d'un locuteur à l'autre. Cette variabilité se traduit par une variabilité du spectre des voyelles et des consonnes nasales, aussi bien au niveau des antiformants et des formants d'origine nasale, qu'au niveau de l'influence du couplage nasal sur les formants d'origine orale. Ces différences anatomiques ont l'avantage de ne pas varier à court terme et de ne pas être modifiables consciemment (imitations).

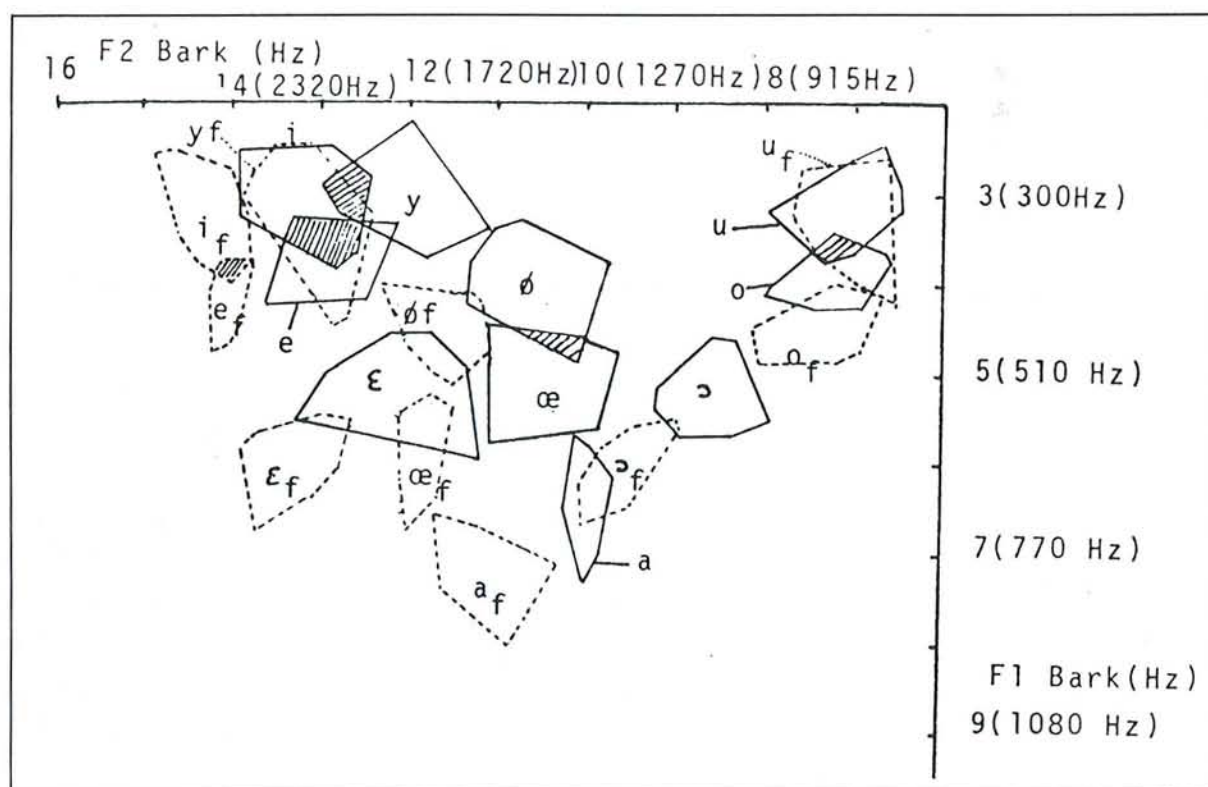


Figure A.50. Zones de dispersion des voyelles orales du français dans le plan  $F_1$ - $F_2$  (échelle Bark), en traits pleins pour les locuteurs, en traits pointillés pour les locutrices. Les hachures délimitent les zones de recouvrement entre voyelles pour un même sexe, d'après [Lonchamp 87b].

En revanche, elles ont l'inconvénient d'être sensibles à l'état pathologique du locuteur (coryzas, sinusites, allergies, ...). La figure A.52, issue d'une étude réalisée par M.R. Sambur [Sambur 75], permet de comparer les spectres de deux [ɲ], l'un est prononcé dans un état normal, l'autre avec un léger rhume.

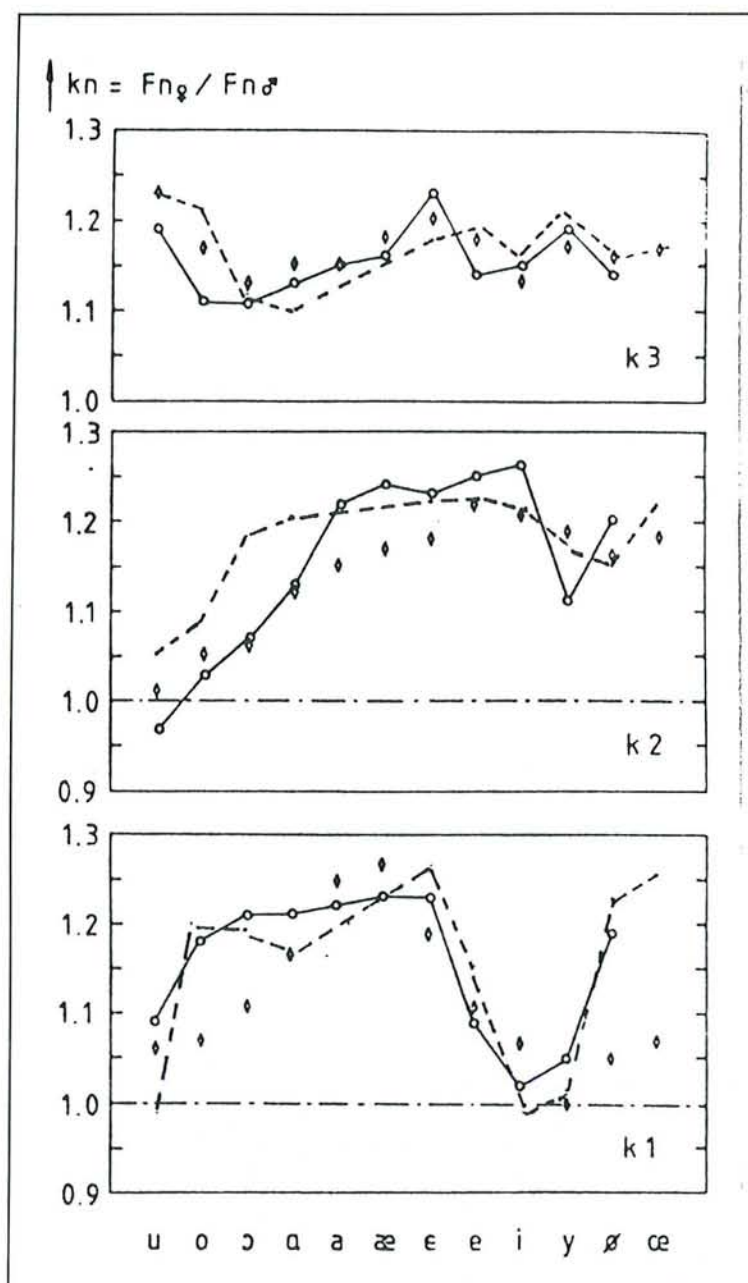


Figure A.51. Les facteurs d'échelle entre les fréquences formantiques féminines et masculines des voyelles orales ; en traits pointillés pour l'étude de F. Lonchamp, avec des losanges pour les données de G. Fant et en traits pleins pour la simulation de H. Traunmüller ; d'après [Lonchamp 87b].

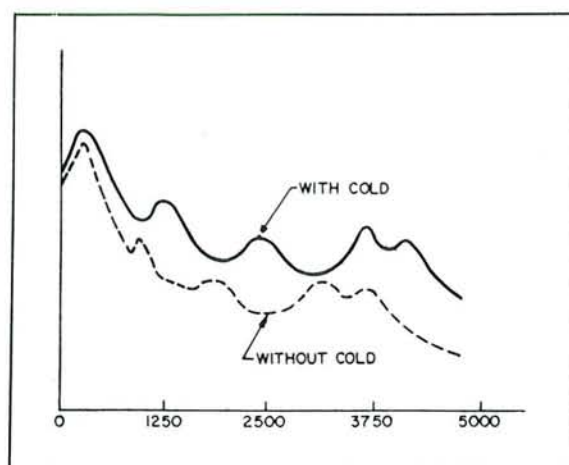


Figure A.52. Comparaison des spectres LPC de deux [n], l'un est prononcé dans un état normal, l'autre avec un léger rhume, d'après [Sambur 75].

#### d) Conclusion

Toutes les manifestations acoustiques des différences anatomiques, que nous avons mentionnées tout au long de cette partie, ne sont pas exploitables facilement pour la reconnaissance du locuteur. Nous avons déjà souligné le fait que ces manifestations acoustiques sont plutôt des domaines de variation dans lesquels le locuteur va positionner ses paramètres acoustiques en fonction de ses habitudes linguistiques mais aussi en fonction d'informations paralinguistiques conscientes ou inconscientes, et qui font intégralement partie de l'acte de communication. Il faut également tenir compte de la plasticité de l'appareil vocal, sauf au niveau des cavités nasales et des sinus. Cette plasticité permet au locuteur de produire des paramètres acoustiques presque normalisés malgré sa spécificité anatomique. Il peut par exemple obtenir une valeur de  $F_1$  moins élevée en abaissant son larynx. Toutefois, l'appareil vocal étant constitué d'entités anatomiques interdépendantes, le locuteur ne peut agir de façon indépendante sur tous les paramètres. Dans le dernier exemple, l'abaissement du larynx entraîne aussi une baisse de  $F_0$ , ce qui montre que toutes les manifestations acoustiques des différences anatomiques ne sont pas simultanément imitables ou modifiables.

#### 4.2.2. Idiomes et habitudes linguistiques

Les membres d'une communauté linguistique ne parlent pas tous de façon identique. Sans aller jusqu'à la notion de dialecte, il existe entre les parlers des locuteurs des divergences linguistiques qui n'affectent pas leur compréhension mutuelle mais qui reflètent leur localisation géographique, leur origine socioculturelle et leurs habitudes linguistiques personnelles. Ces variations idiomatiques s'appliquent à tous les niveaux de la communication orale et notamment à la prosodie, aux variantes lexicales, aux variantes phonologiques et à leurs réalisations phonétiques.



a) Variabilité interlocuteur des paramètres prosodiques

Le débit d'élocution et les variations de  $F_0$  ne sont pas seulement fonction du style de parole et de l'anatomie du locuteur mais aussi de sa personnalité. Certains locuteurs parlent naturellement plus vite ou plus lentement que d'autres. Les manifestations de la variabilité du débit d'élocution ont été présentées au paragraphe 3.3. Ainsi que nous pouvons le vérifier sur la figure A.53, G. Fant constate qu'en parole lue, le taux de pause à long terme (pourcentage de la durée des pauses sur la durée totale d'un énoncé) se stabilise rapidement et constitue un bon indice du locuteur. A. Lacheret-Dujour constate le même phénomène dans son étude sur quatre locuteurs (3F et 1H). Pour un débit donné (lent, normal ou rapide), le temps d'articulation d'un énoncé est beaucoup moins variable d'un locuteur à l'autre que le temps de pause (8% versus 70% en débit normal) [Dujour 90].

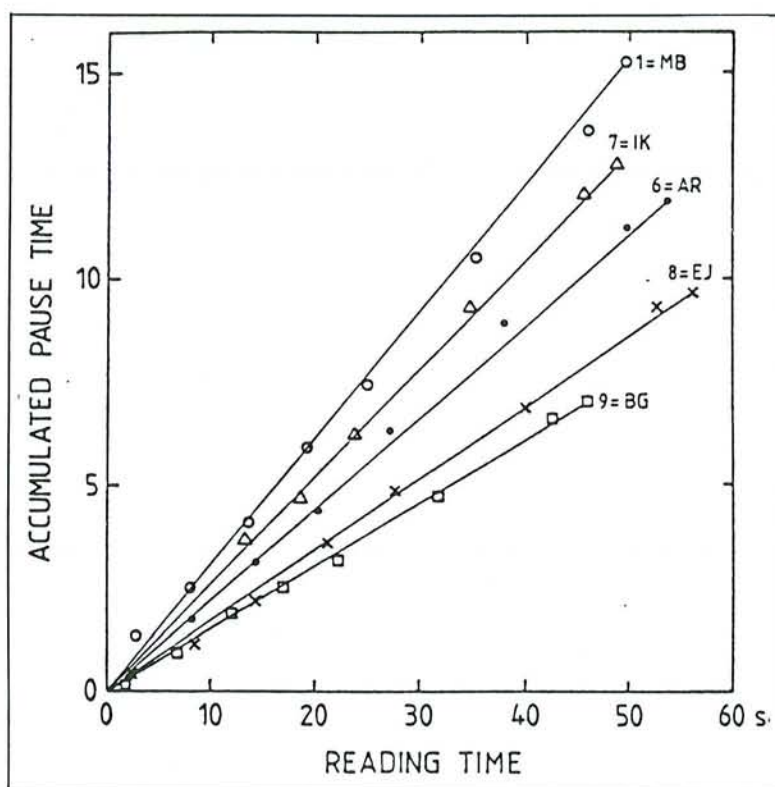


Figure A.53. Temps de pause accumulé par rapport au temps de parole lue pour cinq locuteurs, d'après [Fant 90b].

La fréquence fondamentale est le paramètre prosodique le plus sensible à la variabilité idiomatique. Ainsi, il existe des sujets qui s'expriment naturellement avec un contour mélodique moyen relativement plat alors que, pour d'autres, il est beaucoup plus nuancé et emphatique [Vaissiere 84]. Par ailleurs, les contours mélodiques associés aux mots prosodiques (cf. paragraphe 3.1) ne sont pas les mêmes pour tous les locuteurs [Vaissiere 88]. Ces variabilités interlocuteur de la durée et de la fréquence fondamentale, auxquelles s'ajoute la variabilité de l'énergie articulatoire, se retrouvent aussi dans la réalisation de l'accent [Nolan 83].

Comme le montre la figure A.47, le degré de contraste entre les syllabes accentuées et les syllabes inaccentuées varie selon le locuteur : l'articulation du troisième / g / semble complètement relâchée alors que, perceptivement, il est accentué.

Plus généralement, dans la plupart des langues, la manière de réaliser une frontière syntaxique change avec le locuteur : insertion d'une pause, allongement de la dernière syllabe du groupe de phonation, diminution de la fréquence de vibration des cordes vocales qui peut aller parfois jusqu'à la friture vocale, ou tout simplement rien.

#### b) Différences lexicales

Les locuteurs ne prononcent pas de façon identique tous les mots du dictionnaire. Plusieurs de ces différences présentent une certaine régularité et peuvent s'exprimer sous la forme de règles phonologiques, ou tout du moins, des recherches tentent de le faire [Dujour 90] [Perennou 89]. Ces différences sont présentées dans les paragraphes suivants. D'autres sont singulières, elles ne concernent que quelques mots particuliers. Voici quelques-unes de ces particularités :

- la gémiation des consonnes en milieu de mot :  
"immortel" peut se prononcer [ **imɔRtɛl** ] ou [ **immɔRtɛl** ] ;
- la prononciation facultative de certaines consonnes en fin de mot :  
"fourmil" peut se prononcer [ **fuRnɪl** ] ou [ **fuRni** ] ;
- la prononciation des mots d'origine étrangère :  
"joker" peut se prononcer [ **ʒɔkɛR** ] ou [ **dʒɔkɛR** ].

#### c) Différences phonologiques

Le système phonologique d'une langue se fonde sur un ensemble maximal de phonèmes correspondant à un certain nombre d'oppositions phonologiques. Mais toutes les personnes parlant cette langue n'utilisent pas nécessairement toutes ces oppositions. Ainsi, un locuteur écossais peut n'avoir qu'un seul / u / alors que l'anglais standard (*Received Pronunciation*) possède l'opposition / u : / ~ / / ("good" ~ "food") [Nolan 83]. De même, le français standard comporte l'opposition / ɛ̃ / ~ / œ / alors que la plupart des Parisiens ne la réalisent plus.

Par ailleurs, même lorsqu'un phonème est réalisé par un ensemble de locuteurs, ceux-ci peuvent se différencier dans la manière d'utiliser ce phonème. Tout d'abord, ils n'emploient pas toujours le même phonème dans le même contexte phonologique. Ainsi, selon la variété d'anglais, le phonème / r / est prononcé ou ne l'est pas lorsqu'il précède une consonne ou une pause ("fierce", "car"). De la même façon, en français, le triplet / occlusive-R-ə / situé devant une pause se réalise, selon les locuteurs, par / occlusive /, / occlusive-R / ou / occlusive-R-ə /. Nous pouvons également classer dans cette catégorie l'élision du schwa dont le taux sur un texte dépend plus du locuteur que du débit d'élocution imposé [Dujour 90].

Enfin, il existe pour chaque individu un degré de fréquence d'emploi d'un phonème par rapport à l'ensemble du vocabulaire (tous les mots, certains mots savants ou étrangers, mots possédant une graphie particulière, ...) [Martinet 73] [Walter 76]. Par exemple, certains locuteurs parisiens n'utilisent le phonème / œ / que dans des noms scientifiques comme "tungstène" ou "alun".

Les locuteurs se différencient aussi par leur stratégie d'utilisation de ces variantes entre deux occurrences rigoureusement identiques de la même situation phonologique. Cette stratégie



peut être répétitive, varier selon le style ou le débit d'élocution ou bien être complètement aléatoire [Dujour 90].

Pour terminer ce paragraphe, nous indiquons quelques résultats que nous avons extraits d'une étude réalisée par Henriette Walter sur la dynamique du système phonologique du français contemporain [Walter 76]. Elle est issue d'une enquête effectuée en 1971 auprès de 17 locuteurs parisiens âgés de 22 à 73 ans qui proviennent d'un milieu socioculturel élevé. Elle montre que sept oppositions ne sont pas utilisées unanimement par ces locuteurs. Avant de détailler les résultats de cette étude, nous donnons quelques précisions concernant leur établissement. Rappelons tout d'abord qu'il y a neutralisation d'une opposition entre deux phonèmes, soit lorsqu'un des deux phonèmes disparaît au profit de l'autre, soit lorsque l'un ou l'autre phonème est employé indifféremment dans un mot sans que cela modifie le sens du mot. La neutralisation (ou l'existence) d'une opposition est un phénomène qui n'est ni binaire ni statique. Pour chaque opposition, l'auteur effectue le calcul des pourcentages, indiqués dans les résultats, sur l'ensemble des mots pour lesquels au moins un locuteur a utilisé le phonème qui tend à disparaître. Les locuteurs qui réalisent l'opposition sont ceux qui ont employé ce phonème dans plus d'un tiers des mots de cet ensemble. Réciproquement, les locuteurs qui ne réalisent pas l'opposition sont ceux qui ont employé ce phonème dans moins d'un tiers des mots de l'ensemble.

Les sept oppositions étudiées sont les suivantes :

- l'opposition / a / ~ / ɑ / ("*patte*" ~ "*pâte*"). Elle est employée essentiellement dans des mots monosyllabiques. Dans ce cas, 82% des locuteurs la réalisent mais ils ont une moyenne d'âge de 49 ans. Elle tend à disparaître pour 18% des locuteurs dont la moyenne d'âge est 30 ans. Seuls quatre mots sont prononcés unanimement avec un / ɑ / : âtre, bât, mâle et pâte ;
- l'opposition / ẽ / ~ / œ / ("*brin*" ~ "*brun*"). Elle est conservée par la moitié des locuteurs mais ceux-ci ont une moyenne d'âge de 51 ans alors que l'autre moitié a une moyenne d'âge de 38 ans. Il faut noter une très grande stabilité des locuteurs qui ont tendance à prononcer tous les mots communs soit avec un / ẽ / soit avec un / œ /. En revanche, aucun de ces mots n'a été prononcé unanimement avec / œ /.
- l'opposition / ɛ / ~ / ɛː / ("*faite*" ~ "*faîte*"). Elle n'est plus réalisée qu'en syllabe finale fermée<sup>1</sup> par 30% des locuteurs dont la moyenne d'âge est de 53 ans.
- l'opposition / œ / ~ / ø / ("*veulent*" ~ "*veule*", "*jeune*" ~ "*jeûne*"). L'opposition n'existe plus que pour ces deux paires de mots et pas pour tous les locuteurs. Notons qu'en syllabe non finale ouverte ("*malheureusement*"), l'opposition n'est pas pertinente et que la réalisation de l'archiphonème en / œ / ou / ø / dépend de la prononciation du radical ("*malheur*", "*malheureux*") [Lonchamp 87a], du locuteur et de l'harmonisation vocalique.
- l'opposition / o / ~ / ɔ / ("*maure*" ~ "*mort*"). Son statut dépend du contexte :
  - en syllabe finale ouverte, elle est neutralisée en / o /. Les locuteurs ne font plus la distinction entre "*pot*" et "*peau*",
  - en syllabe finale fermée par / z /, elle est neutralisée en / o /. Les mots "*morose*" et "*pause*" possèdent le même / o / final,
  - en syllabe finale fermée par une autre consonne, l'opposition est conservée sauf pour les syllabes terminées par / m /, / n /, / r / ou / s /, pour lesquelles les locuteurs ont des

<sup>1</sup> Une syllabe ouverte est une syllabe qui ne se termine pas par une ou plusieurs consonnes : dans / *sistematik* / les deuxième et troisième syllabes sont des syllabes ouvertes, les autres des syllabes fermées.



prononciations divergentes. Pour prononcer *“amazone”*, onze locuteurs emploient / o / et six / ɔ /,

– en syllabe non finale, l'opposition est encore bien établie. Il y a peu de mots dont les prononciations diffèrent selon le locuteur. Parmi ceux-ci, figurent les mots à préfixe (*“rhino-”, “socio-”, ...*) dont la prononciation est / o / si le locuteur a tendance à séparer le préfixe du radical. Cette tendance est une fonction de la modernité du mot et de son caractère scientifique. Par ailleurs, trois locuteurs réalisent dans certains mots comme *“jolie”* une voyelle plus antérieure que [ ɔ ] et qui se note [ ɔ̃ ], pour indiquer que c'est un [ ɔ ] centralisé ;

• l'opposition / e / ~ / ɛ / (*“gué”* ~ *“gai”*). Cette opposition dépend également du contexte phonologique dans lequel elle intervient :

– en syllabe finale ouverte, l'opposition se maintient contrairement aux oppositions précédentes. Les locuteurs respectent la graphie (*“ai”* et *“et”* → / ɛ / ; *“e”, “er”* et *“es”* → / e / ). Toutefois, les réalisations des locuteurs divergent sur quelques mots comme *“quai”* et *“gai”*. Les locuteurs parisiens prononcent / e / la conjonction de coordination *“et”*, alors que ceux qui sont originaires du Sud de la France la prononcent / ɛ / [Walter 76] [Dujour 90],

– en syllabe finale fermée, l'opposition se neutralise en / ɛ / (/ ləkɛl / ),

– en syllabe non finale ouverte, les graphies en *“ai”, “ei”* et *“ê”* sont prononcées majoritairement en / e / , ce qui semble indiquer une forte harmonisation vocalique (*“abêtissant”* → / abetisã / ),

– en syllabe non finale fermée, les résultats sont surprenants. Si les syllabes terminées par / R / et / l / se prononcent avec un / ɛ / , celles terminées par une autre consonne sont articulées avec un / e / par 40% des locuteurs avec une moyenne d'âge de 40 ans (*“escargot”* → / eskaRgo / ). Ceci indique un recul de la neutralisation en / ɛ / ;

• l'opposition / p / ~ / n̥j / (*“pagne et”* ~ *“panier”*). A. Martinet, co-auteur avec H. Walter du dictionnaire dont est issue cette étude [Martinet 73] a examiné plus précisément l'état de cette opposition. Il considère qu'en position intervocalique (*“agneau”*) seuls cinq locuteurs, dont l'âge moyen est 51 ans, conservent cette opposition. Tous les autres réalisent la graphie *“gn”* par / n̥j / . D'après lui, cette neutralisation s'explique en partie par l'isolement de la consonne nasale qui ne correspond à aucune consonne palatale orale. 65% des locuteurs prononcent encore / p / , en finale de mot (*“charogne”*), et 96%, avant une consonne (*“éloignement”*).

#### d) Habitudes articulatoires

Même si les systèmes phonologiques des locuteurs sont identiques, chaque locuteur dispose d'une certaine latitude articulatoire pour produire les réalisations phonétiques associées à chaque phonème. Cette latitude permet au locuteur d'adapter les possibilités de son appareil vocal aux habitudes linguistiques de son groupe sociogéographique, afin de produire un son que l'auditeur ne risque pas de confondre avec la réalisation phonétique d'un autre phonème dans le même contexte. Les résultats de cette adaptation constituent ce que nous appelons les *“habitudes articulatoires”*.

Nous avons déjà rencontré des exemples de telles habitudes dans les chapitres et les paragraphes précédents, comme la réalisation des voyelles nasales (cf. paragraphe III.2.2) ou les

variantes du phonème / r / (cf. paragraphe III.3.5). De même, la variabilité des fréquences formantiques des voyelles orales, présentée au paragraphe c, résulte des différences anatomiques des locuteurs mais aussi de la latitude que possède chaque locuteur pour "placer ses voyelles" dans le trapèze articulatoire.

Les habitudes articulatoires interviennent également dans la réalisation des phonèmes appartenant à des oppositions qui tendent à se neutraliser. Ainsi, le trait phonétique permettant d'opposer / ɑ / à / a / peut être, selon les locuteurs, une durée plus longue, une postériorisation de l'articulation, une plus grande ouverture de la mâchoire, une combinaison de plusieurs de ces traits ou bien encore une accentuation de la syllabe incluant le phonème / ɑ / [Walter 76].

Les phénomènes de coarticulation sont aussi une source de différences entre les locuteurs. Des études qui mettent en œuvre la variabilité interlocuteur de la réduction du geste articulatoire et de la superposition de mouvements articulatoires seront présentées dans le chapitre III de la partie B. Nous présentons ici quelques exemples qui concernent les phénomènes d'assimilation. Dans son étude sur les variantes phonologiques des 30 locuteurs qui ont lu le texte "La bise et le soleil" du corpus BDSONS, A. Lacheret-Dujour remarque que certains locuteurs affectionnent particulièrement l'harmonisation vocalique et l'assimilation de nasalisation, alors que d'autres ne les utilisent jamais [Dujour 90]. Par ailleurs, Henriette Walter [Walter 76] relève, dans le corpus de phrases lues par 17 locuteurs, 76% d'assimilation régressive de sonorité (voisement ou dévoisement) lorsque les deux consonnes sont adjacentes. Lorsqu'elles sont séparées par un schwa, elle note 32% d'assimilation vers le voisement et 27% vers l'assourdissement. De plus, elle remarque 10% d'assimilation régressive dans des cas où elle ne devrait pas exister, c'est-à-dire lorsque la deuxième consonne est une sonante. En revanche, le taux d'assimilation de nasalisation se limite à 11%. Parmi ces assimilations de nasalisation, figurent deux exceptions à la règle énoncée page 65 : / ɛdəmɛmwaR / > [ ɛnmɛmwaR ] et / ɛdɛmɪ / > [ ɛnmɪ ]. Globalement, la faiblesse des taux d'assimilation obtenus en présence d'un schwa est certainement due au fait que le corpus est lu.

#### e) Conclusion

Une conclusion facile du paragraphe a) est que les faits prosodiques sont de bons indices pour l'identification du locuteur. Malheureusement, les mêmes raisons font qu'ils sont aussi mal connus et mal structurés et donc actuellement peu exploitables pour la reconnaissance du locuteur. En outre, comme nous le verrons ultérieurement, les faits prosodiques sont sujets à la variabilité intralocuteur et sont imitables.

A l'instar des faits prosodiques, certaines variantes lexicales et phonologiques paraissent très variables d'un locuteur à l'autre. Elles semblent moins sensibles à la variabilité intralocuteur sauf peut-être à une variabilité stylistique qui traduirait l'intention sociale du locuteur (vouloir adopter le parler d'une certaine classe socioculturelle). Leurs principaux défauts sont qu'elles exigent un vocabulaire particulier et qu'elles sont facilement imitables.

En revanche, les habitudes articulatoires décrites dans le paragraphe d) sont très difficilement imitables ou modifiables consciemment. Malheureusement les différences acoustiques qui en résultent sont plus faibles et donc plus sensibles à la variabilité intralocuteur ou à la précision de la mesure.



### 4.3. Variabilité intralocuteur

#### 4.3.1. Introduction

La variabilité intralocuteur est une association de sources de variabilité qui devient de plus en plus complexe au fur et à mesure que le style de parole devient de plus en plus naturel.

Dans le cas de la parole lue, les principaux facteurs susceptibles de modifier les paramètres acoustiques du signal de parole sont les changements d'états physique et psychique du locuteur, auxquels il faut ajouter une variabilité minimale naturelle.

Dans le cas du dialogue homme-machine, le style reste en général assez contraint. Mais il faut prendre en compte la variabilité due à l'influence que peut avoir sur l'énoncé d'un locuteur le fait qu'il soit absorbé dans une autre tâche.

Dans le cas du dialogue réel, l'éventail des sources de variabilité intralocuteur devient plus important. Tout d'abord, le locuteur souhaite transmettre avec son message linguistique des sentiments et une attitude vis-à-vis de ce message. Ceci peut se traduire par exemple par une modification des paramètres suprasegmentaux ou bien par une nasalisation volontaire. Par ailleurs, le locuteur prend en compte dans son discours le contexte social engendré par ses interlocuteurs, la manière dont il se situe dans ce contexte et l'image qu'il veut transmettre de lui-même. Ainsi, un locuteur ne parle pas de la même façon à un ami, à son directeur ou à son subordonné, mais cette relation dépend aussi d'un contexte plus général. Dans un pays lointain, deux Français se considèrent dans le même contexte social même si leurs origines socioculturelles sont très éloignées. Enfin, il est possible que le locuteur modifie certains paramètres acoustiques dans le but de gérer le dialogue. Un test de perception a permis à des auditeurs de reconnaître si une phrase extraite d'un texte se situait en début ou en fin de paragraphe [Nolan 83].

L'étude des conséquences acoustiques de toutes ces sources de variabilité intralocuteur est très délicate voire impossible puisque, lors des expérimentations, les sources sont très difficiles à quantifier de manière objective. Étant donné ces difficultés, nous nous limiterons dans les paragraphes suivants à ne citer que les résultats de quelques études se rapportant à la variabilité intralocuteur.

#### 4.3.2. Variabilité minimale

Lors de la production d'un son stationnaire voisé, l'onde glottale n'est pas vraiment périodique mais fluctue légèrement d'une période à l'autre. Ces fluctuations sont regroupées sous l'appellation de "*shimmer*" lorsqu'elles concernent l'amplitude et sous celle de "*jitter*" lorsqu'elles portent sur la période. Ces variations, qui dans le cas d'une voix normale atteignent au maximum 1% de la fréquence fondamentale pour le jitter et 0.2 dB pour le shimmer, ne sont pas perceptibles [O'Shaughnessy 87].

Tous les auteurs sont unanimes sur l'existence d'une autre variabilité intralocuteur minimale. En effet, les écart-types observés lors des répétitions d'un son par un locuteur au cours d'une même session, dans un contexte et un style identiques, sont de l'ordre de 5 à 10 ms pour les durées [Lehiste 75] et de 50 à 100 Hz pour les fréquences des formants F1, F2, et F3 [O'Shaughnessy 87]. Toutefois, sachant que les estimations des erreurs de mesure sur ces paramètres ont le même ordre de grandeur [Monsen 83], on peut se demander si ces variations ne sont pas simplement le reflet de ces erreurs de mesure.



### 4.3.3. Influence de l'âge

L'examen par W. Endres et al. [Endres 71] d'une série d'enregistrements de personnalités allemandes (4 hommes et 2 femmes) effectués sur une période de 29 ans, révèle une évolution avec l'âge des caractéristiques acoustiques des locuteurs. Il montre, notamment, un abaissement des fréquences formantiques des voyelles, un affaiblissement de la fréquence fondamentale moyenne pouvant aller jusqu'à 40 Hz ainsi qu'une diminution de son domaine de variation. Les auteurs corroborent leurs résultats expérimentaux par des conclusions d'études médicales.

Les amygdales palatines, linguales et pharyngiennes ainsi que les ganglions lymphatiques présentent une forte involution après la période de 35 à 40 ans, produisant un élargissement de l'oropharynx. De plus, le larynx des personnes âgées a tendance à s'affaïsser et les muscles à se relâcher, ce qui entraîne également un allongement du conduit vocal et donc un abaissement de ses fréquences de résonance.

Par ailleurs, les effets du vieillissement sur les cordes vocales se manifestent par une diminution de leur élasticité et par une augmentation de leur épaisseur suite à l'agglomération de substances lipidiques autour de leurs fibres musculaires.

Enfin, les auteurs de ces études médicales notent un accroissement avec l'âge de la raideur des articulations laryngiennes, de l'ossification de certains cartilages et de l'hypotonie des muscles du larynx, provoquant une diminution de la force vocale et de sa dynamique.

Signalons pour terminer ce paragraphe, le résultat d'une étude sur la perception de la classe d'âge de 60 locuteurs par 40 auditeurs qui met en évidence une bonne corrélation entre la classe d'âge choisie et l'âge réel du locuteur, même si celle-ci est un peu moins bonne pour les locutrices [Tielen 90].

### 4.3.4. Variabilité due à l'état émotionnel du locuteur

La difficulté de connaître l'état émotionnel réel d'un locuteur rend problématique l'étude de son influence sur les paramètres acoustiques du signal de parole. Dans ces conditions, les auteurs de recherches sur ce sujet ont recours à une (ou une combinaison) des méthodologies suivantes :

- l'analyse, à partir de suppositions sur l'état émotionnel du locuteur, de dialogues réels non reproductibles ;
- l'étude de simulations jouées par des acteurs professionnels ;
- l'étude perceptive de parole synthétisée, de dialogues réels ou de simulations.

En se fondant sur les simulations de quatre acteurs et sur un cas réel, C.E. Williams et K.N. Stevens ont étudié les répercussions sur le signal de parole de trois situations émotives, la colère, la peur et le chagrin [Williams 72].

Par rapport à une voix neutre, la voix coléreuse se caractérise par un spectre moyen plat ou montant, reflet d'une voix intense, par une fréquence fondamentale moyenne plus élevée (d'au moins une demi-octave) associée à un plus grand domaine de variation avec en particulier des pics très proéminents sur les voyelles les plus intenses. Les auteurs notent également mais de façon moins systématique : des gestes articulatoires plus marqués, des conduits vocaux plus ouverts lors de la prononciation des voyelles et un débit d'élocution plus lent.

Dans le cas de la peur, le contour mélodique se rapproche de celui de la voix neutre mais présente des pics plus importants et surtout des irrégularités au voisinage de ces pics. Comme

pour la colère, l'articulation des sons est plus précise. La vitesse d'élocution est plus lente et l'onde glottale présente des irrégularités d'une période à l'autre qui proviendraient à la fois d'une salivation excessive (symptôme de la bouche sèche) et de tremblements d'origine neuromotrice.

Quant au chagrin, il se manifeste par une voix plus grave que la normale avec un domaine de variation de  $F_0$  très faible, ainsi que par une pente spectrale moyenne négative. De plus, tous les locuteurs utilisent un débit d'élocution deux fois plus lent que dans les autres situations émotionnelles et les spectrogrammes de leurs énoncés révèlent une onde glottale irrégulière et bruitée.

La figure A.54 illustre l'exemple du contour mélodique sur une phrase.

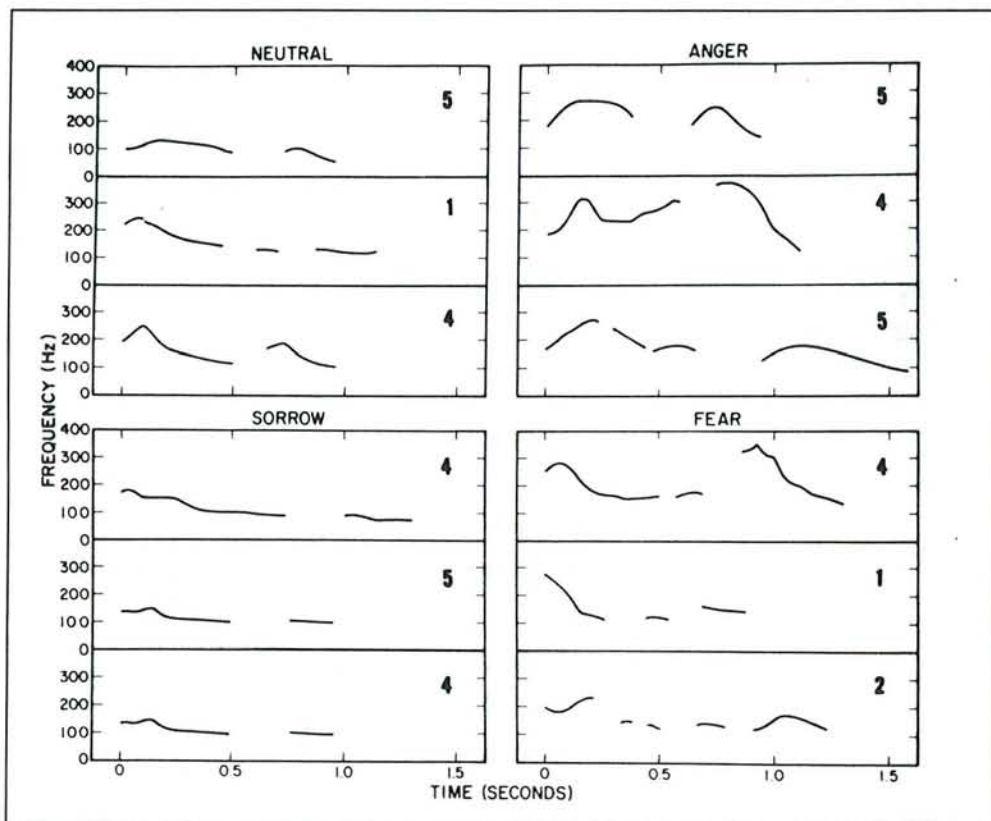


Figure A.54. Contours mélodiques sur une phrase pour un locuteur donné suivant quatre états émotionnels, l'état normal, la colère, la peur et le chagrin ; d'après [Williams 72].

D'autres études sur les dialogues réels concluent que les indices acoustiques caractérisant l'état émotionnel ne sont pas les mêmes pour tous les locuteurs [Streeter 83] [Broeders 90]. En revanche, toutes les études perceptives montrent qu'une augmentation de la fréquence fondamentale, de son domaine de variation ou de l'intensité acoustique indique un état émotionnel de stress. En outre, plus ces paramètres sont élevés plus le stress est perçu comme important [Streeter 83].

#### 4.4. Conclusion sur la variabilité liée au locuteur

Sans être exhaustive, nous avons mis en évidence l'existence de différences entre les locuteurs à tous les niveaux du processus de communication orale. Nous avons également mentionné quelques conséquences acoustiques de ces différences. Cette analyse permet de justifier les études sur la reconnaissance du locuteur, en général, et la recherche de paramètres pertinents dans la caractérisation du locuteur, en particulier. Nous avons aussi montré que les conséquences acoustiques de la variabilité interlocuteur sont sensibles à la variabilité intralocuteur. Aussi, après avoir établi des systèmes de reconnaissance du locuteur ou après avoir déterminé des caractéristiques du locuteur, il faudra vérifier leur robustesse, selon l'application souhaitée, vis-à-vis du temps, des changements d'état physique et psychique du locuteur, du style de parole et du contexte du dialogue lorsqu'il y en a un.

### 5. Conclusion

Nous avons exposé dans ce chapitre tous les faits qui sont susceptibles de participer à l'établissement de la valeur à un instant donné d'un paramètre acoustique du signal de parole. Nous avons également essayé de montrer les interactions qui existent entre ces faits. Cette complexité du processus de communication qu'est la parole explique celle des tâches comme la reconnaissance automatique de la parole continue ou la reconnaissance automatique du locuteur.

Une conclusion évidente du dernier paragraphe est que les études sur la reconnaissance automatique du locuteur doivent exploiter les paramètres acoustiques dont la variabilité interlocuteur est maximale et la variabilité intralocuteur est minimale. Comme nous le verrons dans la partie suivante, ceci n'est pas simple. Outre l'abondance de paramètres potentiellement pertinents, les problèmes d'extraction et de complexité calculatoire, il faut tenir compte de la difficulté à maintenir constantes les autres sources de variabilité. Par ailleurs, la variabilité interlocuteur dépend du nombre de locuteurs étudiés et de la dispersion de leurs caractéristiques anatomiques, sociogéographiques et linguistiques. Quant à la variabilité intralocuteur, elle est liée au nombre de répétitions et de situations considérées pour chaque locuteur.



## BIBLIOGRAPHIE



# BIBLIOGRAPHIE

## Abréviations employées dans la bibliographie :

- **JASA** : Journal of the Acoustical Society of America,
- **ICASSP** : International Conference on Acoustics, Speech and Signal Processing,
- **JEP** : Journées d'Etude sur la Parole,
- **EUROSPREECH** : European Conference on Speech Communication and Technology,
- **ASSP** : Acoustics, Speech and Signal Processing,
- **ICSLP** : International Conference on Spoken Language Processing.

- [Atkinson 78] J.E. Atkinson.  
Correlation Analysis of the Physiological Factors Controlling Fundamental Voice Frequency.  
*JASA*, 63(1):211–222, January 1978.
- [Bolt 70] R.H. Bolt, F.S. Cooper, E.E. David, P.B. Denes, J.M. Pickett et K.N. Stevens.  
Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes.  
*JASA*, 47(2):597–612, 1970.
- [Bonneau 89] A. Bonneau, D. Fohr et F. Lonchamp.  
Normalisation formantique des locuteurs féminins et masculins à l'aide de connaissances et d'un apprentissage minimal.  
*Séminaire sur la variabilité et la spécificité des locuteurs*, pages 188–191, Marseille Luminy, juin 1989. GRECO PRC Communication Homme-Machine.
- [Bouchet 83] A. Bouchet et J. Guilleret.  
*Anatomie topographique, descriptive et fonctionnelle*.  
SIMEP, Lyon, 1983.
- [Broeders 90] A.P.A. Broeders et A.C.M. Rietveld.  
The Effect of Cognitive Stress on Pitch and Duration.  
*Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pages 72–77, Edinburgh, June 1990. European Speech Communication Association.
- [Calliope 89] Calliope.  
*La Parole et son Traitement Automatique*.  
CNET-ENST, Masson, Paris, 1989.
- [Cooper 83] W.E. Cooper, C. Soares, A. Ham et K. Damon.  
The Influence of Inter- and Intra-speaker Tempo on Fundamental Frequency and Palatalization.  
*JASA*, 73(5):1723–1730, May 1983.



- [Duchet 86] J.L. Duchet.  
*La Phonologie.*  
Collection Que sais-je ?. Presses Universitaires de France, Paris, 1986.
- [Dujour 90] A. Lacheret Dujour.  
Contribution à l'Analyse de la Variabilité Phonologique pour le Traitement Automatique de la Parole Continue Multilocuteur.  
Thèse de l'Université de Paris VII, juin 1990.
- [Endres 71] W. Endres, W. Bambach et G. Flösser.  
Voice Spectrograms as a Function of Age, Voice Disguise and Voice Imitation.  
*JASA*, 49(6.2):1842–1848, 1971.
- [Eskenazi 88] M. Eskenazi, F. Lonchamp et J. Vaissière.  
Cours sur les Indices Acoustiques du Français.  
GRECO - Communication Parlée, octobre 1988.
- [Fant 60] G. Fant.  
*Acoustic Theory of Speech Production.*  
Mouton, The Hague, 1960.
- [Fant 90a] G. Fant.  
The Speech Code. Segmental and Prosodic Features.  
*ICSLP, Kobe, Japan*, pages 1389–1397, October 1990.
- [Fant 90b] G. Fant, A. Kruckenberg et L. Nord.  
Prosodic and Segmental Speaker Variations.  
*Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pages 106–120, Edinburgh, June 1990. European Speech Communication Association.
- [Ladefoged 76] P. Ladefoged, I. Kameny et W. Brackenbridge.  
Acoustic Effects of Style of Speech.  
*JASA*, 49(1):228–231, 1976.
- [Laprie 88] Y. Laprie.  
SNORRI : un système d'étude interactif de la parole.  
*17e JEP*, pages 71–76, Nancy, septembre 1988. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Lehiste 75] I. Lehiste.  
Suprasegmental Features of Speech.  
N.J. Lass, éditeur, *Contemporary Issues in Experimental Phonetics*, chapitre 7, pages 225–239. Academic Press, 1975.
- [Lienard 89] J.S. Lienard.  
Variabilité, contraintes et spécification de la parole : un cadre théorique.  
*Séminaire sur la variabilité et la spécificité des locuteurs*, pages 1–10, Marseille Luminy, juin 1989. GRECO PRC Communication Homme-Machine.
- [Lofqvist 84] A. Lofqvist, N.S. McGarr et K. Honda.  
Laryngeal Muscles and Articulatory Control.  
*JASA*, 76(3):951–953, September 1984.

- [Lonchamp 79] F. Lonchamp.  
Analyse acoustique des voyelles nasales françaises.  
*Verbum*, II(1):9-54, 1979.  
Institut de Phonétique de Nancy.
- [Lonchamp 87a] F. Lonchamp.  
La transcription phonétique du français.  
Document interne, Institut de Phonétique de Nancy, 1987.
- [Lonchamp 87b] F. Lonchamp.  
Les sons du Français - Analyse acoustique descriptive.  
Document interne, Institut de Phonétique de Nancy, 1987.
- [Lonchamp 88] F. Lonchamp.  
Etudes sur la Production et la Perception de la Parole. Les Indices Acoustiques de la Nasalité Vocalique - La Modification du Timbre par la Fréquence Fondamentale.  
Thèse de Doctorat d'Etat ès Lettres et Sciences Humaines, Université de Nancy II, avril 1988.
- [Malmberg 74] B. Malmberg.  
*Manuel de Phonétique Générale*.  
Collection Connaissance des Langues. Picard, Paris, 1974.
- [Malmberg 79] B. Malmberg.  
*La Phonétique*.  
Collection Que sais-je ?. Presses Universitaires de France, Paris, 1979.
- [Marchal 80] A. Marchal.  
*Les Sons et La Parole*.  
Collection Langue et Société. Guérin, Montréal, 1980.
- [Markel 76] J.D. Markel et A.H. Gray.  
*Linear Prediction of Speech*.  
Springer-Verlag, New-York, 1976.
- [Martinet 73] A. Martinet et H. Walter.  
*Dictionnaire de la Prononciation Française dans son Usage Réel*.  
France Expansion, Paris, 1973.
- [Martinet 80] A. Martinet.  
*Eléments de Linguistique Générale*.  
Colin, Paris, 1980.
- [Monsen 83] R.B. Monsen et A.M. Engebretson.  
The Accuracy of Formant Frequency Measurements: A Comparison of Spectrographic Analysis and Linear Prediction.  
*Journal of Speech and Hearing Research*, 26:89-97, 1983.
- [Nolan 83] F. Nolan.  
*The Phonetic Bases of Speaker Recognition*.  
Cambridge University Press, Great Britain, 1983.
- [O'Shaughnessy 87] D. O'Shaughnessy.  
*Speech Communication: Human and Machine*.  
Addison Wesley, Reading, Massachusetts, 1987.

- [Perennou 89] G. Perennou, M. de Calmès, I. Ferrané et J. Tihoni.  
Idiolecte et phonologie. Incidence sur la transcription automatique adaptée au locuteur par le système GEPH.  
*Séminaire sur la variabilité et la spécificité des locuteurs*, pages 68–77, Marseille Luminy, juin 1989. GRECO PRC Communication Homme-Machine.
- [Rabiner 78] L.R. Rabiner et R.W. Schafer.  
*Digital Processing of Speech Signals*.  
Prentice-Hall, New Jersey USA, 1978.
- [Sambur 75] M.R. Sambur.  
Selection of acoustic features for speaker identification.  
*I.E.E.E. Transactions on ASSP*, pages 176–182, April 1975.
- [Shoup 75] J.E. Shoup et L.L. Pfeifer.  
Acoustic Characteristics of Speech Sounds.  
N.J. Lass, éditeur, *Contemporary Issues in Experimental Phonetics*, chapitre 6, pages 171–224. Academic Press, New-York, 1975.
- [Stevens 77] K. Stevens.  
Sources of Inter- and Intra-speaker Variability in the Acoustic Properties of Speech Sounds.  
*7 th International Congress of Phonetics Sciences*, pages 206–232, 1977.
- [Streeter 83] L.A. Streeter, N.H. Macdonald, W. Apple, R.M. Krauss et K.M. Galotti.  
Acoustic and Perceptual Indicators of Emotional Stress.  
*JASA*, 73(4):1354–1360, April 1983.
- [Tielen 90] M.T.J. Tielen.  
Perception of the Voices of Men and Women in Relation to their Profession.  
*Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pages 192–197, Edinburgh, June 1990. European Speech Communication Association.
- [Vaissiere 80] J. Vaissiere.  
La structuration acoustique de la phrase française.  
*Annales de l'Ecole Normale Supérieure de Pise*, X.2(série III), 1980.
- [Vaissiere 83] J. Vaissiere.  
Language-independent Prosodics Features.  
A. Cutler et D.R. Ladd, éditeurs, *Prosody: Models and Measurements*, chapitre 5, pages 53–67. Springer-Verlag, New-York, 1983.
- [Vaissiere 84] J. Vaissiere.  
Speech Recognition: A Tutorial.  
F. Fallside et W.A Woods, éditeurs, *Computer Speech Processing*, pages 191–236. Prentice Hall, Englewood Cliffs, New-Jersey, 1984.
- [Vaissiere 87a] J. Vaissiere.  
Effect of Phonetic Context and Timing on the F-Pattern of the Vowels in Continuous Speech.  
*11 th International Congress of Phonetics Sciences*, pages 43–46, Tallinn, U.S.S.R., August 1987.



- [Vaissiere 87b] J. Vaissiere.  
The Use of Allophonic Variations of /a/ in Automatic Continuous Speech Recognition of French.  
Speech Communications Group, Working Papers, March 1987.
- [Vaissiere 88] J. Vaissiere.  
The Use of Prosodics Parameters in Automatic Speech Recognition.  
H.Nieman, M.Lang et G.Sagerer, éditeurs, *Recents Advances in Speech Understanding and Dialog System*, volume 46, série NATO ASI series, pages 71–99. Springer-Verlag, New-York, 1988.
- [Walter 76] H. Walter.  
*La Dynamique des Phonèmes dans le Lexique Français Contemporain*.  
France Expansion, Paris, 1976.
- [Whitehead 84] R.L. Whitehead, D.E. Metz et B.H. Whitehead.  
Vibrations Patterns of the Vocal Folds During Pulse Register Phonation.  
*JASA*, 75(4):1293–1296, April 1984.
- [Williams 72] C.E. Williams et K.N. Stevens.  
Emotions and Speech: Some Acoustical Correlates.  
*JASA*, 52(2):1238–1250, 1972.
- [Zwicker 80] E. Zwicker et E. Terhardt.  
Analytical Expressions for Critical-band Rate and Critical-bandwidth as a Function of Frequency.  
*JASA*, 68(5):1523–1525, 1980.
- [Zwicker 81] E. Zwicker et R. Feldtkeller. Traduit par C. Sorin.  
*Psychoacoustique : l'Oreille Récepteur d'Information*.  
CNET-ENST. Masson, Paris, 1981.

## **PARTIE B**

# **RECONNAISSANCE ET CARACTERISATION AUTOMATIQUES DU LOCUTEUR**





## TABLE DES MATIERES DE LA PARTIE B

INTRODUCTION	1
<b>I GENERALITES</b>	<b>3</b>
I.1 Introduction	3
I.2 Vérification et identification	4
I.3 Les grands axes de recherche en reconnaissance du locuteur	5
I.4 La dépendance et l'indépendance vis-à-vis du texte prononcé	6
I.5 Evaluation des systèmes de reconnaissance du locuteur	7
I.5.1 Introduction	7
I.5.2 Les taux d'erreur	8
I.5.3 Comparaison des systèmes de laboratoire	9
I.5.4 Evaluation dans des conditions réelles	10
I.5.4.1 Un exemple de système de vérification opérationnel	10
I.5.4.2 Approche des conditions réelles dans les études "de laboratoire"	10
I.5.5 Conclusion	11
I.6 Conclusion du chapitre	11
<b>II TECHNIQUES DE RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR</b>	<b>13</b>
II.1 Introduction	13
II.2 La quantification vectorielle	13
II.2.1 Introduction	13
II.2.2 Application à la reconnaissance automatique de la parole	14
II.2.3 Application à la reconnaissance automatique du locuteur	15
II.2.3.1 Introduction	15
II.2.3.2 Les études	15
II.3 Les modèles de Markov	20
II.3.1 Introduction et application à la reconnaissance de la parole	20
II.3.2 Application à la reconnaissance automatique du locuteur	21
II.4 Les réseaux neuronaux	28
II.4.1 Introduction et application à la reconnaissance de la parole	28
II.4.2 Application à la reconnaissance automatique du locuteur	28
II.5 Les autres méthodes de R.A.L.	31
II.5.1 Introduction	31
II.5.2 Les études plus récentes	31
II.6 Conclusion	38

<b>III CARACTERISATION AUTOMATIQUE DU LOCUTEUR</b>	39
<b>III.1 Introduction</b>	39
<b>III.2 Les méthodologies</b>	39
III.2.1 Le F-ratio	40
III.2.2 L'analyse discriminante linéaire	40
<b>III.3 Les études</b>	41
III.3.1 Introduction	41
III.3.2 Les études sur la langue anglaise	42
III.3.3 Les études sur la langue allemande	53
III.3.4 Les études sur la langue française	54
<b>III.4 Conclusion</b>	63
<b>BIBLIOGRAPHIE</b>	65

## Liste des figures

Figure B.1	Erreur d'identification en fonction du chiffre utilisé, d'après . . . . .	16
Figure B.2	Représentation articulatoire des voyelles de l'anglais britannique d'après . Les sons [ e ], [ a ] et [ o ] ne sont pas des voyelles à part entière mais des début de diphtongues. . . . .	41
Figure B.3	Représentation des voyelles et des trajectoires des diphtongues de l'anglais américain dans le plan $F_1$ , $F_2-F_1$ , d'après . . . . .	42
Figure B.4	Les meilleurs paramètres caractérisant le locuteur d'après l'étude de M.R. Sambur. . . . .	45
Figure B.5	Représentation des échantillons des spectres instantanés de [ m ] et [ n ] selon les deux premiers axes principaux, d'après L.S. Su. . . . .	47
Figure B.6	Evolution des trois premiers formants dans / A r A / en anglais, d'après et dans / A r A / en français, d'après les données du GRECO Communication Parlée . . . . .	49
Figure B.7	Représentation dans le plan des articulations des occlusives vélaires selon leur nombre et les locuteurs d'après Pérennou et al. . . . .	57



## Liste des tables

Table B.1	Comparaison de plusieurs systèmes de transcription phonétique des voyelles anglaises, d'après . Les symboles utilisés par P. Ladefoged sont ceux de l'A.P.I.. . . . .	43
Table B.2	Les six phrases du corpus utilisé par J.J. Wolf et M.R. Sambur. Les phonèmes soulignés sont ceux étudiés par M.R. Sambur. . . . .	44
Table B.3	Les dix meilleures fréquences formantiques d'après U. Goldstein. . . . .	48
Table B.4	Tableau synthétique de l'étude de K.K. Paliwal comprenant : les voyelles étudiées, les mots prononcés, les taux de reconnaissance pour chacune des distances, le classement de chacune des fréquences formantiques selon le F-ratio. . . . .	50
Table B.5	Les triplets / I V C / et / r V C / étudiés par F. Nolan. . . . .	51
Table B.6	Pourcentages moyens d'identification correcte calculés à partir des résultats fournis par F. Nolan. . . . .	52
Table B.7	Pourcentages moyens d'identification correcte calculés à partir des résultats fournis par F. Nolan sur l'identification à l'aide du degré de coarticulation de chaque locuteur. . . . .	53
Table B.8	Les six phrases du corpus utilisé par P. Corsi. Les phonèmes soulignés sont ceux dont la durée a été analysée par P. Corsi. . . . .	54
Table B.9	Les paramètres sélectionnés par P. Corsi répartis en 9 groupes. . . . .	55
Table B.10	Le texte lu par cinq locuteurs masculins dans l'étude de G. Pérennou et al. . . . .	56
Table B.11	Les allophones de / a / étudiés par Pérennou et al. . . . .	56
Table B.12	Les variantes coarticulatoires du couple / ti / de "matinée" dans l'étude de Pérennou et al. . . . .	58
Table B.13	Efficacité des durées d'origine prosodique dans la discrimination des cinq locuteurs dans l'étude de Pérennou et al. . . . .	59
Table B.14	Efficacité des différentes mesures prosodiques de $F_0$ dans une phrase dans la discrimination des cinq locuteurs dans l'étude de Pérennou et al. . . . .	60
Table B.15	Les quinze phonèmes ou ensembles de phonèmes présélectionnés pour les analyses discriminantes. . . . .	60
Table B.16	Résultats des deux analyses discriminantes sur les 15 "phonèmes" présélectionnés dans l'étude de Pérennou et al. . . . .	61

## INTRODUCTION

Dans cette partie, nous souhaitons dans un premier temps définir le sous-domaine du traitement automatique de la parole qu'est la reconnaissance automatique du locuteur, avant de présenter les différentes démarches qui ont été abordées dans ce domaine de recherche, en développant plus particulièrement les approches voisines de notre étude.

Le premier chapitre présente les différents concepts et définitions relatifs au domaine de la reconnaissance automatique du locuteur. Ceux-ci permettent de définir des classifications orthogonales des recherches dans ce domaine, selon que l'on s'attache à la fonction du système de RAL (vérification versus identification), à la plus importante de ses qualités (reconnaissance dépendante du texte versus reconnaissance indépendante du texte) ou à la démarche qui a permis d'aboutir à sa conception (techniques de reconnaissances des formes versus recherche de paramètres linguistiques caractéristiques du locuteur). Nous introduisons également dans ce chapitre, les différentes manières d'évaluer les performances d'un système de RAL et nous terminons sur la difficulté de comparer les performances de plusieurs systèmes qu'ils soient réels ou de laboratoire.

Nous avons choisi de présenter dans les deux chapitres suivants, une partie des nombreuses recherches effectuées en RAL en adoptant une classification fondée sur les deux grandes démarches, qui, à notre avis, régissent la recherche en reconnaissance automatique du locuteur. La première démarche consiste à appliquer et à adapter à la reconnaissance du locuteur des techniques de reconnaissance de formes conçues et validées pour la reconnaissance automatique de la parole. La seconde démarche essaie d'exploiter explicitement les variabilités interlocuteur et intralocuteur de la parole en recherchant des paramètres acoustiques ou phonétiques qui caractérisent au mieux le locuteur.

Le second chapitre de cette partie propose donc une revue des travaux effectués en reconnaissance automatique du locuteur qui relèvent de la première démarche mais qui concernent plus particulièrement les techniques récentes comme la quantification vectorielle, les modèles de Markov ou les réseaux neuronaux,... Avant de détailler ces travaux, les concepts sous-jacents à chacune des techniques sont introduits et quelques lignes situent leur utilisation en reconnaissance automatique de la parole.

Le dernier chapitre est consacré à l'étude bibliographique des recherches entreprises dans le domaine de la caractérisation du locuteur à l'aide de paramètres acoustiques et phonétiques. Sans pouvoir être exhaustif, ce chapitre essaie de décrire tous les paramètres traités dans ces études et d'en dégager les résultats caractéristiques.





## CHAPITRE I GENERALITES

### 1. Introduction

En reconnaissance automatique de la parole, les variations du signal de parole qui résultent de la prononciation par différents locuteurs d'un même énoncé sont considérées comme du bruit qu'il faut essayer d'éliminer, soit en utilisant de nombreuses références pour cet énoncé, soit en normalisant le segment de parole du locuteur, soit encore en adaptant le système de reconnaissance de parole au locuteur. L'objectif de la reconnaissance automatique du locuteur est au contraire d'utiliser au mieux ce "bruit" pour identifier ou vérifier l'identité du locuteur.

Les recherches en reconnaissance du locuteur ont été – et sont encore – beaucoup moins nombreuses que les recherches en reconnaissance de la parole. Les principales causes de cette "pénurie" ne sont pas indépendantes : le nombre insuffisant d'applications du point de vue économique, la présence sur le marché de systèmes plus fiables que la reconnaissance de la voix et surtout le manque de connaissance sur la manière dont sont codées dans le signal de parole les informations concernant le locuteur.

Le principal domaine d'application de la reconnaissance automatique du locuteur est la vérification de l'identité d'une personne pour lui permettre l'accès à un site protégé (zones militaires, centrales nucléaires, salles machines, ...) ou à un service réservé (transactions bancaires, bases de données, ...). L'autre domaine, plus restreint, est l'identification d'une personne à des fins militaires ou judiciaires. Cette dernière application de la reconnaissance du locuteur, qu'elle soit automatique ou humaine, a suscité de tout temps de vives controverses [Bolt 70] [Nolan 83].

Dans le cadre de l'accès à un site protégé ou celui de l'identification judiciaire, des techniques biométriques d'identification plus fiables que la reconnaissance de la voix sont le plus souvent employées, comme les empreintes digitales, l'exploration rétinienne ou l'analyse génétique. Ainsi, les empreintes digitales sont difficilement falsifiables ; certains détails restent inchangés toute la vie ; leur identification est hiérarchisée et l'interprétation des différences et des similarités est fiable. La probabilité que deux empreintes de doigt soient identiques est estimée à  $10^{-42}$  [Bolt 70]. Ces techniques sont plus fiables parce qu'elles exploitent des caractéristiques statiques de la personne alors que des techniques comme la reconnaissance de la voix ou l'analyse de signature se fondent sur des caractéristiques dynamiques qui sont sujettes à une plus grande variabilité.

S'il est actuellement possible d'associer des caractéristiques spectrales ou temporelles du signal de parole à des phonèmes et même à des entités linguistiques de plus haut niveau, on ne sait pas du tout le faire en ce qui concerne les caractéristiques du locuteur. Les aspects acoustiques qui différencient les locuteurs sont intimement liés à ceux qui différencient les entités linguistiques et, comme nous l'avons vu dans la première partie de ce mémoire, s'expriment en général sous la forme de variations de ces derniers.

Du point de vue linguistique, les locuteurs se différencient sur trois plans, tout d'abord au niveau physiologique (cordes vocales, conduit vocal, ...), puis dans leur manière de réaliser les différentes unités linguistiques (variations dans les positions articulatoires des phonèmes, dans les phénomènes de coarticulation ou dans la prosodie, ...) et enfin dans le choix de ces unités linguistiques (emploi de certains mots, de certaines tournures, ...). Les systèmes de reconnaissance du locuteur n'exploitent que les deux premières sources de variabilité et encore de façon incomplète en ce qui concerne la prosodie. Celle-ci, ainsi que la dernière source de variabilité, sont trop difficiles à quantifier et à contrôler lors d'expérimentations.

Toutefois, l'utilisation de plus en plus intensive du téléphone dans le domaine tertiaire (transactions bancaires, services téléphoniques de réservations ou de commandes) et le développement de la communication homme-machine (nécessité de réaliser des systèmes de reconnaissance de la parole indépendants du locuteur) ont relancé les recherches en reconnaissance automatique du locuteur et en caractérisation du locuteur.

## 2. Vérification et identification

Le terme "reconnaissance automatique du locuteur" (RAL) est un terme générique qui recouvre deux tâches distinctes : la vérification automatique du locuteur (VAL) et l'identification automatique du locuteur (IAL).

Le système de vérification, comme son nom l'indique, vérifie l'identité annoncée par un locuteur inconnu. Pour cela, il détermine si les informations extraites du message qu'il demande au locuteur de prononcer sont suffisamment proches des informations stockées en mémoire pour ce locuteur. La décision prise est donc binaire et la performance d'un système de vérification ne dépend pas du nombre de locuteurs à vérifier : un système aléatoire aurait un taux de réussite de 50% à condition qu'il y ait autant de "bons locuteurs" que d'imposteurs. En revanche, la principale difficulté réside dans l'établissement du seuil de similarité (ou de vérification), ce qui suppose d'avoir préalablement effectué une étude statistique fine sur la variabilité intralocuteur et interlocuteur des informations à comparer. Le domaine d'application de la VAL est essentiellement celui du contrôle d'accès à un site ou à un service.

Le système d'identification, quant à lui, a pour but de déterminer parmi les  $N$  locuteurs de sa base de données celui qui vient de prononcer le message vocal. L'identification d'un locuteur est donc une tâche similaire à la reconnaissance automatique de la parole, il faut comparer une forme inconnue à  $N$  formes de références et prendre  $N$  décisions. Par conséquent, la performance d'un système d'identification est une fonction du nombre de locuteurs dans la base. Plus ce nombre est grand, plus il est probable que deux locuteurs possèdent des caractéristiques ayant des distributions voisines. Un système aléatoire aurait un taux de réussite de  $(100/N)\%$  en supposant que tous les locuteurs testés soient dans la base et soient équiprobables.

Le processus d'identification, qui vient d'être décrit, réalise une identification sur un ensemble fermé de locuteurs ("closed-set identification"). Un système d'identification sur un ensemble ouvert de locuteurs ("open-set identification") peut décider que le message émis ne correspond à aucun des locuteurs de sa base mais à un imposteur. Ce type de système est le plus complexe puisqu'il cumule la complexité des  $N$  comparaisons à la nécessité d'une étude statistique fine des paramètres. Du point de vue pratique, il correspond au domaine sensible de l'identification à des fins judiciaires ou militaires.



Bien qu'il constitue le mode d'identification le plus éloigné de la réalité, le système d'identification sur un ensemble fermé de locuteurs est le plus utilisé au niveau de la recherche. En effet, comme nous le détaillerons dans un prochain paragraphe, il a l'avantage de ne comporter qu'un type d'erreur et qui n'est fonction d'aucun seuil. Il est donc plus simple pour comparer les différentes techniques de reconnaissance ou les divers paramètres susceptibles de caractériser les locuteurs. Toutefois, il peut, à notre avis, avoir une application pratique dans un système de reconnaissance automatique de la parole. Situé en amont de celui-ci, il permettrait l'adaptation de la reconnaissance au locuteur ou à une classe de locuteurs.

Comme nous le verrons dans la suite de ce mémoire, la vérification et l'identification utilisent des techniques d'analyse et de décision similaires. Pour ces raisons, nous avons jugé inopportun de faire une présentation des recherches en reconnaissance automatique du locuteur en séparant les deux types de systèmes. Nous avons préféré adopter une présentation qui reproduise la démarche suivie, à notre avis, dans ce domaine de recherche et qui fait l'objet du paragraphe suivant.

### **3. Les grands axes de recherche en reconnaissance du locuteur**

Tout au long des trente années que compte la recherche en reconnaissance automatique du locuteur, deux démarches "orthogonales" ont été adoptées.

La première consiste à appliquer les nouvelles méthodes découvertes en reconnaissance automatique de la parole à celle du locuteur. Dans une première étape, des études établissent la faisabilité de ces nouvelles techniques à la reconnaissance automatique du locuteur (programmation dynamique, quantification vectorielle, réseaux neuronaux, ...). Viennent ensuite d'autres études qui essaient d'améliorer les performances obtenues en recherchant les meilleurs paramètres au sens de la paramétrisation du signal acoustique (coefficients LPC, coefficients cepstraux, ...) et les meilleures distances qui leur sont associées. Remarquons que dans ce type de démarche, les sources de variabilités interlocuteur et intralocuteur sont utilisées de façon implicite et sans "grand discernement" au travers des paramètres acoustiques précédemment cités. De ce fait, le choix des corpus mis en œuvre dans les méthodes relevant de cette démarche n'est pas guidé par des considérations linguistiques ou paralinguistiques. Ce sont le plus souvent des corpus créés pour la RAP (chiffres, expressions usuelles, ...) [Soong 85] ou facilement segmentables (mots monosyllabiques) [Doddington 85].

L'autre démarche cherche, au contraire, à extraire du signal de parole des paramètres, qui caractérisent au mieux le locuteur, en exploitant explicitement toutes les sources de différences entre les locuteurs. Ces paramètres sont soit d'ordre statistique (estimation à partir de longs énoncés de parole des densités de probabilité de paramètres comme le spectre à long terme) soit d'ordre linguistique (certains paramètres acoustiques d'un phonème donné dans un contexte donné, prosodie, ...). Pour ces derniers, le choix du corpus est primordial.

Comme nous le verrons, notre travail de recherche se situe dans le cadre de cette dernière démarche.

L'objet de cette partie étant une revue des recherches entreprises en reconnaissance automatique du locuteur, nous avons décidé de les classer selon ces deux démarches dans deux



chapitres distincts qui s'appellent respectivement "Techniques de reconnaissance automatique du locuteur" et "Caractérisation automatique du locuteur".

Mais auparavant nous allons préciser les notions de dépendance et d'indépendance vis-à-vis du texte et développer les problèmes rencontrés dans la comparaison des divers systèmes de reconnaissance du locuteur.

## 4. La dépendance et l'indépendance vis-à-vis du texte prononcé

Les systèmes de reconnaissance du locuteur sont qualifiés de "dépendant du texte" ou d'"indépendant du texte" selon qu'ils connaissent ou non le texte que va prononcer le locuteur.

Le cas le plus typique de dépendance par rapport au texte est obtenu lorsque le système de reconnaissance demande au locuteur de prononcer le même énoncé que celui qu'il a prononcé lors de la phase d'apprentissage. Mais c'est aussi le cas lorsque, par exemple, le système demande au locuteur de prononcer une suite de mots isolés issus ou non de l'ensemble des mots ayant servi à l'apprentissage. L'utilisation du même texte pour l'apprentissage et la reconnaissance simplifie le processus de comparaison entre la forme de référence et la forme inconnue et augmente les performances des systèmes de reconnaissance. En revanche, celles-ci sont fortement corrélées au vocabulaire choisi. Par ailleurs, les systèmes dépendants du texte sont beaucoup moins sécuritifs (imitations, duplications des enregistrements d'apprentissage, etc.) et nécessitent la coopération du locuteur. C'est pour ces raisons que la dépendance par rapport au texte se rencontre plus souvent dans les systèmes de vérification que dans les systèmes d'identification. Toutefois, pour améliorer la sécurité, l'énoncé de test peut être composé de mots tirés aléatoirement par le système de vérification.

La notion d'indépendance par rapport au texte est moins clairement définie puisque les auteurs d'études en RAL qualifient de la même manière un système qui reconnaît un locuteur à partir d'un énoncé quelconque et celui qui attend un énoncé de test qu'il ne connaît pas à l'avance mais qui est un sous-ensemble de l'énoncé d'apprentissage. L'exemple le plus communément utilisé est celui où le locuteur doit énoncer une suite quelconque de chiffres. Pour différencier ces deux types de systèmes, il serait utile d'introduire les notions supplémentaires d'indépendance et de dépendance par rapport au vocabulaire.

Les techniques de reconnaissance indépendantes du texte et surtout celles qui sont aussi indépendantes du vocabulaire requièrent de plus longs énoncés que les techniques dépendantes du texte, aussi bien pendant la phase d'apprentissage (quelques dizaines de secondes à plusieurs minutes par locuteur) que pendant la phase de reconnaissance (plus de 5 secondes par locuteur). De ceci, découle leur principal inconvénient, celui de ne pas pouvoir effectuer la reconnaissance en temps réel.

Comme D. O'Shaughnessy [O'Shaughnessy 87] et G.R. Doddington [Doddington 85], nous pensons qu'il existe une troisième méthodologie située à mi-chemin entre la dépendance et l'indépendance par rapport au texte. Celle-ci consiste à utiliser lors de la phase de reconnaissance des phrases engendrées aléatoirement mais contenant un certain nombre d'événements acoustico-phonétiques spécifiques caractéristiques du locuteur. Toutefois, cette nouvelle méthodologie suppose d'une part de connaître ces événements et d'autre part de

maîtriser leur localisation automatique dans le flux de la parole continue. En effet, si les formes de référence peuvent être extraites d'un corpus d'apprentissage étiqueté manuellement, les formes de test doivent l'être par segmentation automatique.

## 5. Evaluation des systèmes de reconnaissance du locuteur

### 5.1. Introduction

Un bon système de reconnaissance du locuteur doit savoir exploiter la variabilité interlocuteur tout en étant insensible à la variabilité intralocuteur. Une évaluation correcte des systèmes de RAL doit donc comporter le test de ces deux qualités.

Les tests, qui mettent à l'épreuve la pertinence discriminatoire d'un système de RAL, sont ceux qui mettent en œuvre un nombre important de locuteurs, des groupes de locuteurs homogènes, et qui emploient éventuellement des jumeaux ou des imitateurs.

Le test de la robustesse du système de RAL vis-à-vis des différentes sources de variabilité intralocuteur, mises en évidence dans le chapitre V de la partie A, est encore plus complexe. Celui-ci dépend à la fois du domaine d'application du système de reconnaissance du locuteur et de la méthode employée pour faire la reconnaissance.

Dans le domaine de la vérification automatique du locuteur ou de l'identification dans le cadre d'un dialogue homme-machine, le style de parole est assez contraint (l'énoncé de test est lu ou répété après écoute). De plus, le locuteur est coopératif et a tendance à adapter son style à celui de l'ordinateur, de manière à être reconnu le plus rapidement possible. Pour les mêmes raisons, le "vrai" locuteur ne cherche pas à contrefaire sa voix. Dans ce cas, il est donc inutile de tester la résistance du système à la variabilité intralocuteur due au contexte social ou à la gestion d'un dialogue, aux sentiments et aux attitudes que le locuteur souhaite communiquer, etc.

En revanche, de telles sources de variabilité doivent être impérativement testées lorsque le système est destiné à réaliser une identification du locuteur dans des situations réelles (parole spontanée, situations affectives et sociales différentes entre les énoncés de référence et de test, ...), comme c'est le cas dans l'identification judiciaire. Une étude réalisée en 1978 par E.T. Doherty et H. Hollien, citée par F. Nolan [Nolan 83], effectue une identification sur un ensemble fermé de 25 locuteurs à l'aide du spectre à long terme. Le taux de reconnaissance passe de 100% en parole normale à 72% lorsque les locuteurs sont soumis à des décharges électriques et à 24% si les locuteurs modifient leur voix.

Par ailleurs, pour certaines applications, il est nécessaire de vérifier la robustesse du système par rapport à son environnement acoustique (microphone, canal téléphonique, bruit).

Toutes ces remarques sont valables pour les deux grands axes de RAL que nous avons définis au paragraphe 3. Toutefois, quelques remarques supplémentaires s'imposent dans le cadre des études sur la caractérisation du locuteur.

Lorsqu'une étude cherche à mettre en évidence le pouvoir discriminatoire d'un paramètre, l'interprétation de celui-ci est presque toujours reliée à une source de variabilité interlocuteur bien définie. Il faut donc contrôler les autres sources de différences entre locuteurs pour que l'étude soit réellement significative. En effet, qu'elles soient d'origine articulatoire, combinatoire,



phonologique, syntaxico-sémantique ou paralinguistique, les sources de variabilité s'expriment dans les mêmes paramètres acoustiques : le fondamental, les formants, la durée, l'intensité, etc. Donnons tout de suite deux exemples. Lorsqu'on souhaite démontrer la pertinence de la forme du contour de  $F_0$  sur une syllabe accentuée, il faut s'assurer que tous les locuteurs accentuent la même syllabe. Si les paramètres extraits sont liés à l'articulation d'un son (source phonétique), il faut être sûr qu'un locuteur n'a pas prononcé un autre son dans le même contexte (source phonologique).

Tout ce que nous venons de décrire constitue l'évaluation idéale des systèmes de reconnaissance du locuteur. Celle-ci est très éloignée de leur évaluation réelle que nous allons aborder dans la suite de ce paragraphe.

La plupart des études en RAL sont réalisées dans des conditions dites "de laboratoire" : peu de distorsions du signal de parole, corpus pré-enregistrés de phrases lues par des locuteurs familiarisés avec l'entrée/sortie vocale du système. Ces conditions sont souvent éloignées des véritables conditions opérationnelles.

Aussi distinguerons-nous trois étapes dans l'étude de l'évaluation d'un système de RAL. Nous détaillerons d'abord divers taux d'erreur utilisés en reconnaissance du locuteur avant d'exposer les difficultés rencontrées dans la comparaison des performances des études effectuées dans des conditions dites "de laboratoire". Puis, nous évoquerons l'évaluation de la robustesse des résultats d'un système de RAL soit directement dans des conditions réelles d'utilisation soit en se rapprochant de ces conditions.

## 5.2. Les taux d'erreur

La performance d'un système d'identification simple (ensemble fermé de locuteurs) est directement fournie par le taux de confusion entre locuteurs.

En revanche, pour tester la performance d'un système de vérification du locuteur, il faut tenir compte de deux éléments :

- **le taux de faux rejet** (False Reject Rate) qui comptabilise les rejets de locuteurs autorisés,
- **le taux de fausse acceptation** (False Acceptance Rate) qui comptabilise les acceptations d'imposteurs.

Souvent, selon l'application visée, les systèmes de VAL ajustent le seuil de similarité (ou seuil de vérification) de manière à minimiser une fonction des deux taux précédents :

- **EER** (Equal Error Rate) : le taux de fausse acceptation doit être égal au taux de faux rejet,
- **MAFRA** (Minimum Average False Reject and False Acceptance Rate) : le seuil de vérification est fixé de manière à rendre minimale la somme des deux taux d'erreur,
- **CFA** (Constant False Acceptance) : le seuil de vérification est établi de manière à obtenir un taux de fausse acceptation prédéfini. Cette contrainte privilégie la sécurité de l'accès au site ou au service au détriment du confort de l'utilisateur,
- **(FAXFR)<sup>1/2</sup>** : G.R. Doddington [Doddington 83] propose comme bon critère de comparaison entre des systèmes différents la racine carrée du produit des deux taux d'erreur.



### 5.3. Comparaison des systèmes de laboratoire

Comparer les performances des différentes études réalisées en RAL relève presque de la gageure. Comme nous l'avons déjà mentionné, certains vérifient, d'autres identifient, certains sont dépendants du texte à des degrés divers, d'autres sont indépendants du texte, enfin plusieurs taux d'erreurs sont utilisables. Mais, surtout, les conditions d'expérimentation sont rarement semblables.

La plus variable d'entre elles est constituée par l'ensemble des locuteurs qui sert à établir le système et à le tester. Selon l'étude, le nombre de locuteurs varie de cinq à cent ou même plus, comme c'est le cas pour le système de Texas Instruments qui fonctionne depuis plus de dix ans [Doddington 85]. Certaines études mélangent les locuteurs des deux sexes, d'autres se limitent à un sexe et à une tranche d'âge. Par ailleurs, il est évident que plus le corpus est hétérogène (origines socio-géographiques des locuteurs variées) meilleurs sont les résultats. Mais comment mesurer l'homogénéité d'un corpus ?

Un autre paramètre crucial dans l'établissement des taux d'erreur est l'intervalle de temps qui sépare les énoncés des locuteurs de la phase d'apprentissage de ceux de la phase de test. Hormi le cas biaisé où le même corpus sert à la fois à l'établissement du système et à son test, les meilleurs résultats sont obtenus lorsque les corpus d'apprentissage et de test ont été enregistrés lors de la même session. Au fur et à mesure que l'intervalle de temps augmente les performances se dégradent. Ainsi l'étude de F.K. Soong et al. [Soong 85] portant sur 100 locuteurs obtient un taux de 100% d'identification pour la même session et chute à 96% pour un intervalle d'un mois.

De tels résultats mettent en évidence la double nécessité d'utiliser plusieurs répétitions lors de la phase d'apprentissage et de mettre à jour de façon régulière la base de données des locuteurs, afin de prendre en compte une partie de la variabilité intralocuteur.

La composition des corpus d'apprentissage et de test est aussi un paramètre qui intervient dans la comparaison des recherches en RAL. Même si ceux-ci sont relativement contraints (phrases ou listes de mots lues), leur contenu linguistique influe sur les performances de la reconnaissance du locuteur. Dans le cadre de la caractérisation du locuteur, cette influence est primordiale aussi bien dans l'interprétation des paramètres pertinents que dans la comparaison des études.

Enfin, l'ultime variable dans l'évaluation des systèmes est la procédure de test elle-même. Si dans le cas de l'identification, la procédure est toujours la même et consiste à comparer chacun des locuteurs à tous les locuteurs de la base, dans le cas de la vérification, le taux de fausse acceptation peut être obtenu de plusieurs façons. La moins réaliste mais la plus systématique consiste à se servir, pour chacun des locuteurs autorisés, des autres locuteurs de la base comme imposteurs. D'autres études utilisent des locuteurs supplémentaires comme imposteurs, D.K. Burton [Burton 87] emploie 111 imposteurs pour tester la vérification de 16 locuteurs autorisés. Le fait d'utiliser de nombreux imposteurs augmente les chances d'avoir deux voix aux caractéristiques voisines mais cela ne traduit pas la situation réelle où un imposteur s'exerce à imiter au mieux un locuteur particulier. Pour résoudre ce problème, certains laboratoires de recherche ont fait appel à des imitateurs professionnels. Dans une étude réalisée aux laboratoires Bell, quatre d'entre eux ont ainsi réussi à obtenir un taux de fausse acceptation de 27% [Doddington 85], essentiellement parce que les paramètres mis en œuvre étaient de type prosodique.

## 5.4. Evaluation dans des conditions réelles

Nous nous limitons dans ce paragraphe à mettre en lumière les connaissances que peut apporter le test d'un système opérationnel et à décrire brièvement comment les études "de laboratoire " approchent les conditions réelles de fonctionnement.

### 5.4.1. Un exemple de système de vérification opérationnel

A notre connaissance, le seul système testé dans des conditions opérationnelles est le système de vérification que Texas Instruments a conçu en 1974 pour contrôler l'accès à sa salle d'ordinateurs. L'interface utilisateur se compose d'une cabine dotée d'une entrée/sortie vocale et d'un clavier à l'aide duquel l'utilisateur précise son identité. Le sol de la cabine est une balance qui permet au système de connaître le nombre de personnes présentes, les poids des locuteurs étant mémorisés avec leurs références vocales [Doddington 85].

Après plus de dix ans d'utilisation, les performances de ce système atteignent un taux de fausse acceptation de 0,7% et un taux de faux rejet de 0,9% dont 20% sont attribuables à une mauvaise utilisation du système.

Mais son grand intérêt est qu'il met en évidence certaines propriétés de la reconnaissance du locuteur comme :

- l'influence de l'adaptation du locuteur au système. Lors des quatre premières tentatives, le taux de rejet est de 10%, puis il chute aussitôt à 1%, palier qu'il quitte aux environs des 1000 essais pour atteindre moins de 0,25% après 10 000 vérifications ;
- l'influence du stress sur les performances du locuteur. Le taux de rejet de 0,5%, lorsque la personne est seule dans la cabine, passe à 1,8% lorsqu'elle est accompagnée. De plus, Texas Instruments note qu'il y a quatre fois plus de rejets entre 21h et 3h du matin qu'entre 9h du matin et 15h. Est-ce dû à la fatigue ? On sait que "l'état vocal" d'un locuteur évolue au cours de la journée ;
- l'inégalité des locuteurs face au processus de reconnaissance. L'examen de la distribution des taux de faux rejet en fonction du locuteur montre que la valeur médiane du taux de rejet est égale à la moitié du taux moyen de 0,9% et que seulement un quart des locuteurs testés présentent un taux supérieur au taux moyen de 0,9%.

### 5.4.2. Approche des conditions réelles dans les études "de laboratoire"

Malgré le manque de systèmes opérationnels, certaines recherches en RAL ont examiné de manière plus spécifique l'influence des paramètres "environnementaux" sur la reconnaissance du locuteur, aussi bien du point de vue acoustique (canal téléphonique, canal radio bruité, ...), qu'en ce qui concerne l'état du locuteur. Par ailleurs, d'autres études ont essayé d'évaluer les performances de leurs systèmes dans des conditions reflétant des environnements proches de la réalité.

Le test de robustesse le plus communément expérimenté est le passage de l'un des corpus ou des deux corpus à travers le canal téléphonique [Li 83] [Soong 85] [Chi-Shi 90].

Dans certaines applications, les données de test et d'apprentissage proviennent de canaux de communications différents. Certaines recherches simulent cette dissemblance en faisant passer l'un des corpus dans un filtre numérique [Soong 85], d'autres utilisent une procédure de



normalisation des données, appelée "blind deconvolution" (déconvolution aveugle [Li 83]. Celle-ci modifie les deux ensembles de données de manière à ce que leur spectre à long terme soit identique à un spectre donné [Li 83]. Bien sûr, cette méthode est désastreuse pour les méthodes de reconnaissance fondées sur les moyennes à long terme de paramètres.

## 5.5. Conclusion

Nous avons voulu d'esquisser dans ce paragraphe ce que serait une évaluation idéale des systèmes de reconnaissance automatique du locuteur. Remarquons que cette évaluation idéale devrait être également appliquée à la reconnaissance humaine du locuteur, qu'elle soit auditive ou par lecture de spectrogrammes. Puis, nous avons montré ce qu'était l'évaluation réelle en mettant l'accent sur la difficulté de comparer les performances des diverses études sur la reconnaissance automatique du locuteur. Les deux chapitres suivants, dans lequel nous présentons les diverses techniques mises en œuvre dans ces études, fournira une démonstration plus complète de ces difficultés.

Depuis quelques années est apparue la nécessité d'établir des corpus nationaux et internationaux comme TIMIT aux Etats-Unis dans le cadre du projet DARPA [Zue 88], EUROM en Europe dans le celui du projet ESPRIT SAM [Wells 88] et BDSONS en France dans le cadre du GRECO-PRC Communication Homme-Machine [Carre 84]. D'une façon générale, ces corpus comprennent des mots isolés et de la parole continue multi-locuteurs enregistrés dans des conditions variées (parole lue ou spontanée, bruitée ou non), afin de tester les systèmes de reconnaissance automatique de la parole.

De la même manière, il faudra établir des bases de données de sous-ensembles multilingues de locuteurs homogènes ou présentant des caractéristiques particulières (jumeaux, imitations) et dans des environnements acoustiques divers (téléphone, bruit), si l'on veut un jour pouvoir comparer efficacement les différentes méthodes utilisées en reconnaissance du locuteur.

## 6. Conclusion du chapitre

Nous avons défini dans ce chapitre les différentes formes de reconnaissance automatique du locuteur ainsi que les deux grands axes de recherche dans ce domaine. Nous avons présenté des notions qui sont plus ou moins communes à toutes les études relevant de ces deux démarches, comme l'indépendance ou la dépendance par rapport au texte, les taux d'erreur, l'évaluation des performances des études dans des conditions dites "de laboratoire" et dans des conditions opérationnelles,

Nous allons maintenant quitter le domaine des généralités pour entrer dans celui, plus précis, de la description des études qui ont été réalisées dans chacune des deux démarches. Ce sera l'objet des deux derniers chapitres de cette partie.





## CHAPITRE II

# TECHNIQUES DE RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR

## 1. Introduction

Nous avons introduit dans le chapitre précédent notre vision de la recherche en reconnaissance automatique du locuteur sous la forme de deux démarches. Rappelons que la première regroupe les études qui appliquent et adaptent à la RAL des techniques mises en œuvre pour la RAP et que la seconde rassemble les études plus spécifiques qui cherchent à extraire du signal de parole des paramètres acoustiques ou phonétiques qui caractérisent le locuteur.

Ce chapitre traite de la première de ces démarches. Toutefois, notre travail relevant de la seconde, nous n'allons pas faire une étude exhaustive de toutes ces techniques. Dans la suite, nous passons en revue la bibliographie récente sur le domaine, en retenant pour chacune des études les éléments qui nous paraissent fondamentaux. Pour ce faire, nous avons adopté une classification suivant les grands principes adoptés, bien que des recoupements puissent s'opérer entre les techniques. En effet, la programmation dynamique peut être utilisée pour fabriquer des formes de référence utilisées pour l'apprentissage de réseaux neuronaux, de même que la quantification vectorielle peut être antérieure à tout type de processus de reconnaissance. Notre classification se fonde donc sur le principe dominant adopté en phase de reconnaissance tel qu'il est mis en valeur par les auteurs de la publication. Nous abordons successivement :

- les études se servant directement pour la reconnaissance d'un codebook obtenu par quantification vectorielle des données,
- les études fondées sur les modèles de Markov,
- les études faisant appel aux réseaux neuronaux,
- les études faisant appel à des méthodes de classification autres que les précédentes.

Nous préciserons pour chaque étude si le système mis au point effectue une reconnaissance dépendante ou indépendante du texte.

## 2. La quantification vectorielle

### 2.1. Introduction

La quantification vectorielle (QV) est une méthode de compression de données utilisée en codage d'informations afin de réduire la quantité de données à transmettre, dans le cas de la transmission, ou l'espace mémoire, dans le cas de la mémorisation.

Elle associe à un vecteur de  $N$  composantes un scalaire qui est l'index dans un dictionnaire (*codebook*) du mot (*codeword*) représentant le mieux ce vecteur. Dans le cas de la parole, le

vecteur à quantifier est soit une succession d'échantillons du signal de parole, soit un ensemble de paramètres qui codent déjà un intervalle du signal de parole, comme les coefficients issus d'une analyse LPC.

La taille du dictionnaire, qui est toujours une puissance de 2, est un paramètre de la quantification vectorielle.

Les deux principales difficultés rencontrées en quantification vectorielle sont l'établissement du dictionnaire et la recherche dans le dictionnaire du mot représentant un vecteur donné.

L'objectif fondamental de la quantification vectorielle est de réduire le taux de codage tout en garantissant une bonne fidélité par rapport au signal initial. Étant donné une taille prédéfinie, le meilleur dictionnaire est donc celui qui minimise la distorsion moyenne à long terme donnée par :

$$\Delta_m = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} d(V_i, \hat{V}_i)$$

où  $V_i$  est un vecteur codé par le vecteur  $\hat{V}_i$  du dictionnaire et  $d$  est une mesure de distorsion entre les deux, comme la distance euclidienne ou la distance d'Itakura-Saito lorsque les composantes des vecteurs sont des coefficients LPC [Itakura 75b].

Dans la pratique,  $\Delta_m$  n'est pas calculable. Aussi le dictionnaire est-il déterminé en minimisant la distorsion moyenne sur un corpus d'apprentissage et en supposant qu'elle restera minimale pour les données ultérieures. La taille du corpus dépend donc de la spécificité du dictionnaire mais doit être suffisante pour obtenir un codage fidèle et représentatif. Elle varie de quelques répétitions, lorsque le dictionnaire représente un mot isolé, à plusieurs minutes de parole, lorsque le dictionnaire doit être capable de coder un énoncé quelconque indépendamment du locuteur. Il existe des algorithmes permettant dans une première phase d'obtenir un dictionnaire initial puis d'optimiser (optimum local) de manière itérative ce dictionnaire [Gray 84]. Le plus connu d'entre eux est l'algorithme de Y. Linde, A. Buzo et R.M. Gray [Linde 80], souvent appelé algorithme LBG.

Le quantifieur vectoriel que nous venons de décrire est un quantifieur simple et sans mémoire mais d'autres modèles ont été définis comme le quantifieur matriciel [Burton 85a], le quantifieur à mémoire ou le quantifieur adaptatif [Gray 84].

## 2.2. Application à la reconnaissance automatique de la parole

Du point de vue de la reconnaissance globale, la quantification vectorielle permet de résoudre les problèmes d'alignement temporel par des calculs qui sont moins coûteux que ceux de la programmation dynamique. Du point de vue de la reconnaissance analytique, elle permet de s'affranchir de la segmentation. C'est aussi une méthode efficace pour représenter des références multiples dans les systèmes indépendants du locuteur (clustering).

Prenons comme exemple la reconnaissance de mots isolés, où la démarche est la suivante. Un dictionnaire est construit pour chaque mot de référence du vocabulaire. Le mot inconnu est codé avec chacun des dictionnaires en déterminant pour chaque intervalle d'analyse son représentant dans le dictionnaire. Les distorsions minimales sont cumulées sur le mot. Le dictionnaire qui obtient le plus petit cumul fournit le mot reconnu [Burton 85b].

Si, en codage de la parole, un dictionnaire de 1024 entrées est nécessaire pour coder correctement les spectres à court terme des différents sons, en reconnaissance de la parole, des dictionnaires de 64, 128 ou 256 entrées suffisent [O'Shaughnessy 87].



## 2.3. Application à la reconnaissance automatique du locuteur

### 2.3.1. Introduction

En RAL, la quantification vectorielle présente les mêmes avantages qu'en reconnaissance automatique de la parole (RAP). Selon le degré d'indépendance par rapport au texte, il est construit un dictionnaire par locuteur pour tout le corpus d'apprentissage, ou bien un dictionnaire par mot. Les dictionnaires d'un locuteur sont construits indépendamment des autres locuteurs, sauf dans la première étude où les auteurs essaient d'extraire les segments acoustiques qui sont les plus communément émis par un locuteur et qui le distinguent des autres locuteurs. Lors de la phase de reconnaissance, l'énoncé inconnu est codé à l'aide de chacun des dictionnaires. En général, les distorsions minimales entre chacun des vecteurs de l'énoncé et son représentant dans le dictionnaire sont moyennées sur l'énoncé. La plus petite de ces moyennes donne le locuteur reconnu. Dans les descriptions qui suivent, lorsque rien n'est précisé, c'est cette distance qui sert à discriminer les locuteurs. La taille des dictionnaires varie de 16 à 256 selon le degré d'indépendance par rapport au texte. Comme nous allons le voir, de nombreux paramètres acoustiques ont été testés ainsi que de nombreuses distances entre ceux-ci.

### 2.3.2. Les études

Nous avons retenu les études qui nous ont paru les plus pertinentes. Nous les présentons dans l'ordre chronologique des publications sauf pour la seule étude concernant la langue française que nous avons citée en dernier.

**Li et Wrench [Li 83].** A partir de 100 secondes de parole pour chacun des 11 locuteurs masculins, K.P. Li et E.H. Wrench établissent d'abord un dictionnaire général de 1000 entrées. Pour cela, ils utilisent la distance d'Itakura calculée sur 12 coefficients LPC. Puis, ils extraient les 400 entrées qui sont les plus représentatives du corpus d'apprentissage ; elles codent 90% de ce corpus. De ce sous-ensemble, 40 entrées sont sélectionnées pour constituer le dictionnaire de chaque locuteur. Une entrée est choisie si elle code plus souvent les données de ce locuteur que celles des autres locuteurs et si les deux fréquences d'apparition sont stables lorsqu'on augmente la taille du corpus d'apprentissage. Finalement, les 40 entrées caractérisant un locuteur représentent 25 à 40% de son corpus d'apprentissage et seulement 7 à 12% de ceux des autres locuteurs.

Lors de la phase de reconnaissance, une semaine plus tard, une phrase quelconque est codée avec chacun des modèles des locuteurs. La comparaison, entre les locuteurs, des distributions des distorsions minimales calculées au cours de la phrase fournit le locuteur reconnu. Le taux d'identification varie avec la longueur de la phrase de test : de 79% pour 3 secondes à 96% pour 10 secondes. Les auteurs ont terminé leur étude en faisant subir aux données deux types de distorsion. Tout d'abord, l'application aux deux corpus d'une déconvolution aveugle (cf. page 10) double le taux d'erreur. Enfin, lorsque seul le corpus de test subit la distorsion due au téléphone mais que les deux corpus sont adaptés par la déconvolution, le taux d'erreur est multiplié par trois.

**Shikano [Shikano 85].** L'auteur réalise un système d'IAL indépendant du texte comme composante d'un système de reconnaissance de parole indépendant du locuteur. Lorsqu'une phrase inconnue est prononcée, le système d'IAL reconnaît le locuteur et sélectionne le dictionnaire

de phonèmes associé à ce locuteur pour le transférer au système de RAP qui reconnaît alors la phrase prononcée. 9 locuteurs prononcent chacun 10 phrases différentes de 2,5 secondes, 5 d'entre elles servent à construire le dictionnaire associé au locuteur. L'auteur propose une distance qui combine les coefficients d'autocorrélation aux coefficients cepstraux issus d'une analyse LPC d'ordre 12, et qui favorise les pics spectraux. Le taux de reconnaissance est étudié en fonction de la taille du dictionnaire et de la longueur de l'énoncé de test. Pour une durée de 7,5 secondes, 32 entrées suffisent (100%) alors que, pour une phrase de 2,5 secondes, il faut des dictionnaires de 128 ou 256 entrées pour atteindre 98%. Afin de tenir compte des phénomènes de coarticulation, l'auteur remplace la quantification simple, qui code un seul vecteur à la fois, par une quantification matricielle, qui code en une seule fois une séquence temporelle de vecteurs de longueur fixe ou variable. Les résultats obtenus sont moins bons qu'avec la quantification vectorielle simple. La raison en est peut-être la petite taille du corpus d'apprentissage.

**Soong et al. [Soong 85] [Rosenberg 86] [Soong 86].** Ces trois publications correspondent à un travail de fond sur l'application de la quantification vectorielle à l'identification du locuteur.

Dans une première étude [Soong 85], 50 locuteurs (H) et 50 locutrices (F) ont enregistré, via un poste téléphonique, 20 répétitions des dix chiffres, en 5 sessions réparties sur une période de 2 mois. Chaque énoncé comprend les dix chiffres isolés prononcés dans un ordre aléatoire. La paramétrisation du signal de parole est réalisée par une analyse LPC d'ordre 8 et la distance est encore une des variantes de la distance d'Itakura-Saito. Un dictionnaire est associé à chacun des locuteurs à partir des 10 premières répétitions de tous les chiffres. Pour un énoncé de dix chiffres distincts, le taux d'identification passe de 66% à 98,5% lorsque le nombre d'entrées des dictionnaires passent de 2 à 64. Pour des dictionnaires de 64 vecteurs, le taux passe de 76% pour un chiffre à 98,5% lorsque l'énoncé contient les dix chiffres. La figure B.1 compare les taux d'erreur obtenus avec des énoncés contenant 10 répétitions du même chiffre. Notons la très bonne performance du chiffre "nine" (2% d'erreur), suivie de celles de "two" et "five". Le taux de reconnaissance varie également avec le temps séparant les sessions. De 100% pour des énoncés appartenant à la même session, il chute à 96% pour des énoncés enregistrés à un mois d'intervalle, ce qui montre la nécessité de mettre à jour régulièrement les dictionnaires des locuteurs.

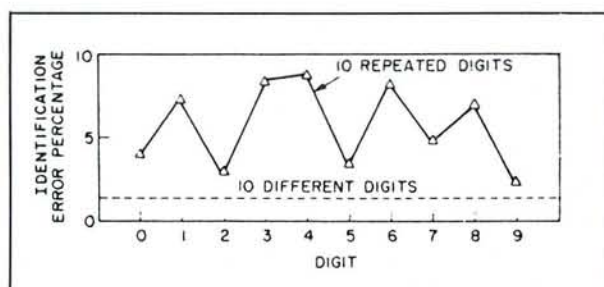


Figure B.1. Erreur d'identification en fonction du chiffre utilisé, d'après [Soong 85].

Dans une seconde étude [Rosenberg 86], les auteurs changent de paramètres. Les locuteurs possèdent maintenant deux dictionnaires, un pour les valeurs instantanées des coefficients cepstraux, l'autre pour leur évolution temporelle. F.K. Soong et A.E. Rosenberg montrent la nécessité de pondérer les coefficients cepstraux par l'inverse de leur écart-type. Les valeurs dynamiques des coefficients cepstraux sont moins pertinentes pour la RAL que leurs valeurs



instantanées mais améliorent la reconnaissance et sont moins sensibles à l'hétérogénéité des canaux entre la phase d'apprentissage et la phase de test. Toutefois, l'utilisation de ces nouveaux paramètres n'augmente pas le taux d'identification (98% pour 10 chiffres et 64 entrées). En vérification, le taux d'erreur EER est de 2% [Soong 86].

A notre avis, ces deux premières études ne sont pas indépendantes du texte puisque le locuteur doit prononcer uniquement une suite de chiffres même si l'ordre est aléatoire. Les auteurs, qui eux les considèrent comme indépendantes du texte, réalisent, dans une troisième étude [Soong 86], une vérification entièrement dépendante du texte mais complexe. Cinq références par chiffre sont codées avec les dictionnaires de chacun des locuteurs. Le chiffre prononcé par le locuteur à vérifier est comparé aux codages des cinq références d'un chiffre, par un algorithme de programmation dynamique. Par rapport à la méthode "indépendante du texte", le taux d'erreur diminue seulement de 1% pour 10 chiffres mais de 3% pour 5 chiffres. Enfin, une double adaptation temporelle, du dictionnaire d'un locuteur et d'une des cinq références d'un chiffre améliore légèrement le taux de vérification. Chacune des entrées du dictionnaire d'un locuteur est remplacée par une combinaison linéaire entre cette entrée et la moyenne des vecteurs qu'elle code dans les dix derniers énoncés qui ont permis de bien reconnaître le locuteur.

**Burton et al. [Buck 85] [Burton 87].** Après avoir étudié trois méthodes de quantification vectorielle pour la reconnaissance de chiffres isolés indépendante du locuteur [Burton 85a] [Burton 85b], les auteurs les appliquent dans une première étude à l'identification du locuteur puis, dans une seconde, à la vérification du locuteur.

La première méthode est une QV classique établissant un dictionnaire par mot et par locuteur. La seconde, appelée QV multisection, construit K dictionnaires ordonnés dans le temps par mot et par locuteur. Pour cela, le mot –ici un chiffre– est découpé en K sections contenant le même nombre de fenêtres d'analyse (analyse LPC d'ordre 10). Les énoncés des dix chiffres, qui ont des durées voisines, sont ramenés au même nombre de fenêtres en jouant sur le recouvrement de ces fenêtres. La dernière méthode, par quantification matricielle, associe un seul dictionnaire à chaque mot et à chaque locuteur, mais une entrée du dictionnaire est une séquence temporelle de N vecteurs codant en une seule opération N fenêtres d'analyse. L'objectif de ces deux dernières techniques est de prendre en compte des phénomènes de durée, de coarticulation et de contexte, ce que ne fait pas la méthode classique.

La première étude [Buck 85] réalise une identification doublée d'une vérification à l'aide de la QV multisection avec 3 dictionnaires de 8 entrées par chiffre pour chacun des 16 locuteurs autorisés (8 H et 8 F). Pour la vérification, un seuil de rejet est établi pour chacun des locuteurs à partir des distributions des distorsions intralocuteur et interlocuteur. Ces distributions sont obtenues en codant la totalité du corpus d'apprentissage (17 répétitions des locuteurs et une répétition de 109 imposteurs (54 H et 55 F)) avec les dictionnaires calculés sur une partie du corpus (9 répétitions). Le seuil est choisi de manière à rendre égales les probabilités de fausse acceptation et de faux rejet en supposant que les distributions suivent des lois normales. L'énoncé de test étant composé des 10 chiffres, le système obtient 0% de confusion et 0,8% de faux rejet pour 8 répétitions des 16 locuteurs et 1,8% de fausse acceptation pour 2 répétitions de 111 autres imposteurs (54 H et 57 F).

En conservant les conditions précédentes, la deuxième étude [Burton 85b] compare les performances des trois techniques de quantification dans le cas de la vérification. Alors qu'elles



ont le même objectif, la QV multisection obtient les meilleurs taux de vérification et la quantification matricielle les plus mauvais. Nous pensons que ces mauvais résultats de la quantification matricielle, obtenus à la fois par D.K. Burton et K. Shikano, sont peut-être la conséquence d'un manque de répétitions et de codes pour coder des diphones ou des triplets phonétiques.

**Xu et al. [Xu 89a] [Xu 89b] [Xu 91].** Comme H. Hermansky l'a fait dans le cadre de la RAP [Hermansky 90], les auteurs examinent la pertinence de l'analyse PLP (Perceptual Linear Predictive) dans le cadre de la RAL. Cette analyse transforme le spectre à court terme du signal de parole en un "spectre auditif" en lui appliquant les transformations non linéaires effectuées par l'oreille humaine au niveau de la fréquence et de l'amplitude.

Dans un premier temps, ils comparent l'efficacité des coefficients cepstraux issus de l'analyse PLP à ceux issus de l'analyse LPC. L'étude porte sur l'identification de 10 locuteurs (5 H et 5 F) à partir d'un seul chiffre. Un dictionnaire est construit pour chaque locuteur, tous chiffres confondus. Les auteurs testent également trois distances qui servent au codage et à l'identification : la distance euclidienne, la distance pondérée par le numéro d'ordre du coefficient et la distance pondérée par l'inverse de l'écart-type du coefficient. L'analyse PLP donne de meilleurs résultats que l'analyse LPC à condition que les coefficients cepstraux soient pondérés. Les deux pondérations se révèlent équivalentes. Dans tous les cas, les taux de reconnaissance se stabilisent à partir de 128 entrées par dictionnaire.

Un prolongement de l'étude montre que les 8 premiers coefficients d'une analyse PLP d'ordre 14 sont les plus pertinents et que la prise en compte de l'évolution temporelle des coefficients (un dictionnaire supplémentaire par locuteur) améliore le taux d'identification.

Les auteurs terminent leur exploration de l'analyse PLP dans le cadre de la reconnaissance du locuteur en faisant varier les paramètres de l'analyse d' Hermansky. En conclusion générale, la meilleure identification du locuteur (97%) est obtenue à partir des coefficients cepstraux pondérés par leur numéro d'ordre et issus d'une analyse PLP d'ordre 14, identique à celle d'Hermansky mais comprenant plus de filtres auditifs (64 au lieu de 17).

**Ren-Hua et al. [Ren-hua 90].** L'objectif des auteurs est de trouver une distance plus efficace entre les coefficients cepstraux issus de l'analyse LPC. Cette distance est une variante de la distance pondérée par l'inverse de l'écart-type. La variance d'un coefficient estimée globalement sur tout le corpus d'apprentissage d'un locuteur est remplacée par 16 variances calculées sur les 16 partitions du corpus représentées par les 16 entrées du dictionnaire de ce locuteur. Le corpus d'apprentissage est constitué des 10 chiffres chinois complétés par les noms des locuteurs (30H et 20 F). L'identification d'un locuteur est faite à partir des vecteurs de 10 coefficients cepstraux issus d'un énoncé de 4 chiffres et d'un nom.

Quelle que soit la distance utilisée, les résultats sont très bons : 96,4% pour la distance euclidienne, 98,2% pour la distance pondérée par l'inverse de l'écart-type et 99,1% pour la nouvelle distance. Mais il est difficile de faire une comparaison avec les études précédentes car la langue n'est pas la même, le nombre de locuteurs non plus, et l'intervalle de temps séparant les répétitions n'est pas précisé.

**Chi-Shi et al. [Chi-Shi 90]** Les fréquences LSP (Line Spectrum Pair) dérivées de l'analyse LPC, introduites en 1975 par F. Itakura [Itakura 75a], ont déjà fait leurs preuves en codage de la parole et en RAP. Dans cette étude, en mandarin, plusieurs combinaisons linéaires de ces fréquences sont comparées aux coefficients cepstraux. Quatre d'entre elles sont plus

performantes que le cepstre dans l'identification de 20 locuteurs (16 H et 4 F). La méthodologie employée est celle de F.K. Soong et al. [Soong 85], sauf pour la distance qui est euclidienne. Les meilleurs paramètres sont les différences entre les fréquences LSP adjacentes qui, d'après les auteurs, sont corrélées aux largeurs de bande des formants. Le passage de tout le corpus par le canal téléphonique met en évidence l'invariance de ce paramètre (toujours 100% pour 10 chiffres) mais aussi la relativement bonne robustesse des coefficients cepstraux dont le taux d'identification ne passe que de 97,5% à 95% pour 10 chiffres et pour des dictionnaires de 32 entrées.

**Matsui et al. [Matsui 90] [Matsui 91]** L'objectif de ces travaux est de créer un système d'identification et de vérification indépendant du texte qui soit robuste vis-à-vis de la variabilité temporelle intralocuteur des paramètres. Les auteurs augmentent la robustesse de la reconnaissance en jouant sur trois points, la pondération des paramètres, le calcul de la distance entre les locuteurs et la construction d'un dictionnaire par grande classe phonétique.

En plus des distributions intralocuteurs de chaque composante du vecteur de paramètres, le coefficient de pondération prend en compte les distributions interlocuteurs.

Nous rappelons qu'habituellement la distance entre l'énoncé inconnu et la référence du locuteur est donnée par la moyenne sur l'énoncé des distorsions minimales entre un vecteur de l'énoncé et un mot du dictionnaire du locuteur. Pour s'affranchir de l'indépendance par rapport au texte et de la variabilité temporelle des paramètres, les auteurs préconisent de ne prendre en compte que les vecteurs de l'énoncé qui ne sont pas plus éloignés des mots du dictionnaire que ceux du corpus d'apprentissage.

Au lieu d'avoir un seul dictionnaire par locuteur, les auteurs associe à chacun des locuteurs un dictionnaire (de taille inconnue) par grande classe phonétique. Dans la première étude, la classification phonétique s'effectue sur le voisement. Elle est remplacée dans la seconde étude par une classification plus générale en K grandes classes phonétiques en utilisant un modèle HMM ergodique à K états et en construisant un dictionnaire par classe pour chaque locuteur. Les paramètres du réseau sont estimés par l'algorithme de Baum-Welch et la classification phonétique est réalisée à l'aide de l'algorithme de Viterbi. Le modèle à deux états améliore le taux d'identification par rapport à la classification voisé/non voisé mais l'augmentation du nombre d'états nécessite un plus grand corpus d'apprentissage.

Les quatre répétitions du corpus, enregistrées par 9 locuteurs, ont été réparties sur 3 ans. Dans la première étude, les paramètres, extraits d'une analyse LPC d'ordre 16, sont les coefficients cepstraux, la fréquence fondamentale et l'évolution temporelle de tous ces coefficients. L'énoncé de test dure 30 secondes. L'identification des locuteurs par l'association des valeurs instantanées aux valeurs transitoires du cepstre donne de moins bons résultats que dans l'étude similaire, mais dépendante du texte, de F.K. Soong et al. (92% vs. 98%). Le fondamental seul est un très mauvais paramètre (80% d'erreur, alors qu'une identification aléatoire conduirait à 88% d'erreur). Mais il améliore les performances du cepstre. Toutefois, il a été abandonné dans la seconde étude.

**Thévenaz et al. [Thevenaz 90].** Nous terminons cette revue des études sur l'emploi de la quantification vectorielle dans la reconnaissance du locuteur par la seule étude connue sur le français. En fait, celle-ci combine trois méthodes fondées sur la QV à une méthode statistique pour obtenir un système de vérification indépendant du texte. Le seuil de vérification est établi de manière à obtenir un EER. Les paramètres de base sont 20 coefficients cepstraux issus



d'une analyse LPC. 10 locuteurs (9 H et 1 F) prononcent lors d'une même session 8 phrases différentes de 15 secondes construites à partir de 20 nombres différents. A notre avis, il demeure une certaine ambiguïté sur les termes "d'indépendance par rapport au texte" et de "phrases différentes" employés par les auteurs, sachant que le vocabulaire est strictement numérique. Pour toutes les méthodes, le corpus d'apprentissage se limite à une seule phrase. La méthode statistique consiste à calculer la moyenne sur une phrase de chacun des coefficients cepstraux. La première méthode de QV construit un dictionnaire de 32 entrées par locuteur et obtient de piètres résultats. Les 80% de bonne vérification obtenus sont à comparer aux 96% obtenus par D.K. Burton [Burton 87], dans le cas de la vérification sur un énoncé de dix chiffres, et aux 100% obtenus par K. Shikano [Shikano 85], dans le cas de l'identification indépendante du texte sur un énoncé de 7,5 secondes. La seconde méthode effectue le codage de la différence de deux cepstres adjacents. La dernière engendre un dictionnaire universel de 256 mots et compare les distributions des fréquences d'apparition de chacune des 256 entrées obtenues sur la phrase de référence, d'une part, et sur la phrase de test, d'autre part.

La meilleure méthode utilisée seule est la méthode statistique. La combinaison des distances des quatre méthodes, selon l'algorithme d'analyse discriminante linéaire de Fisher, engendre un taux de vérification de 94%.

### 3. Les modèles de Markov

#### 3.1. Introduction et application à la reconnaissance de la parole

Les modèles de Markov cachés constituent une technique probabiliste destinée à l'étude de séries temporelles discrètes. Cette technique se fonde sur des méthodes stochastiques : la série temporelle est analysée par un modèle probabiliste paramétrique, c'est-à-dire complètement décrit par une liste finie de nombres réels.

Les premiers travaux sur le sujet datent du début du siècle. La théorie des chaînes de Markov [Markov 13] associée à la théorie de la communication de Shannon [Shannon 48] ont été à l'origine de nombreuses études qui ont conduit notamment aux modèles de Markov cachés. Les champs d'application en sont multiples : reconnaissance de la parole, modélisation du langage, codage, traitement du signal, biostatistiques, finance ...

Dans le domaine de la parole, les publications de F. Jelinek ou S.E. Levinson, par exemple, permettent de se familiariser avec la technique [Jelinek 76] [Levinson 86]. Une chaîne de Markov est composée d'un nombre fini d'états correspondant à des segments stables du signal de parole et des arcs de transition modélisant les variations entre vecteurs de paramètres (des variations spectrales, par exemple) [Haton 91]. Un tel graphe peut être vu comme un modèle de production de la parole, production d'un mot, par exemple. Chaque transition correspond ainsi à la production d'un vecteur de paramètres. Deux distributions de probabilité sont associées à un modèle : les probabilités  $P(a_i/s_j)$  de produire un événement  $a_i$  sur une transition issue d'un état  $s_j$ , et les probabilités  $P(s_j/s_i)$  que le modèle passe de l'état  $s_i$  à l'état  $s_j$ . Les paramètres du modèle sont obtenus par apprentissage. La phase d'apprentissage du modèle peut être effectuée grâce à un algorithme itératif (algorithme de Baum-Welch). En phase de reconnaissance, l'algorithme de Viterbi, par exemple, permet de calculer la probabilité d'émission par le modèle de l'unité (par exemple le mot) à reconnaître.



Dans les problèmes complexes où il existe une hiérarchie de niveaux d'information, comme c'est le cas pour la parole, il a été proposé de construire des réseaux de modèles. Au niveau le plus fin, on trouve un ensemble de modèles élémentaires. Les niveaux supérieurs sont successivement construits par imbrication des modèles d'ordre inférieur. Ce modèle de "réseau intégré" a été appliqué au système HARPY de reconnaissance automatique de la parole [Lowerre 76], dans lequel toutes les phrases du langage peuvent être énumérées, et affiné depuis (cf. les systèmes DRAGON [Baker 75] ou TANGORA [Jelinek 76]).

Nous envisageons dans la suite de ce paragraphe les études faisant appel à la théorie des modèles de Markov cachés en reconnaissance du locuteur. Nous citerons simplement, avant 1988, les travaux de pionnier de A.B. Poritz dans lesquels chacun des locuteurs était représenté par un modèle ergodique (c'est-à-dire autorisant toutes les transitions entre états) à cinq états. A.B. Poritz obtenait de bons résultats dans la discrimination entre dix locuteurs à partir d'un apprentissage de 40 secondes en mode dépendant du texte [Poritz 82].

### 3.2. Application à la reconnaissance automatique du locuteur

Tishby [Tishby 88] [Tishby 91].

Dans le premier de ces articles, l'auteur discute la difficulté de séparer, dans le signal de parole, les propriétés dépendant du locuteur des propriétés "universelles". Dans le cas d'une modélisation statistique, les différentes caractéristiques du signal se trouvent mêlées dans les paramètres d'estimation du modèle. Le problème peut se qualifier de modélisation avec information a priori, cette dernière pouvant relever de la connaissance sur la langue, sur le vocabulaire utilisé, ou même l'âge ou le sexe du locuteur. Une approche possible est la division de la phase de modélisation en deux étapes : dans la première s'effectue une modélisation statistique de l'information a priori, dans la seconde ce sont les écarts par rapport à la modélisation initiale qui sont pris en compte.

Le modèle initial est ainsi chargé de capter les caractéristiques communes à différents locuteurs de même sexe, de même langue, pour un vocabulaire commun. La connaissance initiale est ainsi convertie en une fonction de distribution de probabilité. Une fois cette distribution obtenue, une procédure de minimisation des écarts entre les données observées et la distribution a priori fournit un ensemble de multiplicateurs de Lagrange. L'auteur reprend le concept de minimum discrimination information (MDI) proposé dans [Ephraim 87] en donnant une autre connotation à la notion de distribution a priori et en utilisant les multiplicateurs de Lagrange comme paramètres spécifiques du locuteur.

L'expérimentation a été effectuée à partir d'une base de données de 20000 élocutions des chiffres, de façon isolée, par 100 locuteurs (50 hommes et 50 femmes), par le canal téléphonique. Deux modèles a priori ont été créés par une méthode de clustering, à partir de vecteurs de densité spectrale correspondant aux élocutions de 30 locuteurs d'une part et de 30 locutrices de l'autre. Un test à partir des élocutions des locuteurs n'ayant pas participé à l'apprentissage a permis de mettre en évidence que le sexe du locuteur était correctement identifié pour 90% des locuteurs sur la base de cinq chiffres.

L'apprentissage des modèles dépendants du locuteur a été effectué par clustering des données individuelles avec le modèle initial, selon la méthode de Viterbi (regroupement des prélèvements du locuteur suivant l'état le plus proche avec la distance MDI) puis moyennage des matrices d'autocorrélation à l'intérieur de chaque état et calcul d'une matrice de multiplicateurs

de Lagrange). Le modèle définitif comprend ainsi, pour un locuteur donné, le modèle a priori et un ensemble de multiplicateurs de Lagrange.

Les tests de vérification ont été effectués à partir de 10 locuteurs masculins et 10 locuteurs féminins. Les résultats sont jugés encourageants et ouvrent de nouvelles perspectives, y compris en reconnaissance de la parole.

Dans son article de 1991 [Tishby 91], l'auteur montre qu'avec des modèles de Markov cachés une suite de quatre mots isolés suffit pour vérifier un locuteur avec moins de 3% d'erreur. Le travail de Tishby reprend l'idée originale de A.B. Poritz déjà cité et l'étend à une classe plus riche de modèles, les mixture autoregressive HMM (comme [Savic 90] décrit plus loin). Dans ces modèles, les états sont décrits par une combinaison linéaire (c'est le sens du mot mixture) de sources autorégressives. On peut montrer que les AR-HMM ainsi construits équivalent à des HMM à états simples plus grands avec des contraintes sur les transitions entre états. Le travail ici rapporté concerne la construction d'un HMM utilisant l'algorithme de quantification vectorielle. Le codebook est imbriqué dans les états et les mixtures du modèle de Markov.

Les conditions expérimentales sont celles de l'expérimentation de Rosenberg et al. décrite plus loin [Rosenberg 90a]. L'auteur démontre que des modèles de Markov cachés autorégressifs convenablement entraînés peuvent être utilisés pour caractériser statistiquement des locuteurs, d'une façon indépendante du texte. Comme pour la quantification vectorielle, on peut s'attendre à ce que la modélisation adaptative améliore les résultats mais l'adaptation correcte des modèles reste un problème ouvert.

**Zeng et Yuan [Zheng 88].** Dans cette étude, les auteurs utilisent une classe particulière des modèles de Markov, les modèles de Markov cachés circulaires (CHMMs) dont les propriétés les distinguent des HMMs classiques à propagation de gauche à droite.

6 hommes et 4 femmes de langue chinoise ont participé à l'expérimentation. La parole, filtrée par un filtre passe-bas à 3.4 kHz et échantillonnée à 10 kHz, a ensuite été segmentée en intervalle de 12 ms et analysée par un algorithme de LPC d'ordre 12. Par clustering sur les vecteurs LPC obtenus, un codebook de 64 codes a été obtenu. Puis les matrices des modèles individuels ont été calculées. L'apprentissage a été effectué à partir de la lecture de 7 phrases et le test d'identification à partir de l'énoncé de ces phrases. Les taux de reconnaissance du locuteur obtenus montrent la supériorité des CHMMs (93.7% d'identification) sur les HMMs série (88.7%) ou parallèle (90%).

**Naik et al. [Naik 89].** Le travail présenté s'intéresse aux communications téléphoniques longue distance. Tout d'abord, deux bases de données vocales ont été constituées : une base de quatre phrases différentes enregistrées en laboratoire par 10 hommes et 10 femmes, avec utilisation de dix combinés téléphoniques différents ; une base enregistrée à travers le réseau téléphonique public à partir d'une population de 100 locuteurs, sur une période de quatre mois par locuteur, comprenant trois phrases, le nom du locuteur et le numéro en dix chiffres du téléphone d'appel.

A partir des paramètres LPC, 32 paramètres ont été calculés : niveaux d'intensité, taux de variation spectrale, intensités dans des bandes de fréquences mel, variations temporelles de l'énergie et des sorties de filtres sur une fenêtre de 40 ms. 18 de ces paramètres, contenant 95% de la variance totale des données, ont été retenus à l'aide d'une analyse discriminante.

Les auteurs mettent en comparaison l'utilisation de programmation dynamique et des modèles de Markov. Dans la première méthode, les formes de référence sont créées par



application de la méthode de programmation dynamique qui effectue un alignement temporel entre une forme présentée et une forme de référence et moyennage de ces deux formes en cas de reconnaissance correcte. En phase de vérification du locuteur, des milliers d'essais d'authentification du locuteur et d'essais de détection d'imposteurs ont été effectués, à partir des deux corpus de données vocales, par comparaison dynamique entre références et données de tests. Les scores de mise en correspondance servent d'entrée à un module de décision. Des schémas permettent d'apprécier les taux de faux rejet et ceux de fausse acceptation en fonction de différents ensembles de seuils de décision.

Une expérience de vérification du locuteur a été également menée à partir du premier corpus à l'aide de deux réalisations de modèles de Markov cachés. Les résultats obtenus sont en faveur de ces derniers modèles.

**Noda et Yanagida [Noda 90].** Les HMM sont ici utilisés pour effectuer une segmentation du signal de parole. L'objet de l'étude est la reconnaissance du locuteur indépendante du texte en exploitant les caractéristiques individuelles du locuteur associées aux phonèmes, sans reconnaissance directe des phonèmes. Les auteurs font l'hypothèse que la segmentation peut se faire correctement après estimation des paramètres des modèles de Markov à partir des transcriptions phonétiques d'un grand nombre de données d'entrée. Ils fondent ensuite la reconnaissance sur une mesure de dissemblance entre les vecteurs de paramètres de référence et les vecteurs d'entrée appartenant au même phonème, avec cumul sur l'ensemble des phonèmes du mot.

Dans l'expérimentation, les prononciations de 20 mots par 177 locuteurs masculins via le canal téléphonique ont été utilisées. Parmi ces mots, 10 sont utilisés en référence, les autres servent à la vérification. Les conditions de traitement du signal sont les suivantes : filtrage passe-bas à 4.5 kHz, numérisation à 10 kHz, préaccentuation à l'aide d'un filtre adaptatif du premier ordre, analyse LPC d'ordre 12 sur des fenêtres de Hamming de 25.6 ms toutes les 12.8 ms. Les paramètres retenus finalement sont les coefficients cepstraux.

Les auteurs annoncent 95.4% de taux de vérification contre 89.3% par une méthode faisant appel à la quantification vectorielle. Ils estiment que cette dernière méthode présente l'inconvénient de ne pas utiliser correctement les caractéristiques individuelles du locuteur associées aux phonèmes car un vecteur d'entrée et le centro.de le plus proche n'appartiennent pas toujours au même phonème.

**Rosenberg et al. [Rosenberg 90a] [Rosenberg 90b] [Rosenberg 91].** Les trois articles référencés présentent les résultats de l'application de la représentation d'unités de parole par des modèles de Markov à la vérification automatique du locuteur. Dans les deux premiers articles, les unités de parole considérées sont des unités "sub-lexicales" de deux types : des unités de type phonèmes (PLUs comme phone-like units) qui s'appuient sur la transcription phonétique des mots pour la segmentation des mots avant apprentissage des HMMs, et des segments acoustiques (ASUs comme acoustic segment units) qui sont directement extraits du signal sans utilisation de connaissances linguistiques. Dans le troisième article, les unités de parole modélisées sont des mots extraits d'un corpus de mots enchaînés.

La principale motivation de ces deux études est que la représentation par HMM des sons de la parole doit donner de meilleurs résultats que les méthodes classiques de comparaison entre une forme de test et une ou plusieurs formes de référence. Par ailleurs, l'idée d'utiliser des unités sub-lexicales dans la première étude vient de la possibilité, dans un système de



vérification, d'étendre ainsi de façon significative le champ d'application de ce système. En effet, si un ensemble de modèles robustes de ces sous-unités peut être construit pour chacun des locuteurs, la vérification peut s'effectuer en mode dépendant du texte ou indépendant du texte à partir de phrases sans restriction de vocabulaire. Ce dernier point est un élément important de la discussion engagée dans [Rosenberg 90b] : les notions de reconnaissance dépendante ou reconnaissance indépendante du texte sont plus flexibles lorsque sont utilisées des unités sub-lexicales. En effet, un tel système dans lequel l'apprentissage a été fait de façon dépendante du texte peut être utilisé pour la vérification, soit en mode indépendant en considérant que toutes les unités sont également probables ou que leurs probabilités d'observation reflètent leurs fréquences dans la langue considérée, soit en mode dépendant en prenant pour modèles des mots ou des phrases la concaténation des modèles des unités sub-lexicales.

La première étude comporte deux expériences. Dans la première, à partir d'un corpus de chiffres prononcés isolément, les auteurs évaluent les performances de leur système de vérification en fonction du type d'unités sub-lexicales, du modèle HMM, de la durée de l'énoncé de test et de la dépendance par rapport au vocabulaire. Ils en déduisent ainsi un système de vérification qu'ils appliquent dans la seconde expérience à un corpus de phrases lues. Nous allons détailler ces deux expériences.

La première base de données [Rosenberg 90a] comprend 200 élocutions des dix chiffres par 20 locuteurs (10 H et 10 F), à travers un combiné téléphonique ordinaire (communication locale), obtenues lors de cinq sessions sur une période de deux mois. 80 élocutions sont utilisées pour l'apprentissage, les 120 autres servent aux tests. L'analyse du signal a été effectuée ainsi : filtrage dans la bande 200-3200 Hz, échantillonnage à 6.67 kHz, préaccentuation par un filtre du premier ordre, autocorrélation d'ordre 8 sur des blocs de 45 ms, application d'une fenêtre de Hamming déplacée toutes les 15 ms, conversion de chaque vecteur de coefficients de corrélation en 12 coefficients cepstraux par l'intermédiaire d'une analyse LPC d'ordre 12. Ce dernier vecteur est complété par 12 paramètres fournissant une estimation des pentes cepstrales par une technique de régression. Une pondération de type sinusoïdale fournit le vecteur de 24 paramètres définitifs.

Les mots du corpus d'apprentissage sont segmentés et étiquetés en 20 PLUs et 16 ASUs par des techniques dépendant de ces sous-unités [Rosenberg 90a]. De même, la forme de l'apprentissage est différente selon les unités retenues (dépendant du texte pour les PLUs et indépendant pour les ASUs). A partir de cette segmentation sont déduits les modèles de Markov des unités sub-lexicales. Ces HMMs sont unidirectionnels à 2 ou 3 états et possèdent des densités de probabilité d'observation qui sont des combinaisons linéaires de une, deux ou trois lois gaussiennes multidimensionnelles. Enfin, les tests sont effectués selon deux modes, dépendant du vocabulaire : les modèles sont concaténés en fonction des transcriptions phonétiques, et, indépendants du vocabulaire : les modèles sont équiprobables.

Les résultats fournis sont des moyennes des résultats individuels de chacun des 20 locuteurs. Les auteurs observent tout d'abord une décroissance monotone des taux d'erreur avec le nombre de mots utilisés pour la vérification, les performances s'améliorant très peu au-delà de 5 mots, soit 2.5 secondes de parole environ. L'augmentation du nombre d'états et du nombre de gaussiennes améliore les taux de vérification tant que la taille du corpus d'apprentissage reste suffisante pour évaluer leurs paramètres. Les taux de reconnaissance en mode dépendant du vocabulaire sont toujours les meilleurs. Une expérience de comparaison entre les deux types d'unités retenues plaide en faveur des PLUs mais de façon très significative seulement pour



des essais courts (les taux d'erreur sont de 0.9 % et 1.3 % respectivement pour des essais de 7 chiffres).

La seconde base de données utilisée dans l'étude par unités lexicales est un sous-ensemble d'une base de phrases lues par 9 hommes et 11 femmes pour le projet DARPA [Rosenberg 90b]. Le vocabulaire est d'environ 1000 mots, la durée de l'enregistrement varie selon le locuteur de 110 à 190 secondes. Les conditions d'analyse du signal sont assez proches de celles de l'expérience précédente. Les auteurs ont choisi d'utiliser comme unités sub-lexicales des segments acoustiques (ASUs) déterminés automatiquement (32 et 64 segments de durée moyenne 80 ms). Les auteurs ne précisent pas le nombre d'états et de gaussiennes des HMMs des unités sub-lexicales ; par recoupement avec l'article précédent nous supposons qu'il s'agit de 3 états et de 3 lois gaussiennes par état. L'apprentissage a été effectué à partir de 60 et 90 secondes de parole par locuteur.

Les tests ont porté sur des durées d'élocution de 1 à 5 secondes. La meilleure vérification est obtenue pour 64 ASUs, 90 secondes d'énoncé d'apprentissage et 5 secondes d'énoncé de test, tout en étant moins performante que la meilleure vérification de l'expérience précédente (40 secondes d'apprentissage, 16 ASUs et 7 chiffres isolés). Toutefois, les résultats montrent que l'approche "sub-lexicale" peut être étendue aux grands vocabulaires ou à la parole continue. La durée des phrases d'apprentissage et de test est significativement inférieure à ce qui est généralement requis dans les approches de vérification du locuteur indépendante du texte se fondant sur des statistiques à long terme.

Dans la seconde étude [Rosenberg 91], chacun des dix chiffres est modélisé par un HMM unidirectionnel à 10 états ayant pour densité de probabilité d'observation une combinaison linéaire de lois gaussiennes multidimensionnelles, le nombre de lois variant de 1 à 4.

L'apprentissage des modèles et l'étape de vérification sont effectuées à partir de séquences de 3 chiffres enchaînés enregistrées par 20 locuteurs (10 H et 10 F), en moyenne 150 séquences par locuteur. L'intérêt de ce type d'énoncé est de permettre un assez grand nombre de combinaisons pour la phase de vérification, même avec des vocabulaires limités. En particulier, chaque personne peut être détentrice d'une clé d'identification (Personal Identification Number), ou encore, pour accroître la sécurité, avoir à prononcer une suite de chiffres tirés au hasard. Enfin, avec des modèles des mots indépendants du locuteur, les mêmes phrases peuvent à la fois être reconnues et servir à l'identification du locuteur.

Nous ne détaillons pas les conditions expérimentales de cette étude qui s'apparentent aux précédentes. Il est néanmoins intéressant de noter une discussion sur les techniques de segmentation initiale qui peuvent être utilisées lors de la phase d'apprentissage, qui est effectuée uniquement à partir de chaînes de mots (et pas les mots isolés correspondants). Les auteurs proposent trois techniques : pré-établissement de modèles de mots isolés pour chaque locuteur pour décoder les séquences de mots enchaînés, partition uniforme des chaînes de mots en mots et en états et segmentation par bootstrap faisant appel à un ensemble pré-existant de HMMs des mots connectés indépendants du locuteur. Les auteurs testent les deux dernières techniques, la dernière conduisant à un meilleur taux de vérification. Les résultats montrent également une meilleure vérification lorsqu'il y a au moins 10 exemplaires de chaque chiffre dans le corpus d'apprentissage, lorsque l'énoncé de test comporte 3 ou 4 séquences de 3 chiffres (4 s de parole) et lorsque les densités de probabilité sont modélisées par 4 lois gaussiennes.

Dans une comparaison avec une méthode de vérification se fondant sur des scores de ressemblance entre une forme de test et plusieurs formes de référence, les auteurs concluent à la supériorité de leur démarche. L'écart entre les deux méthodes est moins sensible pour



les mots isolés, pour lesquels les résultats, pour les deux méthodes, restent supérieurs à ceux obtenus avec des mots enchaînés ; cela peut s'expliquer par le fait que la modélisation de ces derniers est plus difficile à cause des effets de contexte entre mots même si elle semble mieux réalisée par les HMMs.

**Savic et Gupta [Savic 90].** Il s'agit ici d'un système de vérification du locuteur indépendante du texte fondée, comme le disent les auteurs, sur un modèle adaptatif du conduit vocal émulant le conduit vocal du locuteur (en se plaçant par opposition aux systèmes dans lesquels un locuteur est représenté par un ensemble de paramètres obtenus par moyennage de caractéristiques spectrales à court terme de la parole, ce qui revient à considérer une configuration moyenne du conduit vocal ...).

Chaque locuteur est ici représenté par un ensemble de vecteurs de paramètres issus de segments appartenant à différentes classes de phonèmes. Une modélisation dite *linear predictive hidden Markov modeling* suivie d'un décodage de Viterbi fondé sur le maximum de vraisemblance permettent d'obtenir un modèle de Markov par locuteur, les vecteurs associés à chaque état représentant une classe phonétique. En phase de vérification, la segmentation phonétique de la phrase inconnue est effectuée par référence au modèle de Markov du locuteur à vérifier. Chaque classe de phonèmes est classée comme appartenant au vrai locuteur ou à un imposteur. Le score final est obtenu par combinaison linéaire pondérée des scores obtenus par chaque catégorie individuellement.

Les conditions expérimentales furent les suivantes : 43 locuteurs, modèles de Markov ergodiques à cinq états, 52 secondes de parole pour l'apprentissage des HMM, 14 secondes de parole pour la vérification, 37 paramètres bruts (12 PARCOR, 12 coefficients cepstraux, 12 coefficients de fonction d'aire, 1 gain), réduction à 3 de la dimension de l'espace des paramètres, seuil de 60% pour le classifieur bayésien. Une table de résultats indique les taux d'erreur (faux rejet et fausse acceptation). Elle montre que la combinaison de catégories phonétiques donne des résultats supérieurs à ceux obtenus à l'aide des segments voisés, des fricatives, des nasales ou des occlusives pris isolément.

**Carey et al. [Carey 90].** La vérification est ici effectuée par comparaison des sorties de deux modèles de Markov (linéaires gauche-droite) de la même unité phonétique. Le premier modèle est spécifique du locuteur à vérifier, le second modélise une grande population de locuteurs. La mise en compétition des deux modèles présentent un double intérêt : tout d'abord, elle permet d'éviter d'avoir à positionner un seuil absolu puisque l'acceptation d'une élocution est fondée sur la différence entre les scores deux modèles ; ensuite, elle offre la possibilité d'améliorer les performances du système en traitant la paire de modèles comme un réseau connexionniste (réseau alpha [Bridle 90]) qui permet un véritable apprentissage discriminant. L'hypothèse sous-jacente est que le score de probabilité d'émission d'une occurrence d'un mot par le modèle individuel sera supérieur à celui fourni par le modèle général, alors que, au contraire, le modèle général fournira des scores plus grands quand on lui présentera un mot prononcé par un imposteur.

La base d'apprentissage était constituée de cinq répétitions des dix chiffres par 50 locuteurs, à travers un combiné téléphonique. La parole a été échantillonnée à un taux de 8 kHz et analysée par un banc de 11 filtres (échelle d'abord linéaire puis logarithmique sur les 6 derniers filtres). La transformation adéquate permet d'obtenir un cepstre Mel toutes les 20 ms dont sont retenus les six premiers coefficients. Le vecteur de paramètres est complété par six pentes cepstrales et



la dérivée de l'énergie dans le temps. Chaque mot est représenté par des modèles individuels à sept états et un modèle général à dix états.

Trois expériences sont rapportées dans l'article. Les résultats des deux premières montrent d'abord que l'adaptation des probabilités associées aux états par rétropropagation des erreurs peut corriger des erreurs de classification. La troisième expérience, quant à elle, a consisté en l'implantation d'un système temps réel sur un processeur DSP32C. Six locuteurs ont été choisis pour son évaluation à l'aide de suites de cinq chiffres. Si l'on accepte un utilisateur lorsque trois chiffres ou plus sont corrects, les résultats montrent une seule erreur sur 600 essais.

Dans le système décrit, seules la moyenne et la variance des distributions de probabilité associées aux états ont été adaptées. Les auteurs se proposent d'aller plus loin en recherchant un espace de paramètres plus performants pour la discrimination. Par ailleurs, ils veulent s'orienter vers des modèles se fondant sur des unités sub-lexicales, par référence à [Bridle 91].

**Vloeberghs et Dupont [Vloeberghs 92].** Cette étude en langue française traite de reconnaissance du locuteur indépendante du texte. A chaque phrase est associé un HMM obtenu par concaténation des modèles propres à chaque phonème. La suite des états dépend donc de la phrase mais les paramètres de ces modèles sont spécifiques du locuteur. On peut donc dire que le terme "reconnaissance indépendante du texte" utilisé par les auteurs est relatif au fait que la phrase prononcée lors de la vérification peut être tout-à-fait différente de celles prononcées en phase d'apprentissage. On retrouve ici l'usage divers fait par les auteurs des notions de reconnaissance dépendante ou indépendante du texte qui doit conduire à être vigilant si l'on veut établir des comparaisons entre les méthodes.

Chaque phonème est traduit par un petit nombre d'états (1 à 3). Les phrases d'apprentissage doivent être choisies de façon à faire apparaître les phonèmes plusieurs fois dans des contextes différents. Lors de l'identification, une phrase est demandée au locuteur. Les modèles correspondants sont construits par concaténation des modèles de phonèmes pour chaque locuteur. Le modèle à probabilité d'émission la plus élevée "identifié" le locuteur, à condition qu'elle dépasse un seuil de rejet, supposé rejeter les imposteurs.

Les conditions expérimentales sont les suivantes : 8 locutrices et 13 locuteurs (élèves-officiers francophones), 50 phrases prononcées en trois séances d'enregistrement ..., échantillonnage à 10 kHz, codage sur 16 bits, segmentation parole-non parole, calcul de 16 coefficients cepstraux toutes les 20 ms, construction des modèles de Markov à l'aide de l'algorithme de Viterbi, détermination du seuil de rejet (ce dernier point est jugé essentiel pour le succès de la méthode).

Les meilleurs résultats des tests, en mode indépendant du texte, sont de 92% d'identifications correctes et 60% de rejet correct d'imposteurs. Un tableau récapitulatif indique les valeurs de ces deux taux suivant les conditions de l'expérience : nombre maximal d'états dans les HMM, nombre maximal de transitions, nombre d'itérations lors de l'apprentissage, nombre de locuteurs utilisés pour le calcul du seuil de rejet ...

## 4. Les réseaux neuronaux

### 4.1. Introduction et application à la reconnaissance de la parole

Les réseaux neuronaux ou neuromimétiques sont des assemblages de processeurs élémentaires reliés entre eux par des connexions affectées de poids. Chaque processeur (ou "neurone formel") réalise une fonction d'intégration de ses entrées pondérées puis une transformation, souvent de type sigmoïdale, qui, en fonction d'un certain seuil, résulte finalement en une réponse du processeur. Certains neurones formels sont en prise directe avec les données, ce sont les neurones d'entrée, d'autres correspondent aux réponses du système, ce sont les neurones de sortie. Les neurones qui n'ont pas de lien direct avec l'environnement sont dits cachés.

Différents modèles de réseaux ont été proposés [Jodouin 90] : perceptrons multicouches, réseaux de Hopfield, machines de Boltzmann, cartes de Kohonen, colonnes corticales ... Les réseaux neuronaux ont d'abord été employés à des tâches de classification : ils sont dans cette tâche particulièrement bien adaptés au cas de données bruitées.

Des articles récents écrits par des chercheurs sur la reconnaissance de la parole permettent de faire le point sur les architectures neuronales [Lippmann 88] et leur intérêt potentiel en reconnaissance de la parole [Huang 88]. Des expériences de classification des sept premiers chiffres monosyllabiques à partir de leur représentation cepstrale ont montré que les taux d'erreur d'un perceptron multicouche entraîné par rétropropagation du gradient d'erreur étaient du même ordre que ceux d'un classifieur gaussien [Lippmann 87].

Les idées présentées dans [Huang 88] au sujet de la reconnaissance de mots laissent entrevoir l'intérêt des réseaux neuronaux pour la reconnaissance du locuteur : leur pouvoir de classification qui permet d'avoir des repères indépendants du locuteur, l'alignement temporel offert par les réseaux de Viterbi [Lippmann 87] ou par les réseaux à retard temporels [Bourlard 89], pour en donner deux exemples. Enfin, en dehors des performances de tels systèmes, certains grands thèmes de la recherche dans le domaine des réseaux neuronaux émergent des travaux menés en vue de l'identification du locuteur : computational tractability, scalability notamment [Oglesby 90].

### 4.2. Application à la reconnaissance automatique du locuteur

**Bennani et al. [Bennani 90] [Bennani 91] [Bennani 92].** L'article de 1990 présente une approche connexionniste fondée sur l'algorithme de quantification vectorielle pour l'apprentissage LVQ (Linear Vector Quantization) [Kohonen 88]. Les classifieurs fondés sur cet algorithme fonctionnent selon le principe du plus proche voisin qui a fait ses preuves en classification de phonèmes notamment [McDermott 89]. L'objet du système mis en place est de comparer différentes méthodes d'analyse (coefficients LPC et MFCC, modèles de phrases divers, classifieurs bayésiens ou LVQ).

La base de données utilisée comprend 10 élocutions de 10 phrases phonétiquement équilibrées par 10 locuteurs français (5 H et 5 F), enregistrées sur bande audio en environnement de bureau, puis filtrées dans la bande 0-4000 Hz, préaccentuées et numérisées à l'aide de la carte OROS à 10 kHz sur 16 bits. Deux méthodes de paramétrisation ont été appliquées : analyse LPC d'ordre 12 ou calcul de 8 coefficients cepstraux sur une échelle Mel.



Le nuage de points correspondant à chacun des mots est modélisé par son point moyen et les deux premiers vecteurs propres de sa matrice de covariance (ce qui, en somme, donne la position du nuage et sa forme), soit par trois vecteurs. Seuls les deux premiers sont utilisés dans cette étude. Les modèles sont comparés par un calcul de distance euclidienne.

Dans l'expérimentation rapportée dans cet article, seule la première phrase "*il se garantira du froid avec ce bon capuchon*", d'une durée de 2.4 à 3 secondes, a été utilisée. Les modèles des 100 mots (dix pour chacun des dix locuteurs) ont été calculés selon la démarche indiquée ci-dessus. La moyenne est supposée se comporter comme le spectre à long terme. Les auteurs concluent que : les coefficients cepstraux Mel donnent de meilleurs résultats que les coefficients LPC quel que soit le classifieur utilisé (bayésien ou LVQ) mais ils demandent bien sûr des temps de calcul plus longs ; la moyenne et le premier vecteur propre semblent suffisants pour l'identification ; les confusions se rencontrent dans l'ensemble des locutrices.

Dans leur article de 1991, les auteurs se servent de la base TIMIT pour tendre vers une identification indépendante du texte d'un locuteur parmi 20. En effet, les 5 phrases MIT, identiques pour tous et renfermant de nombreux phonèmes en différents contextes, servent pour l'apprentissage, alors que les phrases TI et les phrases dialectales, représentant mieux le langage parlé, sont utilisées pour les tests. Une analyse LPC d'ordre 16 est effectuée toutes les 10 ms sur des fenêtres de 15.6 ms après échantillonnage 16 kHz- 16 bits et préaccentuation. Les 200 phrases sont alors représentées par 200 tableaux  $16 \times N_i$ , où  $N_i$  est le nombre de spectres du  $i$ ème mot.

Ces tableaux sont utilisés pour l'apprentissage de réseaux neuronaux à retard temporel (TDNN) par application à leur entrée de fenêtres de 25 spectres tirées aléatoirement. Trois réseaux sont en fait mis en place : un premier réseau M1 est entraîné à faire l'identification du sexe du locuteur, deux réseaux M2 et M3 serviront à l'identification du locuteur ou de la locutrice suivant le cas. C'est l'algorithme de rétropropagation du gradient d'erreur qui est utilisé.

En phase de reconnaissance, une succession de fenêtres de 25 spectres avec décalage temporel d'un spectre est présentée en entrée des réseaux. Les réponses d'activation de sortie sont calculées et l'identification est effectuée par décision majoritaire. Le nombre et le rôle des couches cachées sont décrits. En conclusion, le réseau M1 effectue une identification du sexe à 100%. Les réseaux M2 et M3 donnent des résultats à 98%.

Ce travail trouve son prolongement dans une étude plus vaste utilisant 102 locuteurs de la base TIMIT [Bennani 92]. Il s'agit ici de faire coopérer un ensemble de réseaux : certains d'entre eux sont entraînés à reconnaître un type du locuteur, les étiquettes de la base fournissant une typologie des locuteurs, d'autres sont chargés de l'identification. Les conditions d'analyse sont celles précédemment décrites. La phase d'identification s'effectue en "temps réel".

Les grandes conclusions sont les suivantes : la division de la tâche d'identification en sous-tâches fait que le temps d'apprentissage est une fonction quasi linéaire de la taille des données (cette méthode de partage des tâches est celle antérieurement proposée dans [Oglesby 90] et décrite ci-après). Par ailleurs, la comparaison avec les modèles autorégressifs vectoriels est en faveur de l'approche connexionniste multimodulaire discutée ici. Trois fenêtres suffisent en entrée pour l'identification en mode indépendant du texte.

La progression des auteurs, telle qu'elle apparaît dans les articles ici retenus et dans celui qui sera présenté dans le chapitre consacré à la caractérisation automatique du locuteur (chapitre III) est ainsi la suivante : étude de la variabilité en mode dépendant du texte à partir d'enregistrements de mots par 10 locuteurs français et de l'analyse en composantes principales



sur différents jeux de paramètres [Daudin 89], utilisation de l'algorithme LVQ et tests sur une phrase du français et dix locuteurs, comparaison entre coefficients LPC et coefficients cepstraux, mise en comparaison avec un classifieur bayésien [Bennani 90], mise en place d'une architecture fondée sur les TDNN de façon à aller vers l'identification indépendante du texte, décomposition de la tâche d'identification en sous-tâches grâce à l'utilisation de plusieurs réseaux, tests sur 20 locuteurs de la base TIMIT [Bennani 91] approche connexionniste multimodulaire testée sur 102 locuteurs de la base TIMIT, orientation enfin, dans le futur, vers des systèmes hybrides [Bennani 92].

**Oglesby et Mason [Oglesby 90] [Oglesby 91].** Les articles auxquels nous faisons référence ici font suite à des études préliminaires que les auteurs ont effectuées sur la mise en place d'un système utilisant des réseaux neuronaux à rétropropagation : le système a la particularité de modéliser directement les sons spécifiques du locuteur. Un modèle donné est entraîné à présenter une sortie active pour toute phrase du locuteur donné et une sortie inactive pour tout autre entrée ; l'apprentissage est effectué par minimisation de l'erreur quadratique moyenne entre la sortie désirée et la sortie courante du réseau, en conjonction avec un algorithme du gradient. Chaque réseau est entraîné au même degré de performance sur les données d'apprentissage. De cette façon, les valeurs de sortie d'un réseau donnent une image de la vraisemblance qu'une donnée de test provienne du locuteur correspondant.

L'expérimentation a été menée à partir d'une base de données comprenant au total 500 élocutions des dix chiffres par dix locuteurs différents (100 ont servi à l'apprentissage et 400 aux tests). Une analyse LPC d'ordre 10 sur des fenêtres de Hamming préaccentuées de 256 échantillons fournit les coefficients cepstraux. Ce sont ainsi 2000 formes environ qui servent à l'apprentissage.

Le module d'apprentissage du système est implanté à l'aide d'un réseau de transputers, chacun d'eux étant chargé d'un sous-ensemble des données, ce qui rend les temps d'apprentissage presque linéaires.

Les grandes conclusions sont les suivantes : le choix de l'architecture et la taille des données d'apprentissage ont une influence forte sur le taux de reconnaissance ; les modèles à deux couches cachées sont moins bons que les modèles à couche cachée unique et comprenant moins de poids ; l'approche par réseaux neuronaux équivaut, pour les performances de sortie, à l'approche par systèmes à codebooks personnalisés résultant d'une étape de quantification vectorielle. Les meilleurs résultats obtenus sont de 8% d'erreur avec un réseau à couche cachée unique de 128 nœuds, ce que les auteurs disent équivalent à ceux obtenus à partir d'une approche par quantification vectorielle utilisant un codebook de 64 codes. Pour des tailles de réseau inférieures, l'approche neuronale est supérieure.

Dans leur seconde étude [Oglesby 91], les auteurs proposent une forme modifiée des réseaux à rétropropagation fondés sur des fonctions de base radiales qui fournit des taux d'identification encore meilleurs, en plus des avantages sur le plan de la rapidité d'adaptation du système à l'apprentissage.

**Rudasi et Zahorian [Rudasi 91].** Les auteurs présentent une méthode de classification par partitionnement binaire. Appliquée à l'identification des 47 locuteurs de la base TIMIT par réseaux neuronaux, elle conduit à 100% de taux de bonne identification à partir de 9 à 14 secondes de parole pour la phase d'apprentissage et de 8 secondes de parole pour les tests, en

mode indépendant du texte. Les paramètres retenus sont quinze coefficients cepstraux calculés sur la plage 150-6000 Hz.

Des facteurs statistiques calculés sur l'ensemble du corpus d'apprentissage permettent de normaliser la distribution (moyenne et écart-type) de chacun des paramètres. La solution revient à remplacer un grand réseau permettant de différencier  $N$  locuteurs par  $N(N-1)/2$  réseaux beaucoup plus petits effectuant la classification par paires.

## 5. Les autres méthodes de R.A.L.

### 5.1. Introduction

Un certain nombre de travaux, à partir des années 70, ont utilisé des techniques de classification pour la reconnaissance automatique du locuteur. Nous citerons simplement ici quelques-unes de ces études au travers notamment d'articles de synthèse : [Das 71], [Paul 75], [Sambur 76], [Rosenberg 76], [Atal 76], [Hunt 82], [Shridhar 83], [Wolf 83], [Furui 86].

Nous donnons dans la suite des précisions sur les études les plus récentes.

### 5.2. Les études plus récentes

**Feix et DeGeorge [Feix 85].** Les auteurs ont conçu un système de vérification du locuteur fondé sur une méthode de reconnaissance de mots en parole continue conçue par T.B. Martin. Cette approche dépendante du texte utilise un nombre variable d'élocutions (entre deux et quatre) durant la transaction de contrôle d'accès.

Le signal est filtré dans 16 canaux inspirés d'études sur la perception, dans la bande 200-6700 Hz. Un vecteur de paramètres, comprenant des caractéristiques spectrales et phonétiques, est extrait toutes les 2.5 ms. Seize vecteurs sont retenus par mot, de sorte qu'un mot isolé est représenté par un tableau de taille fixe. Les formes de référence sont construites à partir de ces tableaux avec une résolution de 4 bits pour chaque paramètre. La reconnaissance est effectuée par programmation dynamique [Sakoe 78].

A partir d'une base de données de test, constituée de 16 mots prononcés 10 fois par 15 locuteurs, les auteurs extraient un ensemble de paramètres spectraux qui maximisent le rapport du score moyen d'un locuteur sur ses propres références au score moyen des autres locuteurs du même sexe sur ces mêmes références. Certains paramètres se révèlent intéressants : les maxima du spectre entre 200 et 1100 Hz et entre 3000 et 4000 Hz, les pentes spectrales entre 2000 et 3000 Hz et autour de 1300 Hz. Le vocabulaire utilisé non précisé dans l'article est sélectionné en fonction de règles comme la suivante : choisir, pour des raisons de facilité de segmentation parole-non parole automatique, des mots commençant par une occlusive et finissant par une non occlusive.

Lors de la phase de vérification, un enregistrement comprend 5 répétitions de chaque mot. La décision s'effectue à partir d'un "seuil d'acceptation" diminuant avec le nombre de mots prononcés. La procédure d'amélioration de l'algorithme de décision est décrite avec précision dans l'article. L'intérêt de l'expérience est qu'elle s'est effectuée dans des conditions réalistes sur plusieurs mois avec 53 puis 42 locuteurs.



**Attili et al. [Attili 88].** Les auteurs proposent un système de vérification automatique rapide, fiable, économique, implanté sur le processeur TMS32020. Un IBM PC est utilisé pour la communication avec l'utilisateur. L'algorithme répond à 75% du temps réel (15 ms sont nécessaires pour le traitement complet d'une fenêtre de 20 ms) et demande que soient prononcées deux à trois secondes de parole non contrainte.

Pour mettre au point l'algorithme de reconnaissance du locuteur, les auteurs ont d'abord étudié plusieurs familles de paramètres dans les conditions d'acquisition et d'analyse suivantes : filtrage passe-bas à 4 kHz, conversion analogique-numérique à 10 kHz sur 12 bits, application d'un filtre de préaccentuation et d'une fenêtre de Hamming sur 20 ms sans recouvrement, extraction par analyse LPC (autocorrélation d'ordre 12) des coefficients PARCOR et LPC, transformation en cepstre et en coefficients de fonction d'aire. Un gain normalisé est également calculé.

Ils ont ensuite sélectionné  $N$  paramètres dépendants du locuteur jugés pertinents par une variante de l'analyse discriminante multi-classes (avec une nouvelle définition de la dispersion interclasses). Il en résulte que les paramètres sélectionnés pour un sujet particulier sont indicatifs de ce sujet et que les paramètres utilisés pour l'authentification sont dépendants du locuteur.

La décision d'authentification est effectuée de façon séquentielle ; le système fait répéter le locuteur jusqu'à ce que son degré de confiance devienne suffisamment élevé. Elle s'effectue de la façon suivante : pour la vérification du locuteur  $i$ , le  $i$ -ième vecteur de dimension  $N$  de la fenêtre d'analyse est projeté sur un sous-espace de beaucoup plus petite dimension  $M$  en considérant le modèle de  $i$ . La décision est ensuite prise sur la valeur d'un ratio de vraisemblance de type bayésien. Deux seuils permettent de prendre localement la décision d'accepter le locuteur, de le rejeter, ou de répéter la recherche sur une autre fenêtre. Si ce dernier cas se répète, il peut y avoir rejet au bout d'un moment.

Un certain nombre d'expériences de test d'authentification du locuteur et de rejet d'imposteur ont été effectuées à partir de 90 locuteurs et 15 énoncés de différents textes. Elles ont permis aux auteurs d'établir les conclusions suivantes. Les performances du système sont améliorées grâce à l'utilisation de différentes familles de paramètres non indépendantes (PARCOR, fonction d'aire, LPC, réponse impulsionnelle, fréquence fondamentale, gain, taux de passage par zéro). Par ailleurs, le taux d'erreur combiné (taux de fausses acceptation et taux de faux rejet) ne change pratiquement pas si l'on fait varier la dimension  $M$  du sous-espace de projection au delà de 2 ou 3 ; les résultats obtenus avec  $M=4$  sont favorables à la vérification dépendante du texte (de l'ordre de 1% de taux d'erreur contre 2% en moyenne en vérification indépendante du texte). Enfin, les performances ne sont pas affectées par l'introduction d'un bruit gaussien de rapport signal sur bruit inférieur à 15 dB.

**Li et Porter [Li 88].** Les auteurs distinguent deux situations. Dans le cas de la vérification, il s'agit d'évaluer la véracité des dires d'un locuteur qui se prétend être  $X$  par mise en comparaison de la parole observée à celle de  $X$  par rapport à ce qui est attendu d'une population générale. Trouver un équilibre entre la fausse identification et le faux rejet nécessite alors l'utilisation de seuils. Dans le cas de l'identification, au contraire, le problème des seuils est évité car il suffit de comparer des scores de mise en correspondance de la parole inconnue avec des modèles d'un ensemble fini de locuteurs. Cependant, la méthode est sensible aux biais des modèles.

K-P. Li et J.E. Porter proposent une méthode de normalisation et de sélection pour améliorer le taux de reconnaissance avec des échantillons de parole non imposée très courts. Leur étude a porté sur un ensemble de conversations, prononcées par 26 locuteurs masculins à une semaine



d'intervalle. Après préaccentuation, 10 cepstres (obtenus par LPC) et 10 valeurs de log-area-ratio sont extraits. Une première normalisation est effectuée en fonction des moyennes et des variances des scores d'un court échantillon inconnu avec les modèles de différents locuteurs. Une phase de sélection écarte les portions de parole de faible pouvoir de discrimination. Une seconde normalisation se fonde sur le rang des scores du modèle concerné avec les modèles des autres. Cet ensemble de techniques doit permettre d'ajuster des seuils pour la vérification d'un locuteur au sein d'une population ouverte. Les résultats obtenus en phase d'identification font apparaître des améliorations significatives dues aux normalisations.

Pour la vérification, les auteurs comparent différentes techniques de calcul de scores et de choix des seuils. Avec l'une de ces techniques (faisant intervenir la moyenne des taux de faux rejet et de fausse authentification), les auteurs observent que la distribution des scores des imposteurs est très stable et que chaque modèle a la même moyenne et la même dispersion des scores d'imposteurs ; les seuils ainsi obtenus peuvent de ce fait être utilisés pour une population ouverte.

**Nakasone et Melvin [Nakasone 88].** Le projet, dénommé CAVIS, a pour objectif l'identification du locuteur à partir de données vocales indépendantes du canal de transmission et indépendantes du texte, dans l'esprit des investigations criminelles. Dans ce domaine, trois méthodes sont généralement utilisées : analyse à l'oreille, analyse à l'oreille associée à une analyse spectrographique, analyse par des méthodes numériques. L'étude effectuée doit se transposer au cas d'un environnement "hostile" et de locuteurs par définition non coopératifs.

21 locuteurs masculins ont eu à lire 10 textes de 30 secondes chacun, extraits de journaux et de revues, avec double enregistrement sur bande audio, l'un par l'intermédiaire d'un combiné téléphonique, l'autre à l'aide d'un microphone haute fidélité. Après préaccentuation, conversion analogique-numérique à 10 kHz sur 12 bits, des paramètres spectraux et temporels sont déterminés. Le spectre moyen présentant trop de différences entre les deux médias d'acquisition, les auteurs retiennent un autre spectre (*Intensity Deviation Spectrum*) dont ils extraient des intensités aux centres de certaines bandes de fréquences. Par ailleurs, avec intervention d'un opérateur pour la détection des périodes successives du fondamental, des coefficients traduisant l'évolution dans le temps du contour mélodique sont obtenus.

L'identification se fait en trois étapes. Tout d'abord, deux sous-ensembles de paramètres sont retenus : d'une part, parmi les paramètres précédemment décrits, ceux possédant le plus fort F-ratio (a priori quatre centres de bande fréquence et la fréquence fondamentale moyenne), et d'autre part, les coefficients de corrélation issus directement du spectre. Puis, une distance entre deux locuteurs est appréciée à partir du calcul d'une distance de Manhattan. Enfin, la décision d'identification est prise selon la règle des k plus proches voisins avec k=1 ou k=2. Les résultats sont de 100% d'identification pour k=2, quelles que soient les conditions de transmission.

**Wilbur et Taylor [Wilbur 88].** Contrairement aux précédentes, cette étude ne présente pas une méthode de reconnaissance automatique du locuteur mais propose une nouvelle représentation spectrale des énoncés des locuteurs qui soit moins sensible à la variabilité intralocuteur.

L'estimateur proposé est dérivé de la fonction de distribution de Wigner. Les auteurs considèrent en effet que la FFT présente une grande variabilité pour un même locuteur suivant les conditions d'enregistrement. Une alternative à cette technique est la distribution de Wigner,

intéressante pour des processus non stationnaires. Elle présente l'avantage d'avoir un indice de dispersion réduit, permettant notamment une meilleure résolution pour les pics des formants. Les auteurs utilisent dans cette étude une version fenêtrée et filtrée de la distribution de Wigner discrète (DWD, Discrete Wigner Distribution). L'expérimentation conduit à la mise en place d'une base de données robuste utilisant l'estimateur de Wigner. Toutefois, cette base de données n'a pas été testée du point de vue de la reconnaissance automatique du locuteur.

**Cohen et Froind [Cohen 89].** Il s'agit d'identification du locuteur indépendante du texte. Les auteurs considèrent que l'hypothèse de même matrice de covariance pour tous les locuteurs, qui conduit à un classifieur linéaire, est trop restrictive. Ils se tournent vers l'utilisation de matrices de covariance individuelles (chaque personne est reconnue avec un espace d'indices propre), qui conduit à un classifieur quadratique. Cette méthode présente peu d'exigence supplémentaire en mémoire et fournit de meilleurs résultats. Les paramètres sont sélectionnés par maximisation d'un critère donné de séparation, cette maximisation étant mise en œuvre par programmation dynamique.

Les enregistrements sont effectués dans un environnement de laboratoire normal. Après filtrage passe-bas à 3200 Hz, échantillonnage à 10 kHz, sélection de fenêtres de Hamming de 10 ms, un certain nombre de paramètres sont extraits : sur des élocutions de 15 secondes, segments voisés et non voisés séparés, les auteurs retiennent 8 coefficients d'autocorrélation normalisés, 10 coefficients LPC d'ordre 10, 9 coefficients de corrélation partielle, 9 coefficients spectraux, l'erreur de prédiction, l'énergie moyenne, la fréquence fondamentale moyenne. Pour chacun des locuteurs, les dix meilleurs paramètres sont sélectionnés. Des moyennes et des matrices de covariance sont ensuite calculées.

L'identification est effectuée suivant un algorithme fondé sur le critère de distance minimale. Les résultats présentés, obtenus à partir de 6 locuteurs masculins, montrent la supériorité du classifieur quadratique sur le classifieur linéaire. 4 phrases en hébreu ont été enregistrées lors de sessions espacées de quelques jours. 6 à 8 minutes de parole pour chaque locuteur sont retenues pour l'apprentissage, 2 minutes, de texte différent, servent ensuite aux tests. On trouve dans l'article le détail des matrices de confusion et la liste des paramètres optimaux choisis pour chacun des locuteurs.

Il est intéressant de noter que, dans le cas du classifieur linéaire, 80% des paramètres sont sélectionnés sur des segments voisés, contre 55% pour le classifieur quadratique. La complexité des algorithmes est également discutée.

**Gong et Haton [Gong 90].** Cette étude, issue d'un travail initial sur la reconnaissance de la parole indépendante du locuteur et transposé ensuite à l'adaptation au locuteur, se fonde sur le principe de comparaison d'espaces des trajectoires. Le principe est le suivant : les productions vocales d'un locuteur donné décrivent des trajectoires à l'intérieur d'un sous-espace (sous-volume) de l'espace des paramètres ; si l'espace des paramètres est convenablement choisi, ces sous-espaces sont différents pour deux locuteurs distincts ; au contraire, pour un locuteur donné, les volumes correspondant aux phases d'apprentissage et d'identification sont similaires. Un ensemble d'algorithmes destinés à résoudre le double problème de la représentation des sous-espaces et de leur comparaison est présenté. Ces algorithmes se fondent sur une première étape de quantification vectorielle des fenêtres de parole.

L'expérimentation s'est effectuée à partir du corpus que nous avons mis au point pour notre propre étude et qui est décrit en partie C. 7 locutrices et 16 locuteurs, trois répétitions de 15



phrases ont été retenus. Trois groupes d'apprentissage et de test (tournants) ont été constitués. Le signal est analysé après une numérisation à 16 kHz : préaccentuation de la forme  $(1-0.94z^{-1})$ , fenêtrage de Hamming sur des durées de 25.6 ms avec recouvrement de 10 ms. Un codebook est constitué suivant l'algorithme LBG [Linde 80].

Divers tests comparatifs ont été ensuite effectués sur l'espace des paramètres, les mesures, la dimension de l'espace, le nombre de codes. Les auteurs concluent que coefficients de LPC et taux de vraisemblance d'Itakura d'un côté, coefficients cepstraux dérivés de LPC avec distance euclidienne de l'autre, donnent des résultats similaires.

Pour finir, les algorithmes annoncés plus haut sont comparés à partir de 207 tests utilisant les coefficients du codage LPC par autocorrélation à l'ordre 16 et 64 codes. Un taux de reconnaissance de 99.5% est obtenu. A partir de ces résultats, les auteurs poursuivent leurs travaux vers la mise en place d'un système de compréhension de la parole continue indépendant du locuteur.

**Rose et al. [Rose 91].** Cette étude fait suite à la définition d'un classifieur pour l'identification du locuteur, que l'on cherche à améliorer ici. Elle s'intéresse à des phrases de conversation (extraits de 30 à 60 secondes), énoncées par 10 locuteurs masculins, lors d'appels téléphoniques à longue distance, sur une période de deux semaines. Le canal téléphonique est caractérisé par un bruit à large bande, une distorsion linéaire et un bruit d'impulsions. Le rapport signal sur bruit vaut de 18 à 33 dB.

L'analyse est effectuée par un banc de 20 filtres Mel avec recouvrement dont les fréquences centrales s'échelonnent de 200 à 3000 Hz. Deux techniques pour compenser le bruit pour un classifieur gaussien sont étudiées : intégration directe d'un modèle de bruit de fond dans celui de la parole, prétraitement avant de passer le signal au classifieur.

Dans une première phase s'effectue l'apprentissage de modèles pour chacun des locuteurs, puis l'évaluation des performances. Les résultats montrent l'intérêt d'introduire des techniques de compensation du bruit. Les travaux se poursuivent pour étendre les techniques aux environnements de bruit non stationnaires.

**Gaganelis et Frangoulis [Gaganelis 91].** Il s'agit ici d'une approche par fonctions de Fourier-Bessel pour la vérification du locuteur à travers le canal téléphonique. Ces fonctions transforment le problème initial en un problème de détection multidimensionnelle, ce qui permet de combiner des ensembles de paramètres en un seul test de classification. Le signal est limité à la bande 100-3200 Hz et échantillonné à 10 kHz sur 16 bits. Après préaccentuation (de la forme  $1-0.96z^{-1}$ ) et application d'une fenêtre de Hamming, les coefficients de prédiction linéaire par autocorrélation d'ordre 10 pitch-synchrone sont calculés et transformés en coefficients cepstraux normalisés (par soustraction de leur moyenne sur la durée de la phrase entière). Les fonctions du temps de ces paramètres sur de petits segments sont projetées sur une base de polynômes orthogonaux, de façon à obtenir les coefficients polynomiaux du premier et du second ordres. La valeur du fondamental est également utilisée.

Des formes de référence sont construites à partir du calcul de distances entre répétitions du même mot prises deux à deux par une méthode de programmation dynamique (les distances entre paramètres spectraux sont normalisées par une fonction de l'inverse de la variance).

Dans la phase de classification, l'usage des fonctions de Fourier-Bessel, décrit avec précision dans l'article, permet de combiner efficacement différents paramètres de la parole de



façon à accroître le pouvoir discriminant du système de vérification. Les formes d'apprentissage sont utilisées pour estimer la distribution des dissemblances intralocuteur, quantité utilisée dans la phase de classification. Les tests ont été effectués avec 45 locuteurs (30 hommes et 15 femmes) sur 3 mots.

Les résultats sont comparés avec ceux que fourniraient un classifieur gaussien standard ou un classifieur gaussien multivarié. En ne considérant que les coefficients cepstraux, l'approche par fonctions de Fourier-Bessel conduit à un taux de fausse acceptation de 1.42% contre 4.69% pour un classifieur gaussien et un taux de faux rejet de 1.98% contre 2.46% pour le classifieur gaussien. Enfin, les auteurs concluent que cette différence de performances en faveur de l'approche par fonctions de Fourier-Bessel est significative à 95% et que les effets des différents mots et du sexe du locuteur sont non significatifs pour les performances obtenues.

**Higgins et Bahler [Higgins 91].** Le but de l'étude est de parvenir à la détection d'une personne cible tout en rejetant les autres locuteurs (dits locuteurs de référence). L'apprentissage est effectué à partir de données utilisées pour estimer les paramètres de discriminateurs de locuteurs par paires. Il suppose que les vecteurs de paramètres des deux locuteurs en question sont des distributions gaussiennes multivariées avec matrices de covariance identiques. Le classifieur de Bayes (ou à minimum d'erreur) est implanté en utilisant une fonction de discrimination linéaire.

Les caractéristiques de l'analyse sont les suivantes : 8 kHz et 12 bits pour la numérisation, 14 filtres Mel FIR sur la bande 350-3500 Hz, normalisation des puissances des canaux. Ceci est effectué pour les 50 locuteurs de la base de données "King" dont 25 locuteurs sont pris comme cibles. L'algorithme est entraîné à reconnaître chacun des locuteurs cibles en utilisant les 24 autres comme référence. En phase de vérification, les tests se font en utilisant l'énoncé du locuteur cible et ceux des 25 locuteurs non-cibles. Pour chaque vecteur de test, les discriminateurs fournissent chacun une sortie scalaire positive ou négative. Ces sorties sont passées à la sigmoïde, puis moyennées dans le temps. Un traitement brutal renvoie ensuite 0 ou 1 suivant le signe. Les valeurs sont ensuite sommées et la décision est prise : une identité hypothétique de locuteur est acceptée si le total dépasse un seuil fixé. C'est donc en quelque sorte le nombre de discriminateurs votant en faveur du locuteur qui est recherché.

L'étude a porté à la fois sur de la parole non bruitée et de la parole ayant transité par la voie téléphonique. Les auteurs montrent que leur méthode a comme importante propriété que la distribution des scores des non-cibles soit uniforme, ce qui permet de rendre prédictible le taux de fausse identification à partir de la valeur du seuil. Les résultats sont détaillés et illustrés. Par ailleurs, les auteurs ont implanté leur algorithme à l'aide de perceptrons multicouches avec apprentissage par rétropropagation sans que cela conduise à de meilleures performances.

**Fussel [Fussel 91].** L'originalité de ce travail réside dans le fait qu'il est destiné à l'identification du sexe du locuteur à partir d'un segment de 16 ms centré sur la réalisation d'un phonème. Il trouve ainsi son application dans le domaine de l'identification du locuteur et aussi dans celui de la séparation de voix dans une conversation où joue l'effet "cocktail party".

Les données expérimentales sont issues du corpus TIMIT comprenant des enregistrements provenant de 290 locuteurs et 130 locutrices, les transcriptions orthographiques et les suites d'étiquettes acoustico-phonétiques alignées sur le signal (parole échantillonnée à 16 kHz sur 16 bits). Les locuteurs proviennent des huit régions dialectales les plus importantes. Les cinq premiers locuteurs de chacune de ces régions sont retenus pour l'apprentissage, les cinq suivants pour les tests. Les soixante étiquettes phonétiques utilisées pour étiqueter la base sont



divisées en six catégories (occlusives, fricatives, consonnes nasales, voyelles, liquides et semi-voyelles, et enfin silence et / h / ). Pour chacun des phonèmes d'une phrase, les traitements suivants sont appliqués sur les 256 échantillons centraux de l'occurrence la plus longue : calcul du spectre Mel à partir d'une FFT à 256 points, application d'une FFT inverse sur le logarithme de façon à obtenir le cepstre. Un vecteur de 18 paramètres est finalement retenu comprenant les neuf premiers coefficients cepstraux et les neuf différences entre le cepstre et son voisin (delta-cepstre).

L'auteur utilise un classifieur gaussien à fonction de vraisemblance simplifiée du fait que ce sont des fenêtres uniques qui sont utilisées pour les tests. Pour chacun des phonèmes, il mesure les performances des paramètres pris isolément. Les résultats sont moyennés par classe de phonèmes avant d'être présentés graphiquement dans l'article. Les résultats majeurs sont les suivants. Le premier coefficient cepstral, en relation avec l'énergie dans la fenêtre étudiée, a un pouvoir d'identification de 55% environ pour toutes les classes, ce qui conforterait d'autres résultats selon lesquels l'énergie moyenne (RMS) est supérieure de 4 dB pour les locuteurs masculins. Le cepstre donne de meilleurs résultats que le delta-cepstre ; le meilleur paramètre pour l'identification du sexe est le quatrième coefficient cepstral, ce qui n'est pas expliqué. En ce qui concerne la pertinence des phonèmes pour l'identification du sexe, les voyelles, les consonnes nasales et liquides donnent des résultats supérieurs à ceux des occlusives, fricatives, du silence ou du /h/, ce qui, souligne l'auteur, n'est pas surprenant.

Une expérience complémentaire discute l'intérêt de conserver l'ensemble des dix-huit paramètres ; de véritables conclusions ne pourraient être tirées qu'à partir de plus grands corpus d'apprentissage. L'auteur teste également plusieurs variantes du classifieur gaussien et constate que le classifieur utilisant une matrice de covariance complète conduit à de meilleurs taux de reconnaissance du sexe. Enfin, trois types d'apprentissage sont mis en œuvre : apprentissage sur tout le corpus, apprentissage pour chacune des six classes de phonèmes et apprentissage par phonème. L'insuffisance du nombre d'occurrences de chaque phonème permettrait d'expliquer les moins bonnes performances paradoxalement obtenues par le dernier type d'apprentissage.

Enfin, les erreurs sont analysées. Elles mettent en cause, par exemple, la petite taille du corpus d'apprentissage ou la présence de "moutons noirs" : dix des quatre-vingts locuteurs sont responsables de douze erreurs ou plus alors que vingt-trois d'entre eux ne sont responsables que d'une erreur d'identification ou d'aucune. L'auteur estime, en conclusion, qu'un classifieur gaussien, entraîné à partir d'un corpus d'apprentissage suffisant, sous la forme de vecteurs de paramètres bien choisis, et traitant de simples fenêtres, devrait permettre la construction d'un système automatique d'identification du sexe du locuteur. Enfin, il fait référence à d'autres travaux du même type utilisant un réseau neuronal à rétropropagation [Thomas 90].

**Montacié et al. [Montacie 92].** Après avoir distingué entre les notions de vérification d'une identité prétendue, d'identification du locuteur et de détection des changements de locuteur au cours d'une conversation, les auteurs centrent leur étude sur l'identification. Aucune phrase spécifique n'est obligatoire. L'étude se fonde sur des modèles autorégressifs vectoriels MAV (modèles de l'évolution spectrale) et sur une distance de distorsion interlocuteur DIV (dite Distance d'Itakura Vectorielle).

L'apprentissage d'un modèle pour chaque locuteur et l'identification du locuteur (celui dont le modèle est le plus proche, au sens de la distance) peuvent se faire sous deux modes, discriminant et non discriminant. En identification du locuteur, les MAV sont interprétés comme une représentation des capacités articulatoires du locuteur (vitesse et accélération instantanées

des paramètres spectraux). La difficulté de leur implantation réside dans la difficulté de trouver l'ordre optimal. Les MAVD (MAV discriminants) sont les MAV minimisant les distances du locuteur à son propre modèle et maximisant les distances des locuteurs aux modèles des autres.

En ce qui concerne les distances interlocuteurs, les auteurs présentent trois distances et insistent sur une quatrième plus discriminante, fondée sur l'idée de discrimination sur l'ensemble d'apprentissage. Le coût de calcul de cette distance étant élevé sur une population importante, ils proposent une méthode pour pallier cet inconvénient.

La base de données étudiée est la base TIMIT. Elle comprend 420 locuteurs, de différents accents de l'anglais américain, 10 phrases par locuteur, 2 communes à tous, huit (5 MIT et 3 TI) différentes. Les phrases avec grande variété de contextes phonétiques (les 5 MIT) servent à l'apprentissage. Les 5 autres aux tests. Les paramètres retenus sont les coefficients cepstraux (LPCC), 20 en général, 8 pour les MAVD pour des raisons de complexité de calcul.

L'identification s'effectue en une seule passe, avec des locuteurs coopératifs et sans imposteur. Les résultats sont présentés en fonction du nombre de locuteurs et de la distance choisie. Il est à noter que les erreurs ne se produisent qu'entre personnes du même sexe et pour des phrases généralement courtes. Les très bons résultats obtenus par les auteurs (98.4% de bonne identification pour 420 locuteurs) encouragent ceux-ci à rechercher une technique de normalisation des paramètres spectraux au locuteur.

## 6. Conclusion

De trop nombreux travaux ont été réalisés en identification et en vérification automatique du locuteur pour pouvoir en faire une synthèse complète dans ce chapitre. Aussi avons-nous plutôt tenté de mettre en évidence comment de nouvelles approches de reconnaissances des formes comme les classifieurs gaussiens multivariés, les modèles autorégressifs, les réseaux de neurones ou les modèles de Markov ont été mises en œuvre en reconnaissance du locuteur.

Les études sur la caractérisation du locuteur étant beaucoup moins nombreuses et plus proches de nos propres travaux, nous serons dans le chapitre suivant beaucoup plus exhaustive dans la description de ces recherches.



## CHAPITRE III CARACTERISATION AUTOMATIQUE DU LOCUTEUR

### 1. Introduction

Ce chapitre est consacré à une étude bibliographique des études ayant pour objet la recherche de paramètres acoustiques et phonétiques susceptibles de caractériser au mieux les locuteurs.

J.J. Wolf [Wolf 72] et F. Nolan [Nolan 83] donnent des définitions similaires du paramètre qui caractérise un locuteur. En résumé, celui-ci doit :

- être présent naturellement et souvent dans la parole normale,
- se mesurer facilement,
- être très différent d'un locuteur à l'autre,
- être consistant pour chaque locuteur : ne pas varier avec le temps, avec l'état de santé du locuteur ou le contexte de la communication,
- ne pas être sensible au bruit et aux distorsions des canaux de transmission,
- ne pas être modifiable consciemment par le locuteur.

Si nous confrontons ces propriétés aux différentes manifestations de la variabilité de la parole, mises en évidence dans la partie A, il apparaît, d'une part qu'un tel paramètre a peu de chances d'exister dans la réalité, d'autre part que la vérification de l'ensemble de ces propriétés pour un paramètre prédéfini représente un travail considérable de plusieurs années. Aussi, les recherches en caractérisation du locuteur se sont-elles contentées d'émettre des hypothèses sur la pertinence de certains paramètres segmentaux ou suprasegmentaux de la parole, puis de vérifier s'ils satisfaisaient une infime partie de ces propriétés.

Après avoir introduit quelques notions sur les méthodologies employées dans ces recherches, nous présentons les principales d'entre elles dans les paragraphes suivants.

### 2. Les méthodologies

Deux méthodologies sont appliquées pour mettre en relief la pertinence de certains paramètres acoustiques ou linguistiques pour la caractérisation du locuteur. Soit les auteurs mettent en œuvre des techniques fondées sur l'étude des distributions intralocuteur et interlocuteur des paramètres, soit ils sélectionnent les meilleurs paramètres en les testant directement dans des procédures de reconnaissance comme celles que nous avons décrites dans le chapitre précédent.

Nous explicitons dans la suite de ce paragraphe quelques-unes des techniques de la première catégorie.

## 2.1. Le F-ratio

Le F-ratio est défini par

$$F = \frac{\frac{N}{L-1} \sum_{j=1}^L (\bar{x}_j - \bar{x})^2}{\frac{1}{L(N-1)} \sum_{i=1}^N \sum_{j=1}^L (x_{ij} - \bar{x}_j)^2}$$

où  $x_{ij}$  est la valeur du paramètre  $x$  pour la  $i^e$  répétition du locuteur  $j$ ,  $\bar{x}_j$  la moyenne du paramètre  $x$  pour le locuteur  $j$ ,  $\bar{x}$  la moyenne des  $\bar{x}_j$  sur tous les locuteurs,  $L$  le nombre de locuteurs et  $N$  le nombre de répétitions pour chaque locuteur.

Le F-ratio est donc proportionnel au rapport de la variance des moyennes du paramètre pour chacun des locuteurs à la moyenne des variances du paramètre pour chacun des locuteurs. Plus les distributions pour chacun des locuteurs sont étroites et plus elles sont éloignées les unes des autres, plus ce rapport est élevé et meilleur est le paramètre pour discriminer les locuteurs.

L'inconvénient du F-ratio est qu'il peut avoir une valeur élevée uniquement parce qu'un ou deux locuteurs sont très différents des autres. Il sert donc surtout à éliminer les paramètres les moins discriminants et il faut vérifier la pertinence des autres paramètres par une autre méthode.

Le F-ratio teste la pertinence de chacun des paramètres pris individuellement. Par conséquent, si plusieurs paramètres obtiennent un F-ratio élevé, il faut encore estimer leur interdépendance avant de les mettre en œuvre dans une procédure de reconnaissance.

Le coefficient de corrélation est la première façon de tester la dépendance de deux paramètres. Mais celui-ci indique seulement la présence ou l'absence d'une relation linéaire entre les deux paramètres. J.J. Wolf [Wolf 72] propose une autre estimation de l'indépendance de deux paramètres en estimant la probabilité de confusion entre les locuteurs pris deux à deux. Soient  $P_x$  et  $P_y$  les probabilités obtenues pour chacun des deux paramètres  $x$  et  $y$  et  $P_{xy}$  la probabilité obtenue en utilisant conjointement pour les deux paramètres. Selon la définition probabiliste de la notion d'événements indépendants, l'expression  $(P_{xy} - P_x P_y)$  doit valoir 0, ou tout du moins s'en approcher, si les paramètres sont indépendants. Il reste alors à déterminer de façon heuristique le seuil mesurant ce voisinage.

## 2.2. L'analyse discriminante linéaire

Soit un espace de dimension  $N$  dans lequel les locuteurs sont représentés par des vecteurs de  $N$  paramètres. L'analyse discriminante linéaire multidimensionnelle recherche un nouvel espace de dimension  $M$  ( $M \leq N$ ) qui maximise le rapport de la variance interlocuteur à la variance intralocuteur. Les nouveaux paramètres sont les combinaisons linéaires des paramètres d'origine qui rendent maximum le rapport

$$\frac{u_m^t B u_m}{u_m^t W u_m}$$

où  $B$  est la matrice de covariance interlocuteur,  $W$  la moyenne des matrices de covariance intralocuteur et  $u_m$  la combinaison linéaire recherchée.

## 3. Les études

### 3.1. Introduction

La plupart des études sur la caractérisation du locuteur se rapportent à la langue anglaise et plus particulièrement à l'anglais américain. Cela pose deux problèmes.

Le premier est de savoir à quelles conditions les résultats obtenus dans ces études sont transposables à la langue française. Non spécialiste de la phonétique française, nous le sommes encore moins de la phonétique anglaise. La figure A27 et les figures suivantes permettent de situer les voyelles anglaises par rapport à celles du français. La figure B.2 présente les voyelles de l'anglais britannique sous la forme d'un trapèze articulatoire similaire à celui du français. La figure B.3 présente les voyelles américaines et les trajectoires des diphtongues dans un repère acoustique ( $F_1$ ,  $F_2-F_1$ ).

Le deuxième problème se situe à la fois dans les divergences de prononciation entre l'anglais américain, et l'anglais britannique et dans la diversité des notations employées par les phonéticiens et les chercheurs de langue anglaise. Cette diversité concerne essentiellement les voyelles et les diphtongues. Par ailleurs, comme plusieurs auteurs n'ont pas inclus dans leurs articles la transcription orthographique des mots prononcés par les locuteurs, il nous est difficile de connaître avec certitude les phonèmes étudiés et d'effectuer des comparaisons entre les études.

Aussi mentionnerons-nous, lorsqu'elles sont connues, la transcription orthographique du corpus et la nationalité des locuteurs. Enfin, nous conserverons les notations phonétiques des auteurs, lorsqu'elles existent. La table B.1, qui fournit une correspondance entre les notations phonétiques de plusieurs phonéticiens de langue anglaise [Ladefoged 75], permettra de retrouver à quoi correspondent ces notations et d'établir une relation entre certaines recherches.

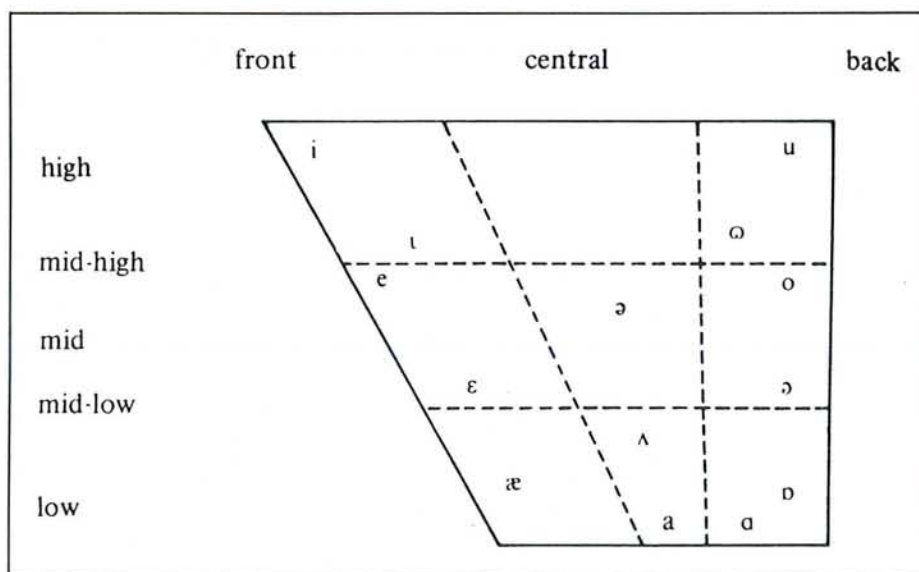


Figure B.2. Représentation articulatoire des voyelles de l'anglais britannique d'après [Ladefoged 75]. Les sons [ e ], [ a ] et [ o ] ne sont pas des voyelles à part entière mais des débuts de diphtongues.





A comparison of some systems for transcribing vowel sounds. Where two symbols are shown corresponding to the vowel in a single word, the first is appropriate for most speakers of American English, and the second for most speakers of British English.

This book		Kenyon & Knott	Trager & Smith	Prator & Robinett	Jones	Webster's
i	beat	i	iy	iy	i:	ē
ɪ	bit	ɪ	i	ɪ	i	ɪ
eɪ	bait	e	ey	ey	eɪ	ā
ɛ	bet	ɛ	e	ɛ	e	e
æ	bat	æ	æ	æ	æ	a
ɑ	father	ɑ	a	a	ɑ	ā
ɒ	bother	ɑ	a	a	ɔ	ā
ɔ	bought	ɔ	oh	ɔ	ɔ:	ō
oʊ	boat	o	ow	ow	ou	ō
ʊ	put	ʊ	u	ʊ	u	ù
u	boot	u	uw	uw	u:	ü
ə	butt	ʌ	ə	ə	ə	ə
aɪ	bite	aɪ	ay	ay	aɪ	i
aʊ	bout	aʊ	aw	aw	au	au
ɔɪ	boy	ɔɪ	oy	oy	ɔi	oi
ɜ	bird	ɜ	ər	ər	ə:	ər

Table B.1. Comparaison de plusieurs systèmes de transcription phonétique des voyelles anglaises, d'après [Ladefoged 75]. Les symboles utilisés par P. Ladefoged sont ceux de l'A.P.I..

Comme le souligne très justement F. Nolan [Nolan 83], la phrase est syntaxiquement et sémantiquement inhabituelle. Aussi, n'est-il pas possible de savoir si les différences entre les locuteurs résultent de la façon de réaliser l'accent de phrase (contours de  $F_0$  des syllabes accentuées) ou bien de l'emplacement de cet accent dans la phrase, c'est-à-dire de la compréhension de la phrase.

**Wolf [Wolf 72].** Les six phrases courtes de la table B.2 ont été prononcées dix fois par 21 locuteurs américains lors d'une seule session. Dans son étude, J.J. Wolf étudie la pertinence d'événements acoustiques calculés en des endroits localisés manuellement dans ces phrases :

- une dizaine de valeurs instantanées de  $F_0$ ,
- deux variations de  $F_0$  entre une syllabe accentuée et une syllabe non accentuée adjacente,
- les sorties d'un banc de 36 filtres (à espacement linéaire de 150 à 1650 Hz et logarithmique au-dessus) au centre d'un / n / et d'un / m /, après normalisation par rapport à l'énergie au centre de la voyelle suivante,
- les 2<sup>e</sup> et 3<sup>e</sup> moments centraux des zones situées entre 1500 et 4500 Hz du spectre instantané de / i / dans "need",

- les 2<sup>e</sup> et 3<sup>e</sup> moments centraux des zones situées entre 500 et 1500 Hz du spectre instantané du premier / a / dans "papa",
- une estimation par analyse-synthèse des formants F1 et F2 de / ə / dans "the",
- une estimation par analyse-synthèse des formants F1 et F2 du premier / a / dans "papa",
- une estimation par analyse-synthèse des formants F1 et F2 de / æ / dans "cash",
- une estimation de la pente spectrale de l'onde glottale par la différence des amplitudes des formants F1 et F3 de / u / dans "cool",
- la forme du spectre du / ʃ / dans "cash",
- la durée du mot "bought",
- le début de voisement dans la tenue d'un / b /.

- 1 *Cool shirts please me.*
- 2 *Pay the man first, please.*
- 3 *I cannot remember it.*
- 4 *Papa needs two singers.*
- 5 *A few boys bought them.*
- 6 *Cash this bond please.*

Table B.2. Les six phrases du corpus utilisé par J.J. Wolf et M.R. Sambur. Les phonèmes soulignés sont ceux étudiés par M.R. Sambur.

Au sens du F-ratio, les meilleurs paramètres sont toutes les valeurs instantanées de  $F_0$  alors que la variation de  $F_0$  au niveau des syllabes accentuées est un très mauvais indice. Ce mauvais résultat s'expliquerait *a posteriori* par l'article de M.R. Sambur [Sambur 75] qui précise que les phrases ont été écoutées avant d'être prononcées afin d'uniformiser le placement de l'accent. Viennent ensuite  $F_2$  de / æ / et  $F_2$  de / ə /. Puis, figurent, pour / m /, les sorties du banc de filtre situées vers 200, 800, 1800 et 3000 Hz, et, pour / n /, celles situées vers 2000, 1000 et 3000 Hz. Parmi ces paramètres, viennent s'interclasser l'estimation de la pente spectrale de l'onde glottale et les moments du / i /.

Par ailleurs, pour les consonnes nasales, l'étude du F-ratio en fonction du numéro de chacun des filtres d'analyse montre une très forte corrélation entre ses valeurs les plus élevées et les zones où se trouvent les maxima du spectre.

L'auteur termine son étude par un essai d'identification et de vérification du locuteur. Dans ce but, il choisit les neuf meilleurs paramètres qui ne soient pas interdépendants deux à deux. Malheureusement, il ne fournit pas la liste de ces paramètres. L'identification est effectuée par un simple calcul de distance euclidienne pondérée par l'inverse de la variance entre un vecteur inconnu et chacun des 21 vecteurs moyens calculés sur 5 répétitions. Un score d'identification de 98,5% est atteint mais sur les dix répétitions qui avaient servi au calcul des F-ratios.



**Sambur [Sambur 75]** M.R. Sambur reprend le corpus de J.J. Wolf qu'il complète par 4 sessions d'enregistrement où, cette fois, les phrases sont lues. Lors des 3 sessions enregistrées plus de deux ans après, seuls figurent 3 locuteurs du premier corpus auxquels s'ajoute un nouveau locuteur. Lors de la dernière session, quelques mois plus tard, se trouvent les 4 locuteurs précédents plus 7 locuteurs du premier corpus.

Feature	Speech Event	Feature	Speech Event
1. NF2	/n/	20. THISFO	FO
2. UF3	/u/	21. MANP2	/m/ in <u>man</u>
3. IP2	/I/	22. MANB3	/m/ in <u>man</u>
4. K	duration of /k/	23. EEF1	/i/
5. REMF3	/m/ in <u>remember</u>	24. EEF4	/i/
6. NF6	/n/	25. EEF3	/i/
7. REMF4	/m/ in <u>remember</u>	26. SHP2	/sh/
8. CASHFO	FO	27. AEF2	/ae/
9. IF4	/I/	28. AEF4	/ae/
10. AI	F2 slope in /aI/	29. AEF1	/ae/
11. REMFI	/m/ in <u>remember</u>	30. SF2	/s/
12. AVFO	FO	31. UF4	/u/
13. SF3	/s/	32. IF1	/I/
14. UF2	/u/	33. BONDFO	FO
15. EEF2	/i/	34. REMF6	/m/ in <u>remember</u>
16. NF1	/n/	35. IP5	/I/
17. MANP4	/m/ in <u>man</u>	36. MANB4	/m/ in <u>man</u>
18. UF1	/u/	37. AEF3	/ae/
19. NF3	/n/	38. SHP1	/sh/

Figure B.4. Les meilleurs paramètres caractérisant le locuteur d'après l'étude de M.R. Sambur.

L'auteur établit un ordre de pertinence entre 92 paramètres par éliminations successives du plus mauvais paramètre. Pour cela, il teste l'efficacité de chacun des sous-ensembles de  $(n-1)$  paramètres d'un ensemble de  $n$  paramètres. L'étude de l'efficacité est fondée sur le calcul de la probabilité d'erreur d'un classifieur linéaire qui suppose que le vecteur de paramètres représentant un locuteur suit une distribution gaussienne multidimensionnelle. Les 38 meilleurs paramètres sont classés dans la table B.4. Une expérience d'identification, fondée sur les 5 premiers paramètres, conduit à un taux d'erreur de 0,3%. Mais, d'une part, celle-ci est réalisée sur les données ayant servi au classement des paramètres, d'autre part, aucune information n'est fournie sur l'utilisation des différents corpus, hétérogènes au niveau des locuteurs, que ce soit au niveau du classement ou celui de la reconnaissance. Nous allons détailler quelques-uns des paramètres choisis par l'auteur :

- **les voyelles** : A partir des pôles d'une analyse LPC d'ordre 12, l'auteur évalue les formants des voyelles /æ/, /i/, /I/ et /u/, en tenant compte de leurs largeurs de bande et en utilisant les valeurs attendues des formants. Les pôles qui restent sont appelés pôles glottaux.

Les voyelles sont segmentées manuellement et les pôles sont calculés à l'endroit où F2 est stationnaire. L'auteur obtient les résultats suivants :

- les fréquences formantiques sont plus pertinentes que leurs largeurs de bande et que les pôles glottaux,
  - les formants d'ordre élevé présentent la plus grande variabilité interlocuteur,
  - certains locuteurs présente, de manière consistante, un pôle glottal vers 1000 Hz pour / *i* /,
  - comme le montre la table B.4, les meilleurs formants sont F2 et F4 de / *I* /, les trois premiers formants de / *u* / et les quatre premiers de / *i* / ;
- **les consonnes nasales** : M.R. Sambur détermine les formants des consonnes nasales à partir des pôles LPC. A cette fin, il évalue la méthode sur les données d'analyse-synthèse de Fujimura et conclut qu'elle est correcte sauf dans les domaines où pôles et zéros interagissent. Or, d'après [Su 74], Fujimura constate que le zéro attire les formants dans son voisinage constituant ainsi un amas formant-antiformant dans la région du zéro. Par ailleurs, le modèle LPC est un modèle tout-pôle, et nous pensons que 12 coefficients ne suffisent pas à modéliser six formants plus quelques zéros. Il serait donc plus raisonnable de considérer que les résultats de l'analyse sont des pôles et non exactement des formants. Quoiqu'il en soit, les pôles des deux nasales — notamment vers 1700 et 2300 Hz pour / *m* / et vers 1000 Hz pour / *n* / — réalisent dans cette étude la meilleure caractérisation du locuteur. Les moins bons résultats de "man" par rapport à "remember" sont à notre avis dus à sa moindre accentuation ;
  - **les fricatives** : l'auteur modélise 5 pôles de / *s* / et / *ʃ* /, compris entre 0 et 10 kHz, par une analyse LPC d'ordre 10. Très peu d'entre eux figurent dans la table B.4 ;
  - **les paramètres temporels** : la durée de la friction et de l'aspiration du / *k* / varie de 20 à 127 ms selon le locuteur. De même, la pente de F2 dans la diphtongue / *aI* / possède une bonne variabilité interlocuteur. Ceci montre que les paramètres liés aux comportements appris du locuteur sont aussi informatifs sur celui-ci que les paramètres reliés à ses particularités physiologiques ;
  - **F0** : la valeur moyenne de F0 et son évolution temporelle sur la phrase 6 caractérisent beaucoup moins le locuteur que dans l'étude précédente. Cela semble dû à une variabilité intralocuteur importante entre les sessions (jusqu'à 20Hz ), variabilité que J.J. Wolf n'avait pas prise en compte.

**Su et al. [Su 74].** Dans cette étude, les auteurs appliquent à l'identification du locuteur les conséquences acoustiques de l'influence coarticulatoire des voyelles sur les consonnes nasales que nous avons introduite au chapitre V de la partie A.

Dans une première étape, ils mettent en évidence le phénomène en comparant les spectres, sélectionnés manuellement, d'une consonne nasale, suivie d'une voyelle postérieure, à ceux d'une consonne nasale, suivie d'une voyelle antérieure. Dans ce but, 4 locuteurs américains (2H et 2F) prononcent 3 répétitions de / *hə* 'CVd /<sup>1</sup> où C est une des consonnes / *m* / et / *n* / et où V est soit une des voyelles d'avant / *i* /, / *e* /, / *æ* /, soit une des voyelles d'arrière / *u* /, / *o* /, / *a* /. Les auteurs soumettent les sorties de 25 filtres compris entre 250 Hz et 3700 Hz à une analyse en composantes principales. La partie importante de l'information spectrale (énergie

<sup>1</sup> Le symbole ' précise que la syllabe suivante porte l'accent lexical principal.



totale, deux premiers zéros et formants ) se retrouve dans les premiers vecteurs propres. La figure B.5 montre la représentation, pour un locuteur, des spectres des consonnes nasales selon les deux premiers axes principaux. Alors que tous les échantillons de [ n ] sont regroupés en un seul nuage, les échantillons de [ m ] se répartissent en deux nuages distincts, celui de gauche correspondant aux voyelles antérieures, l'autre aux voyelles postérieures. Ceci s'explique par le fait que l'articulation du [ n ] ne subit que la réduction du geste articuloire provoqué par la voyelle adjacente tandis que celle de [ m ] subit également l'anticipation du geste articuloire de la langue pour préparer la production de la voyelle. De plus, les auteurs constatent que l'emplacement et la dispersion des deux nuages varient avec le locuteur.

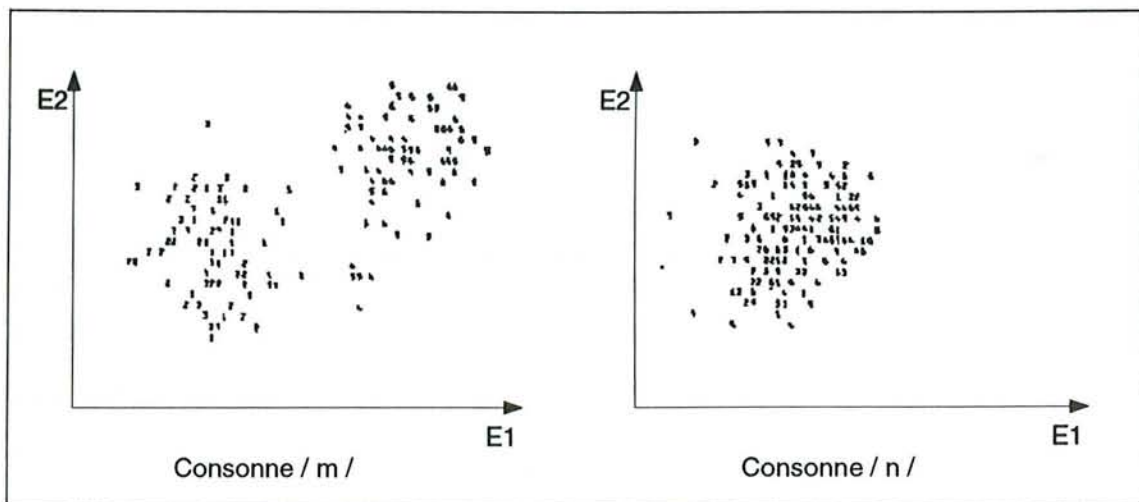


Figure B.5. Représentation des échantillons des spectres instantanés de [ m ] et [ n ] selon les deux premiers axes principaux, d'après L.S. Su.

Les auteurs utilisent donc la différence entre les centroïdes de ces deux nuages des spectres du / m / comme vecteur caractérisant le locuteur. La distance de corrélation entre deux locuteurs est définie comme le produit scalaire des vecteurs normalisés. Ils obtiennent 100% d'identification pour ces 4 locuteurs ainsi que pour 10 autres qui ont prononcé six fois lors d'une même session la suite / hə 'mVd / pour V appartenant à { / i /, / e /, / I /, / u /, / o /, / a / }. L'ajout de 23 locuteurs non avertis fait chuter ce taux à 85%.

Enfin, ils mettent en évidence l'efficacité de la caractérisation du locuteur par ce nouveau paramètre par rapport à celle qui utilise le spectre de / n / indépendamment du contexte dans les mots isolés ou la parole continue. Pour cela, ils étudient la distribution des différences des distances intralocuteur et interlocuteur.

**Goldstein [Goldstein 76].** Ursula Goldstein combine la méthode de M.R. Sambur à celle de J.J. Wolf pour rechercher la pertinence des fréquences formantiques dans un certain nombre de sons choisis pour leur variabilité dialectale. Les phonèmes suivants sont prononcés, dans le contexte / b - d /, cinq fois par 10 locuteurs américains et cinq autres fois, un mois plus tard, par six d'entre eux :

- les voyelles tendues / o /, / e /, / i / et / u /,



- les diphtongues / ɔ I /, / a I / et / a U /,
- la voyelle rétroflexe / ʁ / et les couples rétroflexes / a r / et / r ɛ /<sup>1</sup>.

Les fréquences formantiques sont déterminées entre un point situé à 20 ms du burst du / b / jusqu'au début de la tenue du / d /. Elles sont obtenues à partir des pôles d'une analyse LPC d'ordre 12 qui ont une largeur de bande inférieure à 700 Hz, puis sont vérifiées et corrigées manuellement. Deux cents paramètres sont calculés à partir de celles-ci : minima, maxima, moyennes, etc.

Un premier calcul de F-ratio sélectionne 24 paramètres qui sont ensuite éprouvés selon la probabilité d'erreur de M.R. Sambur. La méthodologie est toutefois différente. Le paramètre, testé seul, qui possède la plus petite erreur est le plus pertinent. Il est couplé aux 23 autres afin de déterminer la meilleure paire et ainsi de suite jusqu'à l'obtention de l'ensemble des dix paramètres présentés dans la table B.3. Le même ensemble est atteint par élimination des paramètres redondants grâce aux coefficients de corrélation et par l'examen des moyennes des paramètres par locuteur afin de rejeter les paramètres ayant un F-ratio artificiellement élevé.

- |    |   |
|----|---|
| 1  | Valeur minimale de F <sub>2</sub> dans / a r /.   |
| 2  | Valeur maximale de F <sub>1</sub> dans / a r /.   |
| 3  | Valeur maximale de F <sub>2</sub> dans / o /.   |
| 4  | Moyenne entre la valeur minimale de F <sub>3</sub> et les deux valeurs adjacentes dans / ʁ /.                 |
| 5  | Valeur maximale de F <sub>2</sub> dans / ɔ I /.   |
| 6  | Moyenne de F <sub>3</sub> sans tenir compte des 20 dernières ms dans / u /.                                   |
| 7  | Moyenne de l'écart entre F <sub>2</sub> et F <sub>3</sub> sans tenir compte des 20 dernières ms dans / r ɛ /. |
| 8  | Valeur maximale de F <sub>2</sub> dans / a U /.   |
| 9  | Moyenne de F <sub>4</sub> sans tenir compte des 20 dernières ms dans / a r /.                                 |
| 10 | Valeur maximale de F <sub>1</sub> dans / ʁ /.   |

Table B.3. Les dix meilleures fréquences formantiques d'après U. Goldstein.

La majorité des fréquences formantiques retenues concernent les sons rétroflexes et en particulier la paire / a r /. Elles traduisent la disparité entre les locuteurs à la fois au niveau de l'influence du / r / sur la voyelle adjacente et au niveau de la forme de langue à l'arrière de l'articulation du / r /. La figure B.6 montre la position des trois premiers formants du / r / par rapport à ceux du / a / en anglais comme en français. La particularité des phonèmes / ʁ / et / r / étant la valeur très basse de F<sub>3</sub>, on peut s'étonner de ne pas trouver ses effets sur les voyelles adjacentes dans la table. Une explication de cette absence serait le peu de fiabilité annoncée par l'auteur pour les suivis des formants F3 et F4.

L'auteur explique les bonnes performances des valeurs maximales de F<sub>2</sub> de / ɔ I / et / a U / par la liberté laissée au locuteur dans le choix de la cible articulatoire constituant

<sup>1</sup> En anglais, le / r / est une consonne vibrante apicale encore appelée rétroflexe (cf. figure A.29). De la même façon, les voyelles rétroflexes sont prononcées avec la pointe de la langue relevée vers le palais et sont dues à la chute du / r / apical, notamment en anglais américain ("girl", "far").

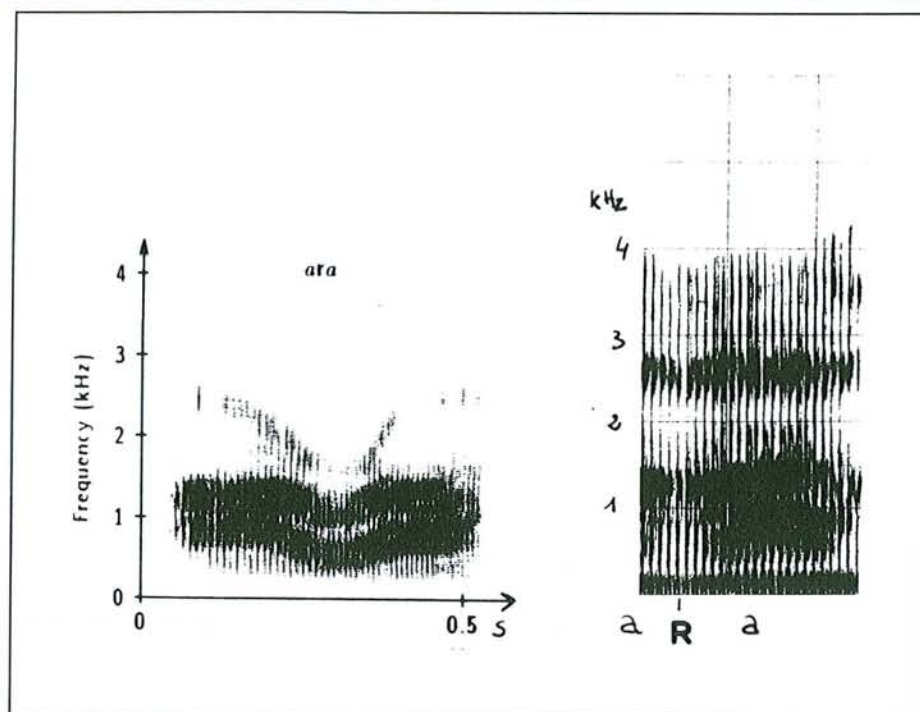


Figure B.6. Evolution des trois premiers formants dans / *a r a* / en anglais, d'après [O'Shaughnessy 87] et dans / *a r a* / en français, d'après les données du GRECO Communication Parlée [Eskenazi 88].

la deuxième partie de la diphtongue. En effet, les traits perceptivement distinctifs des trois diphtongues seraient principalement la vitesse et la direction des trajectoires formantiques. La bonne performance de la valeur maximale de  $F_2$  de / *o* / est due d'après l'auteur à la forme du palais lors de l'articulation de cette voyelle qu'elle considère comme centrale. A notre avis, elle provient plutôt de la variabilité de la coarticulation avec le / *d* / qui suit la voyelle. En effet, les transitions de  $F_2$  et  $F_3$  sont montantes (vers la consonne) pour les voyelles centrales ou postérieures (cf. figures A.41 et B.1). Or, nous pensons que / *o* / correspond au / *ɔ* / de "bought" et donc à une voyelle postérieure centralisée par / *d* /. Par ailleurs, le spectre de la fricative / *ʃ* /, supposé représentatif de la forme du palais du locuteur, a donné de piètres résultats dans les études précédemment décrites. Ce phénomène de coarticulation pourrait expliquer également en partie la pertinence du deuxième formant des diphtongues.

La bonne performance du troisième formant du / *u* / conforte la pertinence de ce phonème, qui avait été établie par ailleurs par M.R. Sambur dans un contexte similaire (/ *bud* / vs. / *tus* /).

**Paliwal [Paliwal 84].** K.K. Paliwal étudie la pertinence des quatre premiers formants de toutes les voyelles de l'anglais britannique dans le contexte / *h - d* /. La table B.4 présente ces voyelles ainsi que les mots les contenant, qui ont été prononcés 5 fois par 10 locuteurs.

Après une numérisation sur 12 bits à 10 kHz, l'auteur extrait manuellement une fenêtre d'analyse de 25,6 ms sur laquelle est calculé un spectre lissé cepstral. Chacune des



fréquences formantiques est obtenue par une interpolation parabolique calculée sur un pic du spectre désigné manuellement.

La pertinence de chacune des voyelles est mesurée en effectuant l'identification du locuteur à l'aide de chacune d'elles. La distance minimale entre le vecteur de test ( $F_1, F_2, F_3, F_4$ ) et la moyenne des quatre autres vecteurs de chacun des locuteurs donne le locuteur reconnu. L'auteur teste deux distances, la distance euclidienne et la distance de corrélation. La table B.4 donne les taux d'identification pour les deux distances. Les deux voyelles / 3 / et / / sont reconnues plus pertinentes au sens des deux distances, suivies de / I /, / u /, / ʌ / et / o /, dans un ordre différent selon la distance. Mais la distance euclidienne est plus performante pour la reconnaissance du locuteur.

Mots prononcés par les locuteurs	Voyelles	Taux d'identification		Classement des fréquences formantiques obtenu par le F-ratio				
		distance eucli- dienne	distance de corrélation	F <sub>1</sub>	F <sub>2</sub>	F <sub>3</sub>	F <sub>4</sub>	Σ
heard	/ 3 /	84	76	14	3	4	12	1
hud	/ ʌ /	72	64	2	9	34	25	3
who'd	/ u /	74	62	32	11	20	38	7
hood	/ /	78	68	21	6	30	31	4
hoard	/ o /	74	44	41	27	24	5	8
hod	/ ɔ /	66	46	18	23	26	19	5
hard	/ a /	72	58	40	43	16	35	11
had	/ æ /	62	56	10	36	22	29	6
head	/ ɛ /	54	46	44	15	13	37	9
hid	/ I /	76	62	28	1	7	17	2
heed	/ i /	60	48	42	8	32	39	10

Table B.4. Tableau synthétique de l'étude de K.K. Paliwal comprenant : les voyelles étudiées, les mots prononcés, les taux de reconnaissance pour chacune des distances, le classement de chacune des fréquences formantiques selon le F-ratio.

L'auteur étend l'étude de la pertinence à chacune des fréquences formantiques en calculant leurs F-ratio. La table B.4 fournit le classement des fréquences formantiques selon un F-ratio décroissant. La pertinence de la voyelle rétroflexe / 3 / est confirmée par la présence de ses quatre formants parmi les 15 premiers paramètres. Rappelons que cette pertinence avait aussi mise en évidence par U. Goldstein. La meilleure caractéristique est F<sub>2</sub> de / I / dont l'efficacité avait déjà été démontrée par M.R. Sambur avec une autre méthodologie, et dans une moindre mesure par U. Goldstein (F<sub>2</sub> de / ɔ I /). En revanche, on ne retrouve pas F<sub>3</sub> de / u / qui avait été bien classé par M.R. Sambur et par U. Golstein. Notons également l'apparition de



/ ʌ / comme nouvelle voyelle caractérisant le locuteur grâce à  $F_1$  et  $F_2$ , qui ont, à notre avis, l'avantage de pouvoir être obtenus de manière fiable par analyse automatique. Si nous ajoutons une dernière colonne à la table B.4, qui donne le classement des voyelles selon la somme des F-ratio des fréquences de leurs formants, nous retrouvons en tête les quatre voyelles classées par l'identification.

**Nolan [Nolan 83].** F. Nolan réalise une étude approfondie de la pertinence des deux phonèmes anglais / l / et / r / situés en initiale de syllabe. Son principal objectif est d'étudier l'influence des faits de coarticulation sur ces deux phonèmes dans la caractérisation du locuteur. Pour cela, il constitue un corpus de mots presque tous monosyllabiques dans lesquels la liquide est suivie par chacune des dix voyelles / i:, I, e, æ, ʌ, ɑ:, , ɔ:, , u: /. La voyelle est elle-même suivie d'une consonne à articulation labiale, dentale-alvéolaire ou vélaire. Lors de deux sessions distantes de trois mois, les mots présentés dans la table B.5 ont été lus deux fois par 15 étudiants âgés de 17 ans, une fois sans article, l'autre fois précédés de l'article indéfini "a".

/ i: /	leap	leak	league	reap	reek	reed
/ I /	lip	lick	lid	rip	rick	rid
/ e /	let	leg	led	wreck	rest	red
/ æ /	lap	lack	lad	rap	rack	rag
/ ʌ /	luck	lust	lug	rut	ruck	rug
/ ɑ: /	lark	lard	large	raft	rasp	raj
/ /	lot	lock	log	rock	rob	rod
/ ɔ: /	lord	lore	law	wrought	roar	raw
/ /	look	look	look	rook	rook	rook
/ u: /	loop	loot	lose	roop	route	ruse

Table B.5. Les triplets / l V C / et / r V C / étudiés par F. Nolan.

Dans la première partie de son étude, l'auteur estime les trois premières fréquences formantiques des réalisations de / l / et / r / manuellement à partir de spectrogrammes et de sections spectrographiques et en rejetant les pics du spectre qui n'apparaissent pas dans les zones fréquentielles attendues. Il en déduit, tous locuteurs confondus, l'influence coarticulatoire de la voyelle sur les trois formants de / l / et / r /, qui est important pour  $F_2$  de / l / et dans une moindre mesure pour  $F_2$  de / r /. De plus, l'influence de la voyelle sur le deuxième formant de / l / varie considérablement avec le locuteur. Par ailleurs, la valeur moyenne calculée sur toutes les voyelles des trois formants des deux liquides montre que les formants  $F_1$  de / l / et / r / varient peu d'un locuteur à l'autre. En revanche,  $F_2$  de / r / et de / l / présentent une plage de variation de 300 Hz et plus, et  $F_3$  de / r / et de / l / une plage de plus de 1000 Hz.

Dans une seconde étude, F. Nolan teste la pertinence de ses paramètres au sens du F-ratio de J.J. Wolf [Wolf 72]. Paradoxalement, les formants de / r / obtiennent un bien meilleur F-ratio, l'auteur conclut à une plus faible variabilité intralocuteur du / r / puisqu'il est moins soumis à la coarticulation.

Dans une troisième partie, l'auteur a mené des expériences d'identification automatique du locuteur à partir des mêmes données : échantillonnage du corpus à 10 kHz et conversion 12 bits, calcul d'un spectre LPC à partir de 14 coefficients et comparaison d'un spectre de test à un spectre de référence en utilisant la distance euclidienne (la distance de corrélation ne donnant pas de meilleurs résultats).

Nous présentons dans la table B.6, les résultats moyens de différentes expériences d'identification qui utilisent toutes une seule occurrence du phonème pour le test et font une moyenne sur vingt ou trente contextes pour la référence. Nous pouvons constater que la variabilité intralocuteur temporelle (deux sessions espacées de 3 mois) est considérable et fait baisser le taux d'identification de plus de 20%, et, que l'influence coarticulatoire de la consonne qui suit la voyelle n'est pas négligeable pour / l /. L'auteur attribue en partie la meilleure pertinence de / r / à des réalisations particulières de trois des quinze locuteurs.

utilisation des données pour l'identification des locuteurs	/ l /	/ r /
même session pour apprentissage et test même mot	56%	62%
session différentes pour apprentissage et test, même mot	35%	36%
même session pour apprentissage et test, mots différents	52%	61%
session différentes pour apprentissage et test, mots différents	35%	36%

Table B.6. Pourcentages moyens d'identification correcte calculés à partir des résultats fournis par F. Nolan.

F. Nolan montre également que le taux d'identification du locuteur augmente –surtout pour / l /– lorsque l'échantillon de test est la moyenne de 2, 5 ou 10 échantillons issus de contextes vocaliques différents. Ceci s'explique, d'après l'auteur, par l'effet coarticulatoire de la voyelle sur le spectre du / l / qui peut entraîner la confusion entre une réalisation du phonème dans un contexte vocalique particulier par un locuteur et la réalisation moyenne d'un autre locuteur.

L'auteur termine son étude sur la pertinence des deux liquides en leur appliquant la méthode mise en œuvre par Su et al. [Su 74] dans leur analyse de la pertinence de l'influence coarticulatoire des voyelles sur / m /. Le degré de coarticulation de chacun des locuteurs est estimé par la distance euclidienne entre le centre d'inertie des spectres des liquides en contexte vocalique antérieur (/ i:, I, e, æ, ʌ /) et celui déterminé en contexte vocalique postérieur / ɑ:, ɔ:, u: /. Pour / l /, il est possible de classer les locuteurs selon cette distance, qui est fortement corrélée à la valeur de F2. Le classement et la corrélation sont moins pertinentes dans le cas de / r /.

Enfin, F. Nolan conduit la même expérience d'identification du locuteur que Su et al. Les résultats sont résumés dans la table B.7. L'évaluation du degré de sensibilité à la coarticulation du / l / ne conduit pas à de meilleurs taux d'identification que la comparaison directe des locuteurs par des spectres moyens des réalisations de / l / sur 5 et 10 échantillons. De plus, ces résultats moins bons que ceux de Su et al. (100% pour 10 locuteurs et une seule session) laissent supposer que le phonème / l / anglais en début de syllabe est moins sensible à la coarticulation que la nasale / m /, ce qui paraît justifié par le fait que, lors de l'articulation du



/ l /, la langue n'est pas libre. Les taux d'identification pour / r / sont médiocres et nettement moins bons que ceux obtenus par la comparaison des spectres moyens.

utilisation des données pour l'identification des locuteurs	/ l /	/ r /
session différentes pour apprentissage et test, centre d'inertie calculé sur 30 mots pour le test et la référence	61%	39%
même session pour apprentissage et test, centre d'inertie calculé sur 20 mots pour la référence et 10 autres pour le test	58%	32%
session différentes pour apprentissage et test, centre d'inertie calculé sur 20 mots pour la référence et 10 autres pour le test	51%	19%

Table B.7. Pourcentages moyens d'identification correcte calculés à partir des résultats fournis par F. Nolan sur l'identification à l'aide du degré de coarticulation de chaque locuteur.

**Autres études citées par F. Nolan.** Nous allons citer ici quelques études pour lesquelles nous n'avons pas eu accès directement aux publications correspondantes mais qui sont présentées dans les travaux de F. Nolan [Nolan 83].

En 1968, J.W. Glenn and N. Kleiner ont effectué une expérience d'identification de 30 locuteurs à partir de spectres de / n / dont le contexte n'était pas le même entre le mot de référence et le mot de test. Pour chacun des locuteurs, les auteurs ont calculé un vecteur représentant la moyenne de trois spectres instantanés compris entre 1000 Hz et 3000 Hz et obtenus au centre de la consonne. Selon que le locuteur était représenté directement par ce vecteur ou par la moyenne de ces vecteurs sur dix mots, le taux d'identification est passé de 43% à 93%. Ce résultat confirme ceux de L.S. Su et al. en ce qui concerne l'importance en caractérisation du locuteur des effets de la coarticulation sur les consonnes nasales.

Une autre étude réalisée en 1975 par J.E. Paul et al. a tenté de classer 13 phonèmes américains dans une expérience semi-automatique de reconnaissance du locuteur. Le classement obtenu est le suivant : / , u, i, m, I, a, ɔ, n, η, 3<sup>r</sup>, ʌ, α, ε /. Notons les mauvaises performances des phonèmes / 3<sup>r</sup> / et / ʌ / par rapport à celles obtenues dans les études de U. Golstein et K.K. Paliwal. En revanche, les bons résultats de / / et / I / obtenus par ces auteurs et par M.R. Sambur sont confirmés.

### 3.3. Les études sur la langue allemande

C'est également dans le livre de F. Nolan que nous avons trouvé la seule référence concernant une étude de la caractérisation du locuteur en allemand. Il s'agit d'une étude réalisée en 1977 par U. Höfker sur la pertinence de 24 phonèmes allemands prononcés isolément pour la discrimination de 12 locuteurs. Les trois premiers phonèmes sont / n /, / η / et / m /.



### 3.4. Les études sur la langue française

**Corsi [Corsi 79].** P. Corsi teste l'efficacité de paramètres temporels pour la reconnaissance du locuteur. Au cours d'une même session, 12 locuteurs d'origine géographique variée, lisent 10 répétitions d'un texte administratif d'une durée moyenne d'une minute, ainsi que 10 répétitions des six phrases de la table B.8.

- |   |  |
|---|--|
| 1 | <i>Il rap<u>a</u>, c'est bien connu.</i>       |
| 2 | <i>Il le ma<u>t</u>a, c'est bien connu.</i>    |
| 3 | <i>L'av<u>o</u>cat, il est bien connu.</i>     |
| 4 | <i>Il l'ag<u>a</u>ca, c'est bien connu.</i>    |
| 5 | <i>C'est un <u>ch</u>at, c'est bien connu.</i> |
| 6 | <i>Quelle cac<u>o</u>phonie.</i>               |

Table B.8. Les six phrases du corpus utilisé par P. Corsi. Les phonèmes soulignés sont ceux dont la durée a été analysée par P. Corsi.

De ce corpus, l'auteur extrait un ensemble de 37 paramètres répartis en 9 groupes selon la table B.9.

Dans une première étape, il recherche des sous-groupes décorrélés de paramètres afin de sélectionner dans chacun d'eux le meilleur paramètre au sens du F-ratio. Finalement, il conserve les groupes initiaux, malgré une corrélation entre les durées des tenues des occlusives et leurs durées totales et une corrélation entre la durée des groupes de phonation dans le texte et les durées des six phrases. Il obtient donc les neuf paramètres suivants (un par groupe) classés dans l'ordre du F-ratio décroissant :

- fréquence fondamentale moyenne sur le texte,
- durée du / **k** / dans la phrase 3,
- durée de la phrase 3,
- durée de la tenue du / **p** / dans la phrase 1,
- débit sur le texte,
- taux de voisement sur le texte,
- durée du premier groupe de phonation dans la troisième phrase du texte,
- durée de la phrase 5,
- durée de la pause dans la phrase 3.

	Fondamental moyen sur le texte.
	Fondamental moyen sur chacune des trois premières phrases du texte.
	Taux de voisement sur le texte.
	Taux de voisement sur chacune des trois premières phrases du texte.
	Débit sur le texte.
	Durée des quatre groupes de phonation de la troisième phrase du texte.
	Durée des pauses entre les quatre groupes de phonation.
	Durée de chacune des six phrases.
	Durée de la tenue de / p / dans la phrase 1.
	Durée de la tenue de / t / dans la phrase 2.
	Durée de la tenue de / k / dans la phrase 3.
	Durée de la tenue de / k / dans la phrase 6.
	Durée complète de / p / dans la phrase 1.
	Durée complète de / t / dans la phrase 2.
	Durée complète de / k / dans la phrase 3.
	Durée complète de / k / dans la phrase 6.
	Durée complète de / s / dans la phrase 4.
	Durée complète de / ʃ / dans la phrase 5.

Table B.9. Les paramètres sélectionnés par P. Corsi répartis en 9 groupes.

Dans une seconde étape, il effectue plusieurs analyses discriminantes pas à pas sur l'ensemble des paramètres avec deux critères de fin possibles. L'une d'elles recherche les 10 meilleurs paramètres sur tout le corpus. Les autres arrêtent la sélection de paramètres sur huit répétitions lorsque le taux d'identification atteint 100% sur les deux répétitions restantes. Les cinq premiers paramètres de la première étude se retrouvent en tête dans toutes les analyses en revanche le taux de voisement n'est plus discriminant. Les 100% d'identification sont atteints avec seulement 4 ou 6 paramètres selon les deux répétitions qui ont été choisies comme énoncés de test.

**Pérennou et al. [Perennou 82] [Caelen-Haumont 88].** L'étude réalisée entre 1980 et 1982 par une équipe du C.E.R.F.I.A. est la plus importante sur la caractérisation du locuteur français par le nombre de paramètres acoustico-phonétiques, phonologiques et prosodiques envisagés. Lors de séances échelonnées sur neuf mois, 5 locuteurs, dont un seul ne serait pas originaire du Sud, lisent 3 répétitions du texte présenté dans la table B.10. Ce corpus est numérisé à 15 kHz avec une conversion sur 12 bits. Puis, il est analysé par un modèle d'oreille qui fournit la fréquence fondamentale de chaque prélèvement de 8,5 ms ainsi que les énergies de 24 canaux répartis de 200 à 6000 Hz.

*Paul Dupont demeure à Blagnac. Lundi sept décembre dans la matinée, il a rendez-vous chez son banquier pour négocier un emprunt. Il désire en effet acquérir un fonds de commerce d'appareils ménagers. Le magasin situé en plein centre de Toulouse, sera tenu par sa femme, Madame Dupont.*

Table B.10. Le texte lu par cinq locuteurs masculins dans l'étude de G. Pérennou et al.

Deux analyses sont mises en œuvre à partir de ce corpus. Dans un premier temps, les auteurs effectuent une étude analytique de plusieurs paramètres. Puis, ils appliquent une analyse discriminante sur certains paramètres issus de la première étude.

Les paramètres étudiés lors de l'étude analytique peuvent se regrouper en trois catégories, qui sont les paramètres acoustico-phonétiques spécifiques à certains phonèmes, les paramètres qui mesurent la dynamique du locuteur au niveau de la fréquence fondamentale, de l'intensité et de la durée, et les paramètres d'ordre prosodique. Ces travaux étant les plus proches des nôtres, nous allons détailler la plupart de ces paramètres.

• **Les paramètres acoustico-phonétiques spécifiques**

Une grande partie de ces paramètres est constituée par les réalisations allophoniques de certains phonèmes. Remarquons tout de suite que, dans ce cas, les auteurs ne précisent pas si les différents allophones sont identifiés à l'écoute ou par une analyse du signal. Les autres paramètres sont d'ordre fréquentiel.

Voici les principaux résultats concernant tous ces paramètres :

- parmi les dix-huit occurrences du phonème / a / présentes dans le texte, onze ont été prononcées différemment selon les locuteurs. Parmi ces prononciations, les auteurs distinguent quatre allophones dont les traits distinctifs sont indiqués dans la table B.11. Les cinq locuteurs sont parfaitement discriminés par les rapports entre les effectifs des différents allophones ;

Allophones	F <sub>1</sub>	F <sub>2</sub>	Articulation
[ a ]	700 Hz	1200 Hz	ouverte / moyenne
[ æ ]	600 Hz	1450 Hz	fermée / postérieure
[ ʌ ]	600 Hz	1250 Hz	fermée / moyenne
[ ɑ ]	600 Hz	1100 Hz	fermée / antérieure

Table B.11. Les allophones de / a / étudiés par Pérennou et al.

- en revanche, les réalisations allophoniques des archiphonèmes (/ ẽ /, / œ /) et (/ o /, / ɔ /) ne permettent pas de discriminer correctement les locuteurs ;
- les auteurs étudient également les différentes réalisations du "e muet", [ œ ], [ ø ] et [ ə ]. Celles-ci ne discriminent pas efficacement les cinq locuteurs ;



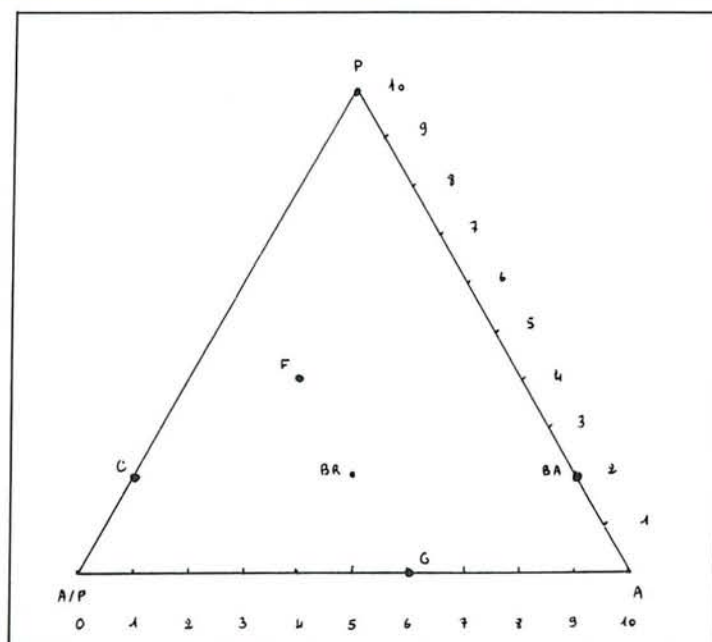


Figure B.7. Représentation dans le plan des articulations des occlusives vélaires selon leur nombre et les locuteurs d'après Pérennou et al.

- la figure B.7 montre que le lieu d'articulation des occlusives vélaires du corpus est un paramètre pertinent pour la caractérisation des locuteurs ;
- les durées, normalisées par rapport au débit, de l'amortissement du voisement et du premier formant des voyelles pendant le début des consonnes sourdes permettent de séparer les locuteurs en trois classes. Mais les auteurs ne précisent pas comment sont déterminées ces durées ni sur quels critères ;
- les paramètres symétriques, qui sont les délais d'établissement du voisement et du premier formant, séparent les cinq locuteurs en quatre classes ;
- le calcul sur l'ensemble des consonnes nasales du pourcentage de prélèvements dont les fréquences formantiques se situent dans certaines bandes de fréquences montre que la bande comprise entre 2000 et 3400 Hz ( $F_3 / F_4$ ) discrimine les cinq locuteurs ;
- le même calcul appliqué aux consonnes constrictives donne aussi de bons résultats, toujours pour la bande comprise entre 2000 et 3400 Hz, alors que le pourcentage de prélèvements ayant de l'énergie au dessus de 4000 Hz est un mauvais paramètre. En revanche, ce dernier paramètre est pertinent lorsqu'il est établi sur des phonèmes qui ne sont pas phonologiquement fricatifs comme / p / et / k / (quatre classes de locuteurs) ;
- la position de F1 pour les consonnes voisées est un mauvais paramètre ;
- les auteurs examinent également le phénomène de coarticulation entre l'occlusive / t / et la voyelle / i / dans "matinée". Les auteurs distinguent quatre variantes :
  1. aucun bruit de friction entre les deux phonèmes, [ tɪn ],
  2. un bruit de friction suit l'explosion du [ t ], [ tsɪn ],

Locuteurs	Répétition 1	Répétition 2	Répétition 3
BA	V2	V3	V3
BR	V2	V3	V2
C	V1	V1	V1
G	V2	V1	V2
F	V1	V1	V1

Table B.12. Les variantes coarticulatoires du couple /tĩ/ de "matinée" dans l'étude de Pérennou et al.

3. le [ĩ] est remplacé par un son intermédiaire entre [ĩ] et [s],
4. le [ĩ] est remplacé par un [s].

La table B.12 montre que seules les trois premières variantes sont réalisées par les cinq locuteurs. De plus, trois des locuteurs présentent des réalisations variables selon les répétitions. Il serait intéressant de savoir s'il existe une corrélation entre cette variabilité intralocuteur et le débit ;

– les auteurs ont également étudié le phénomène de coarticulation qui intervient entre la fin d'une voyelle nasale brève et l'occlusive sonore suivante. Ils distinguent trois variantes :

1. aucune coarticulation entre les deux phonèmes,
2. l'occlusive est nasalisée,
3. un segment consonantique nasal apparaît entre la voyelle et l'occlusive.

Si les cinq locuteurs réalisent la variante 2, deux d'entre eux articulent la variante 1 et n'articulent jamais la variante 3 alors que les trois autres locuteurs ont exactement le comportement réciproque.

#### • Dynamique des paramètres suprasegmentaux

Les paramètres qui suivent sont des mesures de la dynamique de l'énergie, de la durée et de la fréquence fondamentale effectuées sur certains phonèmes indépendamment de leur qualité linguistique, qu'elle soit phonologique ou syntaxico-sémantique. Selon les auteurs, ces paramètres fournissent une information sur le timbre de la voix du locuteur :

- la dynamique de l'énergie vocalique (écart entre l'estimation de l'énergie la plus faible des voyelles à forte énergie et l'estimation de l'énergie la plus élevée des voyelles à faible énergie) discrimine les cinq locuteurs considérés ;
- la dynamique de l'énergie consonantique est trop faible pour pouvoir être exploitée ;
- les dynamiques des durées vocalique (voyelles nasales incluses) et consonantique présentent une variabilité interlocuteur satisfaisante ;
- le rapport de la fréquence fondamentale des phonèmes les plus graves à la fréquence fondamentale des phonèmes les plus aigus constitue un mauvais paramètre pour différencier les cinq locuteurs.

Paramètre de durée	nombre de classes
Durée de phonation	5
Durée phonémique minimale	1
Durée phonémique maximale	3
Nombre de syllabes / temps de phonation	4
Nombre de syllabes / temps de parole	4
Durée cumulée des pauses	2
Rapport de la durée des pauses à la durée de phonation	3
Durée moyenne d'une pause	3
Durée de la voyelle en initiale de phrase	1
Durée de la voyelle finale d'un groupe	1 mais 2 locuteurs s'opposent
Durée de la voyelle finale d'une phrase	1 mais 2 locuteurs s'opposent
Ecart entre la durée d'une voyelle en finale de groupe de phonation et celle en initiale de phrase	3

Table B.13. Efficacité des durées d'origine prosodique dans la discrimination des cinq locuteurs dans l'étude de Pérennou et al.

#### • Les paramètres prosodiques

Dans ce cas, la position syntaxique de la voyelle sur laquelle sont effectuées les mesures est primordiale :

- les différentes mesures de l'énergie sur une phrase (énergie moyenne, en début ou en fin de phrase, ...) donnent de mauvais résultats dans la caractérisation des cinq locuteurs. Seul l'écart entre l'énergie en début de phrase et celle en fin de phrase conduit à une discrimination assez efficace ;
- la table B.13 fournit la séparation des locuteurs en classes disjointes selon les différents paramètres de durée étudiés. Il faut toutefois remarquer qu'une seule mesure par locuteur est effectuée pour les huit premiers paramètres de la table. Pour les paramètres suivants, des écarts types sont calculés mais deux autres remarques s'imposent. Les voyelles en début et en fin de phrase ne sont pas les mêmes pour les quatre phrases du texte et la durée intrinsèque des voyelles peut influencer les résultats. En ce qui concerne la voyelle finale de groupe, la manière dont a été déterminé le groupe de phonation (a priori ou a posteriori) n'est pas indiquée. Il n'est donc pas possible de savoir si la source de différence entre les locuteurs est la façon de réaliser une frontière syntaxique ou l'emplacement de cette frontière dans la phrase,
- comme pour la durée, plusieurs mesures prosodiques de la fréquence fondamentale sont testées. Les résultats sont présentés dans la table B.14. Les deux résultats qui concernent les valeurs minimale et maximale de  $F_0$  ne prennent en compte qu'une valeur par locuteur. La remarque précédente sur la voyelle située en fin de groupe reste valable dans le cas de  $F_0$ . La comparaison par les auteurs des fréquences minimale, moyenne et maximale



Mesure de F <sub>0</sub>	nombre de classes
F <sub>0</sub> moyen sur une phrase	1
F <sub>0</sub> maximale sur une phrase	3
F <sub>0</sub> minimale sur une phrase	2
Ecart entre F <sub>0</sub> maximale et F <sub>0</sub> minimale	3
F <sub>0</sub> en début de phrase	1 mais 2 locuteurs s'opposent
F <sub>0</sub> en fin de groupe	3 classes
F <sub>0</sub> en fin de phrase	2 classes
Ecart entre F <sub>0</sub> en début de phrase et F <sub>0</sub> en fin de phrase	4
Ecart entre F <sub>0</sub> en fin de groupe et F <sub>0</sub> en fin de phrase	5

Table B.14. Efficacité des différentes mesures prosodiques de F<sub>0</sub> dans une phrase dans la discrimination des cinq locuteurs dans l'étude de Pérennou et al.

montre que les locuteurs centrent différemment leur registre phonatoire par rapport à leur fréquence fondamentale moyenne.

Après cette première étude, les auteurs appliquent deux analyses discriminantes de Fisher sur les phonèmes qui respectent les contraintes de validité de ces analyses, c'est-à-dire qui possèdent un nombre suffisant d'occurrences dans le corpus. Ces quinze phonèmes ou groupes de phonèmes sont présentés dans la table B.15. La première analyse recherche un axe discriminant entre chaque couple de locuteurs dans un espace de 12 paramètres qui sont des énergies calculées à partir du modèle d'oreille. La seconde le fait dans un espace de 14 paramètres qui sont issus de l'étude analytique :

- pour les phonèmes sourds :
  - la présence et la position des formants,
  - les énergies dans des bandes de fréquences qui ne correspondent pas à des zones de résonance,
  - les paramètres prosodiques, durée, énergie et F<sub>0</sub>,
  - le pourcentage de prélèvements fricatifs,
  - les durées d'établissement de F0 et de F1 ;
- pour les phonèmes sourds, les durées d'amortissement de F0 et de F1.

/ v + z + ʒ /	/ b + g /	/ m /	/ n /	/ m + p /	/ m + ŋ /	/ n + p /	/ ŋ /
/ a /	/ æ /	/ ā /	/ ě + œ /	/ ɔ̃ /	/ ɔ̃ + œ /	/ ɔ̃ + ě /	

Table B.15. Les quinze phonèmes ou ensembles de phonèmes présélectionnés pour les analyses discriminantes.

Un coefficient appelé "pouvoir discriminant" compris entre 0 et 1 est associé à chaque axe qui discrimine un couple de locuteurs. Plus ce pouvoir discriminant est proche de 1, plus les deux locuteurs sont discriminés. Nous avons regroupé dans la table B.16 les résultats obtenus par les auteurs pour les deux analyses discriminantes. Dans cette table apparaissent les phonèmes qui obtiennent les meilleurs pouvoirs discriminants pour le maximum de couples de locuteurs.

	12 canaux d'énergie		14 paramètres spécifiques	
pouvoir discriminant supérieur à 0.9	phonèmes	nombre de couples discriminés sur 10	phonèmes	nombre de couples discriminés sur 10
	/ ɔ̃ /	7 couples	/ ɔ̃ /	5 couples
	/ ā /	6 couples		
pouvoir discriminant supérieur à 0.8	/ ɔ̃ /	10 couples	/ m /	9 couples
	/ ẽ + œ /	10 couples	/ m + p /	9 couples
	/ n /	10 couples	/ ɔ̃ /	8 couples
	/ ā /	9 couples	/ n /	7 couples
	/ n + p /	9 couples	/ ẽ + œ /	5 couples
	/ m /	9 couples		
	/ m + p /	9 couples		

Table B.16. Résultats des deux analyses discriminantes sur les 15 "phonèmes" présélectionnés dans l'étude de Pérennou et al.

Ces résultats suscitent plusieurs remarques :

- afin de respecter les contraintes de validité de l'analyse discriminante, chacun des quinze "phonèmes" a été étudié tous contextes confondus ;
- les douze énergies issues du modèle d'oreille ont un bien meilleur pouvoir discriminant que les paramètres spécifiques issus de l'étude analytique. Ceci peut provenir de la difficulté de déterminer les formants des phonèmes choisis comme les voyelles et consonnes nasales. Par ailleurs, les valeurs segmentales spécifiques des paramètres prosodiques ne fournissaient déjà pas de bons résultats dans l'étude analytique ;
- les allophones du / a / n'obtiennent pas les performances espérées à la suite de l'étude analytique. Il est possible que cela provienne de la prise en compte des occurrences du / a / dans tous les contextes ;
- en revanche, les voyelles nasales / ẽ / et / œ / obtiennent de meilleurs résultats que dans l'étude analytique ;
- des phonèmes performants dans la première étude, comme les occlusives vélaires ou la coarticulation du couple / tɪ /, n'ont pas pu être traités par l'analyse discriminante faute d'un nombre suffisant de répétitions ;
- les auteurs ajoutent que les phonèmes comme / l / et / R / ne sont pas pertinents pour discriminer les cinq locuteurs.



**Daudin et al. [Daudin 89].** Cette étude constitue en fait une préétude des variabilités intralocuteur et interlocuteur, utile à la mise en place d'un système de vérification du locuteur. La base de données utilisée comprend les élocutions de 6 mots différents par 10 locuteurs (5 H et 5 F), répétés 30 fois sur trois sessions espacées de quelques jours et enregistrés sur bande audio puis numérisés à l'aide de la carte OROS à 16 kHz sur 16 bits. Les six mots utilisés sont : *permission*, *télécom*, *identité*, *entrer*, *prononcer* et *essai*. Une analyse LPC d'ordre 18 issue du logiciel de traitement de signal ILS permet d'extraire les paramètres suivants : les trois premiers formants, les six premiers coefficients PARCOR, les six premiers coefficients de fonction d'aire, les six premiers coefficients spectraux, soit un vecteur de 21 paramètres (il est à noter que tous ces paramètres ne sont pas indépendants). Les caractéristiques des spectres à long terme pour chaque mot enregistré sont aussi retenues. Après une première analyse en composantes principales des paramètres par classes, les auteurs combinent les différentes classes de paramètres entre elles. Les conclusions de l'étude portent sur l'intérêt de combiner des coefficients entre eux, sur le bon pouvoir de discrimination certains mots par rapport à d'autres, "*permission*" étant le plus discriminant et "*entrer*" le moins discriminant. En revanche, aucun résultat n'est fourni dans l'article sur la mesure des formants (méthode, emplacement de calcul) et sur leur pouvoir de discrimination par rapport autres paramètres.

**Bonastre et Méloni [Bonastre 92]** Contrairement aux études précédentes, l'objet de cette étude n'est pas directement la recherche de paramètres acoustico-phonétiques caractérisant le locuteur mais l'influence de la prise en compte du contexte phonologique bilatéral d'un phonème sur la performance de l'identification du locuteur. Pour chacun des phonèmes valides du corpus "la bise et le soleil" de BDSONS, prononcé une seule fois par 22 locuteurs, les auteurs ont calculé trois paramètres en contexte et hors contexte :

- le taux d'identification avec une seule occurrence du phonème,
- le rapport de l'écart-type interlocuteur à l'écart-type intralocuteur,
- le nombre d'occurrences du même phonème nécessaire à l'obtention d'une identification avec un coefficient de sécurité de 95,5%.

Pour comparer les locuteurs, les auteurs utilisent une méthode de quantification vectorielle (même corpus pour l'apprentissage et la reconnaissance) dans laquelle les auteurs combinent plusieurs distances associées à plusieurs méthodes d'analyse spectrale. Les spectres sont calculés toutes les 10 ms par trois analyses, une transformée de Fourier, une analyse LPC d'ordre 14 et une analyse LPC d'ordre 21. Ces spectres sont codés selon 128 canaux linéairement distribués ou selon 24 canaux répartis selon une échelle Mel. Les distances entre les locuteurs prennent en compte soit le spectre soit la dérivée du spectre soit encore le signe de cette dérivée.

Les auteurs concluent que les performances d'identification sont nettement meilleures lorsqu'ils comparent les locuteurs en tenant compte du contexte du phonème : le taux d'identification avec une seule occurrence est multiplié par 2,5 et le nombre d'occurrences nécessaire à l'obtention d'une bonne identification est divisé par 5.

L'observation du tableau situé dans la publication et qui regroupe les valeurs des trois paramètres pour chacun des phonèmes suscite quelques conclusions supplémentaires.

Tout d'abord, cette étude met encore une fois en évidence l'insuffisance du rapport de la variabilité intralocuteur à la variabilité interlocuteur (équivalent au F-ratio) pour évaluer la pertinence d'un phonème. En effet, il prend les mêmes valeurs pour des phonèmes pertinents et des phonèmes non pertinents au sens du pourcentage d'identification.



Par ailleurs, / **m** / est le phonème qui obtient le meilleur taux d'identification lorsque les phonèmes sont utilisés indépendamment de leur contexte. Ce résultat peut conduire à deux interprétations opposées. Selon la première, le spectre de / **m** / serait invariant avec le contexte phonologique. D'après la seconde, il varierait au contraire avec le contexte mais cette variabilité intralocuteur serait moins importante que les différences entre les locuteurs dans la manière de réaliser cette influence contextuelle. Les connaissances sur la coarticulation linguale des consonnes bilabiales (cf. paragraphe V.2 de la partie A) et les conclusions de l'étude de Su et al. (cf. page 46) valident la deuxième interprétation.

Néanmoins, la comparaison des performances des phonèmes, notamment au niveau du nombre d'occurrences nécessaires à une bonne identification, est délicate car nous ne connaissons ni le nombre ni le type de contextes (ou plutôt de classes de contexte) considérés pour chacun des phonèmes.

Les phonèmes en contexte qui obtiennent les meilleurs résultats pour l'identification sont dans l'ordre / **j** /, soit / **o** / soit / **ɔ** / (incertitude due à la notation), les voyelles nasales, / **m** /, / **z** / et / **ɛ** /. Mais, il faudrait examiner le corpus pour connaître les contextes qui ont contribué à cette pertinence. A part / **m** /, tous ces phonèmes engendrent de très mauvais taux de reconnaissance lorsqu'ils sont considérés tous contextes confondus

## 4. Conclusion

Nous avons exposé dans ce chapitre les résultats de la majeure partie des études dont l'objet est l'extraction du signal de parole de paramètres linguistiques et plus particulièrement acoustiques et phonétiques caractérisant au mieux le locuteur. Malheureusement, la plupart de ces études portent sur la langue anglaise. Il est donc très difficile de déduire de ces paramètres pertinents les sources de différences entre locuteurs qui pourront être applicables à la langue française et qui permettront d'émettre des hypothèses sur la caractérisation du locuteur français (la caractérisation du locuteur francophone étant encore un projet encore plus vaste).

Nous pouvons remarquer que la majorité des études anglaises ou américaines étudient systématiquement un sous-ensemble de paramètres très spécifiques [Atal 72], [Su 74], [Goldstein 76], [Paliwal 84], [Nolan 83]. En revanche, si nous considérons les études sur le français, seule celle de P. Corsi [Corsi 79] suit cette optique ; les études de Pérennou et al. [Pérennou 82] et Daudin et al. [Daudin 89], bien que n'ayant pas la même ampleur, sont à notre avis plutôt des études de faisabilité. Quant à l'étude de Bonastre et Méloni [Bonastre 92], son objet n'est pas, comme nous l'avons déjà indiqué, la recherche de paramètres pertinents.

Comme nous allons le voir dans la troisième partie de cette thèse, notre étude se rapproche plus des études anglaises avec l'analyse systématique de la pertinence de sous-ensembles de paramètres dans des contextes bien définis nécessitant l'élaboration d'un corpus spécifique.



## BIBLIOGRAPHIE





# BIBLIOGRAPHIE

## Abréviations employées dans la bibliographie

- **JASA** : Journal of the Acoustical Society of America,
- **ICASSP** : International Conference on Acoustics, Speech and Signal Processing,
- **JEP** : Journées d'Etude sur la Parole,
- **EUROSPEECH** : European Conference on Speech Communication and Technology,
- **ASSP** : Acoustics, Speech and Signal Processing,
- **ICSLP** : International Conference on Spoken Language Processing.

- [Atal 72] B.S. Atal.  
Automatic speaker recognition based on pitch contours.  
*JASA*, 52(6):1687–1697, 1972.
- [Atal 76] B.S. Atal.  
Automatic Recognition of Speakers from Their Voices.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 64:349–363, April 1976.
- [Attili 88] J.B. Attili, M. Savic et Jr J.P. Campbell.  
A TMS32020-based Real Time, text-independent, Automatic Speaker Verification System.  
*ICASSP, New-York, U.S.A*, pages 599–602, 1988.
- [Baker 75] J.K. Baker.  
The DRAGON System: an overview.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 23:24–29, 1975.
- [Bennani 90] Y. Bennani, F. Fogelman-Soulié et P. Gallinari.  
A connectionism approach for automatic speaker identification.  
*ICASSP, Albuquerque, U.S.A*, pages 265–268, 1990.
- [Bennani 91] Y. Bennani et P. Gallinari.  
On the use of TDNN-extracted features information in talker identification.  
*ICASSP, Toronto, Canada*, pages 385–388, 1991.
- [Bennani 92] Y. Bennani et P. Gallinari.  
Une architecture connexionniste modulaire pour l'identification automatique du locuteur.  
*19e JEP*, pages 577–582, mai 1992.
- [Bolt 70] R.H. Bolt, F.S. Cooper, E.E. David, P.B. Denes, J.M. Pickett et K.N. Stevens.  
Speaker Identification by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes.  
*JASA*, 47(2):597–612, 1970.

- [Bonastre 92] J.F. Bonastre et H. Méloni.  
Etude de la variabilité spectrale pour la caractérisation du locuteur.  
*19e JEP*, pages 555–559, Bruxelles (Belgique), mai 1992. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Bourlard 89] H. Bourlard et C.J. Wellekens.  
Speech Pattern Discrimination and Multilayer Perceptrons.  
*Computer, Speech and Language*, 3:1–19, 1989.
- [Bridle 90] J.S. Bridle.  
Alpha-Nets: A Recurrent "Neural" Network Architecture with a Hidden Markov Model Interpretation.  
*Speech Communication*, 9:83–92, February 1990.
- [Bridle 91] J.S. Bridle et L. Dodd.  
An alphanet approach to optimising input transformations for continuous speech recognition.  
*ICASSP, Toronto, Canada*, 1991.
- [Buck 85] J.T. Buck, D.K. Burton et J.E. Shore.  
Text-dependent speaker recognition using vector quantization.  
*ICASSP, Tampa, U.S.A*, pages 391–394, 1985.
- [Burton 85a] D.K. Burton.  
Applying matrix quantization to isolated-word recognition.  
*ICASSP, Tampa, U.S.A*, pages 29–32, 1985.
- [Burton 85b] D.K. Burton, J.E. Shore et J.T. Buck.  
Isolated-word speech recognition using vector quantization codebooks.  
*I.E.E.E. Transactions on ASSP*, pages 837–849, August 1985.
- [Burton 87] D.K. Burton.  
Text-independent speaker verification using vector quantization source coding.  
*I.E.E.E. Transactions on ASSP*, pages 133–143, February 1987.
- [Caelen-Haumont 88] G. Caelen-Haumont et G. Pérennou.  
A synopsis of speaker-recognition work done in France. Phonetico-phonological, prosodic and frequential analyses in a both global and local approach to speaker identification and verification.  
*Bulletin du Laboratoire de la Communication Parlée de Grenoble (France)*, 2:425–455, 1988.
- [Carey 90] M.J. Carey, E.S. Parris et J.S. Bridle.  
A speaker verification system using alpha-nets.  
*ICASSP, Toronto, Canada*, pages 397–400, 1990.
- [Carre 84] R. Carre, J. Descout, J.J. Mariani, M. Eskénazi et M. Rossi.  
The French Language Database. Defining, Planning and Recording a Large Database.  
*ICASSP, San-Diego, U.S.A*, pages 42.10.1–42.10.4, 1984.
- [Chi-Shi 90] L. Chi-Shi, W. Wern-Jun, L. Min-Tau et W. Hsiao-Chuan.  
Study of line spectrum pair frequencies for speaker recognition.  
*ICASSP, Albuquerque, U.S.A*, pages 277–280, April 1990.



- [Cohen 89] A. Cohen et I. Froind.  
On Text-independent Speaker Identification Using a Quadratic Classifier with Optimal Features.  
*Speech Communication*, volume 8, pages 35–44, 1989.
- [Corsi 79] P. Corsi.  
Reconnaissance automatique du locuteur : présentation générale, méthodologies et expérimentation, perspectives d'application.  
Thèse de Docteur Ingénieur de l'Institut National Polytechnique de Grenoble, 1979.
- [Das 71] S.K. Das et W.S. Mohn.  
A Scheme for Speech Processing in Automatic Speaker Verification.  
*I.E.E.E. Transactions on Audio and Electroacoustic*, 19:32–43, March 1971.
- [Daudin 89] E. Daudin, Y. Bennani et G. Chollet.  
Simulation de techniques de vérification automatique du locuteur.  
*Luminy*, pages 200–203, Marseille Luminy, juin 1989. GRECO PRC Communication Homme-Machine.
- [Doddington 83] G.R. Doddington.  
Voice authentication gets the go-ahead for security systems.  
*Speech Technology*, pages 14–23, September 1983.
- [Doddington 85] G.R. Doddington.  
Speaker recognition - Identifying people by their voices.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 73(11):1651–1664, November 1985.
- [Ephraim 87] Y. Ephraim, A. Dembo et L.R. Rabiner.  
A minimum discrimination information approach for hidden Markov modeling.  
*ICASSP, Dallas, U.S.A*, pages 25–28, 1987.
- [Eskenazi 88] M. Eskenazi, F. Lonchamp et J. Vaissière.  
Cours sur les Indices Acoustiques du Français.  
GRECO - Communication Parlée, octobre 1988.
- [Feix 85] W. Feix et M. DeGeorge.  
A Speaker Verification System for Access-Control.  
*ICASSP, Tampa, U.S.A*, pages 399–402, 1985.
- [Furui 86] S. Furui.  
Research on Individuality Features in Speech Waves and Automatic Speaker Recognition Techniques.  
*Speech Communication*, 5:183–197, 1986.
- [Fussel 91] J.W. Fussel.  
Automatic Sex Identification from Short Segments of Speech.  
*ICASSP, Toronto, Canada*, pages 409–412, 1991.
- [Gaganelis 91] D.A. Gaganelis et E.D. Frangoulis.  
A Novel Approach to Speaker Verification.  
*ICASSP, Toronto, Canada*, pages 373–376, 1991.

- [Goldstein 76] U.G. Goldstein.  
Speaker-identifying features based on formant tracks.  
*JASA*, 59(1):176–182, 1976.
- [Gong 90] Y. Gong et J.-P. Haton.  
Text-independent Speaker Recognition by Trajectory Space Comparison.  
*ICASSP, Albuquerque, U.S.A*, volume 1, pages 285–288, April 1990.
- [Gray 84] R.M. Gray.  
Vector quantization.  
*I.E.E.E. Magazine on ASSP*, pages 4–28, April 1984.
- [Haton 91] J.P. Haton, J.M. Pierrel, G. Pérennou, J. Caelen et J.L. Gauvain.  
*Reconnaissance Automatique de la Parole*.  
Informatique. Dunod, Paris, 1991.
- [Hermansky 90] H. Hermansky.  
Perceptual Linear Predictive (PLP) analysis of speech.  
*JASA*, 87, 1990.
- [Higgins 91] A.L. Higgins et L.G. Bahler.  
Text-independent Speaker Verification by Discriminator Counting.  
*ICASSP, Toronto, Canada*, pages 405–408, 1991.
- [Huang 88] W. Huang, R.P. Lippmann et B. Gold.  
A Neural Net Approach to Speech Recognition.  
*ICASSP, New-York, U.S.A*, pages 99–102, 1988.
- [Hunt 82] M.J. Hunt.  
Work on Automatic Talker Recognition at JSRU between 1972 and 1978 and  
comments on more recent work elsewhere.  
Rapport, JSRU, 1982.
- [Itakura 75a] F. Itakura.  
Line spectrum representation of line predictive coefficients of speech signals.  
*JASA*, 57, 1975.
- [Itakura 75b] F. Itakura.  
Minimum prediction residual principle applied to speech recognition.  
*I.E.E.E. Transactions on ASSP*, pages 67–72, 1975.
- [Jelinek 76] F. Jelinek.  
Continuous speech recognition by statistical methods.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 64:532–  
556, 1976.
- [Jodouin 90] J-F. Jodouin.  
Présentation des modèles connexionnistes.  
*Intellectica*, 9-10:9–39, 1990.
- [Kohonen 88] T. Kohonen.  
*Self-Organization and Associative Memory*.  
Springer-Verlag, New-York, 2nd edition édition, 1988.

- [Ladefoged 75] P. Ladefoged.  
*A Course in Phonetics*.  
Harcourt Brace Jovanovich, New-York, 1975.
- [Levinson 86] S.E. Levinson.  
Continuously variable duration hidden Markov models for automatic speech recognition.  
*Computer Speech and Language*, 1:29–45, 1986.
- [Li 83] K.P. Li et Jr E.H. Wrench.  
An Approach to Text-independent Speaker Recognition with Short Utterances.  
*ICASSP, Boston, U.S.A*, pages 555–558, 1983.
- [Li 88] K-P. Li et J.E. Porter.  
Normalizations and Selection of Speech Segments for Speaker Recognition Scoring.  
*ICASSP, New-York, U.S.A*, pages 595–598, 1988.
- [Linde 80] Y. Linde, A. Buzo et R.M. Gray.  
An algorithm for vector quantization design.  
*I.E.E.E. Transactions on ASSP*, pages 84–95, January 1980.
- [Lippmann 87] R.P. Lippmann et B. Gold.  
Neural Classifiers Useful for Speech Recognition.  
*1st International Conference on Neural Network*. IEEE, June 1987.
- [Lippmann 88] R.P. Lippmann.  
Neural Nets for Computing.  
*ICASSP, New-York, U.S.A*, pages 1–6, 1988.
- [Lowerre 76] B.T. Lowerre.  
*The HARPY Speech Recognition System*.  
Thèse de Doctorat, CMU- CSD, Carnegie-Mellon University, 1976.
- [Markov 13] A.A. Markov.  
An example of statistical investigation in the text of "Eugene Onyegin" illustrating coupling of "tests in chains".  
*Proc. Acad. Scie. St. Petesbourg*, volume VI, pages 153–162, 1913.
- [Matsui 90] T. Matsui et S. Furui.  
Text-independent speaker recognition using vocal tract and pitch information.  
*ICSLP, Kobe, Japan*, pages 137–140, 1990.
- [Matsui 91] T. Matsui et S. Furui.  
A text-independent speaker recognition method robust against utterance variations.  
*ICASSP, Toronto, Canada*, pages 377–380, 1991.
- [McDermott 89] E. McDermott et S. Katagiri.  
Shift-invariant, Multi-category Phoneme Recognition Using Kohonen's LVQ2.  
*icassp89*, 1989.
- [Montacie 92] C. Montacie, J-L. Le Floch et X. Rodet.  
Modèles autorégressifs vectoriels et reconnaissance du locuteur.  
*19e JEP*, pages 439–443, mai 1992.



- [Naik 89] J.M. Naik, L.P. Netsch et G.R. Doddington.  
Speaker verification over long distance telephone lines.  
*ICASSP, Glasgow, Scotland*, pages 524–527, 1989.
- [Nakasone 88] H. Nakasone et C. Melvin.  
Computer Assisted Voice Identification System.  
*ICASSP, New-York, U.S.A*, pages 587–590, 1988.
- [Noda 90] H. Noda et M. Yanagida.  
Extraction of phoneme-dependent individuality using HMM-based segmentation for text-independent speaker recognition.  
*ICSLP, Kobe, Japan*, pages 129–132, November 1990.
- [Nolan 83] F. Nolan.  
*The Phonetic Bases of Speaker Recognition*.  
Cambridge University Press, Great Britain, 1983.
- [Oglesby 90] J. Oglesby et J.S. Mason.  
Optimisation of Neural Models for Speaker Identification.  
*ICASSP, Albuquerque, U.S.A*, pages 261–264, 1990.
- [Oglesby 91] J. Oglesby et J.S. Mason.  
Radial Basis Function Networks for Speaker Recognition.  
*ICASSP, Toronto, Canada*, pages 393–396, 1991.
- [O'Shaughnessy 87] D. O'Shaughnessy.  
*Speech Communication: Human and Machine*.  
Addison Wesley, Reading, Massachusetts, 1987.
- [Paliwal 84] K.K. Paliwal.  
Effectiveness of different vowels sounds in automatic speaker identification.  
*Journal of Phonetics*, 12:17–21, 1984.
- [Paul 75] J.E. Paul, A.S. Rabinowitz, J.P. Riganati et J.M. Richardson.  
Development of analytical methods for a semi-automatic speaker identification system.  
*Carnahan Conference on Crime Countermeasures*, pages 52–64, 1975.
- [Perennou 82] G. Perennou, G. Caelen, N. Vigouroux et M. Bréant.  
Identification et vérification du locuteur, timbre de la voix.  
Rapport final de la convention DRET numéro 79.34.658.00.470.7501, décembre 1982.
- [Poritz 82] A.B. Poritz.  
Linear predictive hidden Markov model and the speech signal.  
*ICASSP, Paris, France*, pages 1291–1294, 1982.
- [Ren-hua 90] W. Ren-hua, H. Lin-shen et H. Fujisaki.  
A weighted distance measure based on the fine structure of feature space: Application to speaker recognition.  
*ICASSP, Albuquerque, U.S.A*, pages 273–276, April 1990.
- [Rose 91] R.C. Rose, J. Fitzmaurice, E.M. Hofstetter et D.A. Reynolds.  
Robust Speaker Identification in Noisy Environments Using Noise Adaptive Speaker Models.  
*ICASSP, Toronto, Canada*, pages 401–404, 1991.

- [Rosenberg 76] A.E. Rosenberg.  
Automatic Speaker Verification: A Review.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 64:336–348, April 1976.
- [Rosenberg 86] A.E. Rosenberg et F.K. Soong.  
Evaluation of a vector quantization talker recognition system in text independent and text dependent modes.  
*ICASSP, Tokyo, Japan*, pages 873–876, 1986.
- [Rosenberg 90a] A.E. Rosenberg, C. Lee et M.A. McGee.  
Experiments in automatic talker verification using sub-word unit hidden Markov models.  
*ICSLP, Kobe, Japan*, pages 141–144, 1990.
- [Rosenberg 90b] A.E. Rosenberg, C. Lee et F.K. Soong.  
Sub-word unit talker verification using hidden Markov models.  
*ICASSP, Albuquerque, U.S.A*, pages 269–272, April 1990.
- [Rosenberg 91] A.E. Rosenberg, C. Lee et S. Gokcen.  
Connected word talker verification using whole word hidden Markov models.  
*ICASSP, Toronto, Canada*, pages 381–384, 1991.
- [Rudasi 91] L. Rudasi et S.A. Zahorian.  
Text-independent Talker Identification with Neural Networks.  
*ICASSP, Toronto, Canada*, pages 389–392, 1991.
- [Sakoe 78] H. Sakoe et S. Shiba.  
Dynamic Programming Algorithm for Spoken Word Recognition.  
*Proceedings of the Institute of Electrical and Electronic Engineers*, 36(1):43–49, February 1978.
- [Sambur 75] M.R. Sambur.  
Selection of acoustic features for speaker identification.  
*I.E.E.E. Transactions on ASSP*, pages 176–182, April 1975.
- [Sambur 76] M.R. Sambur.  
Speaker recognition using orthogonal linear prediction.  
*IEEE*, 24:283–289, August 1976.
- [Savic 90] M. Savic et S.K. Gupta.  
Variable parameter speaker verification system based on hidden Markov modeling.  
*ICASSP, Albuquerque, U.S.A*, pages 281–284, April 1990.
- [Shannon 48] C.C. Shannon.  
A mathematical theory of communications.  
*Bell Sys. Tech. Journal*, 27:379–423 623–656, 1948.
- [Shikano 85] K. Shikano.  
Text-independent speaker recognition experiments using codebooks in vector quantization.  
Private communication in 109th Meeting of the Acoustical Society of America, April 1985.

- [Shridhar 83] M. Shridhar, N. Mohankrishnan et M.A. Sid-Ahmed.  
A comparison of Distance Measures for Text-Independent Speaker Identification.  
*ICASSP, Boston, U.S.A*, pages 559–562, 1983.
- [Soong 85] F.K. Soong, A.E. Rosenberg, L.R. Rabiner et B.H. Juang.  
A vector quantization approach to speaker recognition.  
*ICASSP, Tampa, U.S.A*, pages 387–390, 1985.
- [Soong 86] F.K. Soong et A.E. Rosenberg.  
On the use of instantaneous and transitional spectral information in speaker recognition.  
*ICASSP, Tokyo, Japan*, pages 877–880, 1986.
- [Su 74] L.S. Su, K.P. Li et K.S. Fu.  
Identification of speakers by use of nasal coarticulation.  
*JASA*, 56(6):1876–1882, 1974.
- [Thevenaz 90] P. Thevenaz et H. Hügli.  
Combining four text-independent speaker recognition methods.  
*Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pages 187–191, Edinburgh, June 1990. European Speech Communication Association.
- [Thomas 90] S.J. Thomas et J.W. Fussell.  
Backpropagation VS. Gaussian Classifiers for Speaker Sex Identification.  
*3rd Annual North Carolina Symposium on Artificial Intelligence*, November 1990.
- [Tishby 88] N. Tishby.  
Information theoretic factorization of speaker and language in hidden Markov models, with application to speaker recognition.  
*ICASSP, New-York, U.S.A*, pages 87–90, 1988.
- [Tishby 91] N. Tishby.  
On the application of mixture AR hidden Markov models to text independent speaker recognition.  
*ieeet*, 39(3):563–570, 1991.
- [Vloeberghs 92] C. Vloeberghs et P. Dupont.  
La reconnaissance du locuteur basée sur des modèles de Markov cachés de phonèmes.  
*19e JEP*, pages 543–548, Bruxelles (Belgique), mai 1992. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Wells 88] J.C. Wells, W. Barry et A.J. Fourcin.  
Methods of Transcription, Labelling and Normative Reference.  
Extract of the definition phase final report of sam project, Projet ESPRIT : Assessment, Methodology and Standardisation in Multilingual Speech Technology, January 1988.
- [Wilbur 88] J. Wilbur et F.J. Taylor.  
Consistent Speaker Identification via Wigner Smoothing Techniques.  
*ICASSP, New-York, U.S.A*, pages 591–594, 1988.



- [Wolf 72] J.J. Wolf.  
Efficient acoustic parameters for speaker recognition.  
*JASA*, 51(6):2044–2056, 1972.
- [Wolf 83] J. Wolf, M. Krasner, K. Karnofsky, R. Schwartz et S. Roucos.  
Further Investigation of Probabilistic Methods for Speaker Identification.  
*ICASSP, Boston, U.S.A*, pages 551–554, 1983.
- [Xu 89a] L. Xu et J.S. Mason.  
Instantaneous and transitional perceptually-based features in speaker identification.  
*EUROSPEECH, Paris, France*, pages 271–274, 1989.
- [Xu 89b] L. Xu, J. Oglesby et J.S. Mason.  
The optimization of perceptually-based features for speaker identification.  
*ICASSP, Glasgow, Scotland*, pages 520–523, 1989.
- [Xu 91] L. Xu et J.S. Mason.  
Optimization of perceptually-based spectral transforms in speaker identification.  
*EUROSPEECH, Genova, Italy*, pages 439–442, 1991.
- [Zheng 88] Y. Zheng et B. Yuan.  
Text-dependent identification using Circular hidden Markov models.  
*ICASSP, New-York, U.S.A*, pages 580–586, 1988.
- [Zue 88] V.W. Zue et S. Seneff.  
Transcription and alignment of the TIMIT Database.  
*Second Symposium on Advanced Man-Machine Interface through Spoken Language*, Oahu, November 1988.



## **PARTIE C**

# **NOTRE CONTRIBUTION A LA CARACTERISATION DU LOCUTEUR**





## TABLE DES MATIERES DE LA PARTIE C

INTRODUCTION	1
<b>I SELECTION DES PARAMETRES SUSCEPTIBLES DE CARACTERISER LE LOCUTEUR</b>	3
I.1 Introduction	3
I.2 Différences physiologiques et habitudes articulatoires	3
I.2.1 Les paramètres fréquentiels	3
I.2.2 Les paramètres temporels	10
I.3 Différences d'origine linguistique et phonologique	11
I.4 Conclusion	12
<b>II ELABORATION ET ETIQUETAGE DU CORPUS</b>	13
II.1 Introduction	13
II.2 Elaboration du corpus	13
II.2.1 Construction des phrases	13
II.2.2 Enregistrement et numérisation du corpus	13
II.2.2.1 Enregistrement	13
II.2.2.2 Première numérisation	14
II.2.2.3 Deuxième numérisation	14
II.3 Généralités sur l'étiquetage	15
II.3.1 Introduction	15
II.3.2 La transcription	16
II.3.2.1 La transcription orthographique	16
II.3.2.2 Les transcriptions phonologiques	16
II.3.2.3 La transcription phonétique	16
II.3.2.4 La transcription acoustique	17
II.3.3 L'alignement	18
II.3.3.1 Alignement manuel et segmentation	18
II.3.3.2 Alignement semi-automatique	20
II.3.4 Conclusion	20
II.4 Etiquetage du corpus	21
II.4.1 Introduction	21
II.4.2 La transcription	23
II.4.3 Les critères de segmentation	25
II.4.3.1 Les voyelles orales	25
II.4.3.2 Les occlusives	27
II.4.3.3 Le phonème / r / en contexte vocalique	29
II.4.3.4 Le phonème / l /	31

II.4.3.5	Les voyelles nasales . . . . .	31
II.4.3.6	Les semi-voyelles et les couples de voyelles . . . . .	33
II.4.3.7	Le schwa bref et le schwa épenthétique . . . . .	33
II.4.3.8	Les phénomènes de “voix craquée” et de “friture vocale” . . . . .	33
II.4.4	Conclusion . . . . .	34
II.5	<b>Conclusion sur l'élaboration et l'étiquetage du corpus</b> . . . . .	34
III	<b>PERTINENCE DES TROIS PREMIERS FORMANTS DES VOYELLES ORALES</b> . . . . .	37
III.1	<b>Introduction</b> . . . . .	37
III.2	<b>Détermination des trois premiers formants des voyelles orales</b> . . . . .	37
III.2.1	Introduction . . . . .	37
III.2.2	Etude préliminaire . . . . .	39
III.2.2.1	Méthodologie de détermination des formants de la voyelle . . . . .	39
III.2.2.2	Vérification des fréquences formantiques . . . . .	41
III.2.2.3	Conclusions de l'étape de vérification . . . . .	42
III.2.3	Affectation des pôles LPC aux formants intermédiaires F1, F2 et F3 . . . . .	45
III.2.4	Détermination des formants finaux F1, F2 et F3 . . . . .	48
III.2.4.1	Détermination des formants intermédiaires à trois emplacements de la voyelle . . . . .	48
III.2.4.2	Détermination des formants finaux . . . . .	50
III.2.4.3	Analyse de la robustesse des formants finaux . . . . .	52
III.2.4.4	Vérification des formants finaux . . . . .	61
III.2.5	Conclusion . . . . .	62
III.3	<b>Etude de la pertinence des trois premiers formants des voyelles orales</b> . . . . .	63
III.3.1	Méthodologie d'étude de la pertinence . . . . .	63
III.3.1.1	Les combinaisons formantiques étudiées . . . . .	63
III.3.1.2	Méthodes de reconnaissance d'un locuteur . . . . .	64
III.3.1.3	Les indicateurs de pertinence . . . . .	65
III.3.2	Résultats et commentaires de l'étude de pertinence . . . . .	66
III.3.2.1	Introduction . . . . .	66
III.3.2.2	Comparaison des deux variantes de la reconnaissance d'un locuteur . . . . .	66
III.3.2.3	Les combinaisons (F1), (F2) et (F3) . . . . .	68
III.3.2.4	Les combinaisons (F2, F3), (F1, F2) et (F1, F3) . . . . .	71
III.3.2.5	La combinaison (F1, F2, F3) . . . . .	75
III.3.2.6	Les combinaisons d'écarts de formants . . . . .	77
III.4	<b>Conclusion sur la pertinence des trois premiers formants des voyelles   orales</b> . . . . .	79
III.4.1	Quelques conclusions sur la pertinence des voyelles . . . . .	79
III.4.2	Quelques réflexions sur la méthodologie d'étude . . . . .	80
	<b>CONCLUSIONS ET PERSPECTIVES</b> . . . . .	83
	<b>BIBLIOGRAPHIE</b> . . . . .	87
	<b>ANNEXE</b> . . . . .	A.1



## Liste des figures

Figure C.1	Spectrogrammes des chaînes [ bVb ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée . . . . .	5
Figure C.2	Spectrogrammes des chaînes [ dVd ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée . . . . .	6
Figure C.3	Spectrogrammes des chaînes [ gVg ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée . . . . .	6
Figure C.4	Evolution des fréquences des formants et des antiformants en fonction du degré de couplage. $F'_1$ et $F'_2$ sont les fréquences formantiques d'origine orale. $F_N^1$ et $A_N^1$ sont respectivement la fréquence du premier formant nasal et celle du premier antiformant nasal ; d'après O. Fujimura dans . . . .	7
Figure C.5	Spectrogrammes des voyelles nasales [ % ], [ @ ] et [ * ] précédées de [ m ] et prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée . . . . .	8
Figure C.6	Spectrogrammes des chaînes [ V I V ] prononcées par un locuteur masculin, où V est une des voyelles [ i ], [ a ], [ u ], d'après les données du GRECO Communication Parlée . . . . .	9
Figure C.7	Spectrogrammes des chaînes [ VRV ] prononcées par un locuteur masculin, où V est une des voyelles [ i ], [ a ], [ u ], d'après les données du GRECO Communication Parlée . . . . .	10
Figure C.8	Les différents niveaux d'étiquetage dans la base de données japonaise du laboratoire ATR, d'après . . . . .	19
Figure C.9	Les informations disponibles sur l'écran lors de l'étiquetage de la phrase "Donne-moi le bocal de cacao !" prononcée par le locuteur jph : le signal temporel (a), le spectrogramme de parole couleur (b), le zoom temporel (c), le zoom spectrographique (d), le résultat de l'étiquetage (e). . . . .	22
Figure C.10	Exemple de voyelle se terminant par un souffle pour le locuteur jlc. . . . .	26
Figure C.11	Exemples de formants qui se prolongent pendant la tenue d'une occlusive sonore pour le locuteur gm. . . . .	28
Figure C.12	Exemple de bruit avant la barre d'explosion d'un [ t ] pour le locuteur df. . .	29
Figure C.13	Exemples de segmentation du triplet [ voyelle-R-voyelle ], du triplet [ p-voyelle-R ] et du [ l ] en début de phrase, dans la phrase "la porte du garage tomba avec lourdeur" prononcée par le locuteur jmp. . . . .	30
Figure C.14	Segmentation des voyelles nasales et du [ R ] en fin de phrase pour le locuteur aq. . . . .	32
Figure C.15	Le phénomène de voix craquée ou de friture vocale à la fin de la phrase " le bateau à vapeur a quitté le port" prononcée par le locuteur jg. . . . .	35
Figure C.16	Différentes réalisations du triplet [ piR ] de la phrase 11. . . . .	55
Figure C.17	Différentes réalisations du triplet [ puR ] de la phrase 08. . . . .	56
Figure C.18	Différentes réalisations du triplet [ buR ] de la phrase 16. . . . .	57
Figure C.19	Spectrogramme du triplet [ buR ] de la phrase 12. . . . .	58
Figure C.20	Spectrogrammes des réalisations des voyelles [ E_09 ] et [ E_07 ]. . . . .	58
Figure C.21	Spectrogrammes des réalisations des voyelles [ a_16 ] et [ a_03 ]. . . . .	59
Figure C.22	Spectrogrammes des réalisations des voyelles [ O_02 ] et [ O_04 ]. . . . .	60
Figure C.23	Influence des voyelles antérieures fermées [ e ] et [ i ] sur les formants du [ R ] dans le mot "péri". . . . .	61

Figure C.24	Domaines de variation de $F_2$ pour chacun des locuteurs pour les trois occurrences de la voyelle [ 0 ]. . . . .	71
-------------	--	----

## Liste des tables

Table C.1	Les dix-sept phrases du corpus. . . . .	14
Table C.2	Les locuteurs qui ont enregistré le corpus rangés dans l'ordre dans lequel ils ont été étudiés. . . . .	15
Table C.3	Les symboles utilisés dans l'étiquetage manuel réalisé avec le logiciel SNORRI. La correspondance avec les symboles de l'Alphabet Phonétique International est également fournie. . . . .	24
Table C.4	Les occurrences des triplets / p-voyelle-R/ et / b-voyelle-R / dans les dix-sept phrases du corpus. . . . .	38
Table C.5	Les domaines de définition D(Fi). . . . .	41
Table C.7	Comparaison des bornes prédéfinies des domaines de définition avec les valeurs minimales et maximales des fréquences des formants F1, F2 et F3 trouvés, ou de celles des pôles bruts lorsque les formants sont considérés comme faux par rapport au spectrogramme. Les valeurs en grisé sont en dehors de leur domaine de définition. Les fréquences représentées en caractères gras ne sont pas vérifiables sur le spectrogramme. . . . .	43
Table C.6	Domaines de définition définitifs et valeurs de référence des quatre premiers formants des voyelles orales pour les locuteurs masculins. . . . .	44
Table C.8	Comptabilisation du nombre de cas où un domaine de définition D(Fi) contient 2 pôles candidats (partie supérieure) et 3 pôles candidats (partie inférieure) lors de l'application de l'algorithme d'affectation à la détermination des 36 formants intermédiaires par locuteur et par voyelle. Les cases vides correspondent à des valeurs nulles. . . . .	49
Table C.9	Détermination d'un formant final à partir de trois formants intermédiaires. . . . .	51
Table C.10	Pourcentages de formants finaux très robustes ( $df_2 = 0$ ) et de formants finaux robustes ( $df_2 = 0$ ou $df_2 = f1$ ). . . . .	52
Table C.11	Effectifs des formants finaux selon leur niveau de robustesse. Le niveau de robustesse est donné par la valeur de $df_2$ , de 0 pour le formant le plus robuste à 5 pour le formant le moins robuste. Effectifs des formants finaux forcés à 0. Les cases vides correspondent à des valeurs nulles. . . . .	53
Table C.12	Nombre de formants finaux considérés comme incorrects lors de l'étape de vérification. Les voyelles absentes du tableau ne possèdent pas de formants jugés incorrects. . . . .	62
Table C.13	Nombre d'échanges des deux premiers locuteurs reconnus "réussis" et "ratés" sur un total de 1800 expériences de reconnaissance par combinaison étudiée et par mode de calcul de la distance. Le total des deux colonnes fournit le nombre d'échanges qui concernent le locuteur à reconnaître. . . . .	67
Table C.14	Pertinence de chacun des trois premiers formants des voyelles étudiées selon l'indicateur Taux pour les distances à pondération multiplicative et pour la distance à pondération perceptive. . . . .	69
Table C.15	Pertinence du couple (F2, F3) selon les indicateurs Taux et Score et pour tous les modes de calcul de distance. . . . .	72



Table C.16	Pertinence du couple (F1, F3) selon l'indicateur Taux pour tous les modes de calcul de distance. . . . .	74
Table C.17	Pertinence du couple (F1, F2) selon l'indicateur Taux pour tous les modes de calcul de distance. . . . .	75
Table C.18	Pertinence de la combinaison (F1, F2, F3) selon les indicateurs Taux et Score et pour tous les modes de calcul de distance. . . . .	76
Table C.19	Meilleures voyelles toutes combinaisons confondues, au sens de Taux, pour la distance euclidienne simple et la distance pondérée par le minimum des deux composantes. . . . .	78
Table C.20	Meilleures voyelles toutes combinaisons confondues, au sens de Taux, pour les distances pondérées par la valeur de référence, la largeur du domaine de définition et la distance perceptive. . . . .	78

## INTRODUCTION

Comme son nom l'indique, cette partie est entièrement consacrée à notre travail. Les trois chapitres qui la constituent suivent donc la progression de celui-ci.

Dans le premier chapitre, nous fournissons une description argumentée des paramètres acoustiques et phonétiques que nous avons sélectionnés en vue d'étudier leur pertinence pour l'identification automatique du locuteur.

Le second chapitre décrit l'élaboration et l'étiquetage du corpus de parole qui nous a permis d'étudier les paramètres sélectionnés. Dans la partie consacrée à l'étiquetage, nous présentons quelques généralités sur l'étiquetage de corpus de parole, avant d'exposer notre vision de sa problématique. Puis, nous détaillons comment nous avons réalisé l'étiquetage de notre corpus. Nous décrivons notamment les critères de segmentation que nous nous sommes définis afin d'obtenir un étiquetage homogène, de pallier certaines carences du logiciel d'étiquetage ainsi que notre manque d'expérience en matière d'étiquetage.

Le troisième chapitre est consacré à l'étude de la pertinence de certains des paramètres sélectionnés qui sont les trois premiers formants des voyelles orales /  $\epsilon$  /, /  $\text{œ}$  /, /  $\text{ɔ}$  /, /  $\text{a}$  /, /  $\text{i}$  /, /  $\text{e}$  / et /  $\text{u}$  / précédées des contextes neutres au sens de la coarticulation linguale /  $\text{p}$  / ou /  $\text{b}$  /, et suivies d'un contexte postérieur allongeant /  $\text{R}$  /. Les deux premiers sous-chapitres correspondent aux deux étapes de cette étude. Dans le premier d'entre eux, nous exposons notre méthodologie de détermination de valeurs robustes des trois premiers formants des voyelles et nous commentons les résultats obtenus. Dans le second, nous décrivons comment nous avons étudié la pertinence des combinaisons de formants et des écarts entre les formants à partir principalement de deux indicateurs issus d'expériences de reconnaissance d'un locuteur parmi dix. Enfin, dans le dernier sous-chapitre, nous développons quelques réflexions sur les résultats obtenus et sur notre méthodologie d'étude.





## **CHAPITRE I**

### **SELECTION DES PARAMETRES SUSCEPTIBLES DE CARACTERISER LE LOCUTEUR**

#### **1. Introduction**

Les industriels et les entreprises du secteur tertiaire attendent des systèmes automatiques de vérification ou d'identification du locuteur qu'ils soient fiables pour un grand nombre de locuteurs (plusieurs centaines) et dans n'importe quelles conditions (téléphone, variabilité intralocuteur, ...). Dans la deuxième partie de ce mémoire, nous avons proposé une classification des recherches actuelles en reconnaissance automatique du locuteur selon deux démarches. L'une, fondée sur des techniques mises au point pour la reconnaissance automatique de la parole (RAP), exploite la variabilité interlocuteur de manière implicite sans se soucier du contenu de l'acte de communication orale. Ainsi, même lorsque la reconnaissance est dépendante du texte, celui-ci a peu d'importance au niveau de son contenu. L'autre exploite les sources de variabilité interlocuteur de façon explicite en essayant d'extraire du signal de parole les paramètres qui caractérisent au mieux le locuteur. Nous pensons qu'un système uniquement conçu selon la première approche ne peut pas atteindre les objectifs souhaités par le monde industriel. La méthode de reconnaissance utilisée, même si elle doit dériver d'une méthode de RAP, devra être appliquée sur un énoncé construit à partir de paramètres pertinents au sens de la discrimination des locuteurs. Or, peu d'études ont été réalisées sur la caractérisation du locuteur de langue française (cf. chapitre III de la partie B). Aussi notre travail se situe-il dans cette optique. Pour cela, un ensemble de paramètres acoustiques, phonétiques et phonologiques, susceptibles d'être pertinents pour la reconnaissance du locuteur, a été sélectionné avec l'aide de F. Lonchamp, Professeur à l'Institut de Phonétique de Nancy. Nous allons présenter ces paramètres en les classant selon les sources de différences entre locuteurs qu'ils traduisent.

#### **2. Différences physiologiques et habitudes articulatoires**

Les paramètres qui suivent exploitent les différences anatomiques des conduits vocaux des locuteurs ainsi que leurs habitudes articulatoires aussi bien au niveau de la réalisation d'une cible articulatoire qu'au niveau des faits de coarticulation. Nous avons classé ces paramètres selon l'analyse qu'ils requièrent, spectrale ou temporelle.

##### **2.1. Les paramètres fréquentiels**

Les paramètres fréquentiels concernent essentiellement l'étude des voyelles et des consonnes à formants.

- Les fréquences formantiques des voyelles /  $\epsilon$  /, /  $\text{œ}$  /, /  $\text{ɔ}$  /, /  $\text{a}$  /, et /  $\text{i}$  /, précédées d'un contexte neutre au sens de la coarticulation linguale, /  $\text{p}$  /, et suivies d'un contexte postérieur allongeant, /  $\text{R}$  /.

Comme nous l'avons souligné dans le chapitre V de la partie A, les habitudes articulatoires des locuteurs et les différences anatomiques de leurs conduits vocaux se répercutent sur les fréquences formantiques des voyelles orales. D'autres exemples de la variabilité interlocuteur des voyelles orales sont fournis indirectement par des expériences de perception.

Lors d'une expérience de reconnaissance perceptive de voyelles anglaises, P. Ladefoged et D.E. Broadbent ont mis en évidence l'adaptation au locuteur réalisée par l'auditeur [Ladefoged 57]. Pour cela, les auteurs ont synthétisé six versions d'une phrase contenant les trois phonèmes /  $\text{i}$  /, /  $\text{a}$  / et /  $\text{w}$  /. Les deux premiers phonèmes et le début du /  $\text{w}$  / correspondent dans le plan ( $F_1$ ,  $F_2$ ) aux trois sommets du triangle acoustique des voyelles anglaises (cf. figure B.3). Chacune des versions représente une position différente du triangle acoustique dans ce plan. Dans un premier temps, les auditeurs ont identifié les six versions de la phrase comme provenant de locuteurs distincts mais appartenant au même groupe sociogéographique. Lors d'un deuxième test, les auditeurs ont reconnu une même voyelle de synthèse différemment selon la version de la phrase qui la précédait.

A l'aide d'un système d'analyse-synthèse fondé sur une analyse pitch-synchrone, H. Kuwabara et T. Takagi [Kuwabara 90] ont modifié la fréquence fondamentale et les fréquences formantiques de cinq voyelles japonaises prononcées par deux locuteurs. Trois auditeurs devaient reconnaître chacun des locuteurs à partir des cinq voyelles modifiées. Les auteurs ont déduit des résultats de ces expériences les conclusions suivantes. La manipulation de  $F_0$  influence moins la reconnaissance du locuteur que celle des fréquences formantiques. Un décalage de toutes les fréquences formantiques de plus de 8% ne permet plus aucune reconnaissance du locuteur alors qu'un décalage de moins de 2% n'affecte pas la reconnaissance. Les décalages de  $F_1$ ,  $F_2$  et  $F_3$  ont plus d'importance que ceux des fréquences des formants supérieurs.

Si les formants des voyelles orales anglaises ont fait l'objet de quelques études dans le cadre de la reconnaissance automatique du locuteur [Goldstein 76] [Paliwal 84] (cf. chapitre III de la partie B), les formants des voyelles françaises ont été peu étudiés. L'une des causes de cette pénurie est sans doute la difficulté de les déterminer automatiquement de façon fiable, c'est-à-dire de numérotter correctement les pôles issus d'une analyse LPC ou bien les pics d'un spectre. Ainsi, K.K. Paliwal a établi manuellement les quatre fréquences formantiques des voyelles anglaises qu'il a étudiées.

Les voyelles que nous avons sélectionnées remplissent deux conditions. Premièrement, ce sont les voyelles dont les fréquences formantiques  $F_1$  et  $F_2$  présentent les plus grands écarts entre les deux sexes (cf. figures A.50 et A.51). Deuxièmement, elles ne risquent pas d'être remplacées par un autre phonème dans le contexte phonologique choisi. Ainsi, les phonèmes /  $\text{o}$  / et /  $\text{e}$  / présentent aussi des écarts importants pour  $F_1$  et  $F_2$ . Mais, ils n'apparaissent pas dans les syllabes fermées par /  $\text{R}$  /, et, lorsqu'ils sont suivis de /  $\text{R}$  / dans des syllabes ouvertes, ils peuvent être remplacés par /  $\text{œ}$  / et /  $\epsilon$  /. Le contexte phonologique a été choisi afin de faciliter la détermination automatique des formants et de limiter les origines des différences entre les locuteurs à la réalisation d'une cible articulatoire isolée ou tout du moins afin de s'approcher de cet objectif.



- Les formants de certaines voyelles orales précédées des occlusives voisées / **b** /, / **d** / et / **g** /.

Dans ce cas, le but recherché est la caractérisation du locuteur grâce aux phénomènes de coarticulation. Deux études sont prévues, celle des fréquences formantiques au centre de la voyelle, qui correspondent à la réduction du geste articulatoire, et celle des trajectoires formantiques, qui correspondent à la transition entre la consonne et la voyelle. Les voyelles étudiées ont été choisies par l'expert phonéticien en fonction de l'occlusive, afin de maximiser les faits de coarticulation :

- pour l'occlusive bilabiale / **b** /, essentiellement / **e** / et / **a** /, auxquelles on peut ajouter / **ə** /, / **ɔ** / et / **u** /,
- pour l'occlusive dentale / **d** /, principalement / **œ** / (ou / **ə** /), / **a** /, / **ɔ** / et / **u** /,
- pour l'occlusive vélaire / **g** /, essentiellement / **e** /, / **ɛ** / et / **a** /, auxquelles on peut ajouter / **ɔ** / et / **u** /.

Les figures C.1, C.2 et C.3 présentent les trajectoires formantiques des voyelles orales lors de la prononciation par un locuteur masculin de la chaîne / **VCV** / où **C** est l'une des trois consonnes / **b** /, / **d** / et / **g** /. Nous pouvons constater par exemple que les formants de la voyelle / **i** / ne présentent pas de transitions. Ils sont parallèles à l'axe du temps pour les trois consonnes, ce qui explique pourquoi cette voyelle n'a pas été retenue.

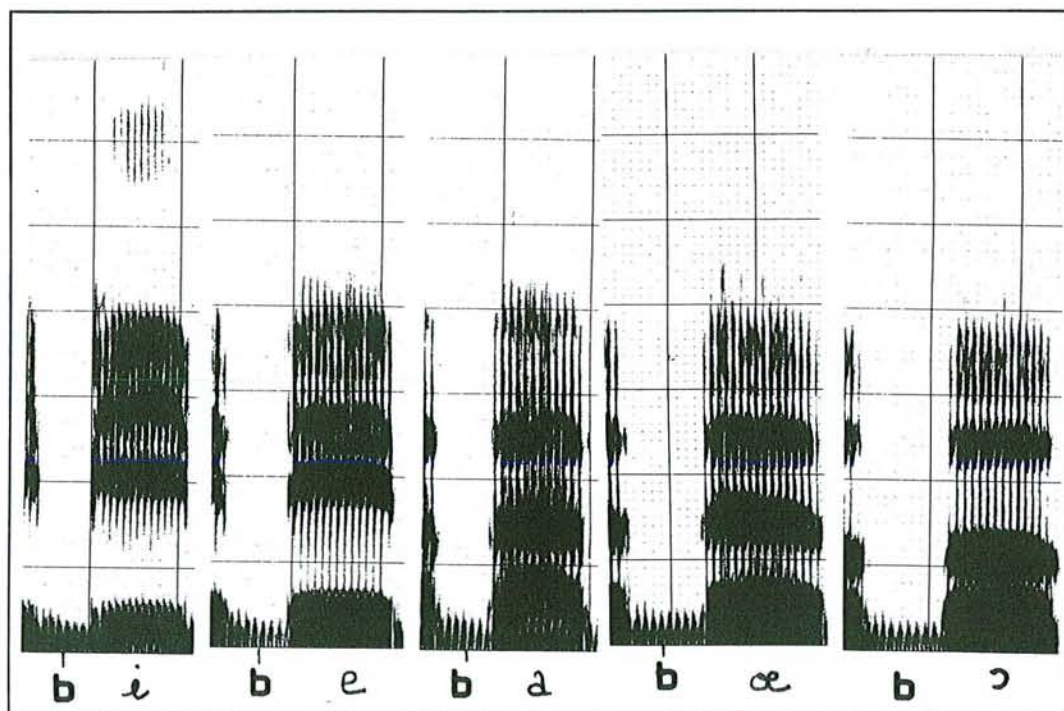


Figure C.1. Spectrogrammes des chaînes [ **bVb** ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée [Eskenazi 88].



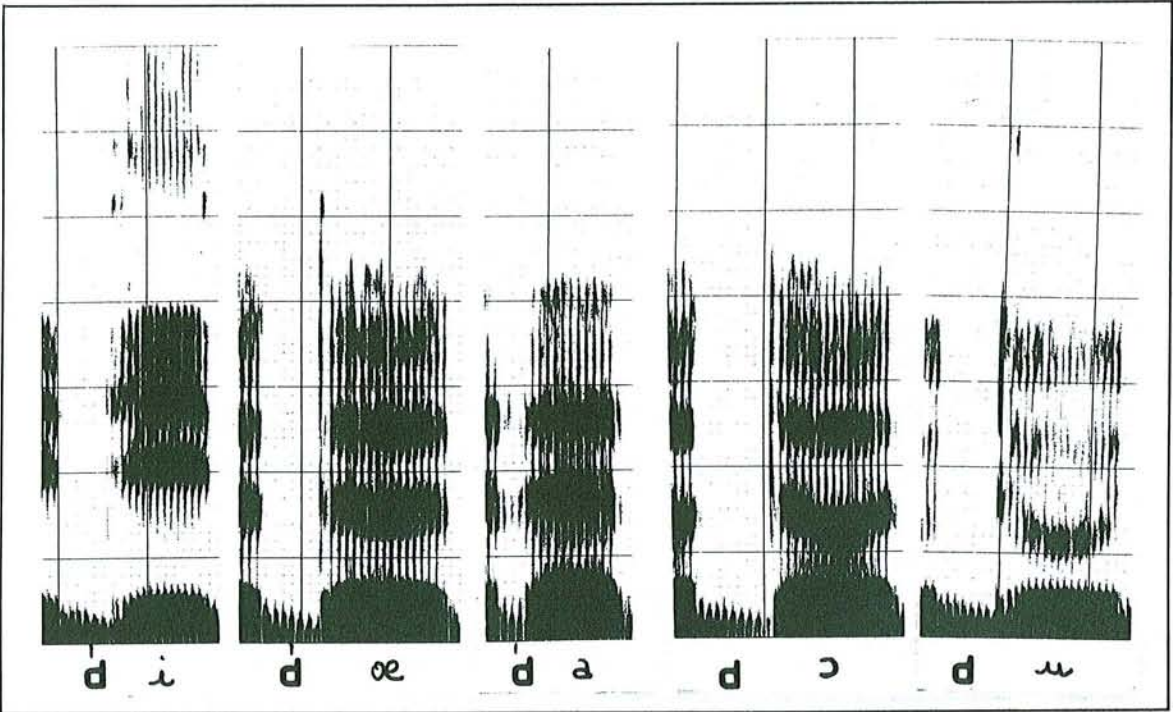


Figure C.2. Spectrogrammes des chaînes [ dVd ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée [Eskenazi 88].

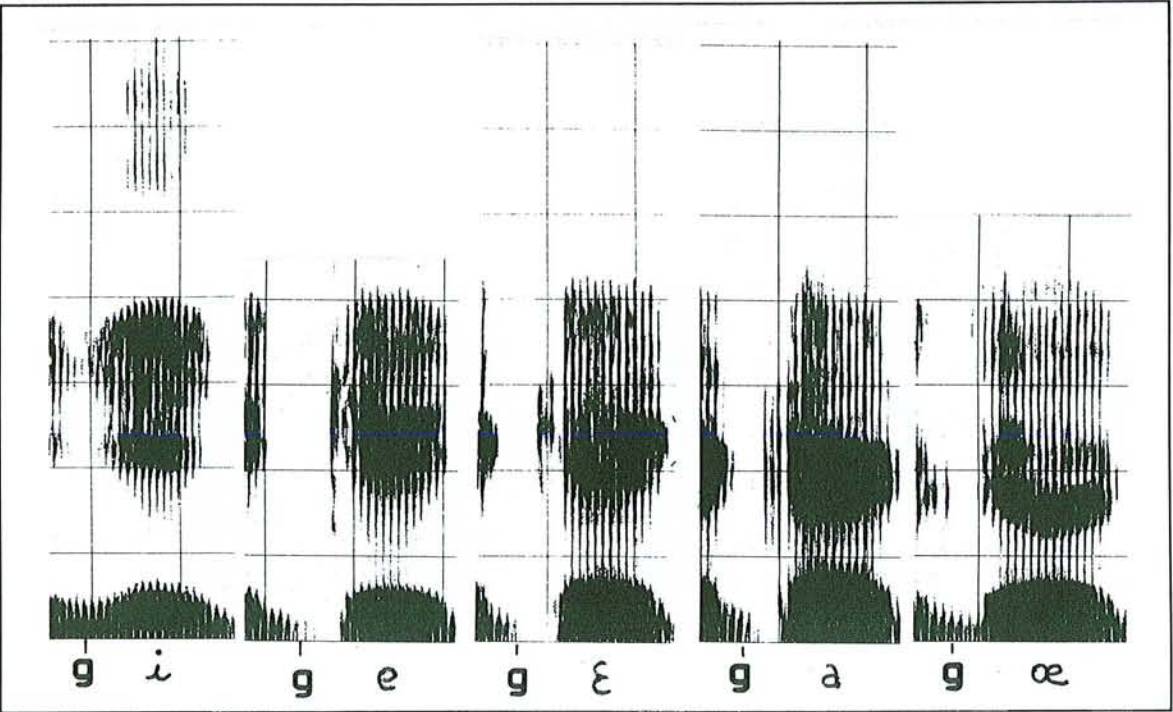


Figure C.3. Spectrogrammes des chaînes [ gVg ] prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée [Eskenazi 88].

- Le spectre des consonnes nasales / **m** / et / **n** /.

Des études sur l'allemand et l'anglais ont démontré la pertinence des consonnes nasales pour l'identification du locuteur [Nolan 83] [Wolf 72] [Sambur 75] [Su 74]. Certaines d'entre elles, comme l'étude de L.S. Su et al., ont plutôt mis en évidence la pertinence des conséquences acoustiques des phénomènes de coarticulation, en particulier l'anticipation de l'articulation linguale de la voyelle lors de la prononciation du / **m** /. D'autres ont plutôt établi la pertinence des consonnes nasales proprement dites. Dans ces études, / **n** / semble légèrement meilleure que / **m** /, alors qu'elle est beaucoup moins sujette au phénomène d'anticipation articulatoire. En français, l'étude réalisée par J.P. Bonastre et H. Méloni [Bonastre 92] a souligné la pertinence de / **m** / pour l'identification de 22 locuteurs, lorsque ce phonème est utilisé tous contextes confondus. De même, l'étude de G. Pérennou et al. [Pérennou 82] a mis en évidence la pertinence de / **n** / et / **m** / pour la discrimination de cinq locuteurs, toujours dans le cas où tous les contextes sont confondus.

- Le spectre des voyelles nasales /  $\tilde{e}$  /, /  $\tilde{a}$  /, /  $\tilde{o}$  / précédées de / **m** / ou / **n** /.

La langue anglaise ne possède pas de voyelles nasales. Il y a donc eu très peu d'études les concernant. A notre connaissance, les seules études en français sont celles de G. Pérennou et al. [Pérennou 82] et de J.P. Bonastre et al. [Bonastre 92]. La première a établi la pertinence des voyelles /  $\tilde{o}$  /, /  $\tilde{e}$  / et /  $\tilde{a}$  / dans la discrimination de cinq locuteurs. Dans la seconde, l'emploi d'une seule occurrence du phonème /  $\tilde{a}$  / engendre un taux d'identification de 64% pour 22 locuteurs alors que la voyelle /  $\tilde{o}$  / conduit à un taux de 61%. Par ailleurs, nous avons montré, dans la première partie de ce mémoire au paragraphe III.2.2, qu'il existait des différences entre les articulations orales des voyelles nasales de deux locuteurs. De telles différences se répercutent sur les fréquences de résonance du conduit oral considéré seul et donc sur les fréquences formantiques correspondantes. Par ailleurs, le couplage du conduit nasal au conduit vocal entraîne une élévation des fréquences formantiques orales. Comme le montre la figure C.4, cette élévation est une fonction du degré de couplage, c'est-à-dire du degré d'abaissement du voile du palais. De plus, le passage de l'air dans le conduit nasal introduit plusieurs paires de pôle-zéro dont les fréquences dépendent de la taille et de la forme des cavités nasales mais aussi du degré de couplage [Calliope 89].

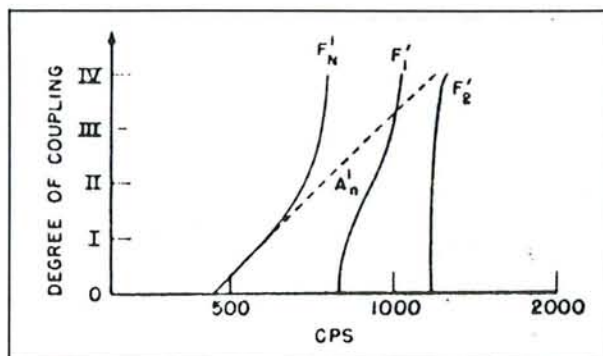


Figure C.4. Evolution des fréquences des formants et des antiformants en fonction du degré de couplage.  $F_1'$  et  $F_2'$  sont les fréquences formantiques d'origine orale.  $F_N^I$  et  $A_N^I$  sont respectivement la fréquence du premier formant nasal et celle du premier antiformal nasal ; d'après O. Fujimura dans [Lonchamp 88].



La variabilité des fréquences des formants d'origine orale et nasale et des antiformants (zéros) d'origine nasale engendrent des spectres de voyelles nasales qui sont eux aussi très variables. En effet, un formant oral et un formant nasal peuvent se regrouper en un seul pic. Dans d'autres cas, la proximité d'un formant nasal peut décaler le formant oral vers les fréquences plus élevées, ou bien la présence d'un antiformant dans le voisinage d'un formant peut faire disparaître le pic spectral correspondant au formant.

Deux études spectrales sont possibles sur les voyelles nasales. La première est une étude statique qui calcule à un instant donné le spectre d'une voyelle nasale et le compare à un spectre de référence calculé au même instant. La deuxième est une étude dynamique qui compare les spectres calculés au début, au milieu et vers la fin de la voyelle afin d'établir comment le locuteur gère l'ouverture du conduit nasal. En effet, celui-ci peut abaisser le voile du palais dès le début de la voyelle ou plus tard, articulant ainsi un début de voyelle orale. De plus, il peut effectuer ce geste articulatoire de deux manières. Soit il abaisse le voile du palais brusquement et le maintient dans cette position jusqu'à la fin de la voyelle, soit il l'abaisse progressivement tout au long de l'articulation de la voyelle. Toutefois, le début de la nasalité est une phase transitoire difficile à détecter. Les voyelles nasales étudiées sont donc précédées d'une consonne nasale afin de reporter cette phase dans la consonne.

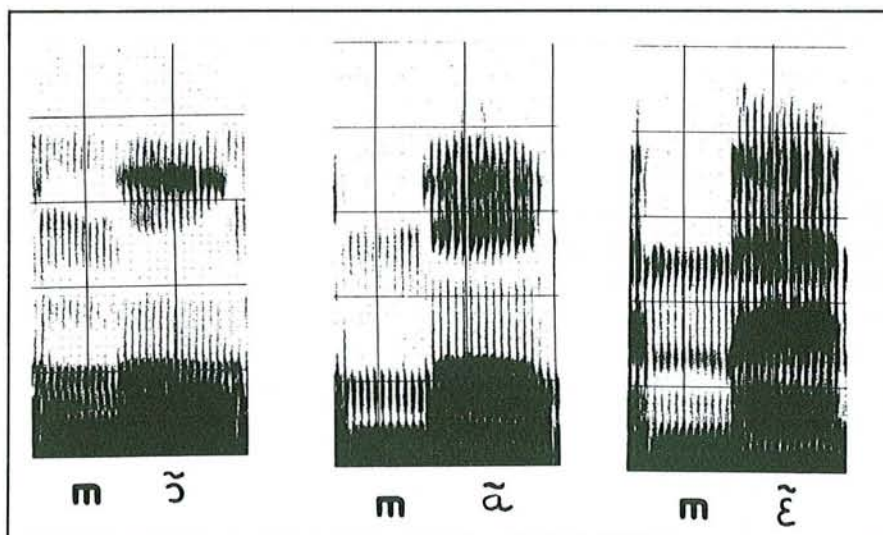


Figure C.5. Spectrogrammes des voyelles nasales [õ], [ã] et [ẽ] précédées de [m] et prononcées par un locuteur masculin, d'après les données du GRECO Communication Parlée [Eskenazi 88].

La voyelle /õ/ n'a pas été sélectionnée car son spectre semblait moins informatif que les trois autres voyelles. Comme le montre la figure C.5, les formants F'1, F'2 et FN1 sont très proches et forment un bloc compact sous 1000 Hz. Mais il est vrai que nous venons de citer la bonne performance de cette voyelle dans l'étude de J.F. Bonastre. De même dans les travaux de G. Pérennou et al. [Pérennou 82], cette voyelle précédée de /p/ (*"Monsieur et Madame Dupont"*) est le phonème le plus pertinent pour discriminer les cinq locuteurs. De même, dans



l'étude de J.P. Bonastre [Bonastre 92], l'identification d'un locuteur parmi vingt-deux à l'aide d'une seule occurrence en contexte du phonème / ʁ / atteint en moyenne un taux de 61%.

- Le spectre des liquides / l / et / R / suivies des voyelles / i /, / a / et / u /.

Du point de vue de la caractérisation du locuteur, la consonne latérale / l / comporte deux aspects intéressants. Tout d'abord, son articulation présente une certaine variabilité aussi bien au niveau du lieu d'articulation (dents ou alvéoles), qu'au niveau de la forme de la langue au point d'articulation et à l'arrière de ce point (l'air passe de chaque côté de la langue ou des deux côtés). De plus, elle est également très sensible à la coarticulation. La fréquence formantique  $F_1$  est une fonction de la longueur du conduit vocal alors que  $F_2$  dépend du degré de constriction pharyngale qui est fortement influencé par la voyelle suivante (cf. figure A.23).  $F_2$  peut varier de 1400 Hz pour / a / à 1800 Hz pour / i /. Quant à  $F_3$ , elle dépend de la longueur du passage latéral et de la cavité labiale. Enfin, la cavité située à l'arrière du point d'articulation engendre un zéro dont la fréquence varie avec la taille de cette cavité, ce qui rend difficile l'affectation des pôles LPC ou des pics du spectre aux formants [Lonchamp 87].

Les bons résultats des études sur le / r / apical anglais [Goldstein 76] sont plus transposables au / l / français qu'au / r / français. En effet, celui-ci est rarement apical mais est soit une vibrante dorso-uvulaire soit, le plus souvent, une constrictive dorso-uvulaire. Toutefois, il présente aussi une grande variabilité acoustique qui résulte de celles du degré de constriction et du lieu de constriction qui est lui-même influencé par l'entourage vocalique. La constriction est dorso-vélaire au contact des voyelles antérieures et dorso-uvulaire au contact des voyelles postérieures.

Les figures C.6 et C.7 montrent la variabilité des trois premiers formants de / l / et / R / dans le contexte / VCV /.

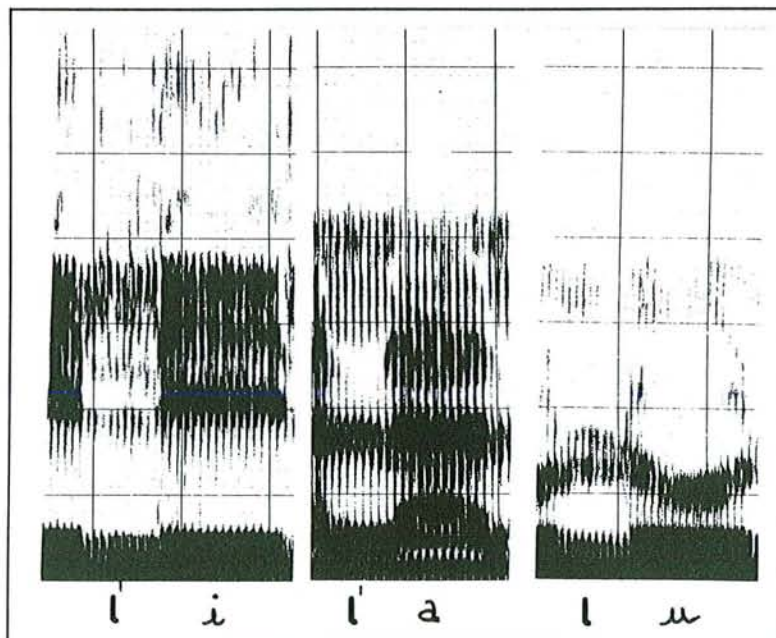


Figure C.6. Spectrogrammes des chaînes [ V l V ] prononcées par un locuteur masculin, où V est une des voyelles [ i ], [ a ], [ u ], d'après les données du GRECO Communication Parlée [Eskenazi 88].

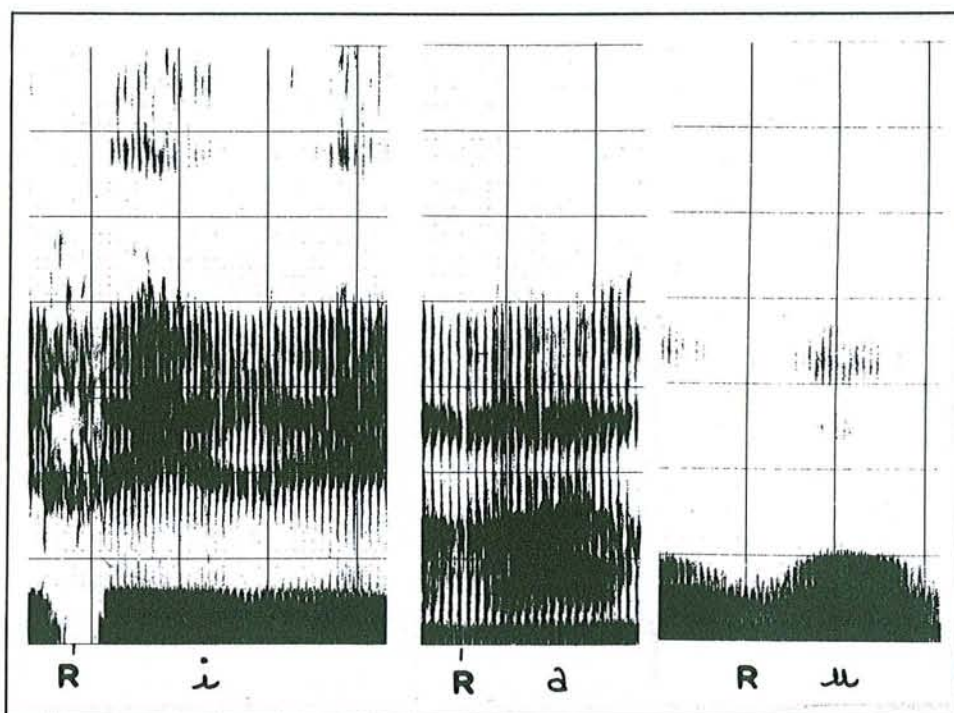


Figure C.7. Spectrogrammes des chaînes [VRV] prononcées par un locuteur masculin, où V est une des voyelles [i], [a], [u], d'après les données du GRECO Communication Parlée [Eskenazi 88].

- *Le spectre des occlusives /k/ et /g/ en fonction de l'entourage vocalique.*

Grâce à l'analyse spectrale, nous souhaitons étudier les habitudes articulatoires des locuteurs en ce qui concerne le lieu d'articulation des occlusives vélaires en fonction de leur contexte vocalique. Plus précisément, nous désirons comparer le spectre de l'explosion de l'occlusive vélaire dans un contexte central /a/, par rapport aux spectres obtenus dans les contextes postérieur /ɔ/ et antérieur /ɛ/. G. Pérennou et al. [Pérennou 82] ont déjà montré que le lieu d'articulation des occlusives vélaires de leur corpus permettait de discriminer les cinq locuteurs de leur étude.

## 2.2. Les paramètres temporels

Les paramètres temporels sélectionnés sont issus d'une analyse temporelle fine des occlusives soit au niveau de leur articulation intrinsèque soit au niveau de leur coarticulation avec des voyelles.

- *Le rapport de la durée de la tenue d'une occlusive sonore à celle de la tenue d'une occlusive sourde.*
- *Le délai d'établissement et le maintien du voisement des occlusives sonores situées en début de groupe de phonation.*



Les indices perceptifs du voisement d'une occlusive sont le délai d'établissement du voisement (*Voice Onset Time*), la durée de la tenue, la durée de la voyelle précédente, l'intensité de la phase d'explosion et la durée des transitions formantiques.

Le délai d'établissement du voisement est l'intervalle de temps qui sépare la phase d'explosion du début de la vibration des cordes vocales. Pour les occlusives sourdes, cet intervalle est compté positivement, les cordes vocales commencent à vibrer après l'explosion. Pour les occlusives sonores, il est compté négativement, la vibration des cordes vocales débute pendant la tenue de l'occlusive. Ce délai d'établissement est une fonction du locuteur en particulier pour les consonnes voisées. Pour des raisons de commande neuromusculaire, le voisement commence plus ou moins tôt selon le locuteur. Par ailleurs, l'égalité des pressions sousglottique et supraglottique peut entraîner l'arrêt du voisement pendant la tenue.

Dans l'étude de G. Pérennou et al. [Perennou 82], le délai d'établissement du voisement des consonnes sourdes calculé globalement sur tout le texte permet de répartir les cinq locuteurs en quatre classes.

La variabilité et la pertinence de la durée d'une occlusive pour la reconnaissance automatique du locuteur ont déjà été établies aussi bien pour la langue anglaise que pour la langue française. La durée du burst du / **k** / américain est classée en troisième position dans l'étude sur la caractérisation du locuteur effectuée par M.K. Sambur (cf. figure B.4). En français, P. Corsi classe dans les neuf paramètres les plus pertinents la durée globale du / **k** / et la durée de la tenue du / **p** /.

- *La durée de la friction qui suit l'explosion de / t / et / d /, lorsqu'elles sont suivies des voyelles / i /, / y / ou / e /.*

Nous avons introduit dans le chapitre V de la partie A le phénomène de palatalisation et plus précisément celui d'affrication des occlusives apico-dentales lorsqu'elles sont suivies des voyelles antérieures fermées. L'observation de spectrogrammes indique que ce phénomène est variable selon les locuteurs. Par ailleurs, G. Pérennou et al. [Perennou 82] ont étudié ce phénomène en distinguant trois variantes dans la prononciation par cinq locuteurs du mot "*matinée*". Sous cet aspect, le phénomène présente une variabilité interlocuteur mais aussi pour trois des locuteurs une variabilité intralocuteur.

### 3. Différences d'origine linguistique et phonologique

Les paramètres qui suivent révèlent les particularités des systèmes phonologiques des locuteurs et leur manière de réaliser certaines fonctions linguistiques comme la prosodie.

- *Le rapport de la durée d'une voyelle située en fin de groupe de phonation à celle d'une voyelle non finale.*

Ce paramètre permet d'établir le degré d'allongement utilisé par le locuteur comme marqueur de frontière syntaxique (cf. paragraphe V.3.1.3 de la partie A).



- *La réalisation phonétique du “e muet”.*

Étant donné un contexte phonologique, le “e muet” peut-être élide ou réalisé en un allophone proche de [œ] ou de [ø]. Dans leur enquête pour l'établissement du “Dictionnaire de la prononciation française dans son usage réel” [Martinet 73], A. Martinet et H. Walter ont noté les réalisations phonétiques du “e muet”. 44% des schwas sont articulés [œ], 33% sont articulés [ø] et 24% comme une voyelle centrale notée [ə]. Ce dernier type d'articulation a été réalisé par les trois locuteurs les plus âgés. Cependant, le schwa est le phonème qui obtient les plus mauvais taux d'identification dans l'étude de J.F. Bonastre

- *L'analyse de la neutralisation de l'opposition phonologique / ẽ / ~ / œ /.*
- *L'analyse de la neutralisation de l'opposition phonologique des couples (/ e /, / ɛ /) et (/ o /, / ɔ /) dans certaines syllabes.*

Ces paramètres n'ont pas donné lieu à un corpus spécifique mais seront étudiés à partir du corpus établi pour les paramètres précédents. C'est pour cette raison que nous n'avons pas cité toutes les oppositions sujettes à neutralisation dont nous avons parlé au paragraphe V.4.2.2 de la partie A.

## 4. Conclusion

Nous venons d'établir la liste des paramètres sélectionnés en vue d'une étude de leur pertinence pour la caractérisation automatique du locuteur. A partir de cette sélection, nous avons élaboré un corpus de phrases qui a été enregistré par un certain nombre de locuteurs avant d'être numérisé et étiqueté. Ceci fera l'objet du prochain chapitre.

La plupart des paramètres choisis sont soit des manifestations acoustiques des différences physiologiques entre les locuteurs soit des conséquences de leurs habitudes articulatoires, notamment dans les faits de coarticulation.

Nous pouvons remarquer que des paramètres comme la fréquence fondamentale moyenne ou le débit ont été volontairement écartés. L'objectif de cette étude n'est pas de réaliser un système d'identification automatique du locuteur mais de mettre en évidence de nouveaux paramètres plus subtils et qui soient moins sensibles à la modification consciente. Dans une étape ultérieure, il sera toujours possible d'associer à ces nouvelles caractéristiques du locuteur des paramètres déjà connus afin de réaliser un tel système.

## CHAPITRE II ELABORATION ET ETIQUETAGE DU CORPUS

### 1. Introduction

Nous développons dans ce chapitre toutes les étapes qui ont été nécessaires à l'établissement des paramètres que nous souhaitons étudier : construction des phrases les mettant en œuvre, enregistrement du corpus par un groupe de locuteurs, numérisation et étiquetage du corpus.

Nous allons tout d'abord décrire rapidement ce qui concerne l'élaboration du corpus, de sa conception à sa numérisation. Puis, nous présenterons la problématique générale de l'étiquetage avant de détailler notre méthodologie d'étiquetage. Enfin, nous évoquerons dans la conclusion l'utilité de notre corpus étiqueté pour les chercheurs en reconnaissance et en compréhension automatiques de la parole au sein du laboratoire.

### 2. Elaboration du corpus

#### 2.1. Construction des phrases

Après avoir sélectionné les paramètres, nous avons construit dix-sept phrases en essayant de minimiser le nombre de phrases à prononcer par rapport au nombre de paramètres retenus. Dans cette phase d'expérimentation, l'emploi de phrases a été jugé préférable à celui de mots isolés. En effet, la lecture de listes de mots hors contexte devient vite fastidieuse pour un locuteur bénévole et reflète peu sa prononciation naturelle. La table C.1 présente les phrases du corpus.

#### 2.2. Enregistrement et numérisation du corpus

##### 2.2.1. Enregistrement

Pour des raisons d'homogénéité de population, dix-huit locuteurs et vingt et une locutrices, tous originaires de Lorraine ou y résidant depuis de nombreuses années, ont été choisis pour lire au cours d'une même session quatre répétitions de chaque phrase. Presque tous ces locuteurs sont des enseignants universitaires ou des chercheurs de notre laboratoire qui n'ont donc pas été impressionnés par le fait de parler dans un microphone. Les répétitions ont été présentées aux locuteurs sous la forme de quatre listes de phrases rangées aléatoirement afin de ne pas toujours reporter sur les mêmes phrases les phénomènes de fin de liste (énergie plus faible, débit plus rapide, modification de la mélodie, ...). Le corpus obtenu comporte donc plus de 2600 phrases.

L'enregistrement a été effectué à l'aide d'un microphone dynamique Shure Unisphère B et d'un magnétophone Revox A77 à 19 cm/s et sur des bandes de très bonne qualité. Il ne s'est pas déroulé dans une véritable chambre sourde mais dans une pièce isolée et à des heures calmes.

- 1 *Guy a péri bêtement du diabète en Italie.*
- 2 *La porte du garage tomba avec lourdeur.*
- 3 *La partie de belote dura toute la matinée.*
- 4 *Un bateau à vapeur a quitté le port.*
- 5 *Le petit gamin traîne un jouet.*
- 6 *Donne-moi le bocal de cacao.*
- 7 *En ski, la godille permet d'éviter les tournants.*
- 8 *Un coq bien dodu pour demain.*
- 9 *Lequel des bandits guette près du repère.*
- 10 *Le trappeur commun redoutait le loup-garou.*
- 11 *Douze nains conspirent derrière le bosquet.*
- 12 *Le soldat brisa la baguette de son tambour.*
- 13 *Goûtez-moi ce cake au beurre.*
- 14 *Le rire de la gouvernante est revigorant.*
- 15 *La cousine du nain soupire dans son délire.*
- 16 *Le départ de la course Strasbourg-Paris aura du retard.*
- 17 *Notre guide charmant quitte la jolie route danoise.*

Table C.1. Les dix-sept phrases du corpus.

### 2.2.2. Première numérisation

Une première numérisation du corpus a été effectuée sur un micro-ordinateur Exormacs à base de MC68000 et équipé d'une carte d'acquisition effectuant un échantillonnage à 10 kHz et une conversion A/N sur 10 bits. Dans une seconde phase, nous avons souhaité étiqueter notre corpus. Comme aucun logiciel interactif de traitement et de visualisation du signal de parole n'était disponible sur ce micro-ordinateur, nous avons entrepris l'écriture de programmes Pascal de traitement du signal mettant en œuvre une carte Sky à base de processeurs vectoriels et un terminal graphique couleur Tektronix 4105.

Entre temps, le micro-ordinateur Exormacs a été remplacé par un mini-ordinateur Masscomp 5600, qui était un matériel plus performant et dédié au temps réel, donc mieux conçu pour l'acquisition de parole dans un mode de fonctionnement multi-utilisateur. Ce mini-ordinateur était équipé d'un système de numérisation comportant un convertisseur A/D sur 12 bits à une fréquence d'échantillonnage de 16 kHz. De plus, il possédait deux écrans Bitmap couleur gérés par une bibliothèque graphique et un gestionnaire de multifenêtrage. Aussi avons-nous décidé de recommencer la numérisation de notre corpus sur cette machine plus performante.

### 2.2.3. Deuxième numérisation

Nous avons donc numérisé la partie du corpus correspondant aux dix-huit premiers locuteurs et aux sept premières locutrices de la table C.2. L'acquisition a été réalisée phrase par phrase



afin d'obtenir une dynamique maximale tout en évitant la saturation du convertisseur. La carte d'acquisition ne comportant pas de filtre passe-bas, nous avons utilisé celui de l'Institut de Phonétique de Nancy réglé sur la fréquence de coupure 6800 Hz.

H	aq	bz	df	gm	jfm	jg	jlc	jmp	jph	ms	am	as	bm	cs	fsc	gh	gyf	rg			
F	cf	gaf	mcp	nc	of	ot	sm	ab	bw	mg	mk	ym	pj	bj	ng	mc	mch	js	hc	mc	fs

Table C.2. Les locuteurs qui ont enregistré le corpus rangés dans l'ordre dans lequel ils ont été étudiés.

C'est seulement après cette seconde numérisation que nous avons pu étiqueter une partie de notre corpus. Avant de détailler cette phase d'étiquetage, nous allons développer quelques généralités sur l'étiquetage, en particulier la définition de ses différentes composantes et la présentation de quelques-uns des problèmes auxquels se trouve confronté tout étiqueteur.

## 3. Généralités sur l'étiquetage

### 3.1. Introduction

L'étiquetage d'un énoncé oral comporte deux phases, une phase de transcription qui associe une suite de symboles à l'énoncé et une phase d'alignement (*labelling*) qui effectue la mise en correspondance temporelle du signal de parole et de la transcription. Cette définition n'est pas unique et dans certains cas, le terme "étiquetage" s'applique seulement à la phase d'alignement, notamment lorsqu'on différencie l'étiquetage manuel de l'étiquetage semi-automatique. La transcription et son alignement sont deux tâches qui peuvent être effectuées séparément par des agents différents, que ce soient des personnes ou des programmes, ou bien simultanément par le même agent.

Comme nous allons le voir, la problématique de l'étiquetage n'est pas simple. Avant de la détailler, nous en proposons un résumé sous la forme d'un extrait d'un article de G. Pérennou et N. Vigouroux [Perennou 88] : *"L'étiquetage se révèle une tâche délicate et coûteuse dès que l'on s'intéresse aux unités inférieures au mot : délicate, si l'on considère les désaccords entre phonéticiens quant au statut des unités phonologiques et phonétiques, à leur polymorphisme et aux difficultés qu'il y a parfois à les localiser dans le signal ; coûteuse car ce type d'étiquetage demande beaucoup de temps à un spécialiste muni d'un poste de travail convenablement équipé"*.

Cet article a été rédigé dans le contexte de l'élaboration de la base de données des sons du français (BDSONS) et dans celui du projet Esprit SAM (Assessment, Methodology and standardisation in multilingual Speech technology) dont l'un des objectifs était d'établir une norme européenne d'étiquetage des corpus multilingues. En effet, au moment où nous avons étiqueté notre corpus, de nombreuses bases de données de parole étaient en cours d'élaboration dans le monde, afin de tester plus efficacement les systèmes de RAP. Tous ces projets ont donc été confrontés aux différents problèmes posés par l'étiquetage.

Nous allons présenter les transcriptions les plus utilisées ainsi que les principales méthodes d'alignement, en prenant pour exemple l'étiquetage des bases de données internationales. Dans

le même temps, nous aborderons les problèmes sous-jacents à ces différentes composantes de l'étiquetage.

### 3.2. La transcription

Selon la nature linguistique des symboles utilisés dans la transcription, celle-ci sera orthographique, phonologique, phonétique, acoustique ou prosodique. Les deux premières peuvent être obtenues sans écoute et sans visualisation du signal, sauf dans le cas de la parole spontanée. Les transcriptions phonétiques et prosodiques peuvent résulter d'une simple écoute ou d'une écoute combinée à l'observation du signal temporel ou du spectrogramme de parole. Nous laisserons de côté la transcription prosodique qui concerne plutôt la parole naturelle.

#### 3.2.1. La transcription orthographique

La transcription orthographique est simplement l'énoncé orthographique usuel de la phrase ou du mot prononcé. C'est ce dont dispose le locuteur dans le cas d'un corpus lu comme le nôtre.

- Exemple : *"les six petites maisons"*.

#### 3.2.2. Les transcriptions phonologiques

Les transcriptions phonologiques (*phonemic transcription*) résultent de l'analyse phonologique manuelle ou automatique [Perennou 89] [Dujour 90] [Haton 91] de la transcription orthographique. Chacune d'elles correspond à l'une des étapes de cette analyse phonologique. Plus une transcription phonologique est proche de la forme syntaxique, plus elle comporte d'unités ou de diacritiques phonologiques comme les frontières (de morphèmes, de mots, de groupes de phonations, ...) ou les marques indiquant la prononciation éventuelle d'une consonne selon son contexte phonologique. Au fur à mesure de l'application de règles phonologiques, la transcription s'affine pour converger vers une transcription qui correspond soit à une prononciation normative soit à une ou plusieurs prononciations dialectales ou idiolectales. Cette transcription n'est constituée que de phonèmes et éventuellement d'archiphonèmes lorsque l'opposition entre deux phonèmes est neutralisée. Cette ultime transcription phonologique est appelée transcription phonotypique par certains auteurs [Perennou 88] [Autesserre 88a].

- Exemples de transcriptions phonologiques :
  - / \$IEz''#sis'#pətit''+ə+z''# mEzõ+s''#\$ /.<sup>1</sup>
  - / IEsipətitəməzõ /.

#### 3.2.3. La transcription phonétique

La transcription phonétique précise la prononciation effective. Elle résulte soit d'une simple écoute soit d'une écoute couplée à une visualisation du signal de parole dans le domaine temporel et/ou dans le domaine spectral. La précision de cette transcription dépend de son mode d'établissement et de l'application visée. Selon le degré de précision, les phonéticiens parlent de transcription large ou étroite [Autesserre 88a].

<sup>1</sup> La marque de latence '' indique que la consonne qui la précède ne se prononce qu'en cas de liaison.  
La marque ' indique que la consonne qui la précède se prononce selon le contexte.  
E est l'archiphonème (e, ε).



Un étiqueteur non expérimenté aura tendance à réaliser une transcription phonétique large, c'est-à-dire qu'il est capable de détecter l'élision de certains phonèmes, certaines substitutions et certaines assimilations totales mais qu'il ne saura pas déterminer le timbre exact d'une voyelle ou les assimilations partielles [Perennou 88]. Par ailleurs, s'il ne se contraint pas à suivre certaines règles, sa transcription risque d'évoluer au fur et à mesure de son apprentissage.

Seul l'expert phonéticien pourra effectuer une transcription phonétique fine. Mais, même dans ce cas, on peut se demander dans quelle mesure cette transcription correspond exactement à ce qui a été prononcé et si elle n'est pas en partie subjective. En effet, quels sont les critères utilisés par le phonéticien ? Connaît-il d'avance ce qui a été prononcé ? Effectue-t-il une écoute globale (groupes de phonation ou mots) ou une écoute partielle (phones ou triphones) ? Utilise-t-il des informations visuelles comme la courbe d'évolution de  $F_0$  ou la barre de voisement pour conforter son jugement ? En ce qui concerne les voyelles, comment détermine-t-il leur timbre ? Procède-t-il perceptivement par rapport à ses propres représentations des voyelles ou en fonction de l'image qu'il s'est construite du triangle acoustique du locuteur ? ou bien en évaluant les fréquences formantiques sur le spectrogramme ?

Prenons l'exemple de la base TIMIT dont l'étiquetage s'est effectué en trois étapes [Zue 88]. Dans un premier temps, elle a été transcrite phonétiquement par un phonéticien à partir de l'écoute attentive de portions de signal et de la visualisation du signal temporel et du spectrogramme. Puis, cette transcription a été alignée automatiquement à l'aide du programme CASPAR [Leung 84]. Enfin, les frontières issues de cet alignement ont été vérifiées visuellement et corrigées par des phonéticiens. Lors d'une étude perceptuelle entreprise quelques années plus tard, seize auditeurs sont amenés à identifier des voyelles et des diphtongues extraites de phrases de cette base. Lorsque les sons sont présentés isolément, seulement 55% des identifications correspondent à la transcription du phonéticien. Ce taux atteint 66% si le son est écouté avec ses contextes phonétiques antérieur et postérieur [Cole 92].

- Exemples de transcriptions phonétiques :
  - transcription large : [ **lEsipætitmEzõ** ].
  - transcription étroite : [ **lɛsɪpøtɪtmezõ** ].

Dans l'étiquetage de certaines bases de données, il est quelquefois réalisé une transcription semi-automatique grossière qui est en fait un découpage automatique du signal de parole en grandes classes phonologiques (voyelles, sonantes, occlusives, fricatives, silences, ...) avec vérification manuelle [Muthusamy 92]. Dans ce cas, l'alignement et la transcription sont simultanés.

### 3.2.4. La transcription acoustique

La transcription acoustique, lorsqu'elle existe, est toujours reliée à une phase de segmentation fine du signal de parole. En effet, elle consiste à repérer sur le signal des événements acoustiques dont la durée est inférieure à celle d'un son, comme la tenue d'une occlusive, sa barre d'explosion, le délai d'établissement du voisement, les parties stables des sonantes ou leurs phases transitoires [Autesserre 85].



### 3.3. L'alignement

Nous avons vu dans la première partie que la parole est engendrée par un continuum de gestes articulatoires qui peuvent se superposer les uns les autres. Dans ces conditions, vouloir aligner une suite de symboles discrets sur le signal de parole a toujours constitué un problème fondamental pour l'étiqueteur. Malheureusement, cet alignement est obligatoire pour pouvoir exploiter automatiquement les bases de données de parole.

La remarque précédente traduit une problématique "philosophique" de l'alignement qui concerne les experts phonéticiens. Les systèmes d'alignement automatique, quant à eux, n'en sont pas encore là, puisque leur principal objectif est encore de faire aussi bien que l'expert humain. En effet, ces systèmes sont plutôt qualifiés de semi-automatiques car ils nécessitent tous une vérification et une correction manuelle de l'alignement.

Nous allons donc tout d'abord examiner cette problématique de l'alignement manuel avant de donner quelques indications sur les méthodes d'alignement semi-automatique.

#### 3.3.1. Alignement manuel et segmentation

Deux démarches s'opposent au sein de la communauté scientifique qui est confrontée à ce problème d'alignement. La première démarche consiste à segmenter malgré tout le signal de parole pour associer un segment acoustique à un symbole. Les partisans de l'autre démarche essaient de résoudre l'antagonisme entre la transcription discrète et le signal continu en ne segmentant pas le signal de parole mais en localisant chaque phonème par son centre sur le signal de parole [Boe 88b] [Erp 89]. Mais ce type d'alignement est inexploitable lorsqu'on souhaite extraire automatiquement le segment de parole à analyser. Sur quels critères simples peut se faire l'extraction automatique d'une certaine quantité de signal de part et d'autre du centre ? Ces frontières calculées ne sont-elles pas plus imprécises que les frontières établies manuellement ? Par ailleurs, nous n'avons trouvé dans aucun des articles qui préconisent ce type d'étiquetage la définition exacte du centre d'un phonème. On peut se demander également si la localisation d'un pseudo-centre est conforme à la démarche naturelle du phonéticien. N'évalue-t-il pas les frontières avant de désigner le centre ? Une comparaison entre un étiquetage aux frontières et un étiquetage au centre, effectuée dans le cadre du projet Esprit SAM, a montré que le pseudo-centre correspond dans la majorité des cas au milieu du segment délimité par les frontières [Autesserre 88b].

Dans la presque totalité des bases de données internationales, l'alignement de la transcription est effectué par la pose de frontières sur le spectrogramme ou éventuellement directement sur le signal de parole. Dans certaines d'entre elles, le problème de l'incertitude des frontières est résolu par une segmentation hiérarchisée dans laquelle chaque niveau de segmentation correspond à un niveau de transcription [Dolmazon 87] [Kuwabara 89] [Hedelin 90].

La figure C.8 fournit un exemple d'étiquetage manuel hiérarchisé adopté par le laboratoire ATR au Japon. Il comprend cinq niveaux, de haut en bas :

- alignement d'une transcription phonologique,
- alignement d'une transcription acoustico-phonétique qui représente ce qui a été réellement prononcé, tout en indiquant des segments infraphonémiques comme les débuts et les fins de voyelle, les tenues d'occlusives sourdes ou sonores, etc.,
- pose de diacritiques comme le dévoisement ou la présence de bruit dans une voyelle,
- marquage des parties insegmentables,

- indication des centres des voyelles.

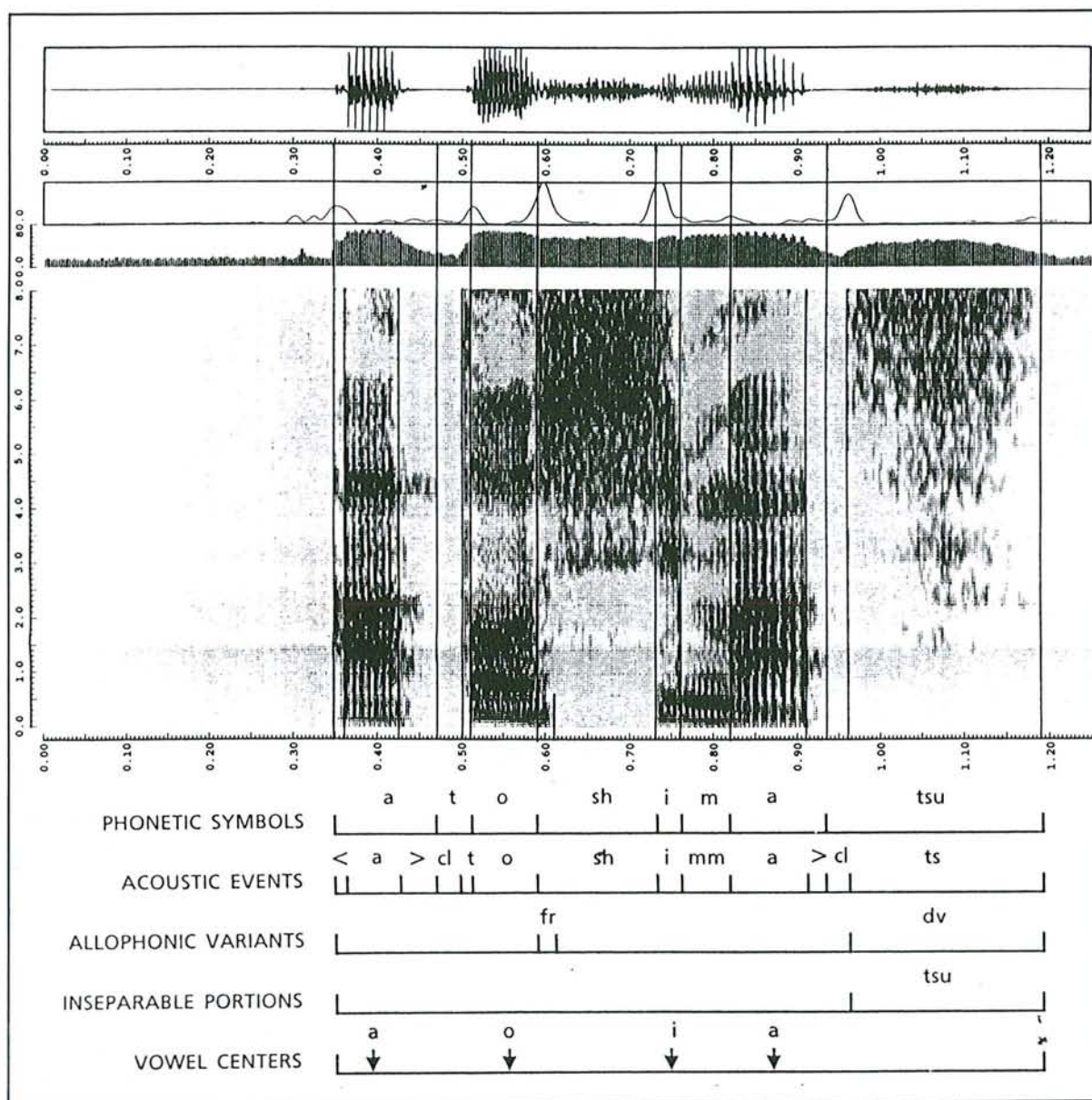


Figure C.8. Les différents niveaux d'étiquetage dans la base de données japonaise du laboratoire ATR, d'après [Kuwabara 89].

Un tel étiquetage est idéal notamment pour le test des systèmes de reconnaissance de parole continue. Il permet la recherche automatique de segments de parole qui se rapportent à une prononciation normative, tout en indiquant la réalité acoustique du signal (au problème près du timbre des voyelles que nous avons déjà évoqué). Malheureusement, pour être valable, cet étiquetage doit être fait par un expert phonéticien, alors que c'est une tâche de longue haleine et



assez ennuyeuse. Ainsi, l'étiquetage de BDSONS en événements acoustiques fins [Autesserre 85] a été évalué à cinq minutes par son [Boe 88a], ce qui à notre avis est insuffisant mais ce qui représente déjà deux heures pour une phrase de vingt-quatre phonèmes (taille de la phrase 15 de la table C.1). Aussi, dans bien des cas, faut-il se contenter d'une seule segmentation.

Une solution plus simple pour résoudre cet antagonisme serait un alignement qui permettrait le chevauchement des frontières. Néanmoins, si aucune de ces solutions n'est disponible lors d'une phase d'étiquetage, nous estimons qu'il est quand même concevable de segmenter le signal de parole en éléments disjoints. Il suffit que l'utilisateur de la base de données soit conscient que les frontières sont incertaines et qu'il y a un recouvrement des sons de part et d'autre. En revanche, l'étiqueteur, qu'il soit ou non expert, doit respecter certaines règles de segmentation afin que l'alignement soit uniforme. Ces règles doivent être transmises aux utilisateurs de la base.

### 3.3.2. Alignement semi-automatique

Depuis quelques années, le développement des bases de données parole de plusieurs milliers de phrases a engendré une explosion des recherches sur l'étiquetage semi-automatique. Une synthèse des diverses méthodes employées nécessiterait du temps et un chapitre entier. Aussi nous limiterons-nous à en présenter les grandes lignes.

Nous pouvons distinguer deux approches. L'approche la plus ancienne suscite toujours des recherches en vue d'améliorer ses performances [Perennou 82][Gong 92]. Elle consiste à aligner, à l'aide d'algorithmes de programmation dynamique, un énoncé non segmenté, soit avec un énoncé étiqueté manuellement, soit avec un énoncé synthétisé, soit encore avec une suite d'échantillons de référence de taille variable. Cette démarche permet l'étiquetage des différentes prononciations d'un même énoncé. Les phénomènes d'élision, d'insertion et de substitution, qui peuvent apparaître d'une répétition à l'autre, sont plus ou moins pris en compte selon la méthode.

Dans la deuxième approche, le signal de parole associé à un énoncé est segmenté en événements acoustico-phonétiques. Puis, ces événements sont regroupés et alignés avec la transcription phonétique de l'énoncé, obtenue manuellement. Cet alignement est effectué par des différentes méthodes qui peuvent être combinées comme la programmation dynamique, les modèles de Markov cachés [Svendsen 90], les réseaux de neurones [Dalsgaard 89] ou les systèmes à base de connaissances [Leung 84] [Kabre 91].

Dans cette dernière approche, certaines méthodes sont directement issues des systèmes de décodage acoustico-phonétique mis au point dans le cadre de la reconnaissance automatique de la parole continue. Le résultat du décodage, final ou partiel (grandes classes phonétiques) est aligné avec la transcription manuelle de l'énoncé [Leung 84] [Carlson 90].

### 3.4. Conclusion

Sans avoir fait une étude exhaustive de la problématique de l'étiquetage, nous avons présenté les principaux problèmes posés par l'étiquetage manuel. Ces problèmes demeurent dans l'étiquetage semi-automatique mais ils sont moins cruciaux que celui de la moins bonne la qualité de l'étiquetage par rapport à ce que ferait un étiqueteur humain.

Dans les grandes bases de données internationales, élaborées pour tester les systèmes de reconnaissance automatique de parole continue et naturelle, la tendance est à l'étiquetage manuel aux frontières, éventuellement hiérarchisé. Par ailleurs, même lorsque l'alignement est



réalisé automatiquement, deux phases sont toujours réalisées manuellement par des experts phonéticiens, la transcription phonétique et la phase finale de vérification des frontières. La raison en est simple, ces bases de données doivent contenir un maximum d'informations et celles-ci doivent être fiables.

Etant donné toutes ces considérations, nous avons choisi d'étiqueter notre corpus manuellement aux frontières.

## 4. Etiquetage du corpus

### 4.1. Introduction

Si nous avons choisi un étiquetage manuel aux frontières de notre corpus, en revanche, la méthodologie pour le réaliser a été fortement influencée par le logiciel disponible. Toutefois, nous avons essayé d'adapter cette méthodologie d'une part en prévision des études envisagées pour les paramètres sélectionnés, d'autre part de manière à noter le maximum de particularités des locuteurs. C'est ce que nous exposerons dans les deux prochains paragraphes.

Nous avons étiqueté manuellement les 680 phrases, qui correspondent aux quatre répétitions des dix premiers locuteurs de la table C.2. Pour cela, nous avons utilisé la version 88 de SNORRI, logiciel interactif d'édition de signal de parole qui a été développé au laboratoire par Y. Laprie avec la collaboration de D. Fohr [Laprie 88]. Ce logiciel possède un module d'étiquetage manuel que nous avons légèrement adapté pour étiqueter notre corpus. Il permet d'effectuer lors d'une seule opération la transcription d'un énoncé et son alignement sur le spectrogramme.

La figure C.9 présente les informations dont nous disposons sur l'écran bitmap lors de la phase d'étiquetage :

- a) la représentation dans le domaine temporel de deux secondes de parole ;
- b) le spectrogramme de parole couleur qui est représenté ici en niveaux de gris. Il a été obtenu à partir d'un calcul de FFT sur 256 échantillons de parole avec un recouvrement de 128 échantillons (8 ms). Dans la suite de la rédaction, nous appellerons prélèvement cet intervalle de 128 échantillons. Au préalable, le signal de parole a été préaccentué et multiplié par une fenêtre de Hamming. La couleur correspondant à l'énergie dans chacune des 128 bandes de fréquence de la FFT a été obtenue de la façon suivante :

si (énergie < 36 dB) ou (énergie < énergie maximale sur toutes les bandes – 32 dB)

alors énergie est codée en blanc

sinon

si (énergie > 73 dB)

alors énergie est codée en noir

sinon énergie est codée dans une couleur "proportionnelle" au rapport :  $\frac{\text{énergie} - 36}{73 - 36}$

fsi

fsi

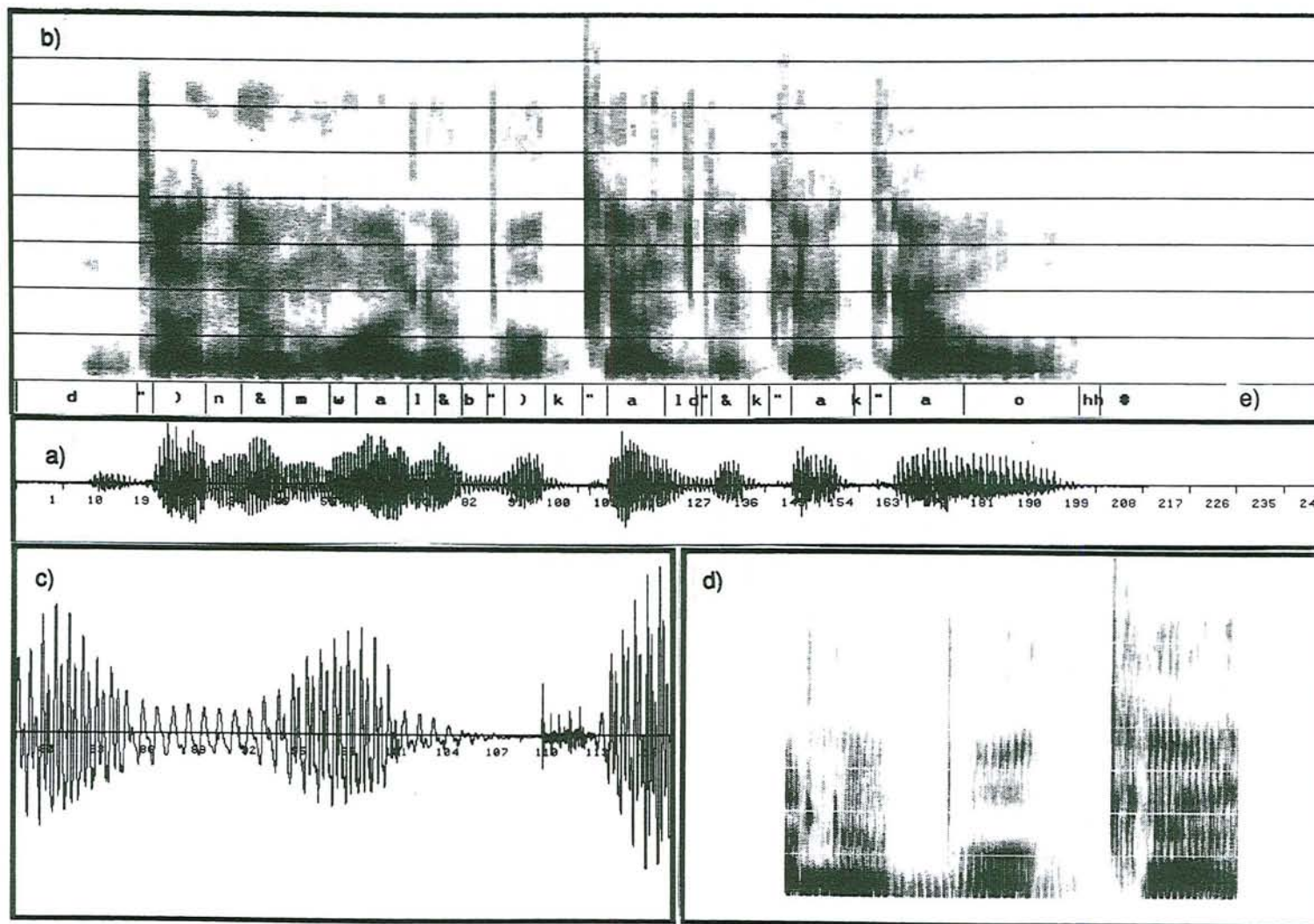


Figure C.9. Les informations disponibles sur l'écran lors de l'étiquetage de la phrase "Donne-moi le bocal de cacao !" prononcée par le locuteur jph : le signal temporel (a), le spectrogramme de parole couleur (b), le zoom temporel (c), le zoom spectrographique (d), le résultat de l'étiquetage (e).



- c) un zoom du signal temporel avec écoute simultanée du segment choisi,
- d) “un zoom spectrographique” obtenu à partir d’un calcul de FFT sur 64 échantillons avec un recouvrement de 32 échantillons. Contrairement à ce qui est présenté sur la figure C.9, les deux zooms n’étaient pas disponibles simultanément sur l’écran.

Le module d’étiquetage de la version actuelle de SNORRI ne propose pas les mêmes informations. Les spectrogrammes de parole qui serviront d’exemples par la suite sont issus de la dernière version de SNORRI. Ils sont plus précis que ceux dont nous disposions lors de l’étiquetage. Ceci explique en partie le léger écart entre la transcription et le signal de parole que l’on pourra observer sur certains d’entre eux. L’autre cause de cet écart provient du cadrage automatique de la frontière sur le début d’un prélèvement.

## 4.2. La transcription

Le logiciel d’étiquetage n’accepte qu’une seule transcription construite à partir des symboles présentés dans la table C.3. Comme nous pouvons le constater, ce sont les phonèmes du français auxquels s’ajoutent quelques symboles acoustiques comme les étiquettes “burst” et “souffle”. Étant donné cette contrainte et notre manque d’expérience, nous avons essayé de réaliser une transcription phonétique large en la complétant par quelques informations acoustiques repérées sur le spectrogramme.

D’une façon générale, notre transcription phonétique large correspond à ce que nous avons entendu et vu, sauf dans les cas ambigus où le symbole est celui de l’élément attendu. Celui-ci est issu d’une transcription de référence que nous avons choisie comme étant celle du dictionnaire “Le Petit Robert”.

Plus précisément, la transcription phonétique a été effectuée selon les principes suivants :

- prise en compte des insertions et des élisions ;
- prise en compte des assimilations consonantiques de sonorité. Comme nous ne disposions pas dans cette première version de SNORRI d’un détecteur de fondamental, le trait de voisement retenu a été la présence d’une barre de voisement sur le spectrogramme (cf. figure C.13) ;
- pas de détermination du timbre exact des voyelles orales ou des voyelles nasales, c’est-à-dire :
  - dans toutes les occurrences, le graphème “un” a été transcrit [ œ̃ ],
  - dans toutes les occurrences, le schwa a été transcrit [ ə ],
  - comme le logiciel ne possédait pas de symboles pour les archiphonèmes ( / ø /, / œ / ), ( / o /, / ɔ / ) et ( / e /, / ε / ), nous avons transcrit les syllabes, pour lesquelles les oppositions n’étaient pas neutralisées en un seul phonème, selon la transcription du “Petit Robert”,
  - pas de prise en compte des harmonisations vocaliques orales ou nasales ;
- quelle que soit sa réalisation allophonique, vibrante ou fricative uvulaire, le phonème / r / a été transcrit [ R ].



Du point de vue de la transcription acoustique, nous avons essayé de coder le maximum de particularités des locuteurs. Puisque la transcription acoustique découle de l'observation du signal de parole, ses éléments sont décrits dans le prochain paragraphe.

VOYELLES			CONSONNES		
Snorri	API		Snorri	API	
/ i /	/ i̥ /	p <u>i</u> e	/ p /	/ p /	p <u>i</u> pe
/ e /	/ e /	p <u>e</u> tale	/ t /	/ t /	t <u>i</u> tre
/ ai /	/ ɛ /	p <u>e</u> re	/ k /	/ k /	k <u>i</u> lo
/ a /	/ a /	p <u>a</u> tte	/ b /	/ b /	b <u>i</u> bble
	/ ɑ /	p <u>a</u> te	/ d /	/ d /	d <u>i</u> x
/ ) /	/ ɔ /	p <u>o</u> rt	/ g /	/ g /	g <u>u</u> i
/ o /	/ o /	p <u>o</u> u	/ f /	/ f /	f <u>i</u> lm
/ u /	/ u /	p <u>o</u> u	/ s /	/ s /	s <u>i</u> x
/ y /	/ y /	p <u>u</u> r	/ ch /	/ ʃ /	ch <u>i</u> chi
/ eu /	/ ø /	p <u>eu</u>	/ v /	/ v /	v <u>i</u> vre
/ œ /	/ œ /	p <u>eu</u> r	/ z /	/ z /	z <u>i</u> zanie
/ & /	/ ə /	p <u>e</u> t	/ gh /	/ ʒ /	g <u>i</u> rafe
/ in /	/ ẽ /	p <u>a</u> in	/ l /	/ l /	l <u>i</u> las
/ an /	/ ɑ̃ /	p <u>e</u> nte	/ R /	/ r /	r <u>i</u> re
/ on /	/ ɔ̃ /	p <u>o</u> nt	/ m /	/ m /	m <u>i</u> me
/ un /	/ œ̃ /	<u>u</u> n	/ n /	/ n /	n <u>i</u> d
SEMI-CONSONNES			/ nj /	/ ɲ /	dign <u>i</u> té
Snorri	API			/ ŋ /	camp <u>i</u> ng
/ j /	/ j̥ /	p <u>i</u> d			
/ w /	/ w /	p <u>o</u> is			
/ ui /	/ u̥ /	p <u>u</u> it			
DIVERS					
Snorri	signification		Snorri	signification	
/ # /	pause		/ !! /	creaky voice	
/ " /	burst		/ ? /	coup de glotte	
/ hh /	souffle		/ ! /	bruit	
/ ?? /	inconnu				

Table C.3. Les symboles utilisés dans l'étiquetage manuel réalisé avec le logiciel SNORRI. La correspondance avec les symboles de l'Alphabet Phonétique International est également fournie.

### 4.3. Les critères de segmentation

Dans SNORRI, l'alignement de la transcription sur le signal de parole s'effectue par l'indication d'une succession de frontières de segments sur le spectrogramme. Une frontière est désignée à l'aide de la souris, puis elle est positionnée automatiquement au début du prélèvement le plus proche ( $\pm 4$  ms).

Bien que manuelle, la segmentation a été conçue de manière à utiliser un nombre maximal de critères définis a priori pour tenter :

- d'obtenir une segmentation répétitive proche de celle des systèmes d'étiquetage semi-automatique. Ceci est impératif lorsqu'il y a une hésitation entre plusieurs frontières ou lorsqu'il n'y a pas de frontière évidente ;
- de remédier aux hésitations dues au manque d'expérience de l'étiqueteuse mais aussi à l'évolutivité de son étiquetage due à l'acquisition progressive d'expérience (si minime soit-elle !) ;
- de faciliter l'extraction automatique des paramètres sélectionnés ;
- d'étiqueter le maximum de particularités des locuteurs ;
- de pallier l'insuffisance de la transcription, aussi bien au niveau de la hiérarchisation qu'au niveau de la variété des symboles. Comme nous le verrons, cela peut conduire dans certains cas à quelques incohérences.

Ces critères ne doivent pas être considérés comme des règles générales d'étiquetage d'une part parce que nous ne sommes pas une spécialiste du domaine, d'autre part parce que nous les avons établis en fonction des paramètres que nous souhaitons étudier. La plupart de ces critères ont été définis par rapport à l'observation du spectrogramme de parole. Ils dépendent donc de l'algorithme d'affichage du spectrogramme couleur (cf. page 21).

Nous n'allons pas décrire tous les critères que nous avons appliqués. Certains phonèmes, comme les consonnes nasales ou les fricatives, ne présentent aucune difficulté d'alignement. Dans ce cas, les critères utilisés se fondent surtout sur la présence de discontinuités majeures et nous n'avons pas jugé nécessaire de les expliciter. Les critères que nous allons présenter sont regroupés selon les grandes classes phonétiques auxquelles ils s'appliquent.

#### 4.3.1. Les voyelles orales

Nous avons considéré quelques particularités de segmentation des voyelles orales.

##### *a) Attaque de la voyelle après une pause*

Nous avons distingué deux cas singuliers :

- le début de la voyelle peut être dévoisé (pas de barre de voisement). Même dans ce cas, il est inclus dans la voyelle ;
- une barre d'explosion peut apparaître au début de la voyelle :
  - si elle est accolée aux formants de la voyelle, elle est étiquetée comme un "coup de glotte",
  - si elle est séparée des formants de la voyelle, elle est étiquetée comme un "bruit".

*b) Fin d'une voyelle orale devant une occlusive*

Cette frontière est importante pour le calcul des rapports entre les tenues des occlusives sourdes et voisées. Même si la mesure des tenues pourra être plus tard automatisée, nous avons essayé de délimiter de la même façon la fin des voyelles orales devant les occlusives sourdes et devant les occlusives sonores. Or, la fin d'une voyelle peut se confondre avec la barre de voisement d'une occlusive sonore.

Cette segmentation prend donc en compte la fin des formants F1, F2 et F3 sur le spectrogramme :

- si F1 se prolonge plus que F2 et F3, ce qui est le cas le plus commun, la fin de la voyelle est positionnée à la fin de F2 et F3. De nombreuses applications de cette règle de segmentation apparaissent sur les figures C.9, C.13 et C.15.

De plus, si la fin de F2 et F3 est suivie d'une barre d'explosion plus ou moins proche mais qui se trouve avant la fin de F1, une étiquette "bruit" est ajoutée à la suite de la voyelle ;

- si F1 se termine avant la fin de F2 et F3, la fin de la voyelle est positionnée à la fin de F1 et le dépassement de F2 et F3 est étiqueté :
  - comme un "bruit", s'il comprend une barre d'explosion,
  - comme un "souffle" sinon, comme le montre la figure C.10.

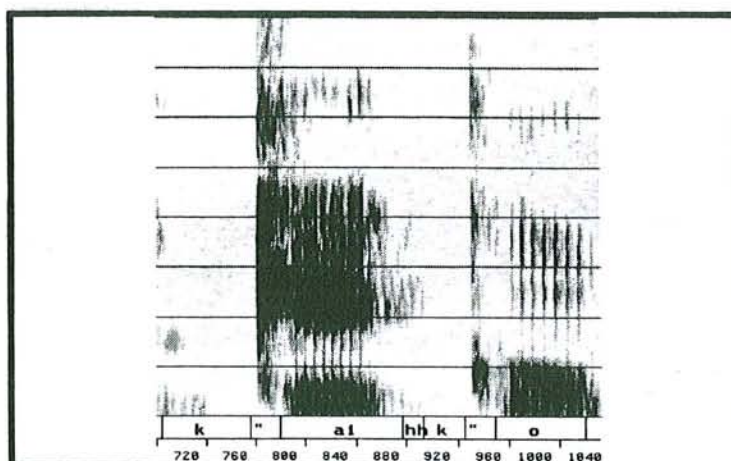


Figure C.10. Exemple de voyelle se terminant par un souffle pour le locuteur jlc.

*c) Fin d'une voyelle avant une pause*

Nous nous limitons ici au cas de la syllabe finale ouverte, celui de la syllabe finale fermée sera envisagé dans le paragraphe consacré à la segmentation du [ R ].

La fin de la voyelle est déterminée par la fin de F1. Si les formants d'ordre supérieur se prolongent, ils sont étiquetés comme du "souffle" (cf. figure C.9).



### 4.3.2. Les occlusives

Du point de vue acoustique, nous avons vu qu'une occlusive se caractérise par une tenue suivie d'une explosion, suivie dans certains cas d'une friction. Le logiciel SNORRI propose deux types de symboles pour étiqueter les occlusives, l'étiquette du phonème ([ b, d, g, p, t, k ]) qui permet d'étiqueter la tenue de l'occlusive et l'étiquette "burst" qui permet d'étiqueter l'explosion et ce qui la suit. Dans certains cas, nous avons ajouté l'étiquette "bruit", pour indiquer ce que nous croyions être des phénomènes particuliers.

Nous allons définir l'alignement de ces trois étiquettes.

#### a) Alignement de l'étiquette "burst"

Nous avons distingué les cas suivants :

- l'occlusive est suivie d'une voyelle.

L'étiquette "burst" comprend la barre d'explosion lorsqu'elle est présente. Elle comprend également le bref silence ou la friction qui peuvent suivre la barre d'explosion, même si cette dernière correspond à une partie dévoisée et bruitée de la voyelle. Les figures C.9, C.10 et C.12 présentent la segmentation de plusieurs "bursts", en particulier avec des voyelles partiellement dévoisées.

Si toute la voyelle est dévoisée, la plus grande partie de celle-ci est affectée à l'étiquette "burst". Le reste permet d'indiquer la présence de la voyelle. Ceci n'est pas satisfaisant car ce dernier critère est incohérent avec le précédent. Mais c'est la seule façon de noter dans une transcription unique la présence d'une explosion, d'un bruit de friction, de la voyelle [ i ] et de son dévoisement partiel ou total. Une transcription à deux niveaux, phonologique et acoustique, nous aurait permis de le faire de manière plus cohérente ;

- l'occlusive est suivie d'une liquide :
  - lorsque la liquide ([ R ] ou [ l ]) est partiellement dévoisée, la partie dévoisée est incluse dans l'étiquette "burst" et la partie voisée est étiquetée [ R ] ou [ l ] (cf. figure C.12),
  - lorsque toute la liquide est dévoisée, une partie du bruit est affectée au "burst" et l'autre partie à la liquide. Bien sûr, nous pouvons faire la même remarque d'incohérence que dans le cas des voyelles.

#### b) Segmentation de la tenue

Pour certains locuteurs, les formants d'ordre supérieur des voyelles orales se prolongent pendant la tenue de l'occlusive sonore suivante. Dans certains cas, comme celui de la figure C.11, ce prolongement atteint même la barre d'explosion. Compte tenu de la règle de segmentation des voyelles orales, la tenue de l'occlusive est réduite au minimum mais elle existe toujours pour signaler la présence de l'occlusive. Deux hypothèses sont avancées pour expliquer ce type de phénomène.

Selon la première, il pourrait être dû à la présence d'un écho dans la salle d'enregistrement. Si cette hypothèse est vraie, le phénomène devrait être indépendant de l'occlusive et de la voyelle. En revanche, cette supposition serait étayée par le prolongement anormalement long de la barre de voisement des voyelles orales pendant la tenue des occlusives sourdes de certains locuteurs. De même, le segment "souffle" qui apparaît à la fin de plusieurs voyelles qui précèdent des consonnes sourdes a peut-être la même origine.

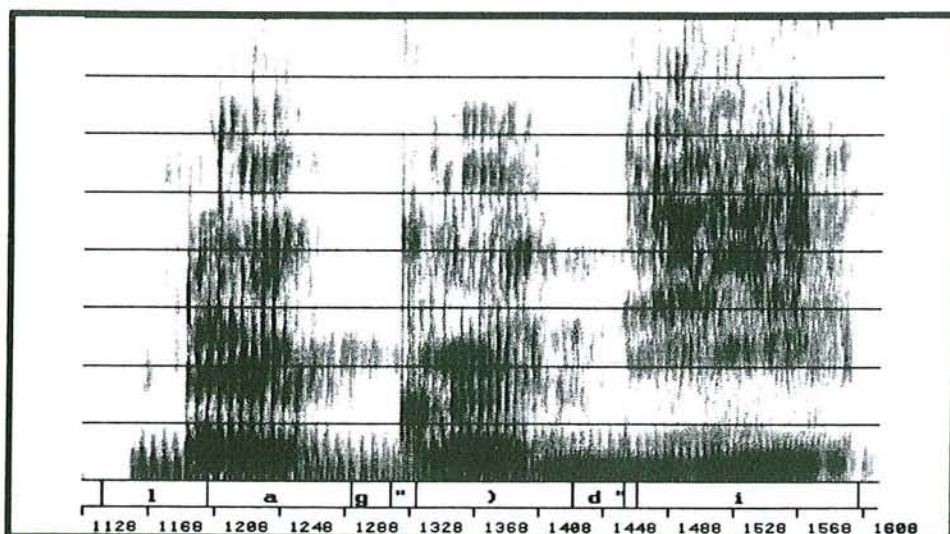


Figure C.11. Exemples de formants qui se prolongent pendant la tenue d'une occlusive sonore pour le locuteur gm.

L'autre hypothèse considère que le conduit vocal conserve la forme adoptée pour la voyelle et résonne encore pendant la tenue de l'occlusive. Cette résonance serait transmise aux molécules d'air par l'intermédiaire des parois. Dans ce cas, le phénomène devrait dépendre des couples [ **voyelle-occlusive** ]. Cette dernière hypothèse pourrait expliquer la présence de formants en début de phrase pendant la tenue d'une occlusive sonore.

Quelle que soit l'hypothèse valide, une autre question demeure. Pourquoi ce phénomène ne s'observe-t-il que chez certains locuteurs ? Est-il une fonction de l'effort vocal, de l'état des parois du conduit vocal, de la position du locuteur par rapport au microphone ? Nous n'avons pas eu le temps d'étayer ces considérations ni de répondre à ces questions. Pour cela, il faudrait au minimum étudier d'autres corpus et enregistrer les mêmes locuteurs dans d'autres conditions.

### c) ajout de l'étiquette "bruit"

Au cours de notre étiquetage, nous avons quelquefois relevé sur le spectrogramme la présence de barres d'explosion supplémentaires pendant la tenue des occlusives. Même si leurs origines sont sûrement très diverses (conduit vocal, bruits d'origine mécanique ou électronique, ...), nous avons décidé de les intégrer dans notre transcription afin de conserver une trace de leur observation. Leur étiquetage a été effectué différemment selon leur emplacement par rapport à la barre d'explosion normale de l'occlusive, c'est-à-dire la plus proche du phonème suivant. Pour cela, nous nous sommes servie du zoom spectrographique. Pour plus de clarté, nous appelons ces barres d'explosion des bruits dans la description suivante :

- si le bruit est éloigné de la barre d'explosion de plus de deux prélèvements, il est étiqueté comme un bruit. L'étiquetage obtenu dépend de l'emplacement exact du bruit, [ a ! t '' ], [ a # ! t '' ] ou [ a t ! '' ] ;
- si le bruit est relié à la barre d'explosion de l'occlusive, il est inclus dans le segment "burst" ;



• si le bruit est très voisin de la barre d'explosion sans y être relié, nous avons examiné deux éventualités :

- si l'occlusive est vélaire, / **k** / ou / **g** /, et si ce phénomène semble correspondre à la double explosion qui est citée dans tout bon cours de phonétique, la double barre d'explosion est incluse dans l'étiquette "burst". Dans les autres cas, lorsqu'il y a par exemple trois barres d'explosion, l'étiquette "bruit" est ajoutée,
- si l'occlusive est bilabiale ou apico-dentale, le bruit est étiqueté "bruit" pour signaler cet épiphénomène. Un exemple est présenté sur la figure C.12.

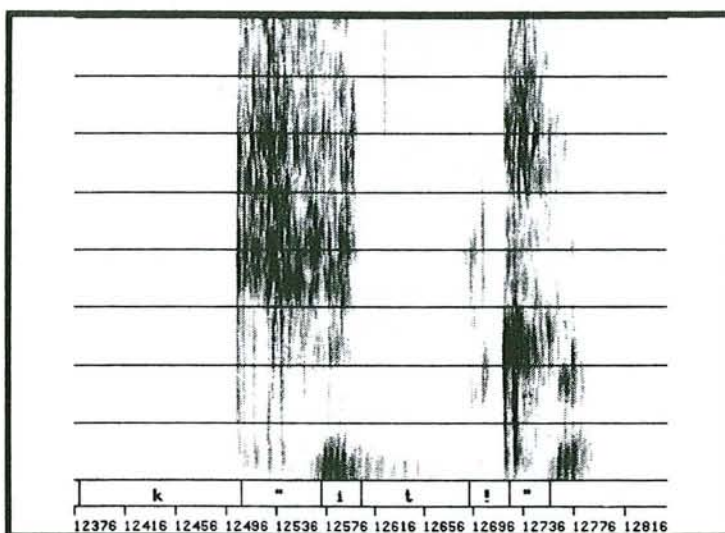


Figure C.12. Exemple de bruit avant la barre d'explosion d'un [ **t** ] pour le locuteur df.

#### 4.3.3. Le phonème / **r** / en contexte vocalique

Nous avons établi des critères de segmentation différents en fonction du contexte phonologique du phonème / **r** / et des paramètres que nous souhaitons étudier. En effet, ce phonème n'a pas la même structure acoustique selon qu'il se trouve en contexte intervocalique, qu'il ferme une syllabe suivie d'une pause ou une syllabe suivie d'une autre consonne. Par ailleurs, dans le premier contexte, c'est le / **r** / qui fera l'objet d'une étude, alors que, dans les autres, c'est la voyelle de la syllabe.

Les figures C.13 et C.14 regroupent les différents cas de segmentation du / **r** / en contexte vocalique.

##### a) *segmentation du triplet [ voyelle **R** voyelle ]*

L'ensemble / **voyelle-R-voyelle** / comporte de longues transitions comparativement aux parties stables des phonèmes. Les frontières peuvent être positionnées au début, au milieu ou à la fin de ces transitions. Ces triplets serviront à l'analyse spectrale du / **R** /, sauf dans de rares cas où ils font aussi partie du triplet [ **p-voyelle-R** ] (phrase 4 par certains locuteurs au débit rapide). Aussi a-t-il été décidé d'inclure les transitions dans les voyelles environnantes et de



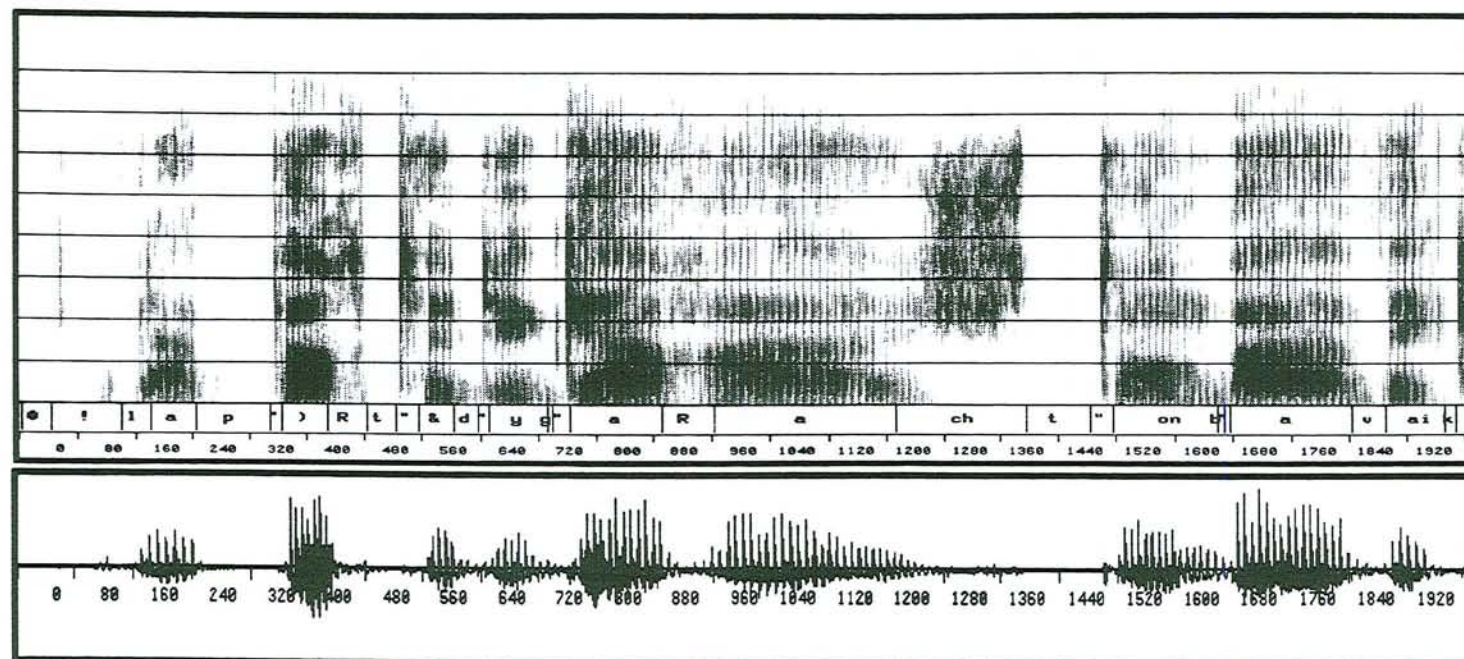


Figure C.13. Exemples de segmentation du triplet [ **voyelle-R-voyelle** ], du triplet [ **p-voyelle-R** ] et du [ **l** ] en début de phrase, dans la phrase "*la porte du garage tomba avec lourdeur*" prononcée par le locuteur jmp.

limiter le [ R ] à sa partie fricative (cf. figure C.13). Il sera toujours possible lors de l'étude du [ R ] d'intégrer automatiquement une partie de la voyelle dans l'analyse. La réciproque serait plus complexe.

*b) la fin d'une voyelle dans les triplets [p-voyelle-R] (ou [b-voyelle-R])*

L'objectif de ce triplet est l'étude des formants des voyelles orales. La plupart des triplets sont suivis d'une frontière syntaxique ou d'une tenue d'occlusive. Nous avons utilisé deux critères de segmentation :

- lorsque le triplet [p-voyelle-R] termine une phrase ou un groupe de phonation, la fin de la voyelle est déterminée par la fin du formant F1 de la voyelle quand ce dernier remonte vers celui du [ R ] et que la barre de voisement disparaît (cf. figure C.14). Préalablement, le segment de parole situé en fin de phrase est écouté pour déterminer s'il s'agit d'un [ R ], d'un souffle ou d'un silence. Il est bien sûr étiqueté en conséquence ;
- dans les autres cas, le [ R ] est souvent vocalique et très difficile à séparer de la voyelle qui le précède. Le positionnement de la frontière s'est surtout fait grâce à l'écoute du segment acoustique, au changement de timbre de la voyelle (cf. figure C.13).

#### 4.3.4. Le phonème / l /

Les contextes phonologiques qui restent à envisager sont la position intervocalique et le début de phrase.

*a) segmentation du triplet [ voyelle-l-voyelle ]*

L'objectif du / l / en position intervocalique étant le même que celui du / r / dans le même contexte, les critères de segmentation sont identiques. Le [ l ] est réduit à sa partie minimale tout en incluant les éventuelles barres d'explosion.

*b) en début de phrase*

La limite déterminée sur le spectrogramme est confirmée par l'écoute. Tout claquement de langue qui est relié au [ l ] en fait partie, sinon il est étiqueté comme un "bruit".

#### 4.3.5. Les voyelles nasales

Si la fin d'une voyelle nasale suivie d'une occlusive sourde correspond simplement à la fin de son premier formant et de la barre de voisement sur le spectrogramme, il est plus difficile de déterminer la limite entre la fin d'une voyelle nasale et la barre de voisement d'une occlusive sonore. L'application du critère établi pour les voyelles orales amputerait la voyelle nasale d'une partie qui pourrait se révéler pertinente pour la caractérisation du locuteur. Par exemple, certains locuteurs terminent leurs voyelles nasales par une consonne nasale. Nous avons donc retenu comme principe d'avoir une voyelle nasale la plus longue possible tout en cherchant à la séparer du murmure de l'occlusive. Pour cela, nous avons essayé de repérer l'arrêt de la décroissance de l'amplitude du signal temporel, lorsqu'il existe. En effet, parfois, cette amplitude décroît constamment jusqu'à l'explosion de l'occlusive. Dans ce cas, le critère des voyelles orales a été appliqué.

La figure C.14 présente la segmentation d'une voyelle nasale dans des contextes sourd et sonore.

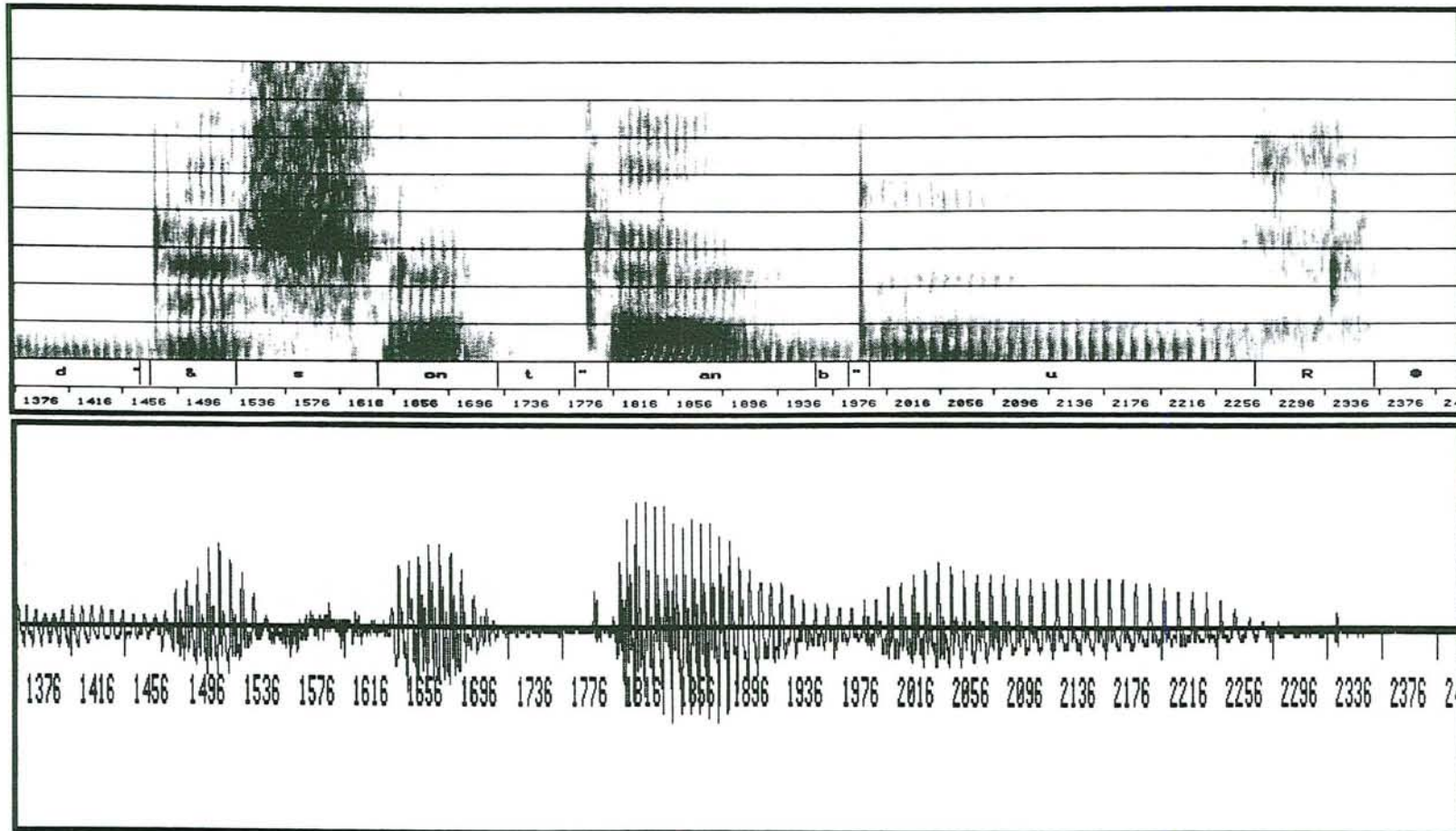


Figure C.14. Segmentation des voyelles nasales et du [ R ] en fin de phrase pour le locuteur aq.



#### 4.3.6. Les semi-voyelles et les couples de voyelles

Les trajectoires formantiques entre une semi-voyelle et une voyelle ne présentent aucune discontinuité susceptible de faciliter la pose d'une frontière. Il nous est difficile de formaliser les critères que nous avons retenus. Souvent, nous avons émis une hypothèse plutôt fondée sur les formants situés au-dessus de 2500 Hz. Puis, nous avons essayé de vérifier cette frontière à l'écoute. Dans sa segmentation automatique, le logiciel CASPAR du MIT [Zue 88] affecte arbitrairement le premier tiers du couple (semi-voyelle, voyelle) à la semi-voyelle et les deux tiers restants à la voyelle. L'observation de la figure C.9 montre que cela correspond approximativement à notre segmentation.

Nous avons adopté le même mode de segmentation pour la pose d'une frontière entre deux voyelles.

#### 4.3.7. Le schwa bref et le schwa épenthétique

Il est possible d'observer sur les spectrogrammes de parole à la fin de certaines consonnes une courte structure vocalique qui ressemble à un [ə]. Selon le contexte, ce petit schwa correspond à un "e muet" qui n'a pas été complètement éliminé ou bien à la vocalisation de la fin d'une consonne. Faut-il étiqueter ce segment de parole à part ou bien l'inclure dans la consonne ? Selon son contexte phonologique, nous avons fait les choix suivants :

- s'il est présent entre deux consonnes susceptibles de faire partie d'une assimilation, il est étiqueté quelle que soit sa durée car sa présence empêche cette assimilation ;
- s'il se situe à la fin d'un [R], il est étiqueté comme faisant partie du [R] ;
- s'il suit une consonne nasale :
  - si sa durée est supérieure à la moitié de la durée de la consonne nasale et s'il est perceptible lors d'une écoute large (plusieurs phonèmes), il est étiqueté séparément comme [ə],
  - sinon il est inclus dans l'étiquette de la consonne.

#### 4.3.8. Les phénomènes de "voix craquée" et de "friture vocale"

Nous avons introduit, au paragraphe II.3.5.3 de la première partie, la voix craquée (*creaky voice*) et la friture vocale (*vocal fry*) qui sont selon les auteurs deux appellations du même phénomène ou deux phénomènes distincts. Dans notre corpus, ces phénomènes semblent se produire dans deux circonstances, soit entre deux voyelles (par exemple, / t̥ɔ̃baavɛk /) soit en fin de phrase lorsque la fréquence fondamentale décroît.

Nous avons signalé ce phénomène lorsqu'il se situe entre deux voyelles grâce à l'étiquette "creaky voice". En revanche, en fin de phrase, nous ne l'avons pas fait apparaître dans la transcription parce que nous nous sommes heurtée à deux difficultés. La première d'entre elles est notre manque de compétence dans la reconnaissance du phénomène et surtout dans la détermination de son commencement. Quel critère faut-il utiliser ? L'apparition des doubles impulsions, la valeur de la période glottale dont l'accroissement semble linéaire, l'écoute ? L'autre difficulté déjà évoquée tient à l'unicité de la transcription. Comment interclasser l'étiquette "creaky voice" avec les étiquettes de phonèmes, alors que le phénomène peut recouvrir tout un

phonème voire plusieurs phonèmes. A notre avis, l'étiquette "creaky voice" correspond à un diacritique qui aurait dû apparaître à un deuxième niveau de transcription.

La figure C.15 présente le phénomène de voix craquée ou de friture vocale à la fin de la phrase 4 prononcée par le locuteur jg.

#### 4.4. Conclusion

Nous avons étiqueté un quart de notre corpus, manuellement et aux frontières, avec un logiciel d'étiquetage qui ne disposait que d'un seul niveau de transcription et de segmentation. A partir de ces conditions, nous avons essayé de trouver un compromis entre un étiquetage normatif, qui facilite l'extraction automatique des paramètres, et un étiquetage fin, qui mémorise les particularités du signal de parole et des locuteurs. Ce compromis a conduit à l'établissement d'un certain nombre de critères de segmentation qui, rappelons-le, ne sont pas des règles générales d'étiquetage.

Nous avons utilisé des étiquettes acoustiques comme le "burst", le "souffle", ou le "bruit" pour marquer certaines spécificités des locuteurs. Ces étiquettes ne gênent en rien l'extraction automatique des segments de parole comme les voyelles orales ou les occlusives. Il suffit lors de cette extraction de leur associer un attribut de transparence. En revanche, la prise en compte dans la transcription de la prononciation réelle et, en particulier, des règles d'assimilation de sonorité oblige la procédure d'extraction à envisager tous les cas possibles. Mais, la prise en compte de la prononciation réelle est de toute façon une nécessité. Ainsi, avant de mesurer la tenue d'une occlusive sourde, il faut déjà s'assurer qu'elle n'a pas subi une assimilation de sonorité.

Avec le recul, nous pensons qu'il aurait été préférable d'effectuer un étiquetage à trois niveaux : un étiquetage normatif, un étiquetage de ce qui a été réellement prononcé et un marquage des particularités acoustiques (dévoisement, bruits, parties insegmentables, creaky voice, ...). Mais, pour être efficace, un tel étiquetage doit être fait par un expert et consomme énormément de temps.

## 5. Conclusion sur l'élaboration et l'étiquetage du corpus

Suite à la sélection des paramètres susceptibles d'être pertinents pour la caractérisation automatique du locuteur, nous avons élaboré un corpus de dix-sept phrases. Ce corpus a été lu quatre fois au cours de la même session par dix-huit locuteurs et vingt-et-une locutrices. Par construction, ce corpus est fortement déséquilibré du point de vue phonétique. Par exemple, il comprend très peu de fricatives et ne contient aucune occurrence du phonème / f /.

Nous avons étiqueté manuellement un quart du corpus correspondant à dix locuteurs masculins. Cet étiquetage a nécessité l'établissement de règles de transcription et de critères de segmentation. Cela nous a permis de nous intéresser à la problématique générale de l'étiquetage ainsi qu'aux bases de données de parole. Du point de vue plus général des sciences de la parole, l'étiquetage manuel est très enrichissant pour un non phonéticien. Son seul inconvénient est d'être long.



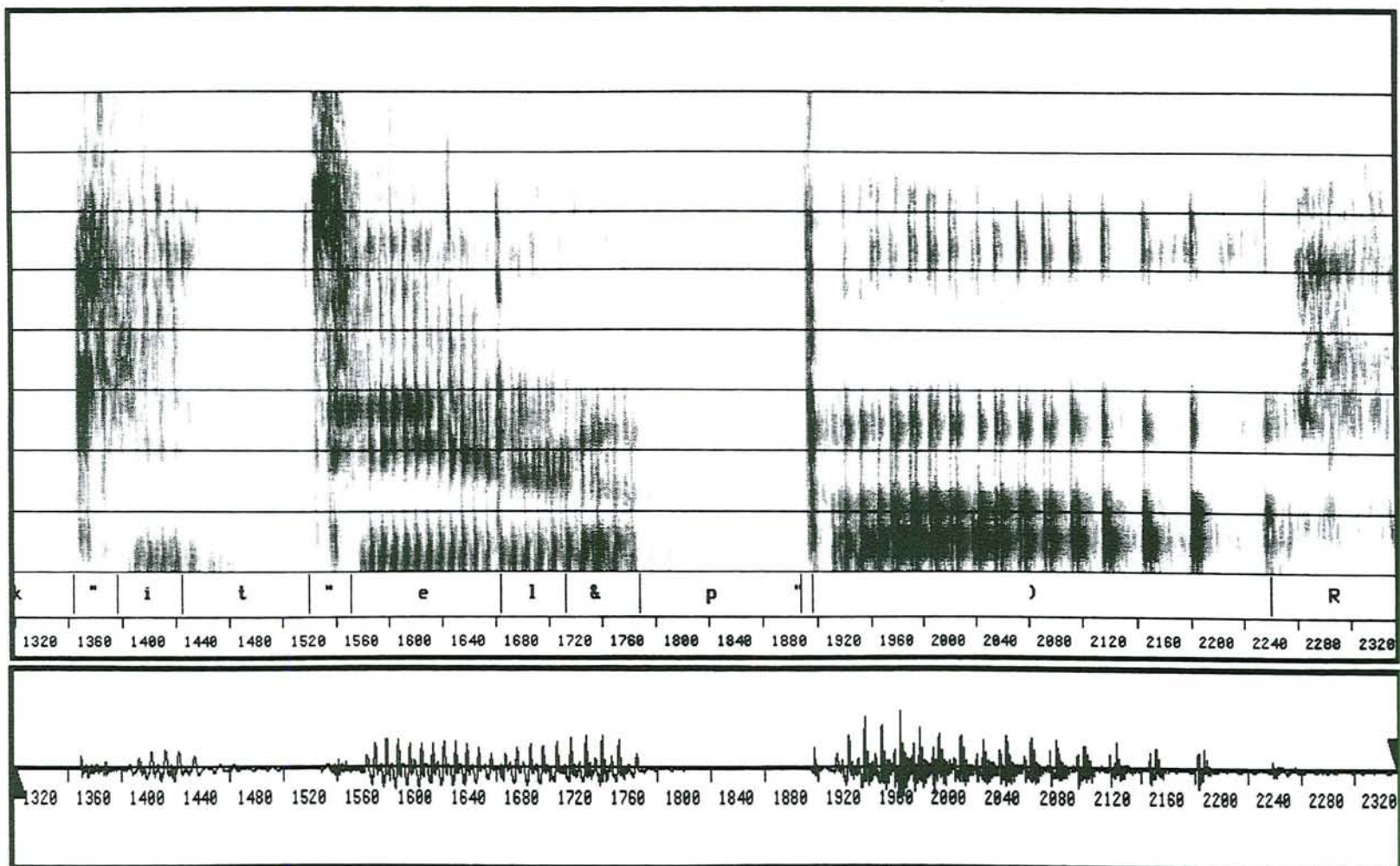


Figure C.15. Le phénomène de voix craquée ou de friture vocale à la fin de la phrase "le bateau à vapeur a quitté le port" prononcée par le locuteur jg.



Dans notre cas, cet inconvénient a été compensé par le fait que notre corpus a été utilisé dans de nombreuses recherches sur la reconnaissance automatique de la parole entreprises au sein de notre laboratoire. En effet, notre corpus a eu le mérite d'être, à une certaine époque, le seul corpus français de parole continue multilocuteur étiquetée, à côté de quelques répétitions du texte "la bise et le soleil" de BDSONS [Carre 84].

Ces recherches concernent aussi bien les systèmes de décodage acoustico-phonétique [Fran-"c -c""ois 90] [Laprie 90] que la reconnaissance de la parole continue [Gong 91], l'application des réseaux connexionnistes à la parole [Dingeon 89] [Fran-"c -c""ois 92], l'étiquetage semi-automatique [Gong 92] ou la reconnaissance automatique du locuteur [Gong 90].

## CHAPITRE III

### PERTINENCE DES TROIS PREMIERS FORMANTS DES VOYELLES ORALES

## 1. Introduction

Après avoir étiqueté une partie de notre corpus, nous avons entrepris l'étude de la pertinence des premiers paramètres fréquentiels sélectionnés, c'est-à-dire les trois premiers formants des voyelles /  $\epsilon$  /, /  $\text{œ}$  /, /  $\text{ɔ}$  /, /  $\text{a}$  /, et /  $\text{i}$  /, précédées d'un contexte neutre au sens de la coarticulation linguale, /  $\text{p}$  /, et suivies d'un contexte postérieur allongeant, /  $\text{R}$  /.

Nous avons décidé d'analyser aussi les triplets /  $\text{peR}$  / et /  $\text{puR}$  / présents dans le corpus, afin de vérifier que les voyelles /  $\text{e}$  / et /  $\text{u}$  / ne sont pas plus discriminantes que celles qui ont été choisies. Nous avons également complété l'ensemble des triplets /  $\text{p-voyelle-R}$  / par les triplets /  $\text{b-voyelle-R}$  / qui étaient disponibles dans les dix-sept phrases. Les occurrences de tous les triplets étudiés sont soulignées dans la table C.4. La voyelle orale du triplet a été étudiée dans tous les cas, même lorsque le /  $\text{R}$  / final a été élide ou remplacé par du souffle.

Cette étude a comporté deux phases qui seront développées dans les deux prochaines sections de ce chapitre. Dans un premier temps, nous avons déterminé automatiquement les trois premiers formants des voyelles orales retenues. Puis, nous avons étudié quelles étaient les voyelles orales les mieux adaptées à la reconnaissance automatique du locuteur, en essayant de déterminer, pour chacune d'elles, quels étaient les formants ou les écarts entre formants les plus discriminants.

## 2. Détermination des trois premiers formants des voyelles orales

### 2.1. Introduction

L'utilisation des formants des voyelles orales en reconnaissance automatique du locuteur impose d'avoir des formants robustes. Pour cela, il est nécessaire, d'une part, que la comparaison des locuteurs s'effectue avec des formants établis dans le même contexte, d'autre part, que la méthode de détermination des formants soit la plus fiable possible. Afin d'atteindre ce dernier objectif, nous avons déterminé un formant, que nous appellerons "formant final", à partir de trois formants, que nous appellerons "formants intermédiaires", obtenus à trois emplacements voisins dans la voyelle. Chacun de ces formants intermédiaires est lui-même déterminé grâce à une méthode automatique d'affectation des pôles issus d'une analyse LPC qui tient compte de la voyelle étudiée. De plus, nous avons associé à chacun des formants finaux  $F_1$ ,  $F_2$  et  $F_3$  un coefficient de défiance mis à jour au cours des diverses étapes de détermination.

- 1 Guy a péri bêtement du diabète en Italie.
- 2 La porte du garage tomba avec lourdeur.
- 3 La partie de belote dura toute la matinée.
- 4 Un bateau à vapeur a quitté le port.
- 5 Le petit gamin traîne un jouet.
- 6 Donne-moi le bocal de cacao !
- 7 En ski, la godille permet d'éviter les tournants.
- 8 Un coq bien dodu pour demain !
- 9 Lequel des bandits guette près du repère ?
- 10 Le trappeur commun redoutait le loup-garou.
- 11 Douze nains conspirent derrière le bosquet.
- 12 Le soldat brisa la baguette de son tambour.
- 13 Goûtez-moi ce cake au beurre !
- 14 Le rire de la gouvernante est revigorant.
- 15 La cousine du nain soupire dans son délire.
- 16 Le départ de la course Strasbourg-Paris aura du retard.
- 17 Notre guide charmant quitte la jolie route danoise.

Table C.4. Les occurrences des triplets / **p-voyelle-R** /  
et / **b-voyelle-R** / dans les dix-sept phrases du corpus.

Nous allons décrire ces étapes dans les paragraphes suivants. Dans un premier temps, nous détaillerons la première méthode d'affectation des pôles LPC aux formants que nous avons conçue et que nous appellerons "étude préliminaire". A l'issue de celle-ci, nous avons vérifié les formants obtenus à l'aide des spectrogrammes de parole des triplets et aux pôles bruts issus de l'analyse LPC. Ceci nous a conduit à l'élaboration d'une méthode similaire mais plus robuste, que nous décrirons également. Puis, nous terminerons par la présentation de l'établissement des formants finaux à partir des trois formants intermédiaires.

Dans toutes ces étapes, un formant  $F_i$ , intermédiaire ou final, est une structure constituée :

- d'une fréquence formantique,  $F_i.fr$ ,
- d'une largeur de bande,  $F_i.bw$ ,
- d'un coefficient de défiance,  $F_i.df$ ,

Le coefficient de défiance  $df$  est composé de trois champs indépendants  $df_1$ ,  $df_2$  et  $df_3$ . Les champs  $df_1$  et  $df_2$  sont mis à jour au cours des différentes phases de détermination des formants alors que  $df_3$  l'est lors du calcul des distances entre locuteurs. Plus le coefficient de défiance est élevé, moins la valeur du formant et la distance calculée à partir de cette valeur sont fiables.



## 2.2. Etude préliminaire

### 2.2.1. Méthodologie de détermination des formants de la voyelle

Lors de cette première étude, nous avons déterminé, pour chacune des voyelles étudiées, les formants du prélèvement situé à 40% du début de la voyelle selon la méthodologie ci-dessous. Nous avons choisi cet emplacement afin de nous situer dans une partie assez stable de la voyelle. Celle-ci doit donc être suffisamment éloignée des deux consonnes pour ne pas être dans une zone de transitions formantiques, tout en demeurant dans une zone assez intense de la voyelle.

Dans un premier temps, une analyse LPC fondée sur la méthode d'autocorrélation de Durbin [MAR, 76] fournit dix-huit coefficients à partir desquels sont extraites toutes les racines LPC ayant une largeur de bande inférieure ou égale à 1000 Hz. Ces racines sont triées dans l'ordre croissant de leurs fréquences.

Puis, un premier filtrage des pôles LPC ainsi obtenus permet d'éliminer ceux dont les fréquences sont trop voisines. Si l'écart fréquentiel entre deux pôles est inférieur à la valeur médiane de la fréquence fondamentale calculée sur la phrase contenant la voyelle, seul est conservé le pôle dont la largeur de bande est la plus petite.

En vue d'affecter les pôles LPC aux formants de la voyelle, un deuxième filtrage des pôles est effectué. Seuls sont conservés les pôles dont la fréquence appartient à l'union de cinq intervalles prédéfinis que nous appellerons "domaines de définition". Ces domaines de définition, notés  $D(F_i)$ , sont présentés dans la table C.5. Les trois premiers d'entre eux sont les intervalles fréquentiels dans lesquels sont censés se trouver les trois premiers formants de la voyelle lorsqu'elle est précédée des contextes /p/ et /b/, suivie du contexte /R/ et prononcée par un locuteur masculin. Ils ont d'abord été établis sur les conseils d'une phonéticienne du laboratoire puis affinés en fonction des premiers résultats obtenus. En ce qui concerne  $D(F_4)$ , nous avons effectué une évaluation très grossière de l'intervalle fréquentiel dans lequel est susceptible de se trouver  $F_4$ . L'existence de  $D(F_5)$  est purement algorithmique et ne sert qu'à définir la limite supérieure des fréquences considérées.

Après ces filtrages, trois des pôles restants sont affectés séquentiellement aux formants  $F_1$ ,  $F_2$  et  $F_3$  selon un algorithme qui tient compte, pour chaque formant  $F_i$ , dans un premier temps du nombre de pôles présents dans le domaine de définition  $D(F_i)$ , puis par ordre de priorité :

- du pôle affecté à  $F_{i-1}$ ,
- du pôle susceptible d'être affecté à  $F_{i+1}$ ,
- des largeurs de bande des pôles candidats.

Cet algorithme est le suivant :

- Procédure générale d'affectation des pôles LPC aux formants intermédiaires :

début

dans le cas

où  $D(F_i)$  ne contient aucun pôle candidat,  
le formant  $F_i$  est mis à zéro ;

où  $D(F_i)$  contient un seul pôle candidat,  
si il a déjà été affecté au formant  $F_{i-1}$ ,  
alors le formant  $F_i$  est mis à zéro,  
sinon il est affecté au formant  $F_i$ ,  
fsi ;

où  $D(F_i)$  contient deux pôles susceptibles d'être affectés au formant  $F_i$ ,  
si l'un des deux pôles a été affecté au formant  $F_{i-1}$ ,  
alors

si c'est le pôle ayant la plus grande fréquence formantique,  
alors le formant  $F_i$  est mis à zéro,  
sinon l'autre pôle est affecté à  $F_i$ ,  
fsi,

sinon

si le pôle ayant la plus grande fréquence formantique est le seul pôle candidat pour  $F_{i+1}$ ,  
alors l'autre pôle est affecté à  $F_i$ ,  
sinon on choisit le pôle qui a la plus petite largeur de bande,  
fsi,  
fsi ;

où  $D(F_i)$  contient trois pôles susceptibles d'être affectés au formant  $F_i$ ,  
un algorithme de choix similaire est appliqué ;

fcas ;

$F_i$  .bw est donnée par la largeur de bande du pôle retenu ;

si  $\left( F_i.bw \geq 250 + \frac{F_i.f_r}{10} \right)$ ,

alors le champ  $df_1$  du coefficient de défiance  $F_i$  .df vaut 1,

sinon le champ  $df_1$  du coefficient de défiance  $F_i$  .df vaut 0,

fsi.

fin.

Voyelle	D(F1)	D(F2)	D(F3)	D(F4)	D(F5)
i	200 - 450	1800 - 2600	2600 - 3400	2800-3800	3500-5500
e	300 - 500	1600 - 2350	2350 - 3250	2800-3800	3500-5500
ɛ	350 - 650	1450 - 2200	2200 - 3100	2800-3800	3500-5500
a	400 - 1000	1050 - 1500	2000 - 2800	2800-3800	3500-5500
ɔ	350 - 600	850 - 1400	2000 - 2700	2800-3800	3500-5500
u	220 - 420	600 - 1200	1800 - 2700	2800-3800	3500-5500
œ	350 - 600	1050 - 1700	2000 - 3000	2800-3800	3500-5500

Table C.5. Les domaines de définition D(Fi).

Une fois les trois premiers formants déterminés, tous les pôles dont la fréquence est comprise entre la borne supérieure du domaine de définition de F3 et 4500 Hz sont conservés.

Pour chacune des voyelles orales étudiées, les résultats de cet algorithme d'affectation sont présentés en annexe sous la forme d'un tableau par voyelle.

Celui-ci comprend pour chacune des répétitions des locuteurs :

- la valeur médiane de  $F_0$  sur la phrase qui contient la voyelle étudiée. Cette valeur est calculée à partir des mesures de la fréquence fondamentale sur tous les prélèvements de la phrase comprises entre 50 et 500 Hz ;
- les valeurs des fréquences et des largeurs de bande des formants F1, F2 et F3 ;
- les valeurs des fréquences et des largeurs de bande des pôles dont la fréquence est comprise entre la limite supérieure de D(F3) et 4500 Hz.

### 2.2.2. Vérification des fréquences formantiques

Toujours dans le but d'avoir des fréquences formantiques les plus fiables possible, nous avons souhaité vérifier les valeurs obtenues. Nous avons pour cela le choix entre deux méthodes :

- une comparaison manuelle avec les fréquences des pics d'un spectre instantané à bande étroite calculé au même emplacement. Cette méthode est la meilleure et la seule véritable vérification des fréquences formantiques mais elle est aussi très coûteuse en temps. De plus, elle revenait à réaliser une détermination manuelle des formants des voyelles ce qui rendait du même coup l'analyse LPC inutile. Par ailleurs, R.B. Monsen et A.M. Engebretson [Monsen 83] ont montré que, pour les voyelles de synthèse, la précision de la détermination spectrographique des fréquences formantiques par des experts — que nous ne sommes pas — n'était, dans la plupart des cas, pas meilleure que l'extraction des racines LPC ;
- une vérification plus globale fondée sur l'observation des spectrogrammes des triplets [ p-voyelle-R ] et [ b-voyelle-R ]. Si elle est moins précise, cette méthode est plus rapide et suffisante pour contrôler les valeurs aberrantes engendrées par l'analyse LPC, la correction de l'algorithme d'affectation et la validité des bornes des domaines de définition.



Nous avons donc choisi la seconde méthode que nous avons complétée en notant la largeur du formant sur le spectrogramme, ce qui donne une information sur sa largeur de bande, et surtout en notant l'évolution du formant au cours de la voyelle, ce qui fournit une indication sur la valeur des fréquences formantiques à un autre emplacement dans la voyelle. Nous avons également ajouté dans le tableau la durée absolue de la voyelle, sa durée relative par rapport à la durée vocalique moyenne et l'évolution de la fréquence fondamentale durant le triplet. Ces dernières informations fournissent une indication sur le caractère plus ou moins accentué de la voyelle.

Les résultats détaillés de cette étape de vérification figurent dans les tableaux que nous avons situés en annexe selon un formalisme dont la description précède les tableaux.

Nous avons complété cette vérification spectrographique par l'examen des pôles bruts issus de l'analyse LPC, notamment dans les cas où les formants trouvés par l'algorithme d'affectation semblaient faux par rapport à ce qu'indiquait le spectrogramme.

### 2.2.3. Conclusions de l'étape de vérification

La vérification des résultats de cette première étude nous a conduite à tirer un certain nombre de conclusions de notre méthodologie de détermination des formants des voyelles orales.

Tout d'abord, l'étape de vérification a mis en évidence le fait que certains pôles LPC qui étaient corrects par rapport au spectrogramme ont été éliminés parce que leurs fréquences formantiques étaient en dehors des domaines de définition. D'une manière plus générale, nous avons comparé les valeurs minimales et maximales des fréquences des formants F1, F2 et F3 trouvés — ou des pôles bruts lorsque les formants étaient faux — avec les bornes des domaines de définition. La table C.7 présente le résultat de cette comparaison pour chacune des voyelles. Toutes les valeurs en grisé correspondent à des fréquences formantiques valides mais qui sont situées en dehors de leurs domaines de définition. Ces derniers sont donc sous-dimensionnés. En revanche, certains autres apparaissent comme inutilement surdimensionnés. Nous avons donc modifié les limites de la plupart des domaines de définition de la table C.5 pour tenir compte de ces observations. Ces nouvelles limites sont présentées dans la table C.6.

Par ailleurs, nous avons pu remarquer sur les spectrogrammes que le prélèvement situé à 40% du début de la voyelle n'était pas toujours le meilleur emplacement pour une détermination optimale des formants de la voyelle. Lorsqu'elle se situe en fin de phrase, la voyelle est très longue et son énergie décroît rapidement. Par conséquent, le prélèvement situé à 40% se trouve déjà dans une zone d'énergie faible et pendant laquelle l'articulation est déjà relâchée. Les formants mesurés sont donc moins robustes et peuvent présenter une certaine déviation par rapport à ceux de la voyelle cible. Même si cette déviation peut constituer une caractéristique du locuteur, nous avons préféré l'éviter dans cette étude. D'un autre côté, lorsque les voyelles sont très courtes et précédées d'un bruit de friction, le prélèvement étudié peut se trouver dans une zone encore bruitée ou présentant des transitions formantiques. Ces derniers cas sont dus à un léger décalage des frontières de la voyelle provoqué par le défaut de précision du spectrogramme utilisé lors de l'étiquetage, associé à un positionnement automatique de la frontière à  $\pm 4$  ms (cf. page 25).

Deux autres conclusions portent sur l'algorithme d'affectation proprement dit.

La plus importante concerne le choix entre deux pôles LPC qui appartiennent au même domaine de définition. Ce choix fondé sur la plus petite largeur de bande peut engendrer des fréquences formantiques erronées lorsque les domaines de définition de deux formants

Voyelle	D(F1)	F <sub>1</sub> minimale	F <sub>1</sub> maximale	D(F2)	F <sub>2</sub> minimale	F <sub>2</sub> maximale	D(F3)	F <sub>3</sub> minimale	F <sub>3</sub> maximale
i_11	200 - 450	241	411	1800 - 2600	1849	2413	2600 - 3400	2683	3567
i_15	200 - 450	213	372	1800 - 2600	1830	2354	2600 - 3400	2503	3439
e_01	300 - 500	341	468	1600 - 2350	1681	2379	2350 - 3250	2380	3032
ɛ_07	350 - 650	352	535	1450 - 2200	1535	1950	2200 - 3100	2223	2748
ɛ_09	350 - 650	406	616	1450 - 2200	1511	1981	2200 - 3100	2298	2980
a_03	400 - 1000	544	725	1050 - 1500	1050	1593	2000 - 2800	2129	2753
a_16	400 - 1000	517	687	1050 - 1500	1105	1407	2000 - 2800	2073	2702
ɔ_02	350 - 600	405	564	850 - 1400	836	1163	2000 - 2700	2018	2703
ɔ_04	350 - 600	417	589	850 - 1400	838	1210	2000 - 2700	2014	2712
u_08	220 - 420	332	454	600 - 1200	607	1001	1800 - 2700	2011	2737
u_12	220 - 420	230	340	600 - 1200	647	816	1800 - 2700	1985	2643
u_16	220 - 420	276	390	600 - 1200	654	896	1800 - 2700	2008	2615
œ_04	350 - 600	405	557	1050 - 1700	1123	1525	2000 - 3000	2154	2661
œ_10	350 - 600	395	549	1050 - 1700	1138	1636	2000 - 3000	2131	2689
œ_13	350 - 600	401	542	1050 - 1700	1139	1671	2000 - 3000	2130	2802

Table C.7. Comparaison des bornes prédéfinies des domaines de définition avec les valeurs minimales et maximales des fréquences des formants F1, F2 et F3 trouvés, ou de celles des pôles bruts lorsque les formants sont considérés comme faux par rapport au spectrogramme. Les valeurs en grisé sont en dehors de leur domaine de définition. Les fréquences représentées en caractères gras ne sont pas vérifiables sur le spectrogramme.



consécutifs ne sont pas disjoints. Aussi avons-nous modifié cet algorithme en introduisant la notion de fréquence formantique de référence. Ces valeurs de référence sont une estimation des fréquences formantiques masculines moyennes pour les voyelles étudiées précédées d'un contexte bilabial et suivies d'un contexte postérieur. Nous avons choisi comme valeurs de référence les valeurs médianes des fréquences formantiques  $F_1$ ,  $F_2$  et  $F_3$  établies par F. Lonchamp dans son étude sur la variabilité interlocuteur des formants d'un corpus de triplets / **p-voyelle-R** / (cf. paragraphe V.4.2.1 de la partie A). Ces valeurs figurent également dans la table C.6.

Voyelle		F1	F2	F3	F4
i	domaine	200 - 450	1800 - 2500	2450 - 3600	3200 - 4200
	référence	308	2064	2976	3407
e	domaine	300 - 500	1600 - 2450	2350 - 3150	3100 - 4200
	référence	365	1961	2644	3362
ɛ	domaine	330 - 680	1450 - 2100	2150 - 3100	3050 - 4200
	référence	530	1718	2558	3300
a	domaine	450 - 800	1000 - 1600	2000 - 2850	2850 - 4200
	référence	684	1256	2503	3262
ɔ	domaine	350 - 650	750 - 1300	1950 - 2800	2700 - 4200
	référence	531	998	2399	3278
u	domaine	200 - 480	500 - 1050	1900 - 2800	2700 - 4200
	référence	315	764	2027	3118
œ	domaine	350 - 600	1050 - 1750	2050 - 2850	2850 - 4200
	référence	517	1391	2379	3353

Table C.6. Domaines de définition définitifs et valeurs de référence des quatre premiers formants des voyelles orales pour les locuteurs masculins.

La deuxième remarque de moindre importance est que le fait de ne conserver que les pôles LPC ayant une fréquence supérieure à la limite supérieure de  $D(F_3)$  entraîne l'élimination, dans quelques rares cas, d'un pôle dont la fréquence est comprise entre celle du dernier  $F_3$  trouvé et cette limite.

A la suite de ces conclusions, nous avons effectué une nouvelle détermination des formants intermédiaires selon une méthodologie qui prend en compte toutes ces modifications et que nous détaillerons dans le prochain paragraphe. Mais auparavant, nous terminons ce paragraphe par une conclusion plus conceptuelle.

En effet, l'examen des pôles d'ordre supérieur, c'est-à-dire ceux dont la fréquence est comprise entre la limite supérieure de  $D(F_3)$  et 4500 Hz, nous a fait abandonner l'idée d'utiliser, dans le cadre de cette étude, le formant  $F_4$  pour discriminer les locuteurs. Considérons, par exemple, le cas des voyelles dont la limite supérieure de  $F_3$  se situe en dessous de 3000 Hz. L'observation des tableaux situés en annexe montre que l'analyse LPC trouve selon les locuteurs



un ou deux pôles ayant une fréquence comprise entre 3000 et 4000 Hz. Y a-t-il effectivement pour certains locuteurs deux formants, qui seraient F4 et F5, dans cette zone fréquentielle ? Dans la négative, d'où provient le deuxième pôle ? A-t-il une origine glottale ou subglottique ou bien est-il dû à un artefact de l'analyse LPC ? Dans la plupart des cas, l'aspect de la zone d'énergie située entre ces deux fréquences sur le spectrogramme ne nous a pas permis de conclure sur la validité ou la non validité de ces deux pôles (fréquences entourées dans les tableaux). En outre, dans le cas de la voyelle [œ] des phrases 4 et 10, nous avons essayé de comparer les résultats de l'extraction des racines LPC avec ceux d'une analyse spectrale manuelle comprenant un spectre FFT à bande étroite, un cepstre et un spectre LPC. Mais souvent, les résultats obtenus étaient différents (un seul pic ou un pic avec "une épaule" ou bien deux pics) selon la méthode d'analyse et selon le cas étudié. Et nous n'avons pas été capable de conclure sur la validité des deux pôles. Nous pensons que la détermination d'un quatrième formant robuste relève d'une expertise phonétique que nous ne possédons pas.

### 2.3. Affectation des pôles LPC aux formants intermédiaires F1, F2 et F3

La méthode qui nous a permis de déterminer les formants intermédiaires des voyelles orales étudiées est directement issue de l'application des remarques du paragraphe précédent à la méthode mise au point dans la première étude.

Cette méthode d'affectation est toujours fondée sur une analyse LPC à dix-huit coefficients dont sont extraits les pôles LPC qui ont une largeur de bande inférieure ou égale à 1000 Hz. Ces pôles sont triés dans l'ordre des fréquences croissantes.

Les deux filtrages sont également conservés. Le premier d'entre eux filtre les pôles aux fréquences trop voisines. Le deuxième ne conserve que les pôles dont la fréquence appartient à l'un au moins des domaines de définition. Les limites inférieures et supérieures des domaines de définition sont présentées dans la table C.6

Trois des pôles restants sont alors affectés séquentiellement aux trois premiers formants de la voyelle selon un algorithme qui tient compte, pour chaque formant  $F_i$ , dans un premier temps du nombre de pôles appartenant au domaine de définition  $D(F_i)$ , puis par ordre de priorité :

- du pôle affecté à  $F_{i-1}$ ,
- du pôle susceptible d'être affecté à  $F_{i+1}$ ,
- des largeurs de bande des pôles candidats,
- de l'écart entre les fréquences des pôles candidats et la valeur de référence qui figure dans la table C.6.

Cet algorithme est présenté sur les trois pages suivantes

Lorsque les trois premiers formants sont déterminés, tous les pôles dont la fréquence est comprise entre la fréquence du pôle affecté à F3 et 4500 Hz sont conservés.

Nous présenterons dans le paragraphe suivant quelques résultats de l'application de cet algorithme d'affectation à la détermination des formants intermédiaires qui ont servi à celle des formants finaux de chacune des voyelles.

- Procédure générale d'affectation :

début

dans le cas

où  $D(F_i)$  ne contient aucun pôle candidat,  
le formant  $F_i$  est mis à zéro ;

où  $D(F_i)$  contient un seul pôle candidat,  
si il a déjà été affecté au formant  $F_{i-1}$ ,  
alors le formant  $F_i$  est mis à zéro,  
sinon il est affecté au formant  $F_i$ ,  
fsi ;

où  $D(F_i)$  contient deux pôles susceptibles d'être affectés au formant  $F_i$ ,  
si l'un des deux pôles a été affecté au formant  $F_{i-1}$ ,  
alors  
si c'est le pôle ayant la plus grande fréquence formantique,  
alors le formant  $F_i$  est mis à zéro,  
sinon l'autre pôle est affecté à  $F_i$ ,  
fsi,

sinon

si le pôle ayant la plus grande fréquence formantique est le seul pôle candidat pour  $F_{i+1}$ ,  
alors l'autre pôle est affecté à  $F_i$ ,  
sinon on choisit entre les deux pôles (appel à choix\_2),  
fsi,

fsi ;

.. / ..

• Procédure générale d'affectation (suite) :

où  $D(F_i)$  contient trois pôles susceptibles d'être affectés au formant  $F_i$ ,  
soit  $P_a$ ,  $P_b$  et  $P_c$  ces pôles rangés dans l'ordre des fréquences croissantes,  
si l'un des trois pôles a été affecté au formant  $F_{i-1}$ ,

alors

dans le cas

où ce pôle est  $P_c$ ,

le formant  $F_i$  est mis à zéro ;

où ce pôle est  $P_b$ ,

$P_c$  est affecté à  $F_i$  ;

où ce pôle est  $P_a$ ,

si  $P_c$  est le seul pôle candidat pour  $F_{i+1}$ ,

alors  $P_b$  est affecté à  $F_i$ ,

sinon on choisit entre  $P_b$  et  $P_c$  (appel à choix\_2),

fsi,

fcas,

sinon

si l'un au moins des trois pôles est candidat pour  $F_{i+1}$ ,

alors

si seul  $P_c$  est candidat

alors on choisit entre  $P_a$  et  $P_b$  (appel à choix\_2),

sinon

si  $P_b$  et  $P_c$  sont candidats,

alors  $P_a$  est affecté à  $F_i$ ,

sinon on choisit entre les trois pôles (appel à choix\_3),

fsi,

fsi,

sinon on choisit entre les trois pôles (appel à choix\_3),

fsi,

fsi ;

fcas ;

$F_i$  .bw, est donnée par la largeur de bande du pôle retenu ;

si  $(F_i.bw \geq 250 + \frac{F_i.f_r}{10})$ ,

alors le champ  $df_1$  du coefficient de défiance  $F_i$  .df vaut 1,

sinon le champ  $df_1$  du coefficient de défiance  $F_i$  .df vaut 0,

fsi ;

fin.



- Procédure choix\_2 (choix entre deux pôles candidats) :

début

si la largeur de bande de l'un des pôles est supérieure au double de celle de l'autre pôle,

alors le pôle ayant la plus petite largeur de bande est affecté à  $F_i$ ,

sinon le pôle dont la fréquence est la plus proche de la fréquence de référence située dans la table C.6 est affecté à  $F_i$ ,

fsi.

fin

- Procédure choix\_3 (choix entre trois pôles candidats) :

début

si la plus grande des trois largeurs de bande est supérieure au double de la plus petite,

alors on choisit entre les deux autres pôles, appel à choix\_2,

sinon le pôle dont la fréquence est la plus proche de la fréquence de référence située dans la table C.6 est affecté à  $F_i$ ,

fsi.

fin.

## 2.4. Détermination des formants finaux F1, F2 et F3

### 2.4.1. Détermination des formants intermédiaires à trois emplacements de la voyelle

Pour chacun des trois premiers formants d'une voyelle, nous avons appliqué la méthode précédemment décrite pour établir trois formants intermédiaires à trois emplacements voisins situés dans la partie la plus stable de la voyelle, un prélèvement central et deux prélèvements situés à 8 ms de part et d'autre de celui-ci. Etant donné les résultats de la première étude, nous avons choisi l'emplacement central en fonction de la durée de la voyelle et donc de sa position syntaxico-sémantique dans la phrase. Si la durée de la voyelle dépasse 160 ms, l'emplacement central se trouve à 80 ms du début de la voyelle, sinon il se situe au milieu de la voyelle.

La table C.8 présente quelques résultats de l'application de l'algorithme d'affectation des pôles LPC à la détermination de ces formants intermédiaires. Sur un total de trente-six déterminations de formants par voyelle et par locuteur, la partie supérieure de la table compabilise le nombre de cas où le domaine de définition contient deux pôles candidats (valeur de gauche, devant le "/") et le nombre de fois où cette situation est résolue par l'application de la procédure choix\_2 (valeur de droite, derrière le "/"). La partie inférieure de la table présente le même type de résultats dans le cas où le domaine de définition contient trois pôles candidats.

Nous pouvons remarquer d'après cette table que plus de 60% des cas où le domaine de définition contient deux pôles candidats concernent le troisième formant du [ i ]. Ceci s'explique par la valeur plus élevée de  $F_3$  pour le locuteur jfm. Pour ce locuteur,  $F_3$  est comprise entre

Voyelle	aq		bz		df		gm		jfm		jg		jlc		jmp		jph		ms		Par voyelle
i_11	7 / 1	F3	12 / 9	F3					3 / 1	F3	9 / 0	F3	11 / 11	F3	5 / 0	F3	6 / 4	F3	3 / 0	F3	56 / 26
i_15	7 / 1	F3	10 / 5	F3	1 / 1	F3				F3	6 / 0	F3	12 / 6	F3	9 / 0	F3	9 / 3	F3	6 / 1	F3	60 / 17
e_01			9 / 0	F2					6 / 5	F3	5 / 2	F3			1 / 0	F2	6 / 2	F2	2 / 0	F2	29 / 9
ε_07					1 / 1	F3															1 / 1
ε_09							3 / 3	F3			1 / 1	F3	1 / 1	F3							5 / 5
a_03	1 / 1	F3	1 / 1	F3	2 / 2	F2	1 / 1	F3									2 / 2	F3	3 / 3	F3	10 / 10
a_16																					0 / 0
ɔ_02									1 / 1	F3											1 / 1
ɔ_04							1 / 1	F3			9 / 6	F3									10 / 7
u_08					1 / 1	F3			1 / 1	F3	1 / 1		1 / 1	F3							4 / 4
u_12											3 / 3	F3									3 / 3
u_16									1 / 1	F3											1 / 1
œ_04	1 / 1	F3																			1 / 1
œ_10					3 / 3	F2															3 / 3
œ_13	1 / 1	F2									3 / 3	F3									4 / 4
Par locuteur	17 / 5		32 / 15		8 / 8		5 / 5		12 / 9		37 / 16		25 / 19		15 / 0		23 / 11		14 / 4		188 / 92

Voyelle	aq		bz		df		gm		jfm		jg		jlc		jmp		jph		ms		Par voyelle
i_11													1 / 0	F3			6 / 0	F3			7 / 0
i_15			2 / 0	F3													3 / 0	F3			5 / 0

Table C.8. Comptabilisation du nombre de cas où un domaine de définition D(Fi) contient 2 pôles candidats (partie supérieure) et 3 pôles candidats (partie inférieure) lors de l'application de l'algorithme d'affectation à la détermination des 36 formants intermédiaires par locuteur et par voyelle. Les cases vides correspondent à des valeurs nulles.



3350 et 3550 Hz, soit 400 à 500 Hz de plus que la valeur de référence. Cette valeur plus élevée a entraîné l'inclusion dans  $D(F3)$  de l'intervalle fréquentiel [3400, 3600] qui est une zone dans laquelle se situe  $F4$  pour d'autres locuteurs (cf. tableaux en annexe). La plupart de ces cas sont résolus par le fait que le deuxième pôle contenu dans  $D(F3)$  est le seul candidat pour  $F4$ . Seuls les cas des locuteurs bz, jlc et jph nécessitent de choisir entre les deux pôles candidats (appel de la procédure choix\_2). En effet, pour ces locuteurs, l'analyse LPC trouve deux pôles dont les fréquences sont comprises entre 3200 et 4200 Hz et donc susceptibles d'être affectés à  $F4$  (cf. page 44). Dans de rares cas, ces deux pôles sont tous les deux inclus dans  $D(F3)$ . Ces derniers constituent donc les cas où  $D(F3)$  contient trois pôles candidats (cf. partie inférieure de la table). Le nombre moins élevé de passages dans choix\_2 pour la phrase 11 est dû, pour le locuteur jlc, à une fréquence du deuxième pôle de  $D(F4)$  légèrement supérieure à 4200 Hz, pour les locuteurs bz et jph, à la présence d'un seul pôle susceptible d'être affecté à  $F4$ .

Pour la voyelle [e], les cas où un domaine de définition contient deux pôles candidats résultent aussi du recouvrement entre certains domaines de définition. Plus précisément, le formant  $F2$  du locuteur jfm a une fréquence voisine de celles des formants  $F3$  des locuteurs bz, jph et ms. Par ailleurs, le locuteur jg possède un formant  $F4$  dont la fréquence est inférieure à la borne supérieure de  $D(F3)$ . En ce qui concerne le formant  $F3$  du locuteur jfm, les paires pôles candidats proviennent toutes de la représentation d'un formant situé vers 2800 Hz par deux pôles, l'un situé vers 2600 Hz et l'autre entre 3000 et 3200 Hz.

Si nous considérons maintenant les résultats par locuteur, les deux locuteurs qui présentent les plus grands nombres de cas où il y a deux ou trois pôles candidats sont jg et bz. Pour les voyelles [i] et [e], ces deux locuteurs possèdent un formant  $F4$  de fréquence faible et donc incluse dans  $D(F3)$ . Pour les voyelles [ɔ], [œ], [ɛ] et [u] du locuteur jg, l'analyse LPC trouve un pôle supplémentaire dont la fréquence se situe entre 2700 et 3000 Hz, qui a une largeur de bande d'environ 450 Hz et qui n'est pas vérifiable sur le spectrogramme.

#### 2.4.2. Détermination des formants finaux

Afin d'obtenir des fréquences formantiques robustes à partir desquelles seront calculées les distances entre les locuteurs, chaque formant final a été établi à partir de trois formants intermédiaires et en fonction de la proximité fréquentielle de ces trois formants intermédiaires. Cette notion de proximité fréquentielle a été définie à l'aide d'un écart fréquentiel maximal  $E_i$  associé à chaque fréquence formantique  $F_i$ . Deux fréquences formantiques intermédiaires sont proches si elles diffèrent d'au plus  $E_i$ . Trois fréquences formantiques intermédiaires sont proches si elles diffèrent deux à deux d'au plus  $E_i$ .

Nous avons utilisé ces écarts pour déterminer la valeur finale du formant et de son coefficient de défiance selon une démarche résumée dans la table C.9. Nous pouvons remarquer que le niveau de robustesse d'un formant est donné par le champ  $df_2$  de son coefficient de défiance mais aussi dans certains cas par la nullité de sa fréquence formantique. Plus la valeur du coefficient  $df_2$  est élevée moins le formant est robuste. Mais lorsque la fréquence formantique semblait trop incertaine, elle a été forcée à zéro pour qu'elle ne soit pas prise en compte dans le calcul de distances entre les locuteurs. Rappelons que la fréquence d'un formant intermédiaire est nulle lorsqu'aucun pôle LPC n'a pu lui être affecté.

Le champ  $df_1$  du coefficient de défiance fournit une information sur les largeurs de bande des formants intermédiaires qui ont servi à établir le formant final.



proximité des formants intermédiaires	formant final fréquence $F_i$ .fr	formant final champ $df_2$	formant final champ $df_1$
3 fréquences proches	moyenne des 3 fréquences	0	somme des 3 $df_1$
3 fréquences proches nulles	0	5	0
2 fréquences proches	moyenne des 2 fréquences	2	somme des 2 $df_1$
2 fréquences proches nulles	0	4	0
2 couples de fréquences proches	moyenne des 2 fréquences ayant les plus petits $df_1$ ou moyenne des trois	1	somme des 3 ou des 2 $df_1$
3 fréquences éloignées non nulles	0	3	somme des 3 $df_1$
2 fréquences éloignées et une fréquence nulle	0	3	somme des 3 $df_1$

Table C.9. Détermination d'un formant final à partir de trois formants intermédiaires.

Pour fixer les valeurs des écarts fréquentiels  $E_i$ , nous nous sommes fondée sur les résultats d'une étude réalisée par R.B. Monsen et A.M. Engebretson [Monsen 83] sur les précisions des mesures des fréquences formantiques obtenues par deux méthodes de détermination. L'une des méthodes consistait en une détermination spectrographique des pics d'un spectre instantané, effectuée par des experts. L'autre méthode était une extraction des racines d'une analyse LPC d'ordre 22. Dans l'étude citée, ces deux méthodes ont été principalement appliquées aux trois premiers formants d'une voyelle [æ] de synthèse dont la fréquence fondamentale pouvait varier de 100 à 500 Hz et la largeur de bande des formants de 50 à 400 Hz.

Lors de la détermination des formants finaux de notre étude, nous avons utilisé les résultats de l'étude de R.B. Monsen et A.M. Engebretson qui étaient présentés dans une publication de F. Lonchamp [Lonchamp 87]. Celle-ci indiquait une précision absolue de  $\pm 60$  Hz pour  $F_1$  et de  $\pm 110$  Hz pour  $F_2$  pour les deux méthodes, ainsi qu'une précision de  $\pm 60$  Hz pour  $F_3$  pour la méthode LPC et de  $\pm 110$  Hz pour la détermination spectrographique. Aussi avons-nous choisi un écart maximal  $E_i$  entre les fréquences formantiques égal à 60 Hz pour  $F_1$  et à 110 Hz pour  $F_2$  et  $F_3$ .

Lors de la rédaction de ce mémoire, nous nous sommes procuré une copie de la publication de R.B. Monsen et A.M. Engebretson. Nous avons alors découvert qu'une erreur de frappe s'était glissée dans la publication de F. Lonchamp et que la précision absolue obtenue par les auteurs étaient de  $\pm 60$  Hz pour  $F_1$  et  $F_2$ , pour les deux méthodes, alors qu'elle était de  $\pm 60$  Hz pour  $F_3$ , pour la méthode LPC, et de  $\pm 110$  Hz, pour la détermination spectrographique.

Sauf mention explicite, les résultats que nous allons présenter dans la suite de ce mémoire correspondent à l'écart maximal entre les fréquences formantiques initialement retenu.

Pour chacune des voyelles orales étudiées, les formants finaux obtenus sont présentés en annexe sous la forme d'un tableau par voyelle à la suite du tableau issu de l'étude préliminaire.

Ce tableau comprend pour chacune des répétitions des locuteurs :

- les valeurs des fréquences formantiques  $F_1$ ,  $F_2$  et  $F_3$ ,
- le champ  $df_1$  du coefficient de défiance,
- le champ  $df_2$  du coefficient de défiance.

2.4.3. Analyse de la robustesse des formants finaux

Nous avons établi quelques statistiques sur la robustesse des formants obtenus que nous allons essayer d'analyser. Ces statistiques, calculées pour chacun des trois premiers formants de chacune des occurrences des voyelles étudiées, sont regroupées dans deux tables.

Pour plus de clarté, nous adopterons dans la suite de ce chapitre une notation simplifiée pour désigner l'occurrence d'une voyelle dans une phrase. Ainsi l'occurrence de [  $\epsilon$  ] dans la phrase 09 sera notée [  $\epsilon_{09}$  ].

Sur un total de quarante formants, la table C.11 comptabilise les effectifs des formants finaux selon leur niveau de robustesse. Ce niveau de robustesse est donné par la valeur de  $df_2$ . Celle-ci varie de 0 pour le formant le plus robuste à 5 pour le formant le moins robuste.

La table C.10, plus générale, présente pour chacune des voyelles les pourcentages de formants que nous pouvons qualifier de "très robustes" et de "robustes". Un formant est considéré comme très robuste s'il est issu de trois formants intermédiaires proches ( $df_2 = 0$ ). Un formant est considéré comme robuste, s'il est issu de trois formants intermédiaires proches ou de deux couples de formants intermédiaires proches ( $df_2 = 0$  ou  $df_2 = 1$ ).

Voyelle	F1		F2		F3	
	très robustes	robustes	très robustes	robustes	très robustes	robustes
i	95%	96%	94%	97%	71%	80%
e	100%	100%	97%	100%	90%	95%
$\epsilon$	80%	94%	84%	91%	94%	95%
a	85%	91%	90%	95%	89%	92%
ɔ	95%	96%	96%	99%	96%	97%
u	89%	90%	87%	92%	64%	77%
œ	97%	98%	93%	97%	96%	98%

Table C.10. Pourcentages de formants finaux très robustes ( $df_2 = 0$ ) et de formants finaux robustes ( $df_2 = 0$  ou  $df_2 = 1$ ).

Nous pouvons remarquer d'après cette dernière table qu'hormis les troisièmes formants de [ i ] et [ u ], tous les formants sont très robustes à plus de 80% et robustes à plus de 90%.

Voyelle	F1.df <sub>2</sub>						F1.fr nulle	F2.df <sub>2</sub>						F2.fr nulle	F3.df <sub>2</sub>						F3.fr nulle
	0	1	2	3	4	5		0	1	2	3	4	5		0	1	2	3	4	5	
i_11	39		1					37	2	1					30	3	6	1			1
i_15	37	1	1			1	1	38	1				1	1	27	4	8			1	1
e_01	40							39	1						36	2	1	1			1
ε_07	27	10	3					28	6	4			2	2	39		1				
ε_09	37	1	1		1		1	39				1		1	36	1	3				
a_03	28	5	6	1			1	33	4	3					32	2	6				
a_16	40							39		1					39	1					
ɔ_02	39	1						37	2	1					39		1				
ɔ_04	37		2		1		1	40							38	1			1		1
u_08	33		5		2		2	31	6	3					17	9	11	2	1		3
u_12	39			1			1	38		1	1			1	26	5	5	4			4
u_16	35	1	3		1		1	36	0	3	1			1	34	1	3		2		2
œ_04	40							40							40						
œ_10	40							37	2	1					39	1					
œ_13	37	1	1		1		1	35	2	2		1		1	36	2	2				

Table C.11. Effectifs des formants finaux selon leur niveau de robustesse. Le niveau de robustesse est donné par la valeur de  $df_2$ , de 0 pour le formant le plus robuste à 5 pour le formant le moins robuste. Effectifs des formants finaux forcés à 0. Les cases vides correspondent à des valeurs nulles.



L'observation des spectrogrammes des triplets [piR] dont sont extraits ceux de la figure C.16, permet d'établir trois causes possibles de la médiocre fiabilité des valeurs de F<sub>3</sub> de [i]. La première résulte de l'influence coarticulaire du [p] qui abaisse le troisième formant du [i] et provoque une transition montante de F<sub>3</sub>. Les deux autres sont la présence d'une friction importante pendant la voyelle pour certains locuteurs (df, gm, jlc, ms) et la proximité de F<sub>3</sub> et F<sub>4</sub> pour d'autres locuteurs (jfm, jlc). La robustesse de F<sub>2</sub>, qui pourtant subit la même influence coarticulaire du [p], nous conduit à penser que les deux dernières causes sont plus importantes.

En ce qui concerne le troisième formant de [u], son manque global de fiabilité est d'abord dû aux très mauvais résultats de l'occurrence de la phrase 08. Dans cette phrase, le mot grammatical et inaccentué *"pour"* est réalisé sous la forme d'un triplet phonétique bref dont la structure acoustique est variable selon le locuteur, souvent bruitée et comportant des transitions formantiques importantes au niveau de F<sub>3</sub>. Cependant, ce manque de fiabilité de F<sub>3</sub> semble aussi corrélé à la faiblesse de l'énergie dans cette zone formantique. En effet, en dépit d'une structure acoustique variable, la voyelle [u<sub>16</sub>] possède un F<sub>3</sub> beaucoup plus robuste que ceux de [u<sub>08</sub>] et [u<sub>12</sub>]. Mais, il est aussi beaucoup plus intense. Sans doute est-ce dû à une meilleure articulation de la voyelle, induite par la difficulté à prononcer le syntagme *"la course Strasbourg-Paris"*. Toutes ces remarques sont illustrées par les exemples des figures C.17, C.18 et C.19.

Nous pouvons également remarquer d'après la table C.10 que les formants des voyelles [œ], [ɔ] et [e] sont globalement les plus robustes.

L'examen de la table C.11 appelle quelques tentatives d'explications à propos de la variabilité de la robustesse des formants entre les différentes occurrences d'une même voyelle. La plupart de ces explications se réfèrent à la durée de la voyelle et à l'évolution des formants le long de la voyelle. Ces données sont spécifiées dans les tableaux situés en annexe.

- **voyelle [ɛ]** : la voyelle finale de la phrase 9 possède des formants beaucoup plus fiables que ceux de la voyelle située en début de verbe dans la phrase 7. Comme le montre la figure C.20, la première est très longue et présente sur le spectrogramme des formants parallèles à l'axe des temps ("formants plats"). Les fréquences de ces formants sont donc robustes. Par ailleurs, la consonne [R] a une structure fricative dévoisée et d'énergie faible. Dans la phrase 7, la structure acoustique du triplet est totalement différente. La voyelle est moins accentuée, beaucoup plus courte et se termine par un [R] plus vocalique et plus bref. Cette occurrence est davantage soumise à la coarticulation. Le formant F<sub>1</sub> est plat ou montant selon les locuteurs alors que F<sub>2</sub> présente pour la plupart des locuteurs une pente descendante. Le formant F<sub>3</sub> reste plat ;

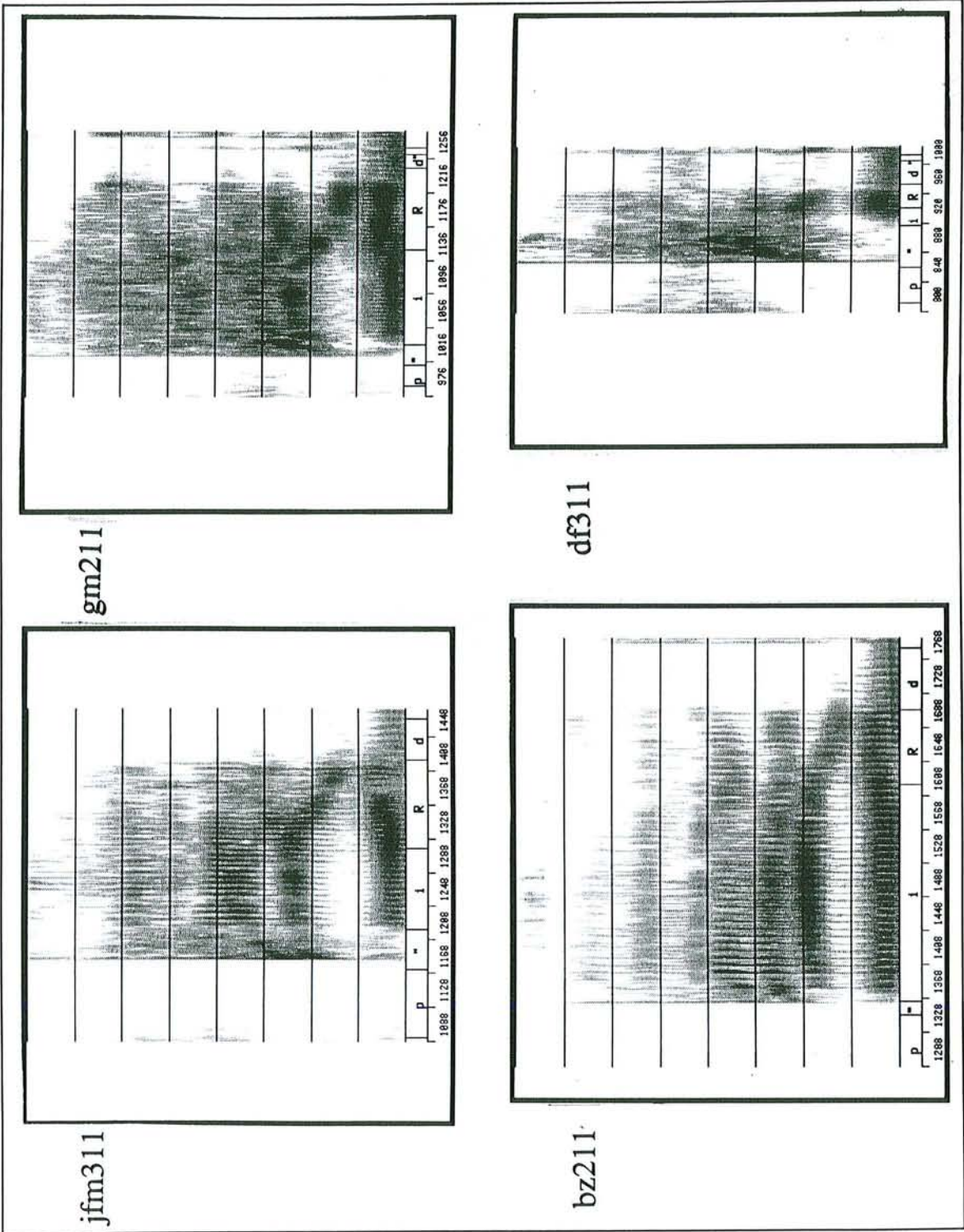


Figure C.16. Différentes réalisations du triplet [ piR ] de la phrase 11.

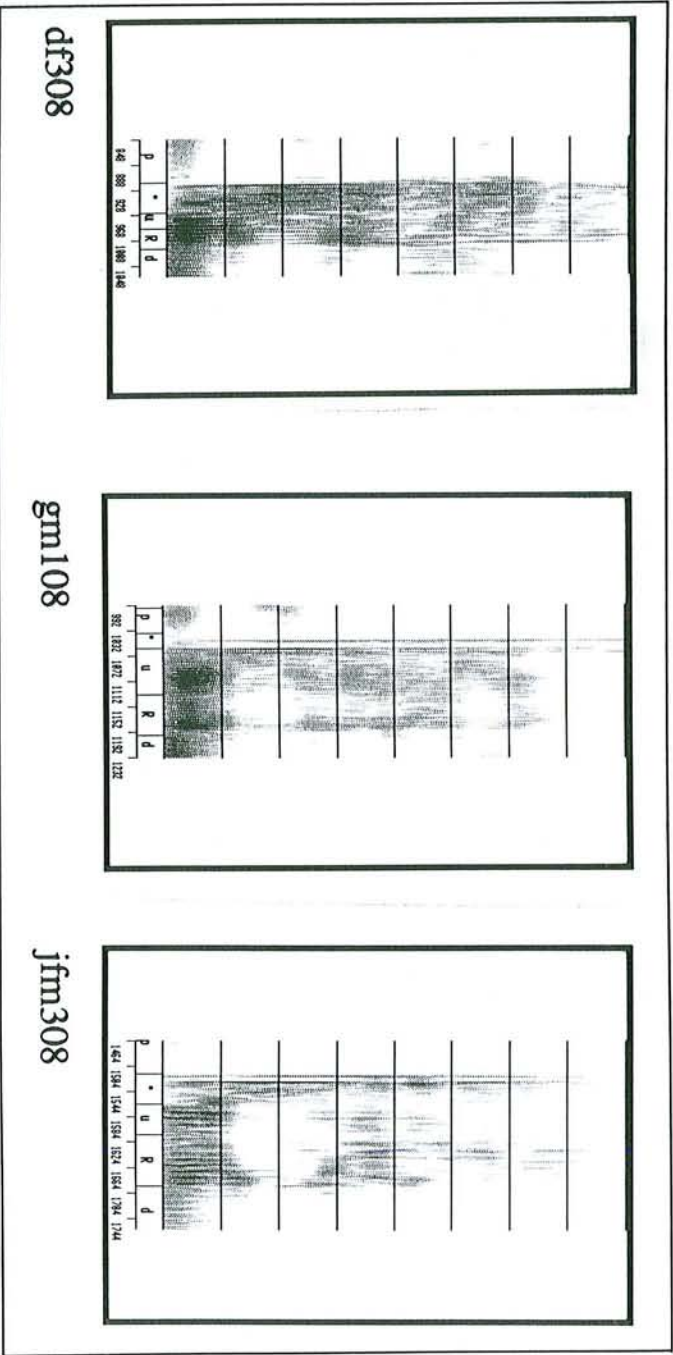


Figure C.17. Différentes réalisations du triplet [ puR ] de la phrase 08.



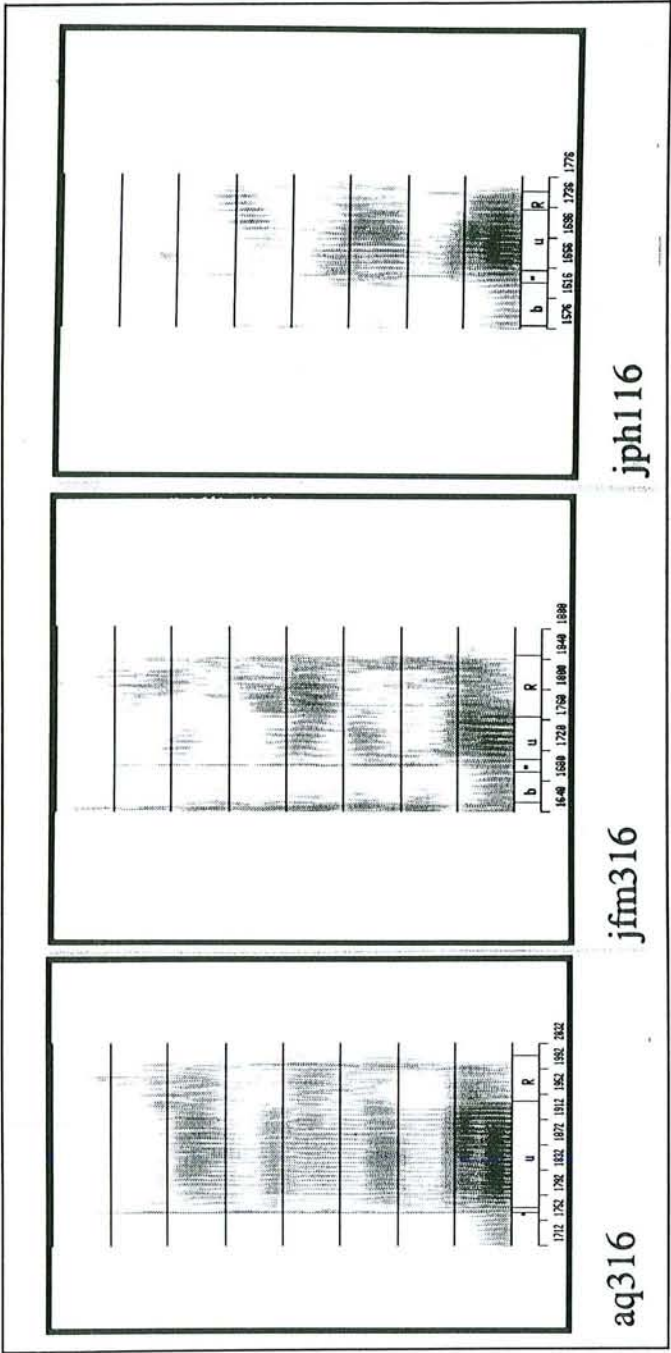


Figure C.18. Différentes réalisations du triplet [ buR ] de la phrase 16.

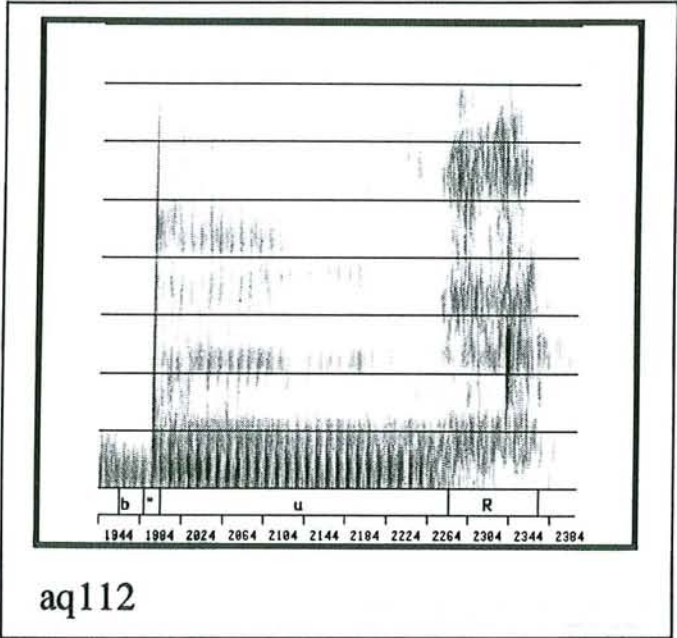


Figure C.19. Spectrogramme du triplet [ buR ] de la phrase 12.

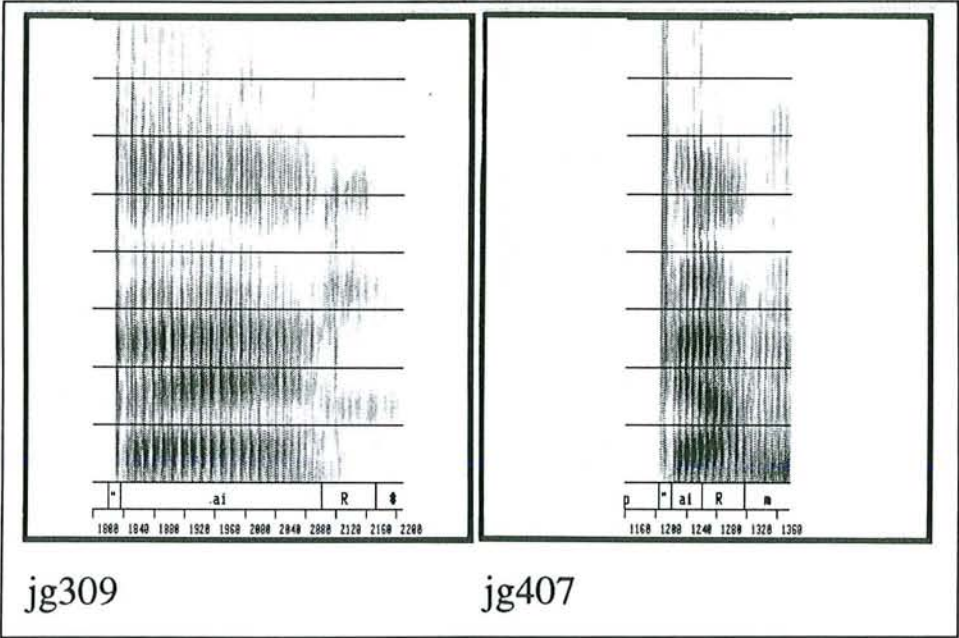


Figure C.20. Spectrogrammes des réalisations des voyelles [  $\epsilon_{09}$  ] et [  $\epsilon_{07}$  ].

- **voyelle [ œ ]** : les différences de robustesse entre les trois occurrences sont minimales, sans doute parce qu'elles sont toutes les trois accentuées (voyelles allongées) même si c'est à des degrés divers. La robustesse globale de la voyelle s'explique par des formants bien séparés (500 Hz, 1400 Hz, 2400 Hz) et peu soumis à l'effet coarticulatoire du [ R ] à cause de l'allongement. Les fréquences formantiques les plus stables sont obtenues pour la voyelle située en fin de syntagme ou en fin de groupe de phonation selon le locuteur ([ œ\_04 ]). Cette stabilité diminue lorsque la voyelle est située en fin de phrase ([ œ\_13 ]), probablement à cause de la baisse d'énergie ;
- **voyelle [ a ]** : comme le montre la figure C.21, les deux occurrences des triplets [ paR ], situées dans des contextes syntaxico-sémantiques différents, ont en commun d'être beaucoup plus intenses que celles des autres triplets. Mais elles diffèrent par la durée de la voyelle et par le type de /R/. La voyelle [ a\_16 ] est assez longue et se prolonge par un [ R ] vocalique qui se distingue difficilement du [ a ]. Le premier formant est légèrement montant au début de la voyelle sous l'influence conjointe du [ p ] et du [ R ]. Sous l'effet coarticulatoire du [ R ], le deuxième formant est soit plat soit légèrement descendant et le troisième formant soit plat soit légèrement montant. Ces faibles transitions formantiques engendrent des fréquences formantiques très fiables pour cette occurrence. La voyelle [ a\_03 ] présente le même type de transitions mais sur une durée beaucoup plus courte (entre 64 et 88 ms), ce qui explique les formants beaucoup plus pentus. Enfin, la consonne [ R ] a une structure beaucoup plus fricative et moins sonore. De plus, pour certains locuteurs, l'imprécision de la segmentation s'ajoute à une durée courte pour engendrer un troisième emplacement de mesure (50% + 8ms) qui se situe au début du [ R ]. Tout ceci explique le peu de robustesse des formants de [ a\_03 ] ;

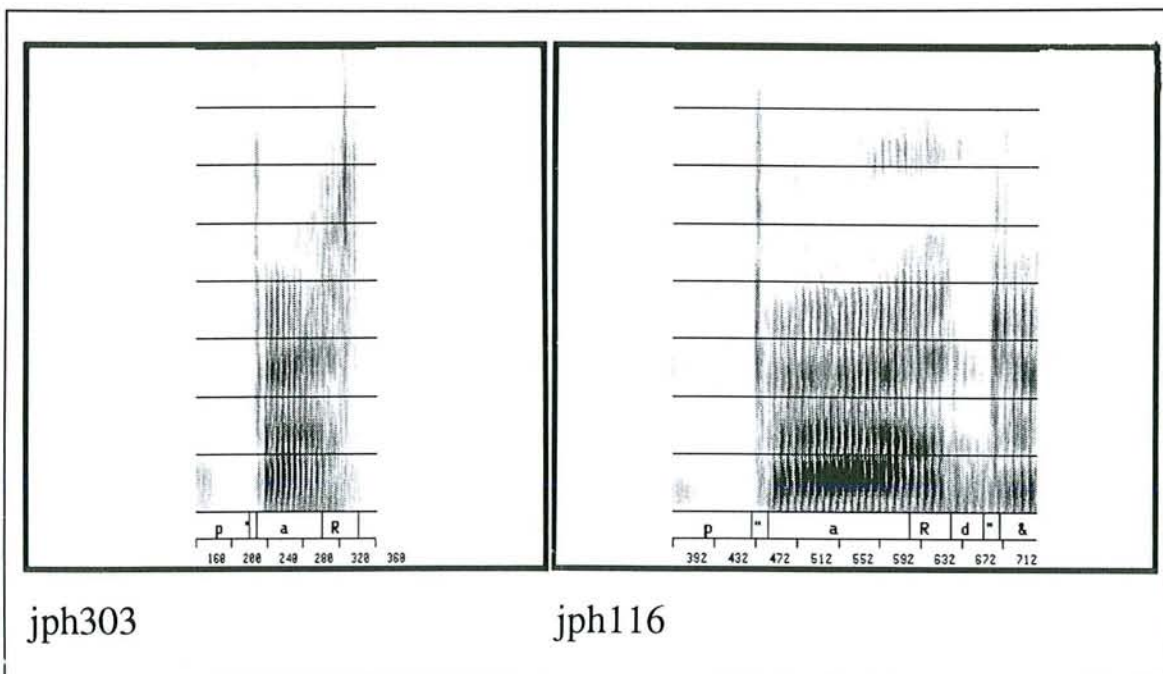


Figure C.21. Spectrogrammes des réalisations des voyelles [ a\_16 ] et [ a\_03 ].



- **voyelle [ɔ]** : les formants des deux occurrences de cette voyelles ([ɔ\_02] et [ɔ\_04]) présentent le même niveau de robustesse, alors que les divergences linguistiques et acoustiques entre celles-ci sont similaires à celles existant entre les deux occurrences des voyelles [ɛ] et [a]. Cette invariance de la robustesse résulte sans doute de la ressemblance spectrale entre [ɔ] et [R]. la figure C.22 présente les deux types d'occurrences de cette voyelle ;

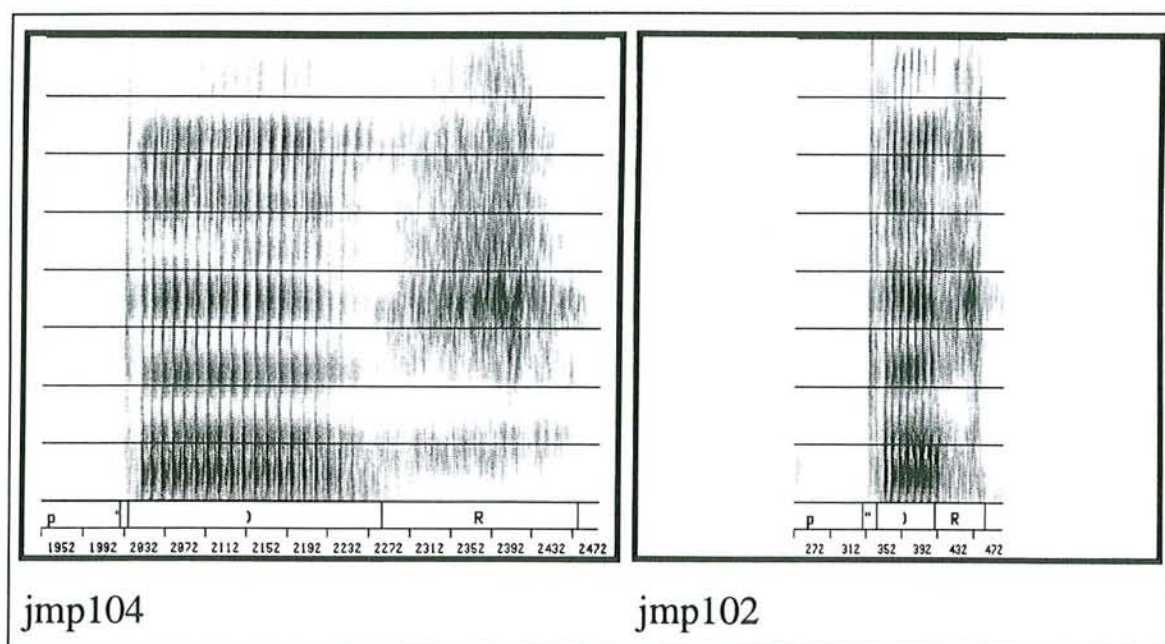


Figure C.22. Spectrogrammes des réalisations des voyelles [ɔ\_02] et [ɔ\_04].

- **voyelle [e]** : la robustesse des formants de la voyelle / e / semble étrange en particulier si nous la comparons à la robustesse globalement moins importante des formants de [ɛ], qui est une voyelle proche dans le triangle articulatoire. L'influence coarticulatoire du [R] devrait être plus marquée pour [e]. Selon ce qui a été dit au paragraphe V.2.4.3 de la partie A, l'influence coarticulatoire du [R] devrait se manifester par des transitions importantes pour les formants F1 et F2. Mais, dans la phrase 1, le [i] de "péri" semble annuler l'influence du [R], qui au contraire subit celle des deux voyelles antérieures fermées. En effet, comme nous pouvons l'observer sur la figure C.23, la fréquence formantique F<sub>2</sub> du [R] atteint presque 2000 Hz.

La table C.11 fournit également le nombre de formants dont la fréquence formantique a été forcée à 0. Ce nombre correspond en fait au cumul des effectifs des formants dont le champ  $df_2$  est supérieur ou égal à 3. Si nous éliminons la répétition  $df_3$  de la phrase 15 où le triplet / piR / n'a pas été prononcé, les fréquences nulles sont rares et concernent principalement la voyelle [u].

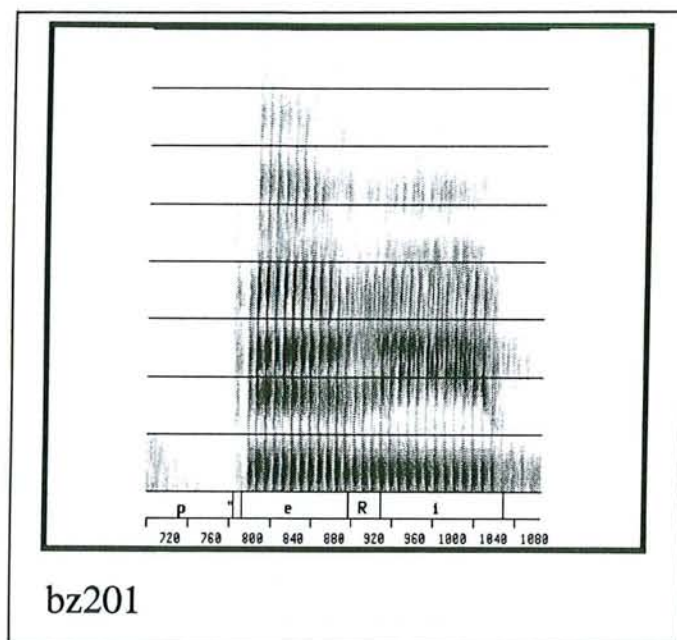


Figure C.23. Influence des voyelles antérieures fermées [ e ] et [ i ] sur les formants du [ R ] dans le mot "péri".

#### 2.4.4. Vérification des formants finaux

Nous avons également vérifié la plausibilité de tous les formants finaux en comparant leurs fréquences aux fréquences formantiques établies dans la première étude. Pour cela, nous avons tenu compte de la durée de la voyelle, du nouvel emplacement de calcul par rapport au prélèvement situé à 40% du début de la voyelle et de l'évolution du formant le long de la voyelle, qui avait été notée lors de la première étude.

Les résultats de cette vérification figurent dans les tableaux présentés en annexe. En particulier, lorsqu'une fréquence formantique semblait incorrecte ou était nulle, nous avons ajouté dans le tableau dans l'ordre des possibilités :

- une fréquence déterminée sur le spectrogramme,
- une fréquence issue de la première étude si le formant était invisible sur le spectrogramme (cas du [ u ]),
- les fréquences des trois formants intermédiaires qui ont servi à établir celle du formant final, lorsqu'aucune des deux possibilités précédentes n'était applicable.

La table C.12 fournit, sur un total de quarante formants par voyelle, le nombre de fréquences formantiques jugées incorrectes lors de cette vérification.

L'ajout de ce nombre de fréquences formantiques jugés incorrectes au nombre de fréquences forcées à 0 par manque de robustesse conduit à un pourcentage vraiment faible de fréquences qui risquent de poser des problèmes lors de l'étude de la pertinence des trois premiers formants.

	i_11	e_01	ε_07	a_03	a_16	ɔ_02	u_08	u_12	u_16	œ_13
F1			2					1		
F2				2	1	3	2	2	1	1
F3	1	2		1			1			

Table C.12. Nombre de formants finaux considérés comme incorrects lors de l'étape de vérification. Les voyelles absentes du tableau ne possèdent pas de formants jugés incorrects.

2.5. Conclusion

Lors de cette première phase, nous avons déterminé les trois premiers formants des voyelles orales / ε /, / œ /, / ɔ /, / a /, / i /, / e / et / u / précédées des contextes neutres au sens de la coarticulation linguale / p / ou / b /, et suivies du contexte postérieur allongeant / R /.

Afin qu'il soit le plus robuste possible, chacun de ces trois formants a été obtenu à partir de trois formants intermédiaires établis à trois emplacements voisins dans la voyelle. Chacun de ces formants intermédiaires résulte lui-même d'une méthode d'affectation des pôles issus d'une analyse LPC. Celle-ci tient compte de la voyelle étudiée et des domaines fréquentiels dans lesquels sont supposés se trouver ses formants. Cette méthode d'affectation a été mise au point et validée à l'aide d'une première étude comportant une étape de vérification des formants sur spectrogramme.

Par ailleurs, nous avons associé à chaque formant final un niveau de robustesse qui traduit la proximité fréquentielle des trois formants intermédiaires et qui s'exprime sous la forme d'un coefficient de défiance. De plus, lorsque le formant a été considéré comme trop incertain, sa fréquence formantique a été forcée à 0, afin qu'elle ne soit pas utilisée pour discriminer les locuteurs.

L'étude globale de la robustesse des formants des voyelles orales a mis en évidence la grande fiabilité des formants des voyelles / ɔ /, / œ / et / e /. Plus généralement, hormis les troisièmes formants de / i / et / u /, chacun des trois formants de chacune des voyelles est "très robuste" à plus de 80% et "robuste" à plus de 90%.

L'examen de la robustesse de chacune des occurrences des voyelles a permis de vérifier que plus la voyelle est en position accentuée, c'est-à-dire intense et allongée, plus ses fréquences formantiques sont fiables. Les fréquences les plus robustes sont obtenues en fin de syntagme ou fin de groupe de phonation alors que les moins robustes sont obtenues en début de mot ou dans des mots grammaticaux monosyllabiques.

Après avoir déterminé des valeurs fiables pour les trois premiers formants des voyelles orales, nous avons pu étudier leur pertinence pour la caractérisation automatique du locuteur. Cette étude fait l'objet de la prochaine section.



### 3. Etude de la pertinence des trois premiers formants des voyelles orales

#### 3.1. Méthodologie d'étude de la pertinence

Le nombre restreint d'échantillons par triplet (quatre par locuteur) ne nous a pas permis d'utiliser une analyse de type analyse discriminante pour déterminer quelles étaient les meilleures voyelles ainsi que les meilleures combinaisons linéaires de leurs formants au sens de la caractérisation du locuteur. Aussi avons-nous décidé de nous limiter à l'étude de la pertinence de certaines combinaisons formantiques et de fonder cette étude sur des expériences de reconnaissance d'un locuteur parmi dix locuteurs de référence.

##### 3.1.1. Les combinaisons formantiques étudiées

Nous avons analysé pour chacune des voyelles sélectionnées la pertinence des combinaisons formantiques suivantes :

- pour les formants : (F1), (F2), (F3), (F1, F2), (F1, F3), (F2, F3), (F1, F2, F3) ;
- pour les écarts entre les formants : (F2-F1), (F3-F2), (F3-F1), (F3-F1, F3-F2), (F3-F1, F2-F1), (F2-F1, F3-F1), (F3-F1, F2-F1, F3-F2).

Lors de l'étude d'une combinaison formantique, pour chaque répétition de la voyelle étudiée, un locuteur est représenté par un vecteur ayant pour composantes les éléments de cette combinaison. Ainsi, lors du test de la pertinence de la combinaison (F1, F2, F3), la répétition  $n$  du locuteur  $k$  est représentée par le vecteur  $V_{k,n}(x_1, x_2, x_3) = V_{k,n}(F1, F2, F3)$ .

Afin de rester conforme à la structure de formant définie au chapitre précédent, une composante de vecteur est constituée d'une fréquence,  $x_i.fr$ , d'une largeur de bande,  $x_i.bw$  et d'un coefficient de défiance,  $x_i.df$ . Lorsque la composante  $x_i$  d'une combinaison est un formant, elle est donnée par les égalités (1). Lorsque la composante  $x_{ij}$  d'une combinaison est un "écart" entre deux formants, elle est donnée par les égalités (2).

$$(1) \quad \begin{cases} x_i.fr = F_i.fr \\ x_i.bw = F_i.bw \\ x_i.df_1 = F_i.df_1 \\ x_i.df_2 = F_i.df_2 \end{cases}$$

$$(2) \quad \begin{cases} x_{ij}.fr = \begin{cases} (F_j.fr - F_i.fr) & \text{si } (F_i <> 0 \text{ et } F_j <> 0) \\ 0 & \text{sinon} \end{cases} \\ x_{ij}.bw = 0 \\ x_{ij}.df_1 = F_i.df_1 + F_j.df_1 \\ x_{ij}.df_2 = F_i.df_2 + F_j.df_2 \end{cases}$$

### 3.1.2. Méthodes de reconnaissance d'un locuteur

L'étude de la pertinence des voyelles et des combinaisons formantiques que nous avons mise en œuvre est fondée sur les résultats d'expériences de reconnaissance d'un locuteur parmi dix locuteurs de référence. Douze expériences par locuteur ont été effectuées en faisant tourner les différentes répétitions des locuteurs (quatre répétitions de référence et, pour chacune d'elles, trois répétitions inconnues).

Pour comparer deux locuteurs, nous avons utilisé la distance pondérée définie par l'égalité (3).

$$D^2(k_n, l_m) = \frac{1}{I} \sum_{i=1}^I \frac{(x_i^{k_n} \cdot fr - x_i^{l_m} \cdot fr)^2}{a_i^2} \quad (3)$$

$D(k_n, l_m)$  est la distance entre le vecteur associé à la répétition  $n$  d'une voyelle par le locuteur  $k$  et le vecteur associé à la répétition  $m$  de la même voyelle par un locuteur  $l$ .

$I$  est le nombre de composantes non nulles de chaque vecteur. Ce nombre varie en fonction de la combinaison formantique étudiée et de la robustesse des formants intervenant dans le calcul de la distance. En effet, puisqu'une fréquence formantique jugée insuffisamment robuste a été forcée à 0, le terme correspondant n'intervient plus dans le calcul de la distance. Remarquons que dans un tel cas, cette distance n'en est plus une au sens mathématique du terme. Ce choix engendre également un léger artefact au niveau de la pertinence d'une combinaison formantique donnée mais nous avons préféré cette méthode à l'établissement d'une pertinence de voyelle sur des formants faux. Par ailleurs, étant donné la faible proportion de formants nuls (cf. table C.11), cet artefact est réellement négligeable.

Le coefficient  $a_i$  est un coefficient de pondération associé à la différence des fréquences des deux composantes de rang  $i$ . Selon les tests, celui-ci pourra prendre les valeurs suivantes :

- *la constante 1.* Lorsque les composantes sont (F1, F2, F3), les locuteurs sont donc comparés directement à partir des différences absolues de leurs fréquences formantiques et ces trois différences ont le même poids dans le calcul de la distance ;
- *le minimum des deux composantes.* Lorsque les composantes sont (F1, F2, F3), les locuteurs sont donc comparés selon les différences relatives à leurs trois fréquences formantiques. Dans ce cas également, la distance entre deux locuteurs n'est plus une distance au sens mathématique du terme ;
- *la valeur de référence du formant de rang  $i$  ou l'écart entre les valeurs de référence lorsque les composantes sont des écarts entre formants.* Lorsque les composantes sont (F1, F2, F3), les locuteurs sont donc comparés en fonction de différences relatives à l'ordre de grandeur des trois fréquences formantiques de la voyelle considérée ;
- *la largeur du domaine de définition du formant de rang  $i$ .* Cette pondération est utilisée uniquement lorsque les composantes sont des formants. La largeur du domaine de définition  $D(F_i)$  est une estimation de l'étendue de la variabilité totale (intra locuteur et inter locuteur) de la fréquence formantique considérée. Les différences entre les locuteurs sont donc exprimées sous la forme de pourcentage des variabilités totales des fréquences formantiques de la voyelle considérée.



Ces différentes valeurs du coefficient  $a_i$  engendrent en fait des combinaisons formantiques supplémentaires mais pour plus de clarté nous avons préféré réserver le terme de combinaison aux combinaisons simples définies page 63 et employer le terme de coefficient de pondération pour  $a_i$ .

Pour ces différentes valeurs du coefficient de pondération, la distance est appliquée à des fréquences formantiques exprimées en Hertz. Nous avons également testé une méthode de classement fondée sur une distance appliquée à des fréquences converties en Bark selon la formule (4) [O'Shaughnessy 87]. Dans ce cas, le coefficient  $a_i$  vaut 1.

$$z = 13 \arctan \left( \frac{0.76f}{1000} \right) + 3.5 \arctan \left( \frac{f}{7500} \right)^2 \quad (4)$$

où  $f$  est exprimée en Hertz et  $z$  en Bark.

Par souci de simplicité, nous qualifierons de "pondération multiplicative", l'ensemble des pondérations des distances exprimées en Hertz et de "pondération perceptive" la pondération de la distance exprimée en Bark.

Quel que soit son mode de calcul, nous avons associé à la distance entre deux répétitions de deux locuteurs un coefficient de défiance  $df$  composé de trois champs. Les champs  $df_1$  et  $df_2$  sont respectivement les sommes des champs  $df_1$  et  $df_2$  des coefficients de défiance des composantes du vecteur. Le champ  $df_3$  indique qu'un ou plusieurs termes intervenant dans le calcul de la distance sont nuls en raison de la nullité d'une ou de plusieurs des composantes. Par conséquent, ce coefficient de défiance rend compte de la robustesse des formants intervenant dans le calcul de distance. Nous avons donc souhaité l'exploiter pour améliorer la reconnaissance du locuteur.

Lorsque deux locuteurs de référence sont susceptibles de correspondre au locuteur inconnu (deux distances très proches), nous avons souhaité favoriser celui dont les formants sont les plus robustes. Aussi avons-nous testé deux variantes de la reconnaissance d'un locuteur :

- dans la première variante, le locuteur reconnu est celui qui minimise la distance au locuteur inconnu,
- dans la seconde, le locuteur reconnu est celui qui minimise la distance au locuteur inconnu, à condition que l'écart entre cette distance minimale et la deuxième distance la plus faible soit supérieur à un seuil. Sinon, parmi les deux locuteurs associés à ces deux distances voisines, le locuteur reconnu est celui dont la distance a le plus petit coefficient de défiance. Nous pensions pouvoir ainsi compenser, au moins dans certains cas, l'artefact causé par les fréquences forcées à zéro. Dans nos tests, nous avons fixé ce seuil à 10% de la distance minimale.

### 3.1.3. Les indicateurs de pertinence

A l'issue de toutes ces expériences de reconnaissance, nous avons calculé trois indicateurs de pertinence de chacune des voyelles et de chacune des combinaisons formantiques, Taux (T), Score (S) et Alpha (A) :

- Taux est le pourcentage de réussite de reconnaissance d'un locuteur parmi dix locuteurs de référence, calculé sur toutes les expériences de reconnaissance. Une valeur importante de T indique une bonne pertinence de la combinaison formantique de la voyelle. Il est possible d'utiliser cet indicateur pour comparer la pertinence de combinaisons de taille et de structure différentes ;



- Score est le rapport du cumul des rangs auxquels les locuteurs sont reconnus au nombre total d'expériences de reconnaissance. Le rang est nul si le locuteur a été bien reconnu, puis il vaut 1 si le locuteur a été reconnu en deuxième, etc. Score fournit donc une information plus complète sur le classement des locuteurs. L'association d'une valeur élevée de taux et d'une valeur faible de Score assure une pertinence robuste d'une combinaison formantique. Comme le précédent, cet indicateur permet de comparer des combinaisons formantiques de taille et de structure différentes ;
- Alpha est un indicateur de type statistique qui estime de façon indirecte le rapport de la variance intralocuteur à la variance interlocuteur. Il est donné par le rapport de la moyenne des distances intralocuteur à la moyenne des distances interlocuteur. Il ressemble au F-ratio et il en possède les inconvénients. Une valeur faible de ce rapport ne permet pas de conclure sur la pertinence de la combinaison formantique considérée. En effet, elle peut résulter uniquement du fait qu'un locuteur est très différent des autres pour cette combinaison formantique. En revanche, une valeur élevée de ce rapport indique une mauvaise pertinence de la combinaison. Par ailleurs, Alpha ne peut pas être employé directement pour comparer des combinaisons formantiques de taille différente et de structure différente.

Nous allons présenter et commenter dans le paragraphe suivant quelques-unes des valeurs prises par ces trois indicateurs au cours des différentes expériences de reconnaissance que nous avons effectuées.

## 3.2. Résultats et commentaires de l'étude de pertinence

### 3.2.1. Introduction

Nous avons effectué un certain nombre de tests dans lesquels nous avons fait varier la voyelle, la combinaison formantique, le mode de pondération de la distance et la méthode de reconnaissance d'un locuteur. Etant donné la masse de résultats recueillis, nous ne pouvons pas commenter toutes les associations possibles de ces paramètres. Au niveau de la pertinence, nous allons plus particulièrement étudier la pertinence des combinaisons de formants des quinze voyelles, en nous fondant principalement sur l'indicateur Taux et en étudiant l'influence du mode de calcul de la distance. Mais nous allons tout d'abord étudier la validité globale de la deuxième méthode de reconnaissance du locuteur.

### 3.2.2. Comparaison des deux variantes de la reconnaissance d'un locuteur

Comme nous l'avons mentionné dans le paragraphe précédent, nous avons testé deux variantes de la reconnaissance d'un locuteur. La première, que nous qualifierons de "sans échange", consiste à choisir le locuteur dont la distance au locuteur inconnu est minimale. La seconde, que nous qualifierons de "avec échange" peut conduire sous certaines conditions à l'échange des deux premiers locuteurs reconnus. Lorsqu'il y a eu échange des deux premiers locuteurs reconnus et que l'un d'eux était le locuteur à reconnaître, nous avons comptabilisé le nombre de fois où cet échange a permis de reconnaître le bon locuteur ("échange réussi") et le nombre de fois où cet échange a empêché la reconnaissance du bon locuteur ("échange raté").

La table C.13 fournit les résultats de cette comptabilisation sur l'ensemble des voyelles, pour chacune des combinaisons de formants étudiées (lignes) et pour chacun des calculs de distance utilisés (colonnes).

Pondé- ration	1		Min de ( $F_i^k, F_i^l$ )		Valeur de référence de $F_i$		Largeur de $D(F_i)$		Echelle Bark	
Echanges	réussis	ratés	réussis	ratés	réussis	ratés	réussis	ratés	réussis	ratés
F1	7	3	7	3	7	3	7	3	6	3
F2	1	3	1	2	1	3	1	3	1	2
F3	4	6	2	6	4	6	4	6	2	7
F1 F2	10	12	7	6	8	9	9	12	9	13
F1 F3	10	9	18	7	12	7	10	8	14	11
F2 F3	15	14	11	7	12	9	13	13	10	13
F1 F2 F3	16	22	16	18	22	14	15	17	19	18
Total	63	60	62	50	66	51	59	62	61	67

Table C.13. Nombre d'échanges des deux premiers locuteurs reconnus "réussis" et "ratés" sur un total de 1800 expériences de reconnaissance par combinaison étudiée et par mode de calcul de la distance.

Le total des deux colonnes fournit le nombre d'échanges qui concernent le locuteur à reconnaître.

Nous pouvons remarquer que le nombre total d'échanges qui concernent le locuteur à reconnaître augmente avec le nombre de formants intervenant dans le calcul de la distance. Ceci peut s'expliquer en partie par le fait que plus la distance prend en compte de paramètres plus elle effectue un moyennage des différents niveaux de pertinence de ces paramètres qui peuvent se compenser. En revanche, le nombre total d'échanges est pratiquement constant entre les différentes pondérations de la distance. Toutefois, quelles que soient la combinaison formantique et la distance testées, ce nombre ne dépasse pas 2% du nombre d'expériences de reconnaissance, ce qui est minime.

La comparaison des effectifs des échanges réussis et des échanges ratés montre que la prise en compte de la robustesse des formants dans la méthode de reconnaissance d'un locuteur n'améliore pas globalement les performances de cette reconnaissance. Dans notre étude, cette inefficacité peut éventuellement s'expliquer par la bonne robustesse des valeurs des fréquences formantiques associée à des critères de robustesse trop sévères qui conduisent à des coefficients de défiance peu réalistes.

Jugeant d'autres tests plus prioritaires, nous n'avons pas essayé d'ajuster le seuil pour améliorer les performances de cette méthode.

Nous avons également analysé l'influence de l'échange des deux premiers locuteurs sur les taux de reconnaissance obtenus. Celle-ci est faible quelles que soient la combinaison de formants et la voyelle. Dans la plupart des cas où le taux de reconnaissance varie, la variation est limitée à 1 ou 2% en valeur absolue et elle ne modifie pas le classement des voyelles. Tous modes de calcul de distance confondus, seules dix variations sont comprises en valeur absolue entre 2,5 et 4% et essentiellement pour la combinaison (F1, F2, F3). La table C.14 fournit un exemple de cette influence dans le cas des combinaisons (F1), (F2) et (F3). Il semble n'y avoir aucun lien de cause à effet entre le type de la voyelle et le sens et la quantité de variation du taux de reconnaissance si ce n'est que, par définition, les échanges concernent surtout les voyelles aux formants peu robustes.



Puisque cette deuxième variante n'améliore pas la reconnaissance d'un locuteur, la plupart des résultats présentés dans la suite de ce chapitre sont issus de l'application de la méthode de reconnaissance "sans échange".

### 3.2.3. Les combinaisons (F1), (F2) et (F3)

Nous nous proposons de commencer l'étude de la pertinence des différentes combinaisons formantiques par celle de chacun des trois formants F1, F2 et F3 des quinze voyelles. Par définition, les valeurs des trois indicateurs de pertinence de ces combinaisons formantiques sont pratiquement identiques pour les quatre distances à pondération multiplicative. La table C.14 présente donc uniquement les valeurs prises par l'indicateur Taux pour l'une des distances à pondération multiplicative et pour la distance à pondération perceptive. De plus, elle présente ces résultats pour les deux variantes de la reconnaissance d'un locuteur. Les voyelles sont classées dans l'ordre de pertinence décroissante dans le cas de la pondération multiplicative et de la variante de reconnaissance "sans échange".

Du point de vue général, nous pouvons constater que les deux types de pondération engendrent les mêmes taux de reconnaissance. De même, les deux méthodes de reconnaissance d'un locuteur produisent les mêmes résultats.

Cette table met également en évidence la pertinence du troisième formant des voyelles orales par rapport aux deux premiers formants. En effet, le taux de reconnaissance atteint pour la meilleure voyelle — considérée seule — près de 50% pour F3 contre 33% pour F2 et 30% pour F1. Nous allons détailler quelles sont ces voyelles pour chacun des formants.

#### a) Le formant F3

En ce qui concerne F3, la voyelle la plus pertinente pour la discrimination des 10 locuteurs est la voyelle [ **u**\_16 ] (*"Le départ de la course Strasbourg-Paris aura du retard."*). De plus, la pertinence du troisième formant de la voyelle [ **u** ] décroît avec la robustesse du formant. En effet, la voyelle [ **u**\_12 ] occupe la troisième place alors que la voyelle [ **u**\_08 ] possède la pertinence la plus faible. Celle-ci n'est d'ailleurs pas améliorée par l'application de la procédure d'échange des deux premiers locuteurs.

Par ailleurs, hormis [ **u**\_08 ], toutes les voyelles labialisées (arrondies) sont situées dans la partie supérieure de la table alors que, excepté les occurrences de [ **a** ], toutes les voyelles non labialisées (non arrondies) sont situées dans sa partie inférieure. La pertinence de F3 semble donc résulter de deux sources de différences entre les locuteurs, le degré d'arrondissement des lèvres et l'influence coarticulatoire du [ **R** ]. En effet, l'accroissement de la labialisation d'une voyelle se répercute dans le domaine fréquentiel par une baisse de  $F_3$  alors que le contexte consonantique uvulaire entraîne une élévation de cette fréquence (cf. paragraphe V.2.4.3 de la partie A).

En revanche, nous n'avons pas d'explication concernant la pertinence du troisième formant de la voyelle [ **a** ] et ceci quelle que soit sa robustesse.



V o y e l l e	Taux (%) pour F3				V o y e l l e	Taux (%) pour F2				V o y e l l e	Taux (%) pour F1			
	Pondération multiplicative		Pondération perceptive			Pondération multiplicative		Pondération perceptive			Pondération multiplicative		Pondération perceptive	
	sans échange	avec échange	sans échange	avec échange		sans échange	avec échange	sans échange	avec échange		sans échange	avec échange	sans échange	avec échange
u_16	49	49	49	49	œ_13	33	33	32	32	a_03	33	33	33	33
œ_10	48	48	48	48	ɛ_09	33	33	33	33	ɔ_04	29	29	28	28
u_12	47	46	47	46	i_11	33	33	32	32	œ_13	28	29	29	30
ɔ_04	46	44	46	44	e_01	32	32	32	32	œ_10	28	28	27	27
a_16	43	43	43	43	i_15	29	29	29	29	ɔ_02	28	28	28	28
œ_04	42	42	42	42	ɔ_02	27	27	27	27	œ_04	26	26	26	26
a_03	40	41	41	42	a_03	25	25	26	26	a_16	26	26	26	26
œ_13	39	40	39	40	œ_04	24	24	24	24	ɛ_09	24	25	23	24
ɔ_02	39	39	38	38	œ_10	22	22	24	24	u_16	22	23	23	23
ɛ_09	37	37	37	37	ɔ_04	22	22	22	22	i_11	21	21	21	21
i_15	36	35	37	34	ɛ_07	17	17	17	17	i_15	20	18	21	19
ɛ_07	27	27	27	27	a_16	17	17	17	17	ɛ_07	19	21	19	21
e_01	26	26	27	27	u_16	16	15	16	14	e_01	19	19	19	19
i_11	22	23	22	22	u_12	15	15	15	14	u_08	12	13	12	13
u_08	19	18	18	17	u_08	12	12	12	12	u_12	12	12	13	13

Table C.14. Pertinence de chacun des trois premiers formants des voyelles étudiées selon l'indicateur Taux pour les distances à pondération multiplicative et pour la distance à pondération perceptive.

b) *Le formant F2*

Les voyelles qui sont pertinentes au sens de F2 sont les deux voyelles antérieures fermées [ i ] et [ e ], et la voyelle antérieure mi-ouverte [ ɛ\_09 ], auxquelles s'ajoute la voyelle [ œ\_13 ].

La particularité des voyelles antérieures est d'avoir une fréquence formantique élevée (autour de 2000 Hz) alors que la présence du [ R ] a tendance à abaisser cette fréquence formantique (cf. paragraphe V.2.4.3 de la partie A). D'après les tableaux situés en annexe, la valeur de F<sub>2</sub> de [ e ] varie selon les locuteurs de 1700 Hz à 2300 Hz, alors que celle de F<sub>2</sub> de [ i ] varie de 1850 Hz à 2400 Hz. Nous supposons que cette variabilité interlocuteur de F<sub>2</sub> n'est pas seulement due à l'influence coarticulatoire du [ R ]. En effet, deux locuteurs au moins (gm et jfm) possèdent des valeurs de F<sub>2</sub> très élevées par rapport à la fréquence moyenne masculine<sup>1</sup>. Mais, il aurait fallu vérifier cette hypothèse en comparant ces valeurs de F<sub>2</sub> avec des valeurs obtenues dans d'autres contextes consonantiques, ce que nous n'avons pas fait par manque de temps.

Afin d'expliquer la différence de pertinence de F2 entre les trois occurrences de la voyelle [ œ ], nous avons représenté sur le graphique C.24 les plages de variation de F<sub>2</sub> pour chacun des dix locuteurs et pour les trois occurrences de la voyelle. Ce sont les plages de variation des valeurs de F<sub>2</sub> déterminées automatiquement. Parmi ces valeurs, seules deux posent problème. F<sub>2</sub> est nulle pour la répétition jmp3 de [ œ\_13 ] donc sa valeur réelle de 1300 Hz n'a pas été prise en compte. Toujours pour [ œ\_13 ], la fréquence F<sub>2</sub> de la répétition jmp4 est surévaluée (1354 Hz au lieu de 1300 Hz).

Nous pouvons donc constater que la meilleure performance de la voyelle [ œ\_13 ] n'est pas due à une variabilité intralocuteur plus faible mais plutôt à un recouvrement moins important des plages de variation de chacun des locuteurs. Cette plus grande dispersion des locuteurs résulte, à notre avis, de la diversité des intonations adoptées par les locuteurs pour prononcer cette phrase au mode impératif (*"Goûtez-moi ce cake au beurre."*).

Cette représentation des plages de variation de F<sub>2</sub> amène d'autres remarques. Indépendamment d'une bonne robustesse de la mesure de F<sub>2</sub> sur une répétition (cf. table C.11), certains locuteurs présentent une variabilité importante entre les répétitions d'une même occurrence du triplet. Par ailleurs, la figure C.24 met en évidence, pour la plupart des locuteurs, une grande variabilité de l'emplacement de la plage de variation de F<sub>2</sub> et de l'étendue de cette plage entre les trois occurrences de la voyelle. En particulier, nous pouvons observer la centralisation des valeurs de F<sub>2</sub> de la voyelle [ œ\_10 ] qui est la moins accentuée. Deux raisons peuvent expliquer cette importante variabilité. L'observation du triangle acoustique (cf. figure A.34) montre que le locuteur dispose d'une certaine latitude pour faire varier F<sub>2</sub> sans que la voyelle produite soit perçue comme une voyelle voisine. Par ailleurs, il semble que le contexte phonologique proche ne suffise pas à expliquer la variabilité formantique intralocuteur. Il faut aussi tenir compte d'un contexte phonologique plus large (syllabes voisines) et du contexte syntaxico-sémantique. Tout ceci rend manifeste qu'il est difficile et dangereux de conclure sur la pertinence générale d'un phonème mais qu'il faut plutôt conclure sur celle d'un phonème énoncé dans un contexte linguistique bien défini.

<sup>1</sup> L'origine de ces fréquences élevées ne semble pas être la même pour les deux locuteurs. En effet, pour le locuteur gm, ces valeurs élevées sont corrélées à une fréquence fondamentale moyenne élevée pour un sujet masculin (190 Hz), alors que le locuteur jfm possède une fréquence fondamentale moyenne de 120 Hz.



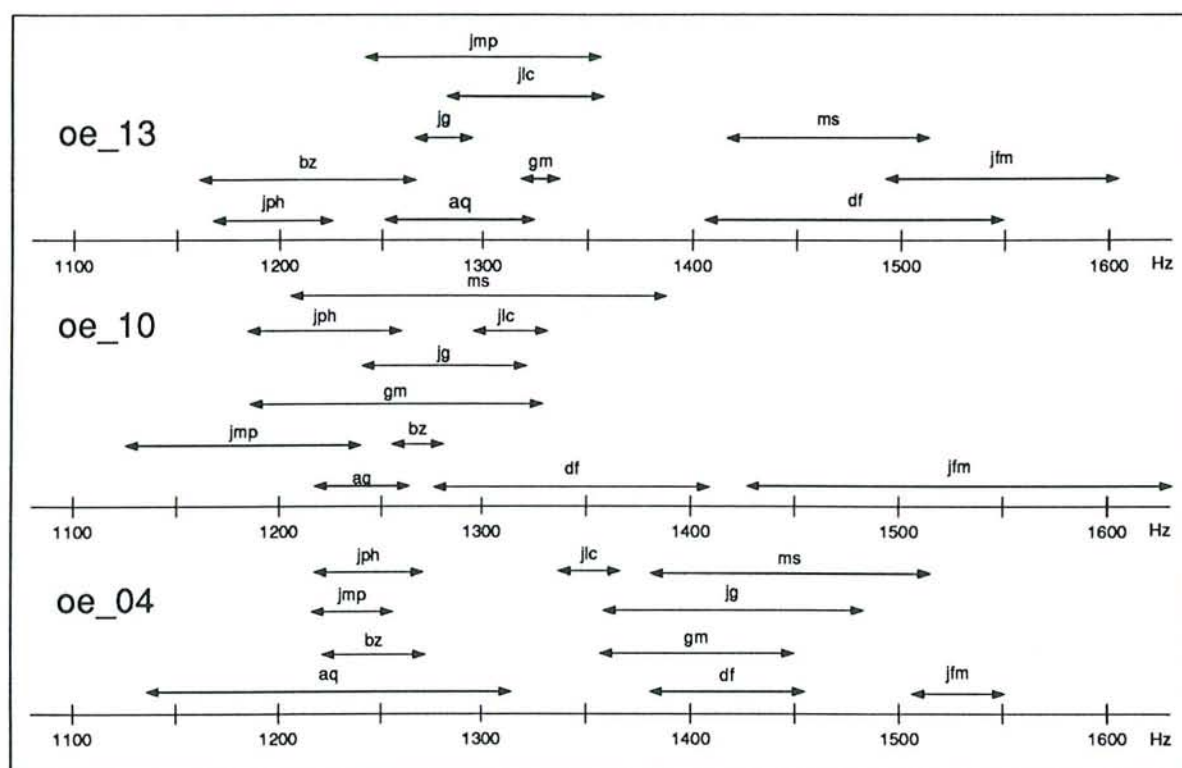


Figure C.24. Domaines de variation de  $F_2$  pour chacun des locuteurs pour les trois occurrences de la voyelle [œ].

### c) Le formant $F_1$

Le classement de voyelles selon le taux de reconnaissance des locuteurs présente un découpage dichotomique entre les voyelles ouvertes [a], [œ] et [ɔ], qui sont pertinentes, et les voyelles fermées comme [i], [e] et [u], qui ne le sont pas, la voyelle [ɛ] se situant à la frontière des deux groupes. Ce sont donc les valeurs élevées de  $F_1$  qui caractérisent mieux le locuteur. Deux raisons peuvent justifier ce phénomène. Tout d'abord, l'observation du triangle acoustique (cf. figure A.34) montre que le locuteur a plus de latitude pour faire varier  $F_1$  vers les valeurs élevées. Par ailleurs, comme nous l'avons indiqué dans le chapitre IV de la partie A, l'analyse LPC a tendance à "caler" les fréquences de ses pôles sur celles des harmoniques du fondamental, ce qui conduit à des valeurs de  $F_1$  peu précises pour les voyelles fermées.

### 3.2.4. Les combinaisons ( $F_2$ , $F_3$ ), ( $F_1$ , $F_2$ ) et ( $F_1$ , $F_3$ )

#### a) Le couple ( $F_2$ , $F_3$ )

La table C.15 présente les résultats obtenus par le couple ( $F_2$ ,  $F_3$ ) pour les deux indicateurs Taux et Score pour toutes les pondérations de la distance. Les voyelles sont classées par ordre de pertinence décroissante au sens de l'indicateur Taux.



1			Min de ( $F_i^k, F_i^l$ )			Valeur de référence de $F_i$			Largeur de $D(F_i)$			Echelle Bark		
Voyelle	T (%)	S	Voyelle	T (%)	S	Voyelle	T (%)	S	Voyelle	T (%)	S	Voyelle	T (%)	S
œ_04	68	0.69	ɛ_09	63	0.62	ɛ_09	63	0.63	œ_04	66	0.71	ɛ_09	62	0.65
ɛ_09	63	0.57	œ_04	61	0.80	œ_04	60	0.79	ɛ_09	63	0.63	œ_13	57	1.02
œ_13	60	0.91	œ_13	59	1.00	œ_13	58	1.00	œ_10	59	0.98	œ_04	56	0.87
œ_10	60	0.89	œ_10	54	1.35	i_15	54	0.67	œ_13	58	0.93	i_15	54	0.69
ɔ_02	57	0.82	i_15	54	0.69	œ_10	54	1.30	i_15	54	0.66	e_01	51	1.58
i_15	55	0.71	e_01	52	1.54	e_01	51	1.55	ɔ_02	53	0.98	œ_10	49	1.46
a_03	52	1.25	a_03	46	1.62	a_03	46	1.60	a_03	51	1.39	a_03	45	1.74
e_01	48	1.52	ɔ_02	44	1.50	ɔ_02	43	1.33	e_01	48	1.51	i_11	44	1.11
u_12	44	1.89	i_11	43	1.12	u_12	42	1.93	u_12	47	1.81	ɔ_02	43	1.62
u_16	43	1.69	ɔ_04	42	1.36	ɔ_04	42	1.39	i_11	43	1.12	u_12	42	1.96
i_11	42	1.26	u_12	41	1.97	i_11	42	1.15	ɔ_04	41	1.30	ɔ_04	41	1.39
ɔ_04	41	1.24	a_16	37	2.21	u_16	38	2.11	u_16	40	1.83	a_16	38	2.38
ɛ_07	38	2.02	u_16	36	2.22	a_16	36	2.19	a_16	37	1.85	u_16	36	2.23
a_16	36	1.67	ɛ_07	32	2.24	ɛ_07	32	2.25	ɛ_07	32	2.24	ɛ_07	32	2.27
u_08	15	3.07	u_08	15	3.27	u_08	14	3.23	u_08	15	3.14	u_08	15	3.25

Table C.15. Pertinence du couple (F2, F3) selon les indicateurs Taux et Score et pour tous les modes de calcul de distance.

Pour les cinq meilleures voyelles, ce classement varie très peu avec le mode de calcul de la distance. En effet, quelle que soit la distance, les voyelles [  $\epsilon_{09}$  ], [  $\text{œ}_{04}$  ], [  $\text{œ}_{10}$  ], [  $\text{œ}_{13}$  ] et [  $\text{i}_{15}$  ] se retrouvent dans les six voyelles les plus pertinentes. En revanche, les taux de reconnaissance diffèrent selon la pondération. Les meilleurs taux sont obtenus avec la distance euclidienne simple (plus de 60% pour [  $\epsilon_{09}$  ] et les trois occurrences de [  $\text{œ}$  ]).

Si nous confrontons ces résultats à ceux obtenus pas les formants pris isolément, nous remarquons que les formants d'une même voyelle ne sont pas indépendants au sens de la reconnaissance du locuteur et que cette dépendance varie avec la voyelle. Selon les voyelles, les divers degrés de pertinence des formants pris isolément se compensent ou se confortent. Ainsi, l'association de F2 à F3 conduit à un meilleur taux de reconnaissance pour la voyelle [  $\text{œ}_{04}$  ] (T [F2] = 24%, T [F3] = 42% et T [F2, F3] = 68%) que pour la voyelle [  $\text{a}_{03}$  ] (T [F2] = 25%, T [F3] = 40% et T [F2, F3] = 52%).

Cette dépendance devrait varier avec la pondération de la distance. Si c'est un peu le cas pour les taux de reconnaissance (Taux de [  $\text{ɔ}_{02}$  ] varie de 43 à 57%), cette variation ne se répercute pas sur le classement des voyelles.

La comparaison des classements respectivement engendrés par Taux et Score montre que l'indicateur S confirme les voyelles pertinentes mises en évidence par T. Pour ces voyelles, le locuteur inconnu est soit reconnu soit pas très loin dans la liste des locuteurs ordonnée par distance au locuteur inconnu croissante. Toutefois, certaines voyelles, comme [  $\text{i}_{15}$  ], [  $\text{ɔ}_{02}$  ] et [  $\text{ɔ}_{04}$  ], sont plus pertinentes au sens de S qu'à celui de T. Par ailleurs, le classement selon l'indicateur Score est encore plus stable que celui de Taux vis-à-vis des différentes pondérations des distances.

#### *b) Le couple (F1, F3)*

La table C.16 présente les résultats obtenus par le couple (F1, F3) selon l'indicateur Taux pour toutes les pondérations de la distance. Les voyelles sont classées par ordre de pertinence décroissante au sens de cet indicateur.

Globalement les taux de reconnaissance sont moins élevés que ceux de la combinaison précédente (entre 50 et 55%). Si les deux voyelles [  $\text{œ}_{10}$  ] et [  $\text{ɔ}_{04}$  ] sont pertinentes pour toutes les pondérations de distance, il n'en est pas de même pour les autres voyelles. Il est normal que l'influence de la pondération soit plus influente pour cette combinaison que pour (F2, F3).

Par exemple, la voyelle [  $\text{œ}_{04}$  ] discrimine bien les locuteurs lorsque les écarts entre les fréquences formantiques sont considérés en absolu mais pas lorsqu'ils sont ramenés à des pourcentages de fréquences ou d'intervalles fréquentiels, ni lorsqu'ils sont ajustés perceptivement. On peut donc conclure que la pertinence de [  $\text{œ}_{04}$  ] pour cette combinaison provient essentiellement de F3.

Réciproquement, la voyelle [  $\text{a}_{03}$  ] est plus pertinente au sens des distances exprimées en pourcentage ou selon une échelle perceptive. Ceci signifie que, par rapport à [  $\text{œ}_{10}$  ] et [  $\text{ɔ}_{04}$  ], elle doit plutôt sa pertinence à F1, ce que confirment les résultats établis au paragraphe 3.2.3.



1		Min de ( $F_1^k$ , $F_1^l$ )		Valeur de référence de $F_i$		Largeur de $D(F_i)$		Echelle Bark	
Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)
œ_10	63	œ_10	54	œ_10	54	œ_10	61	œ_10	57
œ_04	56	ɔ_04	50	ɔ_04	51	i_15	53	ɔ_04	51
ɔ_04	52	a_03	48	a_03	51	ɔ_04	53	a_03	50
a_16	50	ɔ_02	48	u_16	48	ɔ_02	51	i_15	50
œ_13	48	i_15	47	i_15	48	u_16	50	u_16	49
a_03	48	u_16	45	ɔ_02	48	a_03	50	œ_04	48
ɔ_02	47	œ_04	44	a_16	46	a_16	49	ɔ_02	48
i_15	46	a_16	44	œ_13	44	œ_04	48	œ_13	46
ɛ_09	46	œ_13	43	œ_04	44	œ_13	45	a_16	46
u_16	41	ɛ_09	39	ɛ_09	39	ɛ_09	45	u_12	43
u_12	40	i_11	38	i_11	38	u_12	43	ɛ_09	40
ɛ_07	36	u_12	36	u_12	35	ɛ_07	40	i_11	38
e_01	34	ɛ_07	34	ɛ_07	34	i_11	38	ɛ_07	36
i_11	27	e_01	27	e_01	28	e_01	28	e_01	28
u_08	21	u_08	22	u_08	23	u_08	23	u_08	22

Table C.16. Pertinence du couple ( $F_1$ ,  $F_3$ ) selon l'indicateur Taux pour tous les modes de calcul de distance.

c) Le couple ( $F_1$ ,  $F_2$ )

La table C.17 présente les résultats obtenus par le couple ( $F_1$ ,  $F_2$ ) pour l'indicateur Taux pour toutes les pondérations de la distance. Les voyelles sont classées par ordre de pertinence décroissante au sens de Taux.

Du point de vue général, c'est le couple de formants le moins performant pour la reconnaissance du locuteur (entre 45 et 50%).

Quel que soit le mode de calcul de la distance, les six voyelles [  $\epsilon_{09}$  ], [  $i_{11}$  ], [  $i_{15}$  ], [  $\epsilon_{04}$  ], [  $\epsilon_{13}$  ] et [  $\epsilon_{10}$  ] sont situées dans la partie supérieure de la table mais dans un ordre qui varie avec la distance.

La performance intrinsèque des trois occurrences de la voyelle centrale [  $\epsilon$  ] est indépendante du mode de calcul de la distance. En revanche, celle de [  $\epsilon_{09}$  ] augmente notablement lorsque l'écart entre les locuteurs est pondéré par l'inverse de la largeur du domaine de définition du formant (c'est-à-dire en fonction de la variabilité potentielle du formant). Quant à la voyelle [  $i_{11}$  ], qui est pertinente quand le locuteur est représenté par  $F_2$ , elle conserve cette pertinence si l'écart entre les fréquences formantiques n'est pas pondéré par l'inverse de ces fréquences.



1		Min de ( $F_i^k$ , $F_i^l$ )		Valeur de référence de $F_i$		Largeur de $D(F_i)$		Echelle Bark	
Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)	Voyelle	T (%)
i_11	53	ε_09	48	œ_04	47	ε_09	53	i_11	51
ε_09	49	œ_04	47	œ_13	46	i_11	53	ε_09	49
œ_13	44	œ_13	46	ε_09	46	i_15	52	œ_04	48
œ_04	44	i_11	43	i_11	45	œ_04	47	i_15	46
i_15	42	œ_10	43	i_15	45	œ_13	46	œ_13	46
œ_10	41	i_15	42	œ_10	43	œ_10	42	œ_10	41
ɔ_02	40	ɔ_02	38	ɔ_02	38	ɔ_02	39	ɔ_02	40
a_03	38	a_03	38	a_03	38	a_03	38	a_03	38
e_01	37	ε_07	37	e_01	38	e_01	37	e_01	36
ɔ_04	32	e_01	36	ε_07	37	ε_07	33	ε_07	35
u_16	31	ɔ_04	33	ɔ_04	33	ɔ_04	33	ɔ_04	34
a_16	30	a_16	32	u_16	32	u_16	32	a_16	31
ε_07	27	u_16	32	a_16	32	a_16	32	u_16	30
u_12	25	u_12	28	u_12	28	u_12	29	u_12	25
u_08	19	u_08	19	u_08	19	u_08	19	u_08	23

Table C.17. Pertinence du couple ( $F_1$ ,  $F_2$ ) selon l'indicateur  
Taux pour tous les modes de calcul de distance.

Quel que soit le couple de formants considéré, parmi toutes les occurrences de voyelles, deux d'entre elles ne caractérisent absolument pas le locuteur : ce sont les deux voyelles inaccentuées [ ε\_07 ] et [ u\_08 ].

### 3.2.5. La combinaison ( $F_1$ , $F_2$ , $F_3$ )

La table C.18 présente la pertinence des voyelles étudiées lorsqu'un locuteur est représenté par les trois premiers formants de la voyelle qu'il a prononcée. Cette pertinence est donnée pour les deux indicateurs Taux et Score ainsi qu'en fonction de la pondération de la distance.

Du point de vue général, nous pouvons observer que les taux de reconnaissance varient avec la pondération. Les quatre voyelles les plus pertinentes obtiennent les meilleurs taux de reconnaissance pour la distance euclidienne simple (Taux > 60%).

Trois voyelles sont pertinentes pour tous les modes de calcul de la distance. Ce sont [ œ\_04 ], [ ε\_09 ] et [ œ\_10 ]. Selon la pondération, deux autres voyelles viennent compléter ce trio ou s'y intercaler. Ce sont :

- [ œ\_13 ] et [ ɔ\_02 ] pour la distance euclidienne simple,

1			Min de ( $F_i^k, F_i^l$ )			Valeur de référence de $F_i$			Largeur de $D(F_i)$			Echelle Bark		
Voyelle	T (%)	S	Voyelle	T (%)	S	Voyelle	T (%)		Voyelle	T (%)	S	Voyelle	T (%)	S
œ_04	75	0.55	ε_09	62	0.78	œ_10	63	1.12	œ_04	68	0.62	œ_04	66	0.70
ε_09	70	0.47	i_15	59	0.96	ε_09	62	0.78	ε_09	66	0.57	ε_09	64	0.69
œ_13	65	0.81	œ_04	58	0.86	œ_04	59	0.77	œ_10	64	0.92	i_15	61	0.70
œ_10	64	0.83	œ_10	58	1.23	ɔ_04	57	0.85	i_15	63	0.65	œ_10	61	1.21
ɔ_02	58	0.77	œ_13	55	0.91	a_03	55	1.16	œ_13	61	0.83	i_11	58	0.98
a_03	58	1.18	ɔ_04	55	0.80	i_15	54	0.99	ɔ_02	61	0.88	œ_13	57	0.89
i_15	57	0.70	a_03	55	1.17	ɔ_02	54	1.17	i_11	60	0.88	ɔ_04	53	0.98
e_01	56	1.31	ɔ_02	53	1.28	œ_13	53	0.95	a_03	58	1.16	a_03	53	1.26
i_11	53	1.04	u_16	50	1.58	i_11	50	1.45	ɔ_04	56	0.76	ɔ_02	50	1.36
u_16	49	1.51	i_11	50	1.56	u_16	49	1.54	u_16	51	1.40	u_16	48	1.67
ɔ_04	49	1.01	a_16	45	1.53	u_12	46	1.75	u_12	50	1.68	ε_07	47	2.08
u_12	49	1.74	ε_07	45	2.28	a_16	44	1.57	ε_07	48	1.89	e_01	47	1.21
ε_07	45	1.78	e_01	44	1.42	e_01	44	1.50	e_01	45	1.28	u_12	43	1.78
a_16	43	1.45	u_12	43	1.84	ε_07	43	2.22	a_16	43	1.38	a_16	41	1.76
u_08	18	2.83	u_08	18	3.20	u_08	19	3.19	u_08	19	2.93	u_08	19	3.04

Table C.18. Pertinence de la combinaison ( $F_1, F_2, F_3$ ) selon les indicateurs Taux et Score et pour tous les modes de calcul de distance.

- [ i<sub>15</sub> ] et [ ɔ<sub>13</sub> ] pour la pondération de chaque terme de la distance par l'inverse du minimum des deux composantes,
- [ ɔ<sub>04</sub> ] et [ a<sub>03</sub> ] pour la pondération de chaque terme par l'inverse de la valeur de référence des composantes,
- [ i<sub>15</sub> ] et [ ɔ<sub>13</sub> ] pour la pondération de chaque terme de la distance par l'inverse de la largeur du domaine de définition des composantes,
- [ i<sub>15</sub> ] et [ i<sub>11</sub> ] pour la pondération perceptive.

La pertinence des troisièmes formants de [ u<sub>16</sub> ] et de [ u<sub>12</sub> ] et celle du deuxième formant de [ e<sub>01</sub> ] ne se retrouvent pas dans la combinaison globale. En ce qui concerne [ œ<sub>04</sub> ], [ ɛ<sub>09</sub> ], [ œ<sub>10</sub> ] et [ œ<sub>13</sub> ], la pertinence globale des voyelles résulte principalement de celle du couple (F2, F3). Dans le cas de la distance euclidienne simple, la mise en correspondance des tables C.15 et C.18 met en évidence un classement identique pour les cinq premières voyelles et des taux de reconnaissance peu différents. En particulier le formant F1 n'apporte rien à la pertinence de [ ɛ<sub>09</sub> ]. Au contraire, dans les calculs de distance qui le favorisent, il fait baisser le taux de reconnaissance du locuteur.

### 3.2.6. Les combinaisons d'écart de formants

Nous avons également étudié la pertinence des combinaisons construites à partir des écarts formantiques. Sauf pour la voyelle [ œ<sub>04</sub> ], tous les taux de reconnaissance obtenus par les combinaisons d'écarts sont inférieurs à ceux obtenus par les combinaisons de formants.

L'étude des combinaisons limitées à un seul écart (F<sub>j</sub>-F<sub>i</sub>) met en évidence la pertinence de voyelles pour lesquelles les écarts formantiques sont importants et ceci quel que soit le mode de pondération de la distance. Plus précisément, ce sont les voyelles [ i<sub>11</sub> ], [ i<sub>15</sub> ] et [ e<sub>01</sub> ] pour l'écart F<sub>2</sub>-F<sub>1</sub> (entre 30 et 37% pour Taux), et la voyelle [ u<sub>12</sub> ] pour les écarts F<sub>3</sub>-F<sub>2</sub> et F<sub>3</sub>-F<sub>1</sub> (44% et 50%).

Cependant, dès qu'on associe les combinaisons unitaires, la pertinence de ces voyelles disparaît au profit des voyelles qui sont pertinentes pour les combinaisons de formants, [ œ<sub>04</sub> ], [ ɛ<sub>09</sub> ], [ œ<sub>10</sub> ] et [ œ<sub>13</sub> ]. C'est ce qu'illustrent les tables C.19 et C.20 qui fournissent les meilleures combinaisons formantiques au sens de Taux. Nous rappelons que pour les écarts entre les formants, la pondération par l'inverse de la largeur du domaine de définition a été remplacée par la pondération par l'inverse de l'écart entre les deux valeurs de référence (colonne intitulée "Largeur de D(F<sub>i</sub>)" de la table C.20).



1			Min de ( $F_i^k, F_i^l$ )		
Voyelle	Combinaison	T (%)	Voyelle	Combinaison	T (%)
œ_04	F1 F2 F3	75	œ_04	F3-F1 F3-F2 F2-F1	72
œ_04	F3-F1 F3-F2	72	œ_04	F3-F2 F2-F1	70
œ_04	F3-F1 F3-F2 F2-F1	72	œ_04	F3-F1 F3-F2	68
ɛ_09	F 1 F2 F3	70	ɛ_09	F2 F3	63
œ_04	F3-F2 F2-F1	68	ɛ_09	F 1 F2 F3	62
œ_04	F2 F3	68	œ_04	F2 F3	61
œ_13	F1 F2 F3	65	œ_13	F2 F3	59
œ_10	F1 F2 F3	64	i_15	F1 F2 F3	58
œ_04	F3-F1 F2-F1	63	œ_04	F1 F2 F3	58

Table C.19. Meilleures voyelles toutes combinaisons confondues, au sens de Taux, pour la distance euclidienne simple et la distance pondérée par le minimum des deux composantes.

Valeur de référence de $F_i$			Largeur de $D(F_i)$			Echelle Bark		
Voy.	Combinaison	T (%)	Voy.	Combinaison	T (%)	Voy.	Combinaison	T (%)
œ_04	F3-F1 F3-F2 F2-F1	72	œ_04	F3-F1 F3-F2 F2-F1	72	œ_04	F3-F1 F3-F2 F2-F1	72
œ_04	F3-F1 F3-F2	68	œ_04	F3-F1 F3-F2	68	œ_04	F3-F2 F2-F1	69
œ_04	F3-F2 F2-F1	68	œ_04	F3-F2 F2-F1	68	œ_04	F3-F1 F3-F2	68
ɛ_09	F2 F3	63	œ_04	F1 F2 F3	68	œ_04	F1 F2 F3	66
œ_10	F1 F2 F3	63	œ_04	F2 F3	66	ɛ_09	F 1 F2 F3	63
ɛ_09	F1 F2 F3	62	ɛ_09	F1 F2 F3	66	ɛ_09	F2 F3	62
œ_04	F2 F3	60	œ_10	F1 F2 F3	64	i_15	F1 F2 F3	61
œ_04	F1 F2 F3	59	ɛ_09	F2 F3	63	œ_10	F1 F2 F3	61
œ_13	F2 F3	58	i_15	F1 F2 F3	62	i_11	F1 F2 F3	58
ɛ_09	F3-F2 F2-F1	57	œ_10	F1 F3	61	œ_13	F2 F3	57
ɔ_04	F1 F2 F3	57	ɔ_02	F1 F2 F3	61	œ_10	F1 F3	57

Table C.20. Meilleures voyelles toutes combinaisons confondues, au sens de Taux, pour les distances pondérées par la valeur de référence, la largeur du domaine de définition et la distance perceptive.

## 4. Conclusion sur la pertinence des trois premiers formants des voyelles orales

Nous venons d'étudier pour chacune des combinaisons de formants et d'écarts entre les formants, la pertinence des voyelles orales /  $\epsilon$  /, /  $\text{œ}$  /, /  $\text{ɔ}$  /, /  $\text{a}$  /, /  $\text{i}$  /, /  $\text{e}$  / et /  $\text{u}$  / précédées des contextes neutres au sens de la coarticulation linguale /  $\text{p}$  / ou /  $\text{b}$  /, et suivies du contexte postérieur allongeant /  $\text{R}$  /. Nous avons vu que la pertinence de ces triplets dépend de la combinaison étudiée et qu'en particulier les degrés de pertinence des formants isolés se compensent ou se confortent dans les combinaisons plus complexes. Il n'est donc pas facile de résumer ici l'ensemble des résultats trouvés sous peine d'oublier une particularité importante. Nous allons quand même essayer de dégager quelques faits intéressants avant de comparer nos résultats à ceux d'autres études. Puis nous terminerons par une discussion sur la méthodologie d'étude proprement dite.

### 4.1. Quelques conclusions sur la pertinence des voyelles

Dans le cas du formant F3, qui est la combinaison unitaire obtenant le plus fort taux de reconnaissance, nous notons la pertinence des deux occurrences de [  $\text{a}$  ] et surtout des voyelles labialisées, en particulier [  $\text{u}_{16}$  ]. Pour F2, les voyelles pertinentes sont les voyelles fermées [  $\text{i}_{11}$  ], [  $\text{i}_{15}$  ] et [  $\text{e}_{01}$  ], la voyelle mi-ouverte [  $\epsilon_{09}$  ] et la voyelle centrale [  $\text{œ}_{13}$  ]. En ce qui concerne F1, ce sont les voyelles ouvertes [  $\text{a}$  ], [  $\text{ɔ}$  ] et [  $\text{œ}$  ] qui discriminent le mieux les locuteurs.

Dès qu'on associe les formants entre eux, ces voyelles perdent leur pertinence au profit des trois occurrences de la voyelle [  $\text{œ}$  ] et de [  $\epsilon_{09}$  ]. Notons tout de même quelques particularités comme la pertinence de [  $\text{ɔ}_{04}$  ] et [  $\text{a}_{03}$  ] pour la combinaison (F1, F3) et celle de [  $\text{i}_{11}$  ] et [  $\text{i}_{15}$  ] pour le couple (F1, F2).

Du point de vue linguistique, il faut souligner que la pertinence d'une voyelle dépend fortement de son contexte phonologique et surtout de son contexte syntaxico-sémantique. Globalement, ce sont les occurrences des voyelles situées en fin de groupe de phonation comme [  $\text{œ}_{04}$  ] ou en fin de phrase comme [  $\epsilon_{09}$  ] et [  $\text{œ}_{13}$  ] qui sont les plus pertinentes. Nous avons déjà vu que ces voyelles, qui sont plus longues donc mieux articulées et moins sujettes à la coarticulation, sont aussi celles qui avaient les formants les plus robustes. Cette dépendance de la pertinence de la voyelle vis-à-vis de son accentuation est particulièrement évidente pour [  $\epsilon$  ] et [  $\text{u}$  ]. En effet, les occurrences [  $\epsilon_{09}$  ] et [  $\text{u}_{16}$  ] sont les meilleures pour certaines combinaisons, alors que les occurrences [  $\epsilon_{07}$  ] et [  $\text{u}_{08}$  ] sont les plus mauvaises pour toutes les combinaisons. Par ailleurs, deux des occurrences des voyelles les plus pertinentes, [  $\epsilon_{09}$  ] et [  $\text{œ}_{13}$  ], appartiennent à des phrases dont l'intonation a été réalisée diversement par les locuteurs :

- "Lequel des bandits guette près du repère."
- "Goûtez-moi ce cake au beurre."



Considérons maintenant les résultats du point de vue des différentes pondérations utilisées dans le calcul de la distance entre deux locuteurs. La distance euclidienne non pondérée est celle qui fournit les meilleurs taux de reconnaissance quelle que soit la combinaison étudiée. Ainsi, nous obtenons, dans le cas de la combinaison (F1, F2, F3), 75% de reconnaissance pour la voyelle [ œ\_04 ] et 70% pour la voyelle [ ε\_09 ].

En ce qui concerne le classement des voyelles, l'influence de la pondération est une fonction de la combinaison étudiée. Ainsi, pour le couple (F2, F3), toutes les distances sont équivalentes, alors que pour le couple (F1, F3) les écarts entre locuteurs exprimés en absolu augmentent la pertinence de [ œ\_04 ] et ceux exprimés en relatif augmentent celle de [ a\_03 ].

## 4.2. Quelques réflexions sur la méthodologie d'étude

A l'issue des douze expériences de reconnaissance d'un locuteur parmi dix, nous avons déterminé trois indicateurs : Taux qui est le pourcentage de bonne reconnaissance, Score qui est une fonction du rang auquel chacun des locuteurs a été reconnu et Alpha qui est le rapport de la moyenne des distances intralocuteur à la moyenne des distances interlocuteur. En fait, nous avons fondé notre mesure de la pertinence des voyelles essentiellement sur la valeur de l'indicateur Taux parce qu'il est l'évaluateur le plus commun des systèmes de reconnaissance du locuteur. L'indicateur Score a permis dans certains cas de confirmer la pertinence d'une voyelle. Il pourrait éventuellement servir lors de la recherche d'un ensemble de paramètres pertinents pour départager des paramètres équivalents au sens de Taux. Nous n'avons pas utilisé le paramètre Alpha car il permet seulement d'éliminer les voyelles les moins pertinentes et que nous n'avons pas eu besoin de faire un premier tri des voyelles.

Nous avons également testé une méthode de décision du locuteur reconnu qui tient compte de la robustesse des formants intervenant dans le calcul de la distance. Pour cela, nous avons associé un coefficient de défiance à la distance entre deux locuteurs. Lors de la phase de reconnaissance d'un locuteur, lorsque l'écart entre les deux plus petites distances était inférieur à un seuil, nous avons choisi le locuteur dont la distance possédait le plus petit coefficient de défiance. Il s'est avéré que cette variante de la reconnaissance d'un locuteur n'a pas amélioré les taux de reconnaissance. Cette inefficacité est peut-être due à une bonne robustesse globale de tous les formants — ce qui est souhaitable —, associée à des critères de robustesse trop sévères dans l'évaluation des coefficients de défiance.

Nous avons donc mené dans notre étude plusieurs expérimentations à l'issue desquelles nous avons obtenu un certain nombre de résultats sur la pertinence pour la caractérisation de dix des voyelles orales françaises dans un contexte linguistique bien défini. Toutefois, cette étude est incomplète et possède quelques lacunes que nous allons présenter.

Même s'il n'en contient pas beaucoup, notre corpus de formants comprend quelques formants qui ont été "forcés à 0" par manque de robustesse, et quelques formants dont la fréquence semble incorrecte. Pour valider complètement nos résultats, il aurait fallu remplacer ces fréquences incorrectes par les valeurs déterminées manuellement et qui sont situées dans les tableaux en annexe. Nous ne l'avons pas encore fait faute de temps.

Tout d'abord, la variabilité intralocuteur n'a pas été complètement prise en compte dans notre étude. Etant donné le nombre de phrases du corpus et le nombre de locuteurs bénévoles choisis pour enregistrer ce corpus, les quatre répétitions des dix-sept phrases ont été enregistrées par chaque locuteur lors d'une seule session qui s'est déroulée le plus souvent le matin de bonne heure. Pour valider les voyelles pertinentes mises en évidence, il faudrait procéder à



de nouveaux enregistrements des dix locuteurs étudiés, à plusieurs semaines d'intervalle et à plusieurs moments de la journée, ce qui n'est pas facile à réaliser.

De la même façon, il aurait fallu éprouver ces premiers résultats par des tests de reconnaissance sur les huit autres locuteurs de notre corpus. Ceci pourra bientôt être réalisé puisque les phrases de notre corpus correspondant à ces huit locuteurs ont été étiquetées manuellement par un étudiant de l'Institut de Phonétique de Nancy. Celui-ci n'ayant pas utilisé les mêmes critères d'étiquetage que nous, il faudra maintenant harmoniser les deux étiquetages.

Pour chacune des voyelles, les combinaisons formantiques que nous avons étudiées sont des combinaisons simples des trois premiers formants. Ce ne sont donc pas forcément les plus performantes pour l'identification du locuteur. Nous avons vu que dans le meilleur des cas le taux de reconnaissance obtenu par une seule voyelle est de 75%. Si nous avons disposé de plus de répétitions par locuteur, une analyse discriminante aurait permis d'établir des combinaisons linéaires des formants ou des écarts entre formants qui auraient conduit à de meilleurs scores d'identification. Néanmoins, ce type d'analyse a l'inconvénient d'engendrer des combinaisons linéaires qui ne sont pas interprétables du point de vue de la phonétique articulatoire et acoustique. Or, dans le travail que nous avons réalisé, nous avons souhaité ne pas trop nous écarter du processus de production de la parole. L'analyse discriminante pourra être appliquée dans un deuxième temps au sous-ensemble de voyelles pertinentes pour améliorer le taux de reconnaissance.

De la même façon, pour chacune des combinaisons formantiques étudiées, nous avons analysé la pertinence de chacune des voyelles indépendamment de celle des autres voyelles. Nous n'avons donc pas tenu compte dans notre classement de la corrélation qui existe entre les voyelles. Il serait intéressant de compléter cette étude en extrayant de l'ensemble des voyelles pertinentes le plus petit sous-ensemble qui permette d'obtenir un taux de reconnaissance de 100%. Ceci pourrait se faire à partir du calcul des coefficients de corrélation entre les voyelles ou par la méthode proposée par U. Goldstein [Goldstein 76]. Celle-ci consiste à construire de façon incrémentale ce plus petit sous-ensemble de combinaisons pertinentes.



# CONCLUSIONS ET PERSPECTIVES

Nous arrivons ici au terme de notre exposé. Nous nous proposons de repréciser les différentes étapes de notre travail avant de conclure sur les perspectives personnelles et générales de notre travail.

Nous nous sommes intéressée dans ce travail à la caractérisation du locuteur par des paramètres acoustiques et phonétiques en vue de son identification automatique.

Dans cette optique, nous avons établi avec F. Lonchamp, Professeur à l'Institut de Phonétique de Nancy, un ensemble de paramètres acoustiques, phonétiques et phonologiques susceptibles d'être pertinents pour la discrimination des locuteurs. A partir de ceux-ci, nous avons construit un corpus de dix-sept phrases que dix-huit locuteurs et vingt et une locutrices, assez homogènes au niveau de l'origine socio-géographique, ont répété quatre fois.

Afin d'obtenir des résultats fiables, nous avons étiqueté manuellement la partie de ce corpus correspondant à dix locuteurs, soit 680 phrases. Cette étape d'étiquetage nous a permis de découvrir la problématique des deux composantes de l'étiquetage, la transcription et l'alignement de la transcription sur le signal de parole. Pour réaliser cet étiquetage, nous nous sommes défini un certain nombre de critères de transcription et de segmentation. Ces critères ont été établis pour réaliser un étiquetage homogène, pour résoudre arbitrairement nos hésitations d'étiqueteuse et pour adapter les possibilités du logiciel d'étiquetage à nos souhaits. En effet, nous voulions repérer le maximum de particularités des locuteurs tout en préparant l'extraction automatique des paramètres à analyser. Cette étape assez longue de préparation des données a permis aux chercheurs de notre laboratoire travaillant dans le domaine dual de la reconnaissance et de la compréhension de la parole de disposer d'un corpus étiqueté de parole continue multilocuteur.

Ensuite, nous avons débuté l'étude de la pertinence des paramètres sélectionnés par celle des trois premiers formants des voyelles orales /  $\epsilon$  /, /  $\text{œ}$  /, /  $\text{ɔ}$  /, /  $\text{a}$  /, /  $\text{i}$  /, /  $\text{e}$  / et /  $\text{u}$  / précédées des contextes neutres au sens de la coarticulation linguale /  $\text{p}$  / ou /  $\text{b}$  /, et suivies d'un contexte postérieur allongeant /  $\text{R}$  /. Cette étude a comporté deux phases, la phase de détermination des formants et celle de l'étude de la pertinence proprement dite.

Toujours dans le dessein d'obtenir des résultats fiables, nous avons élaboré une méthodologie de détermination de formants robustes. Pour cela, nous avons associé aux constituants usuels d'un formant, qui sont sa fréquence et sa largeur de bande, un coefficient de défiance chargé de coder le degré de robustesse du formant. Un formant "final" a été établi à partir de trois formants "intermédiaires" déterminés à trois emplacements voisins situés dans une partie supposée stable de la voyelle. La fréquence du formant final ainsi que son coefficient de défiance ont été évalués en fonction de la proximité des fréquences des trois formants intermédiaires. De plus, les formants jugés trop peu robustes ont été "marqués" afin qu'ils ne soient pas utilisés dans l'étude de pertinence. Pour déterminer chacun des trois formants intermédiaires, nous avons conçu un algorithme d'affectation des pôles issus d'une analyse LPC du signal de parole aux trois premiers formants d'une voyelle orale. L'algorithme est fondé sur la connaissance de la voyelle dont il recherche les formants et de son contexte. En effet, il



tient compte de l'intervalle fréquentiel dans lequel est supposée se trouver la fréquence formantique recherchée et d'une valeur de référence de cette fréquence pour les locuteurs masculins. Cet algorithme se charge de prendre une décision d'affectation en fonction du nombre de pôles qui pourraient être affectés à ce formant, de leurs largeurs de bande et de l'écart entre leur fréquence et la valeur de référence masculine.

Pour mettre au point cette méthode d'affectation, nous avons effectué une étude préliminaire d'évaluation d'un seul formant intermédiaire par voyelle. A l'issue de celle-ci, nous avons vérifié les formants obtenus en les comparant aux spectrogrammes de parole des triplets et aux pôles bruts issus de l'analyse LPC.

Après avoir établi des formants finaux robustes, nous avons étudié quelles étaient les voyelles orales les mieux adaptées à la reconnaissance automatique du locuteur, en essayant de déterminer la combinaison formantique pour laquelle chaque voyelle était la plus pertinente. Nous avons étudié toutes les combinaisons (au sens combinatoire du terme) des trois formants et des trois écarts formantiques. Leur pertinence a été analysée à l'aide d'expériences de reconnaissance d'un locuteur parmi les dix locuteurs de référence. La méthode de reconnaissance mise en œuvre compare deux locuteurs en calculant une distance entre deux vecteurs dont les composantes sont les différents éléments de la combinaison formantique analysée. Plusieurs modes de calcul de distance ont été testés. Ces modes permettent de faire varier le poids de chacun des éléments de la combinaison qui interviennent dans la comparaison des locuteurs. L'étude de la pertinence de chacune des voyelles et de chacune des combinaisons s'est effectuée essentiellement à partir des valeurs du taux de bonne reconnaissance.

Nous avons effectué l'analyse des résultats pour toutes les combinaisons de formants. Celle-ci nous a permis de mettre en évidence pour chacune d'elles les voyelles qui discriminent au mieux les locuteurs. Cette analyse nous a également permis de montrer que les voyelles pertinentes changent lorsque que la combinaison passe d'un formant unique à un couple ou à un triplet de formants. Nous avons également montré que la pertinence d'une voyelle varie avec le contexte syntaxico-sémantique de ses occurrences dans le corpus et qu'il n'est donc pas possible de conclure sur la pertinence d'une voyelle en général mais sur celle d'une voyelle dans un contexte linguistique bien précis. En revanche, les différents modes de calcul de distance ont plutôt peu influencé les classements des voyelles selon le taux de reconnaissance.

L'analyse des combinaisons d'écarts de formants a fourni peu d'informations supplémentaires.

Simultanément à l'étude de la pertinence, nous avons testé une variante de la méthode de décision du locuteur reconnu qui tient compte d'un coefficient de défiance associé à la distance entre deux répétitions de deux locuteurs. Celui-ci est une fonction de la robustesse des formants qui sont susceptibles d'intervenir dans le calcul de la distance. Cette méthode de décision n'a pas amélioré le taux de reconnaissance des locuteurs. Nous pensons que cela provient de la bonne robustesse de tous les formants, elle-même due aux bonnes conditions d'enregistrement des énoncés des locuteurs. Néanmoins, cette méthode pourrait être valide dans un système de reconnaissance testé dans une situation réelle qui engendrerait des formants de robustesse plus variable.

En ce qui concerne les prolongements immédiats de notre étude, nous avons déjà mentionné à la fin du chapitre III quelques compléments d'étude à effectuer soit pour conforter les résultats établis soit pour concevoir un premier système d'identification automatique du locuteur fondé sur les voyelles orales.



En ce qui concerne les prolongements à plus long terme, nous envisageons de continuer l'étude des paramètres que nous avons sélectionnés. En effet, nous avons défini un nombre important de paramètres susceptibles de caractériser le locuteur, à partir desquels nous avons élaboré et étiqueté un corpus conséquent. Nous n'avons analysé qu'une faible partie des données de ce corpus. Il serait dommage de ne pas exploiter davantage ce corpus. Lorsque nous avons débuté l'analyse de la pertinence des paramètres sélectionnés, nous avons décidé de commencer par l'étude qui nous semblait la plus simple, celle des trois premiers formants des voyelles orales. Nous pensions naïvement que cette étude serait vite terminée et que nous pourrions aborder celle d'autres paramètres qui nous semblaient a priori plus pertinents, les voyelles et les consonnes nasales. Mais en parole rien n'est simple même dans le cas des voyelles orales qui ont une structure acoustique apparemment simple et bien connue.

Par l'intermédiaire de ce travail, nous avons apporté notre contribution à la caractérisation du locuteur de langue française en vue de son identification automatique. Même si cette contribution est peu importante du point de vue du nombre de paramètres étudiés, très peu d'études ont été réalisées jusqu'à maintenant dans ce domaine pour le français. Dans l'introduction générale, nous avons mis l'accent sur la dichotomie de la recherche en reconnaissance automatique du locuteur entre les études qui utilisent implicitement la variabilité interlocuteur et celles qui cherchent à l'extraire explicitement du signal de parole sous la forme de paramètres acoustico-phonétiques caractérisant le locuteur. Toutefois, nous ne considérons pas ces deux orientations comme concurrentes mais au contraire comme complémentaires. Les méthodes de reconnaissance du locuteur conçues autour des modèles de Markov cachés, des réseaux neuronaux ou de la quantification vectorielle doivent utiliser les résultats des études du second type pour améliorer leurs performances afin d'obtenir des systèmes de reconnaissance d'un grand nombre de locuteurs. Dans notre cas, nous avons établi pour chaque combinaison de formants les voyelles orales les plus pertinentes. Ces techniques de reconnaissance peuvent donc utiliser directement ou indirectement les formants de ces voyelles selon que leur méthode d'analyse est fondée sur la détermination de formants ou sur une autre méthode d'analyse du signal de parole. Dans ce dernier cas, les voyelles devront être choisies en fonction de la méthode d'analyse. Par exemple, les méthodes de reconnaissance du locuteur fondée sur une distance d'Itakura utiliseront une voyelle pertinente pour la combinaison (F1, F2, F3) alors qu'une méthode à base de bancs de filtres pourra choisir des voyelles pertinentes pour des formants isolés.

Nous terminons ce mémoire par une remarque optimiste sur l'avenir de la recherche en caractérisation du locuteur. Lors de notre recherche bibliographique sur la caractérisation du locuteur, nous avons constaté que les études sur la caractérisation du locuteur à l'aide de paramètres acoustiques et phonétiques ont été principalement effectuées dans les années 70 ou au tout début des années 80. Lorsque nous avons débuté notre étude, ce type de recherche semblait avoir été complètement abandonné. Depuis quelques temps, comme le laisse apparaître de récentes publications, nous assistons à un regain d'études dans ce domaine [Bonastre 91] [Kraayeveld 91] [Eatock 92]. Nous y voyons plusieurs raisons. Nous venons déjà de mentionner la nécessité d'appliquer les techniques mathématiques de reconnaissance du locuteur à des paramètres pertinents. Une autre raison résulte des progrès réalisés en reconnaissance analytique de la parole continue multilocuteur. Ces progrès permettent d'envisager la segmentation automatique d'un texte quelconque ou plus simplement la recherche de segments acoustico-phonétiques de taille variable qui permettront de faire une reconnaissance automatique du locuteur robuste et indépendante du texte et ceci à partir d'un texte court. Bien que cela semble paradoxal, nous pensons que ce regain provient également de l'insuffisance actuelle des performances atteintes en reconnaissance automatique de la parole continue multilocuteur.

Deux démarches duales sont proposées pour améliorer ces performances, la normalisation du nouveau locuteur par rapport à un locuteur de référence ou l'adaptation du système de reconnaissance au nouveau locuteur. L'identification automatique du locuteur ou d'une classe de locuteurs permet d'adapter automatiquement un système de reconnaissance de parole à un nouveau locuteur.



## BIBLIOGRAPHIE



# BIBLIOGRAPHIE

## Abréviations employées dans la bibliographie :

- **JASA** : Journal of the Acoustical Society of America,
- **ICASSP** : International Conference on Acoustics, Speech and Signal Processing,
- **JEP** : Journées d'Etude sur la Parole,
- **EUROSPEECH** : European Conference on Speech Communication and Technology,
- **ASSP** : Acoustics, Speech and Signal Processing,
- **ICSLP** : International Conference on Spoken Language Processing.

- [Autesserre 85] D. Autesserre et M. Rossi.  
Propositions pour une segmentation et un étiquetage hiérarchisé : Application à la base de données acoustique du GRECO Communication Parlée.  
*14e JEP*, pages 147–151, Paris, 1985. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Autesserre 88a] D. Autesserre, G. Pérennou et M. Rossi.  
Méthodologie de transcription et d'étiquetage de corpus de parole.  
Rapport intermédiaire, Projet ESPRIT : Assessment, Methodology and Standardisation in Multilingual Speech Technology, 1988.
- [Autesserre 88b] D. Autesserre, G. Pérennou, M. Rossi et N. Vigouroux.  
Transcription et étiquetage du corpus de phrases françaises d'EUROM-0 : traduction en code API et en code SAM-PA.  
Rapport intermédiaire, Projet ESPRIT : Assessment, Methodology and Standardisation in Multilingual Speech Technology, 1988.
- [Boe 88a] L.J. Boe, L. Miclet et C. Sorin.  
L'étiquetage de la base de données des sons du français du GRECO Communication Parlée.  
*Premières Journées Nationales du GRECO-PRC "Communication Homme-Machine"*, pages 40–47, octobre 1988.
- [Boe 88b] L.J. Boe et Laurent Miclet.  
*Manuel d'étiquetage large*.  
GRECO numéro 39, juin 1988.
- [Bonastre 91] J.F. Bonastre, H. Méloni et P. Langlais.  
Analytical strategy for speaker identification.  
*EUROSPEECH, Genova, Italy*, pages 435–438, 1991.
- [Bonastre 92] J.F. Bonastre et H. Méloni.  
Etude de la variabilité spectrale pour la caractérisation du locuteur.  
*19e JEP*, pages 555–559, Bruxelles (Belgique), mai 1992. Groupe Communication Parlée de la Société Française d'Acoustique.



- [Calliope 89] Calliope.  
*La Parole et son Traitement Automatique.*  
CNET-ENST, Masson, Paris, 1989.
- [Carlson 90] R. Carlson, B. Granström et L. Nord.  
The KTH Speech Database.  
*Speech Communication*, 9:375–380, 1990.
- [Carre 84] R. Carre, J. Descout, J.J Mariani, M. Eskénazi et M. Rossi.  
The French Language Database. Defining, Planning and Recording a Large Database.  
*ICASSP, San-Diego, U.S.A*, pages 42.10.1–42.10.4, 1984.
- [Cole 92] R.A. Cole et Y.K Muthusamy.  
Perceptual Studies on Vowels Excised Form Continuous Speech.  
*ICSLP, Banff, Canada*, pages 1091–1094, October 1992.
- [Dalsgaard 89] P. Dalsgaard.  
Semi-automatic Phonemic Labelling of Speech Data Using a Self-organising Neural Network.  
*EUROSPEECH, Paris, France*, pages 541–544, 1989.
- [Dingeon 89] C. Dingeon, F. Alexandre, F. Guyot et J.-P. Haton.  
Un autre apprentissage cortical : différencier pour généraliser.  
*Actes Deuxièmes Journées Internationales sur les réseaux Neuro-Mimétiques*, Nîmes, novembre 1989.
- [Dolmazon 87] J.M. Dolmazon, P. Dalsgaard, S. Danielsen, M. Taylor et R. Winski.  
Visit Report of SAM USA Survey Group.  
Rapport, Projet ESPRIT : Assessment, Methodology and Standardisation in Multilingual Speech Technology, May 1987.
- [Dujour 90] A. Lacheret Dujour.  
Contribution à l'Analyse de la Variabilité Phonologique pour le Traitement Automatique de la Parole Continue Multilocuteur.  
Thèse de l'Université de Paris VII, juin 1990.
- [Eatock 92] J.P. Eatock et J.S.D. Mason.  
Phoneme performance in speaker recognition.  
*icslp92*, pages 1411–1414, October 1992.
- [Erp 89] A. Van Erp.  
Manual Labelling of Danish, English and French Speech Material on EUROM-0.  
Extract of the extension phase final report of sam project, Projet ESPRIT : Assessment, Methodology and Standardisation in Multilingual Speech Technology, February 1989.
- [Eskenazi 88] M. Eskenazi, F. Lonchamp et J. Vaissière.  
Cours sur les Indices Acoustiques du Français.  
GRECO - Communication Parlée, octobre 1988.
- [François 90] D. François et D. Fohr.  
Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de parole continue.

- 18e JEP, pages 191–195, Montréal (Canada), mai 1990. Groupe Communication Parlée de la Société Française d'Acoustique.
- [François 92] D. François et D. Fohr.  
Contribution de réseaux neuronaux pour la reconnaissance des occlusives au sein du système expert APHODEX.  
19e JEP, pages 405–408, Bruxelles (Belgique), mai 1992. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Goldstein 76] U.G. Goldstein.  
Speaker-identifying features based on formant tracks.  
*JASA*, 59(1):176–182, 1976.
- [Gong 90] Y. Gong et J.-P. Haton.  
Text-independent Speaker Recognition by Trajectory Space Comparison.  
*ICASSP, Albuquerque, U.S.A*, volume 1, pages 285–288, April 1990.
- [Gong 91] Y. Gong et J.-P. Haton.  
Non-Linear Vector Interpolation by Neural Network for Phoneme Identification in Continuous Speech.  
*ICASSP, Toronto, Canada*, pages 121–124, May 1991.
- [Gong 92] Y. Gong et J.P. Haton.  
DTW-based Phonetic Labeling Using Explicit Phoneme Duration Constraints.  
*ICSLP, Banff, Canada*, pages 863–866, October 1992.
- [Haton 91] J.P. Haton, J.M. Pierrel, G. Pérennou, J. Caelen et J.L. Gauvain.  
*Reconnaissance Automatique de la Parole*.  
Informatique. Dunod, Paris, 1991.
- [Hedelin 90] P. Hedelin et D. Huber.  
The CTH Speech Database: an Integrated Multilevel Approach.  
*Speech Communication*, 9:365–374, 1990.
- [Kabre 91] H. Kabre, G. Pérennou et N. Vigouroux.  
Automatic Labelling of Speech Signal into Phonetics Events.  
12e Congrès International des Sciences Phonétiques, pages 450–453, Aix-en-Provence (France), août 1991.
- [Kraayeveld 91] J. Kraayeveld, A.C.M. Rietveld et V.J. van Heuven.  
Speaker characterization in dutch using prosodic parameters.  
*EUROSPEECH, Genova, Italy*, pages 427–430, 1991.
- [Kuwabara 89] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa et T. Watanabe.  
Construction of a Large-Scale Japanese Speech Database and its Management System.  
*ICASSP, Glasgow, Scotland*, pages 560–563, 1989.
- [Kuwabara 90] H. Kuwabara et T. Takagi.  
Acoustic Parameters of Voice Individuality and Voice Quality Control by Analysis-Synthesis Method.  
*Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pages 140–142, Edinburgh, June 1990. European Speech Communication Association.

- [Ladefoged 57] P. Ladefoged et D.E. Broadbent.  
Information Conveyed by Vowels.  
*JASA*, 29(1):98–104, 1957.
- [Laprie 88] Y. Laprie.  
SNORRI : un système d'étude interactif de la parole.  
*17e JEP*, pages 71–76, Nancy, septembre 1988. Groupe Communication Parlée de la Société Française d'Acoustique.
- [Laprie 90] Y. Laprie, J.-P. Haton et J.-M. Pierrel.  
Phonetic Triplets in Knowledge Based-Approach of Acoustic-phonetic Decoding.  
*ICSLP, Kobe, Japan*, pages 365–368, November 1990.
- [Leung 84] H.C. Leung et V.W. Zue.  
A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech.  
*ICASSP, San-Diego, U.S.A*, pages 2.7.1–2.7.4, 1984.
- [Lonchamp 87] F. Lonchamp.  
Les sons du Français - Analyse acoustique descriptive.  
Document interne, Institut de Phonétique de Nancy, 1987.
- [Lonchamp 88] F. Lonchamp.  
Etudes sur la Production et la Perception de la Parole. Les Indices Acoustiques de la Nasalité Vocalique - La Modification du Timbre par la Fréquence Fondamentale.  
Thèse de Doctorat d'Etat ès Lettres et Sciences Humaines, Université de Nancy II, avril 1988.
- [Martinet 73] A. Martinet et H. Walter.  
*Dictionnaire de la Prononciation Française dans son Usage Réel*.  
France Expansion, Paris, 1973.
- [Monsen 83] R.B. Monsen et A.M. Engebretson.  
The Accuracy of Formant Frequency Measurements: A Comparison of Spectrographic Analysis and Linear Prediction.  
*Journal of Speech and Hearing Research*, 26:89–97, 1983.
- [Muthusamy 92] Y.K. Muthusamy, R.A. Cole et B.T. Oshika.  
The OGI Multi-Language Telephone Speech Corpus.  
*ICSLP, Banff, Canada*, pages 895–898, October 1992.
- [Nolan 83] F. Nolan.  
*The Phonetic Bases of Speaker Recognition*.  
Cambridge University Press, Great Britain, 1983.
- [O'Shaughnessy 87] D. O'Shaughnessy.  
*Speech Communication: Human and Machine*.  
Addison Wesley, Reading, Massachusetts, 1987.
- [Paliwal 84] K.K. Paliwal.  
Effectiveness of different vowels sounds in automatic speaker identification.  
*Journal of Phonetics*, 12:17–21, 1984.



- [Perennou 82] G. Perennou, G. Caelen, N. Vigouroux et M. Bréant.  
Identification et vérification du locuteur, timbre de la voix.  
Rapport final de la convention DRET numéro 79.34.658.00.470.7501,  
décembre 1982.
- [Perennou 88] G. Perennou et N. Vigouroux.  
Préliminaires méthodologiques pour une base de données acoustique  
phonétique.  
*17e JEP*, pages 9–13, Nancy, septembre 1988. Groupe Communication  
Parlée de la Société Française d'Acoustique.
- [Perennou 89] G. Perennou, M. de Calmès, I. Ferrané et J. Tihoni.  
Idiolecte et phonologie. Incidence sur la transcription automatique adaptée  
au locuteur par le système GEPH.  
*Séminaire sur la variabilité et la spécificité des locuteurs*, pages 68–  
77, Marseille Luminy, juin 1989. GRECO PRC Communication Homme-  
Machine.
- [Sambur 75] M.R. Sambur.  
Selection of acoustic features for speaker identification.  
*I.E.E.E. Transactions on ASSP*, pages 176–182, April 1975.
- [Su 74] L.S. Su, K.P. Li et K.S. Fu.  
Identification of speakers by use of nasal coarticulation.  
*JASA*, 56(6):1876–1882, 1974.
- [Svendsen 90] T. Svendsen et K. Kvale.  
Automatic Alignment of Phonemic Labels with Continuous Speech.  
*ICSLP, Kobe, Japan*, pages 997–1000, November 1990.
- [Wolf 72] J.J. Wolf.  
Efficient acoustic parameters for speaker recognition.  
*JASA*, 51(6):2044–2056, 1972.
- [Zue 88] V.W. Zue et S. Seneff.  
Transcription and alignment of the TIMIT Database.  
*Second Symposium on Advanced Man-Machine Interface through Spoken  
Language*, Oahu, November 1988.

# ANNEXE

## RESULTATS DE LA DETERMINATION DES FORMANTS DES VOYELLES ORALES

### 1 Introduction

Pour chacune des quinze voyelles orales étudiées, nous présentons, dans un premier tableau, les résultats de la détermination des formants intermédiaires de la première étude, et, dans un second tableau, les résultats de la détermination des formants finaux  $F_1$ ,  $F_2$  et  $F_3$  qui ont été utilisés pour discriminer les dix locuteurs.

Les deux tableaux associés à chacune des voyelles sont regroupés sur une double page. Celui de gauche concerne les formants intermédiaires de la première étude, celui de droite les formants finaux de la voyelle.

Nous allons tout d'abord décrire les informations contenues dans chacun des tableaux.

### 2 Description du tableau des formants intermédiaires de la première étude

Deux types de résultats sont regroupés dans ce tableau :

- les formants intermédiaires qui résultent de l'application de l'algorithme d'affectation de la première étude aux pôles issus d'une analyse LPC d'ordre 18 appliquée au prélèvement situé à 40% du début de la voyelle ;
- les informations issues de la vérification de ces formants à la fois par l'observation des spectrogrammes des triplets / **p-voyelle-R** / et / **b-voyelle-R** / et par l'examen des pôles bruts issus de l'analyse LPC. Les spectrogrammes de parole ont été calculés à l'aide du logiciel SNORRI et sont du même type que celui de la figure C.13.

Les colonnes du tableau contiennent donc les éléments suivants :

- **colonne Répétition** : les références du locuteur et de la répétition considérés ;
- **colonne  $F_0$**  : la valeur médiane de  $F_0$  sur la phrase. Cette valeur est calculée à partir des mesures de la fréquence fondamentale sur tous les prélèvements de la phrase comprises entre 50 et 500 Hz ;
- **colonnes  $F_1$ ,  $F_2$  et  $F_3$**  : les fréquences des formants intermédiaires  $F_1$ ,  $F_2$  et  $F_3$  issus de l'application de l'algorithme d'affectation aux pôles LPC (cf. page 39). Les résultats de l'étape de vérification visuelle figurent aussi dans ces colonnes sous la forme suivante :
  - lorsque la fréquence formantique est considérée comme fausse par rapport au formant observé sur le spectrogramme, la valeur dans le tableau est transcrite en caractères gras et soulignés. Une autre valeur est estimée à partir du spectrogramme et/ou des pôles LPC. Elle est notée dans la colonne "Remarques",
  - lorsque la fréquence formantique ne peut pas être vérifiée à l'aide du spectrogramme, la valeur dans le tableau est encadrée. C'est le cas notamment lorsque le formant est

invisible sur le spectrogramme, comme les formants d'ordre supérieur du [ u ], ou lorsqu'il est trop diffus,

- lorsque la fréquence formantique est considérée comme correcte par rapport au formant sur le spectrogramme, la valeur dans le tableau figure en caractères normaux sans aucun ajout ;
- **colonnes B<sub>1</sub>, B<sub>2</sub> et B<sub>3</sub>** : les largeurs de bande des pôles affectés aux trois formants ;
- **colonnes F<sub>4</sub> et F<sub>5</sub>** : les fréquences formantiques des deux pôles LPC dont la fréquence est comprise entre la borne supérieure de D(F3) et 4500 Hz. Les résultats de l'étape de vérification figurent aussi dans ces deux colonnes avec le même formalisme que celui employé pour les formants F1, F2 et F3 ;
- **colonnes B<sub>4</sub> et B<sub>5</sub>** : les largeurs de bande de ces deux pôles ;
- **colonnes T(aille)** : une information sur la largeur du formant sur le spectrogramme. Dans un spectrogramme à large bande, comme celui que nous avons utilisé pour effectuer la vérification, un formant correspond à une zone plus sombre, plus ou moins parallèle à l'axe temporel. La largeur de cette zone donne une information sur la largeur de bande du formant. Nous avons retenu les codes suivants :
  - **F** : indique que le formant apparaît sur le spectrogramme comme une bande étroite (200 à 300 Hz) dont la fréquence centrale, pour un prélèvement donné, est facilement déterminable et précise,
  - **M** : indique que le formant apparaît sur le spectrogramme comme une bande de largeur moyenne (400 à 600 Hz),
  - **L** : indique que le formant apparaît sur le spectrogramme comme une large bande (environ 1000 Hz) dont la fréquence centrale, pour un prélèvement donné, est difficilement déterminable et peu précise,
  - la colonne ne contient rien lorsque le formant se voit difficilement ou est complètement invisible sur le spectrogramme ;
- **colonnes A(illure)** : des informations sur l'évolution temporelle des fréquences fondamentale et formantiques notées manuellement lors de l'observation des spectrogrammes mais qui n'ont pas été reportées dans ces tableaux ;
- **colonne Remarques** : deux types d'informations :
  - une évaluation manuelle des fréquences formantiques considérées comme fausses :
    - la forme "**F<sub>i</sub>=**" indique que la nouvelle fréquence a été évaluée directement sur le spectrogramme ou qu'elle provient d'un pôle LPC qui a été éliminé par l'algorithme d'affectation mais dont la valeur est validée par le spectrogramme,
    - la forme "**F<sub>i</sub> brute =**" indique que la nouvelle fréquence formantique est celle d'un pôle LPC qui a été éliminé par l'algorithme d'affectation, qui pourrait correspondre au formant recherché, mais dont la validité est invérifiable sur le spectrogramme,



- quelques informations sur nos observations du spectrogramme du triplet, comme la présence de friction ou l'absence de formants sur le spectrogramme. Ainsi, dans le cas de F5, la remarque "pas de F5" indique que, malgré la présence d'une valeur dans la colonne F<sub>5</sub>, nous n'avons observé aucun formant sur le spectrogramme ayant une fréquence située entre F<sub>4</sub> et 4500 Hz.

Remarque : dans la première étude, nous avons également déterminé la durée de chacune des voyelles et effectué une écoute du triplet isolé. pour des raisons de mise en page, ces données figurent dans le tableau des formants finaux.

### 3 Description du tableau des formants finaux F1, F2 et F3

Dans ce tableau figurent les formants finaux F1, F2 et F3 de chacune des quinze voyelles étudiées ainsi que les coefficients de défiance qui leur sont associés. Chacun de ces formants a été établi à partir de trois formants intermédiaires selon la méthode présentée page 50. Chaque formant intermédiaire résulte de l'application de l'algorithme d'affectation décrit pages 46 et 47 aux pôles issus d'une analyse LPC d'ordre 18. Cette analyse a donc été appliquée à trois prélèvements de la voyelle :

- un prélèvement "central" situé à 80 ms du début pour les voyelles dont la durée dépasse 160 ms et à 50% du début pour les autres voyelles,
- deux autres prélèvements situés à 8 ms de part et d'autre du prélèvement central et qui correspondent aux deux autres formants intermédiaires.

Nous avons également vérifié la plausibilité de tous les formants finaux en comparant leurs fréquences formantiques aux fréquences formantiques résultant de la première étude et situées dans le tableau précédent. Pour cela, nous avons tenu compte de la durée de la voyelle, du nouvel emplacement de calcul et de l'allure du formant précisée dans le tableau précédent.

Les colonnes du tableau contiennent donc les éléments suivants :

- **colonne Répétition** : les références du locuteur et de la répétition considérés ;
- **colonne Durée (ms)** : la durée de la voyelle en millisecondes ;
- **colonne Durée (%)** : la durée de la voyelle en pourcentage par rapport à la durée moyenne des voyelles de la phrase qui ne sont pas suivies d'une pause. Nous avons donc éliminé les cas suivants : / voyelle-# / , / voyelle-hh-# / et / voyelle-R-# / ;
- **colonne F<sub>0</sub>** : la valeur médiane de F<sub>0</sub> sur la phrase. Cette valeur est calculée à partir des mesures de la fréquence fondamentale sur tous les prélèvements de la phrase comprises entre 50 et 500 Hz ;
- **colonnes F<sub>1</sub>, F<sub>2</sub> et F<sub>3</sub>** : les fréquences des formants finaux F1, F2 et F3 :
  - la valeur située dans le tableau est transcrite en caractères gras et soulignés dans le cas suivant : la fréquence formantique est considérée comme fausse par rapport aux valeurs de la première étude mais une valeur correcte a pu être déterminée soit à partir du spectrogramme soit à partir de la comparaison des formants intermédiaires aux résultats de la première étude. Cette nouvelle fréquence formantique figure dans la colonne "Remarques",
  - lorsque la fréquence formantique paraît fausse par rapport aux valeurs de la première étude mais qu'aucune autre valeur n'a pu être estimée, la valeur dans le tableau est encadrée. Dans ce cas, les fréquences des formants intermédiaires dont est issu le formant final figurent dans la colonne "Remarques",

- lorsque la fréquence formantique est considérée comme correcte par rapport au formant sur le spectrogramme, la valeur dans le tableau figure en caractères normaux sans aucun ajout ;
- **colonnes  $df_1$**  : le premier champ du coefficient de défiance associé au formant final. Ce champ donne une information sur les largeurs de bande des formants intermédiaires qui ont servi à établir le formant final. C'est un entier compris entre 0 et 3 inclus. Plus sa valeur est élevée plus les largeurs de bande sont importantes ;
- **colonnes  $df_2$**  : le deuxième champ du coefficient de défiance associé au formant final. Ce champ fournit une information sur la proximité des trois formants intermédiaires dont est issu le formant final et donc sur la robustesse du formant final. C'est un entier compris entre 0 et 5 inclus. Plus sa valeur est élevée moins la fréquence formantique est robuste (cf. table C.9) ;
- **colonne Ecoute** : le résultat de l'écoute du triplet isolé / **p-voyelle-R** / (ou / **b-voyelle-R** /), que nous avons effectuée lors de la vérification des résultats de la première étude. Réalisée par une non-phonéticienne, cette écoute est grossière et subjective. Mais elle permet de vérifier la prononciation de la voyelle et de donner une indication sur la structure plus ou moins vocalique du R ainsi que sur la présence d'un bruit de friction entre l'occlusive et la voyelle ;
- **colonne Remarques** : deux types d'informations :
  - quelques informations complémentaires sur le triplet prononcé, comme le fait qu'il soit suivi ou non d'une pause,
  - un complément d'information lorsque les fréquences formantiques sont considérées comme fausses :
    - lorsqu'une nouvelle valeur a pu être estimée :
      - la forme " **$F_i$  manuelle =**" indique que la nouvelle fréquence a été évaluée directement à partir du spectrogramme,
      - la forme " **$F_i$  ="**" indique que la nouvelle fréquence formantique est issue de la première étude parce qu'elle n'a pas pu être déterminée à partir du spectrogramme,
    - lorsqu'aucune nouvelle valeur n'a pu être estimée, nous avons indiqué, dans l'ordre chronologique, les fréquences des trois formants intermédiaires qui ont servi à établir le formant final.

Remarque : les valeurs de durée en millisecondes et les résultats de l'écoute manquent pour les répétitions  $df_4$ ,  $jfm_4$  et  $ms_4$  car celles-ci ont été enregistrées plus tard à la suite d'une erreur lors du premier enregistrement.

## 4 Tableaux

Les pages suivantes regroupent les quinze couples de tableaux associés aux voyelles orales étudiées.





Les tableaux sont sur des pages numérotées à partir de A.8.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	142		457	42		L	<u>1926</u>	417		L	2465	116		L	3438	458		L					$F_2 = 1750$
aq2	117		419	43		L	1768	160		L	2540	124		M	3596	157		L					
aq3	137		450	30		L	1751	132		L	2538	99		M	3663	160		M					
aq4	123		443	47		L	1718	54		M	2474	102		M	3655	91		L					
bz1	114		423	32		M	1696	90		M	2392	106		M	3563	76		L					
bz2	112		439	52		L	1555	183		L	2223	127		M	3498	131		M					
bz3	112		352	52		M	1755	101		M	2461	79		M	3563	114		L					
bz4	112		429	42		M	1664	147		M	2378	94		M	3656	131		M					
df1	170		438	159		L	1834	44		M	<u>2740</u>	188		L	3786	220		L					
df2	173		400	123		M	1760	169		M	2591	134		M	3541	241		L					
df3	173		397	50		L	1877	43		M	2668	131		M	<u>3662</u>	596		L					
gm1	190		411	41		L	1770	306		M	2397	133		M	3968	381		L	4370	384		L	
gm2	190		432	45		L	1736	187		M	2380	105		M	3466	410		L	<u>4326</u>	495		L	$F_5 = 3900$
gm3	181		508	109		L	1735	171		M	2301	63		M	<u>3631</u>	814		L	<u>3852</u>	476			
gm4	185		486	113		L	1788	91		M	2377	52		L	3651	162		L					
jfm1	117		505	49		L	1950	54		L	2748	348		M	3962	166		L					
jfm2	123		535	40		L	1927	85		M	2734	199		M	3987	151		M					
jfm3	121		521	49		L	1942	64		M	2732	267		L	<u>3516</u>	468		L	<u>4067</u>	110			
jg1	102		497	62		L	1682	173		M	2473	122		M	3439	262		L					
jg2	117		477	70		L	1746	182		M	2495	111		M	3467	259		M					
jg3	126		497	97		L	1742	103		M	2464	61		L	3348	242		M					
jg4	111		479	57		L	1741	219		M	2419	125		M	3376	182		M					
jlc1	123		478	78		L	<u>1844</u>	224		L	2577	102		L	<u>3561</u>	265		L	<u>3945</u>	587			
jlc2	119		432	40		L	1771	190		M	2457	64		L	<u>3138</u>	586		L	<u>3693</u>	354		L	
jlc3	115		471	70		L	1774	191		M	2482	135		L	<u>3445</u>	274		L	<u>4007</u>	421		L	
jlc4	125		413	83		L	1883	223		M	2535	143		L	<u>3405</u>	379		L	<u>3848</u>	342		L	
jmp1	101		503	92		M	1685	215		M	2451	105		M	3629	247		M	<u>3870</u>	791			$F_5 = 4100$
jmp2	101		463	71		L	1574	296		M	2391	102		M	<u>3299</u>	569		M	4136	546		M	$F_4 = 3460$
jmp3	101		478	66		L	<u>0</u>	0		F	2384	146		M	<u>3547</u>	190		L	4129	631			$F_2 = 1535$
jmp4	103		470	140		L	<u>0</u>	0		M	2453	99		L	3532	280		L					$F_2 = 1638$
jph1	137		440	66		L	1767	116		M	2455	73		M	<u>3265</u>	411		L	<u>3613</u>	131		L	$F_4 = 3500$ ; pas de F5
jph2	142		441	65		L	1753	139		M	2443	126		M	<u>3628</u>	308		L					
jph3	140		474	57		L	1670	140		M	2420	92		M	<u>3395</u>	629		L	<u>3577</u>	135			$F_4 = 3577$
jph4	142		446	63		L	1737	95		M	2396	155		L	<u>3565</u>	231		L					
ms1	117		411	78		M	1791	105		M	2471	89		M	<u>3693</u>	171		L					
ms2	111		392	54		M	1741	114		M	2493	100		M	<u>3809</u>	234		L					
ms3	117		416	57		L	1662	102		M	2447	50		M	3634	131		M					

Table 1. Formants intermédiaires de la voyelle /ai/ de la phrase 7 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	72	76	142	474	0	0	1704	2	0	2444	0	0	$\epsilon$	
aq2	64	69	117	459	1	0	1725	0	0	2525	0	0	$\epsilon$	
aq3	72	82	137	466	0	0	1698	0	0	2532	0	0	$\epsilon$	
aq4	72	73	123	474	0	0	1695	0	0	2457	0	0	$\epsilon$	
bz1	88	86	114	445	0	0	1673	0	0	2369	0	0	$\epsilon$	
bz2	64	61	112	451	0	0	1533	2	0	2218	0	0	$\epsilon$	
bz3	64	67	112	369	0	0	1735	0	0	2456	0	0	$\epsilon$	
bz4	64	67	112	444	0	0	1621	0	0	2357	0	0	$\epsilon$	
df1	72	87	170	472	0	0	1808	0	0	2681	0	0	$\epsilon$	
df2	32	50	173	400	0	0	1767	0	0	2574	0	0	$\epsilon$ -e	
df3	48	64	173	421	0	0	1854	0	0	2691	0	0	$\epsilon$	
df4		100	166	<del>546</del>	0	0	1725	1	0	2717	2	1		(533, 554, 551)
gm1	56	74	190	500	2	0	1806	0	0	2370	0	0	$\epsilon$	
gm2	48	57	190	437	2	0	1679	0	0	2354	0	0	$\epsilon$	
gm3	48	58	181	537	0	0	1706	0	0	2327	0	0	$\epsilon$	
gm4	56	56	185	524	1	0	1788	0	0	2372	0	0	$\epsilon$	
jfm1	72	72	117	528	0	0	1914	0	0	2711	0	0	$\epsilon$	
jfm2	72	76	123	557	0	0	1891	0	0	2713	0	0	$\epsilon$	
jfm3	80	77	121	550	0	0	1901	0	0	2698	0	0	$\epsilon$	
jfm4		86	129	597	0	0	1778	0	0	2708	0	0	$\epsilon$	
jg1	56	60	102	511	0	0	1630	0	0	2487	0	0	$\epsilon$	
jg2	72	75	119	494	0	0	1707	0	0	2502	0	0	$\epsilon$	
jg3	56	70	126	516	0	0	1678	1	0	2493	0	0	$\epsilon$	
jg4	40	48	111	500	0	0	1729	2	1	2433	0	0	$\epsilon$	
jlc1	48	55	123	515	1	0	1838	0	1	2578	0	0	$\epsilon$	
jlc2	48	57	119	470	1	0	1741	0	0	2462	0	0	$\epsilon$	
jlc3	32	38	115	467	1	0	1757	1	0	2482	0	0	$\epsilon$	
jlc4	48	62	125	459	1	0	1834	0	0	2512	0	0	$\epsilon$	
jmp1	48	68	101	500	0	0	1632	1	0	2466	0	0	$\epsilon$	
jmp2	40	49	101	479	0	0	1554	0	1	2405	0	0	$\epsilon$	
jmp3	64	80	101	502	0	0	<u>0</u>	5	0	2431	0	0	$\epsilon$	$F_2$ manuelle = 1535
jmp4	40	55	103	487	0	0	<u>0</u>	5	0	2441	0	0	$\epsilon$	$F_2$ manuelle = 1638
jph1	64	67	137	477	1	0	1723	0	0	2431	0	0	$\epsilon$	
jph2	40	38	142	455	0	0	1749	0	0	2454	0	0	$\epsilon$	
jph3	56	60	140	<u>563</u>	2	0	1612	1	0	2419	0	0	$\epsilon$	$F_1$ manuelle = 470
jph4	56	62	142	451	0	0	1749	0	0	2392	0	0	$\epsilon$	
ms1	48	46	117	460	1	0	1731	1	0	2411	0	0	pja-pj $\epsilon$	
ms2	56	63	111	443	1	0	1682	0	0	2435	0	0	pja-pj $\epsilon$	
ms3	32	44	117	412	1	0	1663	0	0	2441	0	0	pja-pj $\epsilon$	
ms4		75	103	508	0	0	1489	2	0	2278	0	0		

Table 1 bis. Formants finaux de la voyelle /ai/ de la phrase 7.  
Analyse LPC à 50% du début de la voyelle.

Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	131		491	57		L	1647	128		L	2358	138		M	3514	419							F4 peu visible
aq2	142		592	109		L	1667	121		L	2400	84		M	<del>3309</del>	800							F4 peu visible
aq3	153		521	42		L	1640	113		L	2393	50		M	<del>3125</del>	828			<del>4189</del>	523			
aq4	150		544	154		L	<del>1739</del>	136		L	2508	77		M	3944	544		M					F4 peu visible
bz1	106		459	113		M	1700	268		M	2463	124		M	3673	140		M					
bz2	102		448	53		M	1653	173		M	2374	41		M	3667	282		M					
bz3	106		526	111		L	1613	180		M	2360	51		M	3837	87		M					
bz4	106		437	90		M	1677	212		M	2395	147		M	3553	212		M					
df1	173		559	207		L	1681	64		M	2939	261		M	<del>4217</del>	777		F					$F_4 = 3800$
df2	173		502	262		L	1726	96		M	2843	132		M	3806	662		F					
df3	170		<del>0</del>	0		F	1750	64		M	2980	230		F	4002	611		F					
gm1	153		510	60		M	1981	79		M	2628	60		M	<del>4418</del>	436		L					
gm2	177		509	47		M	1862	147		M	2541	98		M	<del>4294</del>	228		L					
gm3	173		486	73		M	1849	120		M	2673	119		M	<del>4086</del>	443		L					
gm4	181		518	54		M	1856	73		M	2592	64		M	4135	352		L					
jfm1	115		602	35		L	1935	50		M	2838	184		M	3910	145		M					
jfm2	114		597	40		M	1976	52		M	2847	276		M	<del>3994</del>	436		M					
jfm3	111		616	21		M	1971	38		M	2786	202		M	3958	168		M					
jg1	87		525	123		M	1748	84		M	2471	83		M	<del>3351</del>	549							pas de F4
jg2	101		508	103		M	1736	76		F	2445	89		M	<del>3458</del>	394							pas de F4
jg3	105		534	68		M	1705	120		M	2443	138		M	<del>3251</del>	710							pas de F4
jg4	88		520	46		M	1678	69		F	2412	62		F	<del>3158</del>	316							
jlc1	114		455	194		L	1809	249		M	2535	122		M	3618	282		M	<del>4354</del>	700			pas de F5
jlc2	117		491	62		M	1758	120		M	2459	80		M	3656	262		M					
jlc3	115		455	176		M	1725	293		M	2531	173		M	3849	134		M					
jlc4	111		498	81		M	1739	183		M	2511	123		M	3737	88		M					
jmp1	95		557	51		L	1550	192		M	2417	61		M	3771	333		F					
jmp2	101		561	113		L	<del>2113</del>	948		F	2648	246		F	<del>3622</del>	465		F					$F_2 = 1500$
jmp3	106		541	76		M	1598	196		M	2428	57		M	<del>3672</del>	450		F	<del>4176</del>	463			$F_4 = 3900$ ; pas de F5
jmp4	101		543	99		L	1569	484		M	2457	38		M	3799	140		M					
jph1	137		522	42		L	1511	131		M	2298	97		M	<del>3272</del>	351		M	<del>3478</del>	134			$F_4 = 3478$
jph2	148		535	64		L	1569	172		M	2320	105		M	<del>3136</del>	564		M	<del>3485</del>	128			$F_4 = 3485$
jph3	145		516	29		M	1536	85		M	2316	78		M	3395	173		M					
jph4	150		533	35		L	1593	160		M	2336	103		M	<del>3209</del>	644		L	<del>3418</del>	140			$F_4 = 3418$
ms1	117		406	70		M	1784	105		M	2458	153		M	<del>3698</del>	159		L					
ms2	108		475	162		M	1738	146		M	2394	189		M	3661	366		L	<del>3856</del>	392			pas de F5
ms3	109		417	86		M	1720	159		M	2412	100		M	3739	152		L					

Table 2. Formants intermédiaires de la voyelle /ai/ de la phrase 9 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	264	234	131	516	0	0	1609	0	0	2329	0	0	$\epsilon$	
aq2	280	272	142	539	0	0	1571	0	0	2392	0	0	$\epsilon$	
aq3	272	223	153	519	0	0	1618	0	0	2424	0	0	$\epsilon$	
aq4	264	208	150	562	0	0	1642	0	0	2454	0	0	$\epsilon$	
bz1	296	275	106	456	0	0	1604	0	0	2389	0	0	$\epsilon$	
bz2	312	267	102	468	0	0	1548	0	0	2389	0	0	$\epsilon$	
bz3	328	281	106	499	0	0	1560	0	0	2371	0	0	$\epsilon$	
bz4	384	337	106	458	0	0	1595	0	0	2411	0	0	$\epsilon$	
df1	208	264	173	561	0	0	1690	0	0	2941	0	0	$\epsilon$	
df2	184	215	173	515	2	1	1717	0	0	2851	0	0	$\epsilon$	
df3	160	180	170	514	4	0	1814	0	0	2970	0	0	$\epsilon$	(0, 335, 0)
df4		305	200	609	0	0	1586	0	0	2749	0	0		
gm1	248	216	153	497	0	0	1966	0	0	2583	0	0	$\epsilon$	
gm2	264	267	177	516	0	0	1840	0	0	2533	0	0	$\epsilon$	
gm3	248	241	173	494	0	0	1820	0	0	2625	1	0	$\epsilon$	
gm4	296	280	181	465	0	0	1845	0	0	2574	0	0	$\epsilon$	
jfm1	240	248	115	618	0	0	1914	0	0	2783	0	0	$\epsilon$	
jfm2	232	217	114	589	0	0	1965	0	0	2836	0	0	$\epsilon$	
jfm3	216	197	111	626	0	0	1972	0	0	2769	2	0	$\epsilon$	
jfm4		233	145	594	0	0	1967	0	0	2847	2	0		
jg1	224	231	89	524	0	0	1731	0	0	2476	0	0	$\epsilon$	
jg2	256	217	102	491	0	0	1696	0	0	2460	0	0	$\epsilon$	
jg3	264	243	105	538	0	0	1664	0	0	2445	0	0	$\epsilon$	
jg4	264	270	88	506	0	0	1676	0	0	2407	0	0	$\epsilon$	
jlc1	232	219	115	464	0	0	1717	0	0	2474	0	0	$\epsilon$	
jlc2	208	191	117	512	0	0	1756	0	0	2499	0	0	$\epsilon$	
jlc3	232	202	115	475	0	0	1698	0	0	2532	0	0	$\epsilon$	
jlc4	216	223	111	497	0	0	1754	0	0	2529	0	0	$\epsilon$	
jmp1	184	203	96	552	0	0	1579	0	1	2429	0	0	$\epsilon$	
jmp2	176	196	101	569	0	0	0	4	0	2484	2	0	$\epsilon$	$F_2$ manuelle = 1500
jmp3	176	180	106	503	1	0	1618	0	1	2464	0	0	$\epsilon$	
jmp4	200	213	101	550	0	0	1571	0	1	2452	0	0	$\epsilon$	
jph1	248	256	137	524	0	0	1498	0	0	2273	0	0	$\epsilon$	
jph2	256	270	148	513	0	0	1541	0	0	2330	0	0	$\epsilon$	
jph3	224	212	145	514	0	0	1547	0	0	2314	0	0	$\epsilon$	
jph4	224	214	150	527	0	0	1548	0	0	2317	0	0	$\epsilon$	
ms1	296	220	117	429	0	0	1755	0	0	2404	0	0	eR&	
ms2	232	219	109	469	0	0	1702	0	0	2370	0	0	$\epsilon$	
ms3	264	252	109	443	0	0	1687	0	0	2399	0	0	$\epsilon$	
ms4		274	96	430	0	0	1714	0	0	2360	0	0		

Table 2 bis. Formants finaux de la voyelle /ai/ de la phrase 9.  
Analyse LPC à 80 ms du début de la voyelle.



Répe- tition	F <sub>0</sub> (Hz)	A	F <sub>1</sub> (Hz)	B <sub>1</sub> (Hz)	A	T	F <sub>2</sub> (Hz)	B <sub>2</sub> (Hz)	A	T	F <sub>3</sub> (Hz)	B <sub>3</sub> (Hz)	A	T	F <sub>4</sub> (Hz)	B <sub>4</sub> (Hz)	A	T	F <sub>5</sub> (Hz)	B <sub>5</sub> (Hz)	A	T	Remarques	
aq1	140		455	81		L	1310	365		M	2360	30		M	3598	123		M						
aq2	126		476	72		L	1130	184		M	2335	115		M	3572	94		M						
aq3	133		487	106		L	1245	414		L	2386	51		M	<b>3448</b>	290		L						F <sub>4</sub> = 3600
aq4	126		479	61		L	1166	384		L	2450	232		L	<b>3375</b>	273		L						
bz1	105		414	49		M	1222	183		M	2274	87		M	3564	35		M						
bz2	109		451	57		L	1267	94		M	2342	44		M	3637	124		M	<b>3768</b>	834				pas de F5
bz3	106		441	68		L	1251	154		M	2274	84		M	3581	99		M						
bz4	106		439	38		L	1240	82		M	2284	64		M	3637	56		L						
df1	170		553	100		L	1396	143		L	2625	91		M	3669	120		M						
df2	173		494	117		L	1364	187		M	2661	155		M	3710	133		M						
df3	166		472	119		L	<b>1446</b>	230		L	2635	68		M	3730	463		M						
gm1	190		499	32		L	1461	495		L	2441	37		M	3762	276		F						
gm2	177		482	84		L	<b>1188</b>	489		L	2354	105		M	3447	405		F						F <sub>2</sub> = 1400
gm3	181		472	39		L	1355	214		M	2306	56		M	3657	137		M						
gm4	177		475	38		L	1392	71		M	2273	254		M	3577	90		M						
jfm1	123		544	55		L	1496	121		M	2520	62		M	3795	81		L	<b>4274</b>	625				pas de F5
jfm2	126		554	59		L	1525	149		M	2562	69		M	<b>3858</b>	107		L	<b>4033</b>	612				
jfm3	121		557	55		L	1519	220		M	2562	50		M	<b>3780</b>	163		L	<b>3973</b>	403				
jl1	94		465	80		L	1454	283		M	2410	85		L	3357	293		M						
jl2	102		451	130		L	1406	424		M	2351	55		L	3336	188		M						
jl3	114		491	105		L	1475	274		M	2399	59		M	3302	192		M						
jl4	102		449	65		L	1319	386		M	2383	111		M	3274	211		M						
jl1c1	119		430	51		L	1345	127		M	2430	162		M	<b>3234</b>	188		L	<b>3845</b>	491			L	
jl1c2	121		451	74		L	1326	82		M	2387	44		M	<b>3374</b>	360		L	<b>3952</b>	294			L	
jl1c3	115		438	60		L	1363	49		M	2448	70		L	<b>3198</b>	199		L	<b>3739</b>	186			L	
jl1c4	111		480	140		L	1348	199		M	2472	207		L	<b>3111</b>	807		L	<b>3601</b>	220			L	
jmp1	100		503	94		L	1244	287		F	2243	468		M	3558	159		M						
jmp2	105		466	59		L	1242	208		M	2287	58		M	3518	82		M	<b>4095</b>	640			M	F <sub>5</sub> = 4290
jmp3	105		513	92		L	1246	206		M	2322	88		M	3547	81		M	<b>4458</b>	601			M	F <sub>5</sub> = 4290
jmp4	105		523	61		M	<b>1235</b>	168		L	2319	67		M	3493	125		L	<b>4466</b>	307			L	
jph1	125		485	22		L	1279	118		M	2155	91		M	<b>3323</b>	149		L						
jph2	131		463	59		L	1201	154		M	2230	145		M	<b>3382</b>	144		L						
jph3	133		516	50		L	1206	150		M	2223	66		M	<b>3350</b>	154		L						
jph4	137		504	82		L	1217	131		M	2228	89		M	<b>3298</b>	132		L						
ms1	114		440	63		L	1413	289		L	2154	102		M	<b>3295</b>	849		L	<b>3550</b>	131			L	
ms2	112		405	84		L	1508	122		M	2202	43		M	<b>3459</b>	207		M	3782	435			M	F <sub>4</sub> = 3300
ms3	115		437	71		L	1438	233		M	2231	62		M	3538	110		L						

Table 3. Formants intermédiaires de la voyelle /oe/ de la phrase 4 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.

Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	176	190	140	468	0	0	1312	0	0	2381	0	0	œ	œRa
aq2	280	286	126	489	0	0	1135	0	0	2326	0	0	œ	œR#a
aq3	280	338	135	507	0	0	1249	0	0	2381	0	0	œ	œR#a
aq4	272	313	126	473	0	0	1270	0	0	2411	0	0	œ	œR#a
bz1	272	345	105	414	0	0	1217	0	0	2284	0	0	œ	œR#a
bz2	336	378	109	445	0	0	1272	0	0	2353	0	0	œ	œR#a
bz3	296	401	106	433	0	0	1256	0	0	2320	0	0	œ	œR#a
bz4	328	421	106	431	0	0	1232	0	0	2312	0	0	œ	œR#a
df1	104	143	170	555	0	0	1381	0	0	2659	0	0	œ	œRa
df2	104	140	173	536	0	0	1398	0	0	2652	0	0	œ	œRa
df3	96	127	169	492	0	0	1456	0	0	2638	0	0	œ	œRa
df4		240	153	556	0	0	1389	0	0	2625	0	0		œR#a
gm1	176	196	190	506	0	0	1447	0	2	2455	0	0	œ	œRa
gm2	248	241	177	475	0	0	1389	0	1	2368	0	0	œ	œRa
gm3	184	188	181	491	0	0	1375	0	0	2315	0	0	œ	œRa
gm4	312	359	177	460	0	0	1355	0	0	2296	0	0	œ	œR#a
jfm1	160	154	123	532	0	0	1505	0	0	2506	0	0	œ	œRa
jfm2	144	143	126	542	0	0	1546	0	0	2560	0	0	œ	œRa
jfm3	160	161	121	546	0	0	1552	0	0	2560	0	0	œ	œRa
jfm4		186	137	520	0	0	1511	0	0	2489	0	0		œRa
jg1	224	222	95	477	0	0	1473	0	0	2406	0	0	œ	œRa
jg2	208	193	103	449	0	0	1415	0	1	2360	0	0	œ	œRa
jg3	128	157	117	476	0	0	1483	0	1	2414	0	0	œ	œRa
jg4	144	172	102	457	0	0	1357	0	0	2392	0	0	œ	œRa
jlc1	240	325	119	422	0	0	1364	0	0	2440	0	0	œ-eu	œR#a
jlc2	232	284	123	469	0	0	1346	0	0	2403	0	0	œ-eu	œR#a
jlc3	264	257	117	468	0	0	1365	0	0	2437	0	0	œ-eu	œR??a
jlc4	264	297	111	470	0	0	1335	0	0	2437	0	0	œ-eu	œR#a
jmp1	248	315	100	507	0	0	1255	0	0	2278	0	0	œ	œR#a
jmp2	128	161	105	473	0	0	1241	0	0	2286	0	0	œ	œRa
jmp3	152	177	105	503	0	0	1235	0	0	2336	0	0	œ	œRa
jmp4	264	348	105	525	0	0	1215	0	0	2325	0	0	œ	œR#a
jph1	208	215	125	484	0	0	1270	0	0	2166	0	0	œ	œRa
jph2	128	143	131	485	0	0	1238	0	0	2218	0	0	œ	œRa
jph3	192	196	133	520	0	0	1213	0	0	2225	0	0	œ	œRa
jph4	168	179	137	525	0	0	1251	0	0	2232	0	0	œ	œRa
ms1	240	253	114	428	0	0	1382	0	0	2161	0	0	œ-eu	œR#a
ms2	160	185	112	407	0	0	1513	0	0	2209	0	0	œ-eu	œRa
ms3	240	267	115	440	0	0	1405	0	0	2230	0	0	œ	œR#a
ms4		258	111	438	0	0	1376	0	0	2103	0	0		œR#a

Table 3 bis. Formants finaux de la voyelle /œ/ de la phrase 4.  
Analyse LPC à 50% ou à 80 ms du début de la voyelle, selon la répétition.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	153		442	87		M	<u>1161</u>	351		M	2327	49		M	3555	91		M					$F_2 = 1300$
aq2	150		460	31		M	1287	371		M	2379	111		M	3578	72		M					
aq3	153		422	45		M	1232	143		M	2345	110		M	3511	57		M					
aq4	140		448	67		M	1287	113		M	2384	102		M	3384	185							
bz1	109		455	36		M	1286	147		M	2259	41		M	3501	328		M					
bz2	109		427	59		M	1273	101		M	2309	60		M	3571	163		M					
bz3	111		446	25		M	1270	97		M	2247	59		M	3511	109		M					
bz4	110		452	25		M	1284	105		M	2284	57		M	3591	103		L					
df1	170		523	53		L	<u>1418</u>	297		M	2659	160		M	3726	77		M					$F_2 = 1360$
df2	170		498	105		L	1302	237		M	2689	114		M	3630	153		M					
df3	170		472	72		L	1360	199		M	2684	207		M	3682	296		M					
gm1	163		435	48		L	<u>1256</u>	353		L	2298	145		M	<u>3815</u>	236		L					
gm2	181		463	91		L	1251	327		M	2237	145		M	<u>3427</u>	288		L					
gm3	186		443	45		L	<u>1208</u>	390		M	2302	55		M	3563	53		M					$F_2 = 1330$
gm4	195		488	91		L	1342	353		M	2263	132		M	3469	536		L					
jfm1	123		549	56		L	1636	118		M	2506	75		M	3847	281		L					
jfm2	109		527	63		L	1630	133		M	2548	39		M	3794	177		L	<u>4189</u>	979			pas de F5
jfm3	123		523	35		L	1570	238		M	2557	50		M	<u>3722</u>	187		L	<u>3989</u>	252			
jg1	94		485	74		L	1333	425		M	2420	103		M	3330	129		M					
jg2	96		463	101		L	1346	255		M	2418	100		M	3358	149		M					
jg3	115		470	46		L	1350	219		M	2434	93		L	<u>3229</u>	234		M					
jg4	112		526	57		L	1274	127		M	2436	63		M	3284	61		M					
jlc1	125		436	82		M	1314	106		M	2393	163		M	<u>3235</u>	198		L	<u>3649</u>	106		L	
jlc2	129		421	55		M	1279	126		M	2367	61		M	<u>3142</u>	315		L	<u>3607</u>	81		L	
jlc3	112		427	88		M	1392	204		M	2464	128		L	<u>3118</u>	297		L	<u>3588</u>	159		L	
jlc4	114		416	63		M	1336	157		M	2423	159		L	<u>3118</u>	473		L	<u>3695</u>	193		L	
jmp1	98		491	85		L	1147	278		M	2267	124		M	3481	170		M	<u>3985</u>	666		M	$F_5 = 4300$
jmp2	100		462	128		L	1261	295		M	2283	120		M	3506	148		M	<u>4016</u>	531		L	$F_5 = 4250$
jmp3	97		459	126		L	1138	342		M	2232	94		M	3419	79		M	<u>4089</u>	277		M	$F_5 = 4250$
jmp4	103		444	48		M	1289	378		M	2230	107		M	3544	180		L					
jph1	131		454	40		L	1304	195		M	2336	91		L	<u>3394</u>	237		L					
jph2	137		483	22		L	1236	138		M	2471	247		M	<u>3321</u>	289		L	<u>3664</u>	652		L	
jph3	145		485	22		L	1282	102		M	2419	113		M	<u>3336</u>	270		L	<u>3619</u>	277		L	
jph4	145		492	15		L	1279	105		M	2438	100		M	<u>3248</u>	158		L	<u>3557</u>	349			
ms1	112		395	53		M	1355	176		M	2131	90		M	<u>3168</u>	763		M	<u>3634</u>	167		M	
ms2	114		456	83		M	1356	162		M	2177	35		M	<u>3230</u>	225		M	<u>3908</u>	218		M	
ms3	111		412	38		M	1405	212		L	2237	35		M	<u>3449</u>	297		M	<u>3648</u>	386		M	

Table 4. Formants intermédiaires de la voyelle /oe/ de la phrase 10 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	96	105	153	445	0	0	1217	1	0	2350	0	0	œ	
aq2	152	145	150	458	0	0	1262	0	1	2360	0	0	œ	
aq3	136	139	153	449	0	0	1236	0	0	2325	0	0	œ	
aq4	168	155	140	458	0	0	1258	0	0	2345	0	0	œ	
bz1	256	313	109	449	0	0	1279	0	0	2258	0	0	œ	
bz2	280	228	109	430	0	0	1256	0	0	2306	0	0	œ	
bz3	200	168	111	445	0	0	1267	0	0	2246	0	0	œ	
bz4	176	162	110	449	0	0	1276	0	0	2278	0	0	œ	
df1	128	148	170	510	0	0	1408	0	0	2722	0	0	œ-O	
df2	112	148	170	500	0	0	1275	0	0	2699	0	0	œ	
df3	136	163	170	471	0	0	1385	0	0	2694	0	0	œ	
df4		139	170	532	0	0	1382	0	0	2636	0	0		
gm1	112	116	163	428	0	0	1269	2	1	2295	0	0	œ	
gm2	88	95	181	485	0	0	1185	0	1	2262	0	0	œ	
gm3	88	98	186	477	0	0	1317	0	0	2261	0	0	œ	
gm4	80	93	195	515	0	0	1328	0	1	2244	0	0	œ	
jfm1	128	130	123	554	0	0	1631	0	0	2519	0	0	œ	
jfm2	184	185	109	528	0	0	1613	0	0	2550	0	0	œ	
jfm3	128	129	123	547	0	0	1578	0	0	2555	0	0	œ	
jfm4		112	163	520	0	0	1427	0	0	2509	0	0		
jg1	112	110	95	493	0	0	1321	0	2	2428	0	0	œ	
jg2	104	108	96	487	0	0	1293	0	0	2421	0	0	œ	
jg3	88	98	115	502	0	0	1276	0	0	2444	0	0	œ	
jg4	72	89	112	535	0	0	1241	0	0	2419	0	0	œ	
jlc1	168	165	125	430	0	0	1317	0	0	2394	0	0	œ	
jlc2	120	139	129	430	0	0	1293	0	0	2385	0	0	œ-eu	
jlc3	136	159	112	422	0	0	1331	0	0	2442	0	0	œ	
jlc4	128	150	114	427	0	0	1321	0	0	2369	0	0	œ	
jmp1	104	119	98	496	0	0	1151	0	0	2270	0	0	œ	
jmp2	88	93	100	499	0	0	1239	0	0	2328	0	0	œ	
jmp3	96	118	97	460	0	0	1124	0	1	2240	0	0	œ-O	
jmp4	80	102	103	446	0	0	1236	0	0	2258	0	0	œ	
jph1	104	102	131	491	0	0	1248	0	0	2388	0	0	œ	
jph2	96	108	137	486	0	0	1182	0	0	2487	1	0	œ	
jph3	80	89	145	490	0	0	1258	0	0	2427	0	0	œ-O	
jph4	80	87	145	498	0	0	1247	0	0	2452	0	0	œ	
ms1	152	162	112	409	0	0	1338	0	0	2137	0	0	œ-eu	
ms2	88	103	114	461	0	0	1231	1	0	2199	0	0	œ-eu	
ms3	136	153	111	416	0	0	1388	0	0	2218	0	0	œ	
ms4		106	106	441	0	0	1203	0	0	2161	0	0		

Table 4 bis. Formants finaux de la voyelle /œ/ de la phrase 10.  
Analyse LPC à 50% ou à 80 ms du début de la voyelle, selon la répétition.

Répe- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	150		464	41		L	1341	286		M	2338	115		M	3560	380		M					
aq2	156		458	45		L	1324	170		M	2333	49		M	3638	409		F					
aq3	150		460	60		L	1307	202			2282	96		M	3570	375		M					
aq4	135		456	27		L	1282	94		M	2282	96		M	3595	76		M					
bz1	114		431	43		L	1224	128		M	2238	131		M	3483	53		M					
bz2	108		424	51		L	1226	143		M	2330	129		M	3609	139		L					
bz3	109		445	47		L	1262	105		M	2272	94		M	3580	114		L					
bz4	115		441	42		L	1188	254		M	2214	111		M	3633	247		L	3869	863			pas de F5
df1	181		589	233		L	1656	289		L	2802	191		M	3607	368		L					$F_1 = 470$ $F_2 = 1500$
df2	177		459	292		L	1498	186		M	2563	239		M	3567	306							F4 peu visible
df3	170		491	447		L	1506	124		L	2790	419		L	3659	339							
gm1	173		430	60		L	1327	275		L	2377	167		M	3417	217		M	4232	773		F	
gm2	166		499	59		L	1341	424		M	2993	381		F	4273	147		L					$F_3 = 2256$ $F_4 = 2993$
gm3	163		503	80		L	1377	116		L	2314	312		M	4068	253		L					$F_4 = 2993$ $F_5 = 4068$
gm4	170		494	89		L	1401	71		M	2315	94		M	3023	303		M	4276	251		L	$F_5 = 3800$
jfm1	126		516	41		L	1541	106		M	2542	40		M	3855	75		M					
jfm2	119		538	79		L	1671	567		M	2567	97		M	3907	320		M					
jfm3	121		542	37		L	1524	69		M	2615	80		M	3874	109		M					
lg1	84		481	54		L	1285	86		F	2324	157		L	3179	362							F4 peu visible
lg2	97		471	70		L	1285	100		M	2389	136		M	3419	663							F4 peu visible
lg3	106		510	53		L	1305	160		M	2459	138		L	3101	594							pas de F4
lg4	115		499	73		L	1274	82		M	2420	98		M									pas de F4
jlcl	135		442	85		L	1357	305		M	2359	82		M	3227	270		L	3525	189		L	$F_4 = 3400$
jlcl	126		484	104		L	1326	79		M	2425	103		M	3436	994		L	3575	269			
jlcl	121		431	45		L	1300	332		M	2383	246		M	3373	98		M	4151	485			
jlcl	126		447	53		L	1339	60		M	2520	276		M	3464	75		L	4125	375		F	
jmp1	101		504	112		L	1146	508		M	2301	301		M	3485	221		M	4278	842		M	
jmp2	100		515	110		L	1282	568		M	2321	207		M	3664	520		M					
jmp3	96		515	101		L	0	0		M	2463	172			3592	116		M					$F_2 = 1300$
jmp4	101		480	75		L	1139	638		M	2441	151		F	3632	128		M					
jph1	140		484	66		L	1253	89		M	2179	77		M	3250	140		M					
jph2	137		473	77		L	1244	122		M	2168	347		M	3288	123		L	3549	180			$F_4 = 3549$ ; pas de F5
jph3	142		441	36		L	1177	190		M	2209	143		M	3305	116		L					
jph4	140		492	37		L	1226	100		M	2187	47		M	3277	48		M	3587	818			$F_5 = 4375$
ms1	114		412	74		L	1465	152		M	2163	56		M	3447	118		L	3800	512			
ms2	114		428	139		L	1531	325		M	2163	174		M	3440	320		L	3787	295			
ms3	101		401	95		L	1397	421		M	2157	296		M	3505	212		L					

Table 5. Formants intermédiaires de la voyelle /oe/ de la phrase 13 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	F <sub>0</sub> (Hz)	F <sub>1</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>2</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>3</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	Ecoute	Remarques
aq1	256	295	150	445	0	0	1281	0	0	2331	1	0	œR#	
aq2	296	349	156	452	0	0	1272	0	0	2344	0	0	œR#	
aq3	296	375	150	445	0	0	1322	0	0	2307	0	0	œR#	creaky voice
aq4	272	278	135	471	0	0	1250	0	0	2324	0	0	œR#	
bz1	424	382	114	418	0	0	1265	0	0	2257	0	0	œ#	creaky voice
bz2	304	240	108	427	0	0	1159	0	0	2351	0	0	œ#	creaky voice
bz3	400	326	109	444	0	0	1266	0	0	2317	0	0	œ#	creaky voice
bz4	400	361	115	416	0	0	1214	0	0	2276	0	0	œ#	creaky voice
df1	232	261	181	508	1	0	1450	0	1	2652	0	0	œR#	
df2	184	240	177	444	0	2	1551	1	0	2567	2	0	œR#	
df3	184	284	170	582	2	2	1544	0	0	2786	0	0	œR#	
df4		372	153	<u>0</u>	4	0	1404	0	0	2547	0	0		œ# ; F <sub>1</sub> manuelle = 550
gm1	272	321	173	454	0	0	1322	0	0	2304	0	0	œR#	
gm2	264	262	166	472	0	0	1321	0	1	2265	0	0	œhh#	
gm3	288	294	163	477	0	0	1333	0	1	2296	0	0	œR#	
gm4	336	343	170	492	0	0	1317	0	0	2291	0	0	œR#	
jfm1	232	287	126	525	0	0	1604	0	0	2525	0	0	œR#	
jfm2	280	342	121	521	0	0	1547	0	0	2571	1	0	œR#	creaky voice
jfm3	232	314	121	519	0	0	1533	0	0	2610	0	0	œR#	creaky voice
jfm4		333	163	518	0	0	1492	0	0	2473	0	0		œR#
jg1	280	255	87	502	0	0	1267	0	0	2368	0	0	œR!#	creaky voice ?
jg2	224	222	98	493	0	0	1290	0	0	2442	0	0	œR!#	creaky voice ?
jg3	224	222	123	508	0	0	1292	0	0	2454	0	0	œR!#	creaky voice ?
jg4	248	246	119	505	0	0	1282	0	0	2414	0	0	œR!#	creaky voice
jlc1	240	229	135	428	0	0	1334	0	0	2362	0	0	œR#	
jlc2	256	244	126	457	0	0	1323	0	0	2387	0	0	œR#	
jlc3	280	305	121	447	0	0	1353	0	0	2318	0	0	œR#	
jlc4	272	306	126	448	0	0	1316	0	0	2354	0	0	œR#	
jmp1	224	292	101	495	0	0	1266	1	0	2318	0	0	œR#	
jmp2	216	264	100	506	0	0	1242	2	2	2326	0	0	œR#	
jmp3	240	349	96	539	0	0	<u>0</u>	4	0	2498	2	0	œR#	F <sub>2</sub> manuelle = 1300
jmp4	232	274	101	501	0	0	<u>1354</u>	2	2	2461	0	1	œR#	F <sub>2</sub> manuelle = 1300
jph1	232	261	142	460	0	0	1231	0	0	2183	0	0	œR#	
jph2	256	333	137	457	0	0	1238	0	0	2165	0	0	œR#	
jph3	248	336	142	431	0	0	1167	0	0	2227	0	0	œR#	
jph4	200	236	142	493	0	0	1225	0	0	2191	0	0	œR#	
ms1	304	290	114	446	0	0	1505	0	0	2153	0	0	œR#	
ms2	256	285	114	435	0	0	1512	0	0	2161	0	0	œR#	
ms3	280	289	101	426	0	0	1419	0	0	2156	0	0	œR#	
ms4		302	114	429	0	0	1414	0	0	2126	0	0		œR#

Table 5 bis. Formants finaux de la voyelle /œ/ de la phrase 13.  
Analyse LPC à 80 ms du début de la voyelle.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	125		<del>450</del>	71		L	964	93		F	2438	148		M	3545	139		L					
aq2	114		475	26		M	1051	185		F	2338	132		M	3560	28		M					
aq3	123		461	118		M	1036	198		F	2306	149		M	3511	48		M					
aq4	121		468	119		M	900	129		F	2384	137		M	3477	103		M	4350	463		F	
bz1	106		467	49		M	960	132		F	2300	144		M	3619	71		M					
bz2	105		478	30		M	1013	146		F	2355	50		M	<del>3398</del>	524		L	<del>3744</del>	83			$F_4 = 3744$
bz3	108		478	43		M	987	118		F	2316	57		M	<del>3615</del>	592		L	<del>3729</del>	158			$F_4 = 3729$
bz4	105		494	33		M	1016	115		F	2313	21		M	<del>3515</del>	334		L	<del>3722</del>	120			$F_4 = 3722$
df1	170		<del>564</del>	120		L	994	280		F	<del>0</del>	0		M	<del>2703</del>	67		L	<del>3688</del>	182			$F_3 = 2703$ $F_4 = 3688$
df2	163		<del>527</del>	99		L	916	121		F	2546	94		M	3518	81		L					
df3	166		<del>527</del>	172		L	899	151		F	2640	191		M	3571	104		L					
gm1	177		<del>487</del>	70		L	<del>910</del>	120		F	2374	606			<del>3780</del>	392		L					
gm2	173		<del>453</del>	52		L	<del>0</del>	0		F	2331	93		M	3583	235		M					$F_2 = 836$
gm3	181		473	44		L	900	46		F	2230	156		M	3537	131		M					
gm4	195		457	30		M	885	79		F	2216	164		M	3580	46		L					
jfm1	114		517	58			<del>906</del>	184			2697	120		F	3563	128		M	<del>4127</del>	354			
jfm2	108		519	61			<del>0</del>	0			2617	124		F	3553	125		M	<del>3945</del>	303			$F_2$ brute = 778
jfm3	114		526	64			<del>870</del>	290			2623	115		F	3576	60		M	<del>4169</del>	541			
jl1	102		488	94		L	951	165			2478	117		M	3431	172		M					
jl2	86		457	54			933	221			2438	91		M	3345	282							
jl3	106		487	55			<del>1002</del>	584			2480	118		M	<del>3373</del>	455							
jl4	85		466	70			<del>894</del>	124			2383	225		L	3521	293							$F_2 = 994$
jl1c	119		405	97			<del>886</del>	116		L	2516	95		L	<del>3126</del>	441			<del>3402</del>	374			$F_2 = 986$
jl2c	125		430	134			<del>850</del>	128		L	2537	88		L	<del>3205</del>	153			<del>3565</del>	379			$F_2 = 984$
jl3c	119		420	125			<del>0</del>	0		L	2554	114			<del>3191</del>	562		L	<del>3539</del>	277			$F_2$ brute = 786
jl4c	117		441	81			<del>871</del>	155		M	2566	141			<del>3184</del>	324		L	<del>3592</del>	199			
jmp1	96		476	48		L	990	122		F	2263	59		M	<del>3400</del>	498			<del>3599</del>	179			
jmp2	103		477	52		L	972	115		F	2225	97		M	<del>3569</del>	244			<del>3716</del>	638			
jmp3	100		467	52		L	985	131		F	2243	79		M	<del>3558</del>	110			<del>3780</del>	752			
jmp4	94		482	80		L	954	165		F	2209	205		M	<del>3330</del>	334			<del>3751</del>	270			
jph1	121		485	25		L	1138	67		M	2412	121		M	3358	105		M					
jph2	133		509	23		L	1163	73		M	2410	111		M	3278	140		M	<del>3498</del>	266			$F_4 = 3278$
jph3	135		504	28		L	1141	80		M	2511	65		M	3323	230		M	<del>3621</del>	411			$F_4 = 3323$
jph4	137		502	19		L	1136	118		M	2434	233		M	3328	103		M					
ms1	112		<del>485</del>	129		L	<del>911</del>	119		M	2018	107		M	<del>2993</del>	989		L	<del>3662</del>	339			F4 entre 3000 et 4000 Hz
ms2	112		464	94		L	922	171		M	2126	69		M	3403	297		L	3740	302			
ms3	109		<del>499</del>	111		L	<del>898</del>	118		M	2213	64		M	3274	200		L	3848	198			

Table 6. Formants intermédiaires de la voyelle /O/ de la phrase 2 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.

Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	56	62	125	461	0	0	964	0	0	2462	0	0	O	
aq2	80	74	115	481	0	0	1014	0	0	2343	0	0	O	
aq3	88	83	123	500	0	0	990	0	0	2379	0	0	O	
aq4	72	65	121	491	0	0	876	0	0	2440	0	0	O	
bz1	104	86	106	475	0	0	957	0	0	2298	0	0	O	
bz2	96	101	105	477	0	0	990	0	0	2361	0	0	O	
bz3	96	89	108	484	0	0	978	0	0	2318	0	0	O	
bz4	104	95	105	496	0	0	984	0	0	2306	0	0	O	
df1	80	89	170	559	0	0	984	0	0	2701	0	0	O	
df2	64	69	163	531	0	0	926	0	0	2546	0	0	O	
df3	56	65	166	560	1	0	921	0	0	2603	0	0	O	
df4		69	150	551	0	0	987	0	0	2698	0	0		
gm1	88	97	177	535	0	0	955	0	0	2350	0	0	O	
gm2	96	107	173	457	0	0	833	0	0	2350	0	0	ō-o	
gm3	88	103	181	489	0	0	923	0	0	2251	2	0	o	
gm4	104	115	195	482	0	0	925	0	0	2302	0	0	ō-o	
jfm1	112	100	114	518	0	0	881	0	0	2712	0	0	O	
jfm2	96	93	108	516	0	0	789	0	2	2628	0	0	O	
jfm3	96	80	114	529	0	0	860	0	0	2626	0	0	O	
jfm4		75	129	501	0	0	988	0	0	2662	0	0		
jg1	72	83	103	500	0	0	948	0	0	2502	0	0	O	
jg2	64	73	87	465	0	0	898	0	0	2435	0	0	O	
jg3	64	71	106	505	0	0	944	0	3	2478	0	0	O	
jg4	56	68	86	477	0	0	<u>856</u>	0	0	2416	0	0	O	$F_2$ manuelle = 995
jlc1	72	62	119	419	0	0	<u>859</u>	0	0	2528	0	0	O	$F_2$ manuelle = 986
jlc2	64	63	125	431	0	0	<u>860</u>	0	0	2546	0	0	O	$F_2$ manuelle = 984
jlc3	72	60	119	426	0	0	780	2	0	2553	0	0	O	
jlc4	72	77	117	444	0	0	861	0	0	2572	0	0	O	
jmp1	64	58	96	477	0	0	978	0	0	2278	0	0	O	
jmp2	64	68	103	487	0	0	963	0	0	2257	0	0	O	
jmp3	64	69	100	472	0	0	965	0	0	2248	0	0	O	
jmp4	72	74	94	481	0	0	925	0	0	2237	0	0	O	
jph1	96	77	121	491	0	0	1141	0	0	2460	0	0	O	
jph2	88	85	133	528	0	0	1099	1	0	2465	0	0	O	
jph3	80	77	135	513	0	0	1107	0	0	2563	0	0	O	
jph4	72	70	137	511	0	0	1063	1	0	2426	0	0	O	
ms1	96	85	112	489	0	0	909	0	0	2036	0	0	O	
ms2	88	80	112	490	0	0	928	0	0	2157	0	0	O	
ms3	80	94	109	515	0	0	887	0	0	2232	0	0	O	
ms4		68	106	490	0	0	936	0	0	2073	0	0		

Table 6 bis. Formants finaux de la voyelle /O/ de la phrase 2.  
 nalyse LPC à 50% du début de la voyelle.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	140		450	49		L	1050	178		F	2179	65		M	3465	204		M					
aq2	126		440	70		L	1032	97		F	2224	119		M	3463	48		M					
aq3	133		429	31		L	1089	206		M	2347	42		F	3357	223		F	4417	270			
aq4	126		439	51		L	1011	135		F	2191	62		F	3574	52		M					
bz1	105		426	85		L	861	86		F	2063	151		M	3302	304		L					
bz2	109		437	113		L	913	125		F	2192	137		F	3634	93		M					
bz3	106		429	101		M	913	105		F	2154	109		F	3721	225		M					
bz4	106		425	68		M	886	84		F	2110	84		M	<del>3657</del>	262		L					
df1	170		<del>589</del>	266			1210	300			2698	561		F	3654	951							
df2	173		<u>0</u>	0			<u>1026</u>	509			2616	110		F	<del>3615</del>	168							
df3	166		581	145		L	1116	275		F	<u>0</u>	0		M	<u>2712</u>	206			<u>3542</u>	212			
gm1	190		459	92		L	<del>909</del>	231		F	2230	31		M	2767	218		F	<del>4201</del>	161			
gm2	177		474	75		L	<del>963</del>	190		F	2366	329		M	3047	181		F	<del>4358</del>	70			
gm3	181		488	56		L	<del>965</del>	178		F	2295	85		M	2882	79		F	<del>4190</del>	274			
gm4	177		446	35		L	<del>955</del>	225		F	2176	95		M	<u>2994</u>	837		L	<u>3881</u>	294			
jfm1	123		537	31		M	1168	87		F	2612	53		M	3703	223		M	<del>3901</del>	477			
jfm2	126		544	58		M	1140	163		F	2652	103		M	3560	145		L	<del>4084</del>	340			
jfm3	121		533	82		M	1155	239		F	2643	89		M	3641	124		M	<del>4112</del>	679			
jg1	94		489	71		M	<u>918</u>	239		F	2407	94		M	3843	726							
jg2	102		487	94		M	978	153		F	2421	275		M									
jg3	114		495	13		M	1026	30		F	2369	53		M	<u>2819</u>	199			<del>3983</del>	163			
jg4	102		496	57		M	887	346		F	2263	119		F	<u>2878</u>	580			<del>4070</del>	387			
jlc1	119		462	79		L	<u>0</u>	0		F	2483	150		F	3404	98		M	<del>4200</del>	406			
jlc2	121		450	86		L	941	114		F	2515	309		F	3417	207		M	4347	83		M	
jlc3	115		475	37		L	979	55		F	2491	72		F	<del>3354</del>	137		L					
jlc4	111		464	54		L	912	46		F	2505	121		F	3348	148		M	<del>3687</del>	583			
jmp1	100		490	102		L	959	148		F	2161	91		M	<u>3309</u>	887		M	<u>3533</u>	194			
jmp2	105		482	60		L	915	130		F	2242	230		F	3551	161		L					
jmp3	105		483	96		L	893	141		F	2156	451		F	<del>3501</del>	135		M					
jmp4	105		503	72		L	1001	115		M	2131	148		M	<u>3326</u>	643		L	<u>3589</u>	205			
jph1	125		465	54		L	926	116		F	2230	65		F	3060	261		M	<del>4413</del>	646			
jph2	131		483	20		L	1014	107		F	2274	249		F	3308	66		M					
jph3	133		500	14		L	<u>904</u>	99		F	2284	79		M	3297	85		M	4449	202			
jph4	137		505	36		L	1047	148		M	2301	94		M	3323	39		M	4497	371			
ms1	114		417	116		M	<u>0</u>	0		F	2014	93		F	<del>3351</del>	402		L	<del>3842</del>	101			
ms2	112		417	105		L	885	49		F	2040	37		F	<del>3552</del>	319		L	<del>3879</del>	153			
ms3	115		457	136		L	902	82		F	2029	49		F	<del>3469</del>	120		L	<del>3904</del>	241			

Table 7. Formants intermédiaires de la voyelle /O/ de la phrase 4 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	312	336	140	433	0	0	1035	0	0	2179	0	0	OR#	
aq2	296	303	126	445	0	0	1047	0	0	2244	0	0	OR#	
aq3	296	357	135	438	0	0	1110	0	0	2360	0	0	OR#	
aq4	272	313	126	418	0	0	951	0	0	2203	0	0	OR#	
bz1	336	426	105	418	0	0	881	0	0	2098	0	0	O#	creaky voice
bz2	296	333	109	415	0	0	908	0	0	2253	0	0	O#	creaky voice ?
bz3	256	347	106	436	0	0	931	0	0	2176	0	0	OR#	creaky voice ?
bz4	328	421	106	429	0	0	907	0	0	2158	0	0	OR#	
df1	200	275	170	586	0	0	1185	0	1	2700	0	1	OR#	
df2	184	246	173	□	4	0	1075	0	2	2609	0	0	OR#	(0, 0, 487)
df3	168	222	169	596	2	0	1123	0	0	2653	1	0	OR#	
df4		383	153	535	0	0	1109	0	0	2652	0	0		
gm1	248	276	190	445	0	0	911	0	0	2223	0	0	Ohh#	
gm2	264	257	177	468	0	0	863	0	2	2365	0	0	OR#	
gm3	360	368	181	459	0	0	935	0	0	2269	0	0	OR#	
gm4	344	396	177	478	0	0	952	0	2	2203	0	0	O#	
jfm1	208	200	123	535	0	0	1161	0	0	2603	0	0	OR#	
jfm2	200	199	126	541	0	0	1138	0	0	2650	0	0	OR#	creaky voice
jfm3	168	169	121	529	0	0	1150	0	0	2668	0	0	OR#	
jfm4		284	137	485	0	0	1061	0	0	2632	0	0		
jg1	256	254	95	486	0	0	997	0	0	2426	0	0	OR#	creaky voice
jg2	344	319	103	493	0	0	1029	0	0	2388	2	0	OR#	creaky voice
jg3	288	352	117	498	0	0	1021	0	0	2373	0	0	OR#	creaky voice
jg4	224	267	102	491	0	0	940	0	1	2310	0	0	OR#	creaky voice
jlc1	240	325	119	491	0	0	945	0	0	2488	0	0	OR#	
jlc2	248	303	123	448	0	0	965	0	0	2503	0	0	OR#	
jlc3	240	233	117	467	0	0	950	0	0	2464	0	0	OR#	
jlc4	232	261	111	465	0	0	930	0	0	2482	0	0	OR#	
jmp1	248	315	100	481	0	0	986	0	0	2180	0	0	OR#	
jmp2	184	231	105	481	0	0	950	0	0	2230	0	0	OR#	
jmp3	192	224	105	491	0	0	925	0	0	2164	0	1	OR#	
jmp4	176	232	105	492	0	0	992	0	0	2131	0	0	OR#	
jph1	200	207	125	466	0	0	947	0	0	2225	0	0	OR#	
jph2	200	223	131	485	0	0	1002	0	0	2279	0	0	OR#	
jph3	200	205	133	499	0	0	918	0	0	2282	0	0	OR#	
jph4	216	230	137	505	0	0	1045	0	0	2306	0	0	OR#	
ms1	280	295	114	437	0	0	854	0	0	1990	0	0	OR#	
ms2	264	304	112	443	0	0	893	0	0	2023	0	0	OR#	
ms3	240	267	115	469	0	0	897	0	0	2012	0	0	OR#	
ms4		297	111	432	0	0	893	0	0	<u>0</u>	4	0		$F_3$ manuelle = 1960

Table 7 bis. Formants finaux de la voyelle /O/ de la phrase 4.  
analyse LPC à 80 ms du début de la voyelle.

Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	140		<del>690</del>	125		L	<del>1464</del>	312		L	2409	118		L	<del>2832</del>	737		L	<del>3796</del>	205		L	$F_4$ brute = 3796
aq2	133		603	48		L	1469	435		L	2409	232		L	3774	223		M					
aq3	137		646	193		L	<del>1379</del>	401		L	2451	209		L	3976	194		L					
aq4	137		627	75		L	<del>0</del>	0		L	2502	109		L	3884	191		L					$F_2$ brute = 1593
bz1	109		564	65		L	1112	117		F	2343	59		M	<del>3229</del>	659			3955	193		L	
bz2	102		578	98		L	1155	124		F	2329	55		M	<del>2900</del>	902			3981	180		L	
bz3	109		558	63		L	1163	174		F	2387	121		M	3748	194		L					
bz4	108		544	59		L	1211	113		M	2311	114		M	3859	140		M					
df1	170		592	90		L	1185	129		M	2529	167		L	3568	68		M					
df2	170		<del>559</del>	201		L	<del>0</del>	0		L	2598	100		L	3581	106		L					$F_2$ = 1200
df3	173		586	172		L	1163	268		M	2571	97		L	3663	125		L					
gm1	190		609	59		L	1174	259		M	2157	160		M	<del>3870</del>	147		L					
gm2	195		579	99		L	1050	512		M	<del>2187</del>	397		L	<del>3729</del>	162		L					
gm3	195		561	95		L	1100	380		M	2228	168		L	<del>3643</del>	122		L					
gm4	195		594	42		L	1217	167		M	2188	174		M	<del>3518</del>	69		L					
jfm1	121		725	136		L	1400	317		M	2656	251		M	<del>3403</del>	922		M	<del>3898</del>	168			$F_4$ = 3898
jfm2	119		721	84		L	1326	180		M	2753	169		M	3735	208		M	4353	773			
jfm3	126		691	95		L	1231	175		M	2732	214		M	3737	193		M	<del>4188</del>	942			pas de F5
jg1	108		633	91		L	1169	225		M	2443	88		M	3404	208		M					
jg2	97		620	70		L	1258	169		M	2495	68		M	3423	127		M					
jg3	119		596	116		L	<del>1075</del>	334		F	2391	114		M	3326	204		M					$F_2$ = 1150
jg4	123		665	142		L	1191	472		M	2435	64		L	3404	394		M					
jlc1	117		637	140		L	<del>1099</del>	347		M	2549	98		L	<del>3423</del>	718		L	<del>3633</del>	348			
jlc2	119		631	45		L	1216	160		M	2512	118		L	<del>3285</del>	968		L	<del>3561</del>	185			
jlc3	111		638	95		L	1126	352		M	2491	115		L	<del>3311</del>	223		L	<del>3699</del>	452		L	
jlc4	115		620	32		L	1148	118		M	2504	245		M	<del>3143</del>	474		L	<del>3646</del>	218		L	
jmp1	98		<del>582</del>	91		L	1257	142		M	2218	71		M	3550	129		M	4373	252		L	
jmp2	102		551	122		L	<del>1341</del>	314		M	<del>2308</del>	585		L	3626	200		L					$F_5$ = 4300
jmp3	100		596	100		L	1238	152		M	<del>2425</del>	219		M	3511	128		L					$F_3$ = 2300
jmp4	100		595	98		L	1261	222		M	<del>2443</del>	295		L	3505	253		L	4318	411		L	$F_3$ = 2300
jph1	131		614	51		L	1238	206		M	2389	141		M	<del>3313</del>	608		L	<del>3679</del>	221		L	
jph2	126		632	65		L	1252	134		M	2381	88		M	<del>3297</del>	533		L	<del>3562</del>	180		L	$F_4$ = 3400
jph3	133		<del>650</del>	131		L	1217	283		M	2435	149		M	<del>3564</del>	351		L					
jph4	140		602	115		L	1286	190		M	2412	116		M	<del>3601</del>	440		M					
ms1	112		584	71		L	<del>1104</del>	192		F	2129	411		L	3504	271		L	4275	161			$F_2$ = 1200
ms2	119		601	65		L	<del>1081</del>	237		M	2277	89		M	<del>3583</del>	154		L	4260	341		F	$F_2$ = 1200
ms3	114		613	121		L	1120	294		M	<del>2532</del>	160		L	3605	93		M	4426	108		L	

Table 8. Formants intermédiaires de la voyelle /a/ de la phrase 3 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	72	81	142	656	1	0	1517	0	0	2433	0	0	a	
aq2	80	97	133	608	0	0	1440	0	3	2417	0	0	a	
aq3	88	102	137	0	3	0	1323	1	1	2316	2	0	a	$F_1$ manuelle = 632
aq4	88	95	137	614	0	0	1451	0	1	2550	2	0	a	
bz1	88	107	109	575	0	0	1117	0	0	2309	0	0	a	
bz2	80	92	102	578	0	0	1166	0	0	2316	0	0	a	
bz3	88	92	109	588	0	0	1158	0	0	2332	0	0	a	
bz4	72	82	108	539	0	0	1212	0	0	2326	0	0	a	
df1	80	103	170	594	0	0	1181	2	0	2534	0	0	a	
df2	80	110	170	550	2	0	1194	2	1	2576	0	0	a	
df3	64	100	173	589	0	0	1191	0	0	2576	0	0	a	
df4		102	160	591	0	0	1246	0	0	2591	2	0		
gm1	80	103	190	568	1	0	1176	0	0	2173	0	0	a	
gm2	80	102	195	573	0	0	1027	1	1	2107	1	0	a ; p voisé	
gm3	80	95	195	567	0	0	1079	0	3	2273	1	0	a ; p voisé	
gm4	72	84	195	586	0	0	1248	1	0	2219	0	0	a	
jfm1	96	104	121	730	0	0	1388	0	0	2690	0	0	a	
jfm2	88	100	119	745	2	0	1350	0	0	2754	0	0	a	
jfm3	88	96	126	690	2	0	1248	0	0	2811	0	0	a	
jfm4		98	142	706	0	0	1401	0	0	2827	0	0		
jg1	64	80	109	659	1	0	1189	0	0	2431	0	0	a	
jg2	72	91	97	650	0	0	1266	0	0	2496	0	0	a	
jg3	64	89	119	616	2	0	1128	1	0	2384	0	0	a	
jg4	64	79	123	660	0	0	1203	0	1	2427	0	0	a	
jlc1	80	95	117	643	2	0	1093	2	1	2561	0	0	a	
jlc2	88	104	119	650	2	0	1237	0	1	2540	0	0	a	
jlc3	72	86	111	639	0	1	1099	0	1	2526	0	0	a	
jlc4	80	103	115	634	0	0	1139	0	0	2532	0	0	a	
jmp1	80	119	98	585	0	0	1239	0	0	2251	0	0	a	
jmp2	80	108	102	564	1	0	1354	0	0	2460	2	0	a	
jmp3	72	93	100	588	0	0	1220	0	0	2406	0	0	a	
jmp4	88	118	100	603	0	0	1247	0	0	2413	0	0	a	
jph1	88	107	131	633	0	0	1246	0	0	2417	0	0	a	
jph2	72	90	126	646	0	0	1246	0	0	2442	0	0	a	
jph3	72	90	133	656	0	0	1237	0	0	2447	2	0	a	
jph4	80	101	140	632	1	0	1270	0	0	2451	0	0	a	
ms1	80	100	112	593	0	0	1066	0	0	2560	0	1	a	$F_2$ manuelle = 1229
ms2	80	109	119	608	0	0	1078	0	0	2282	0	0	a	$F_2$ manuelle = 1229
ms3	72	78	114	631	0	1	1132	0	1	2505	0	0	a	(2532, 2462, 2523)
ms4		76	106	604	0	0	1230	0	0	2304	2	0		

Table 8 bis. Formants finaux de la voyelle /a/ de la phrase 3.  
Analyse LPC à 50% du début de la voyelle.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	140		584	43		L	1391	283		L	2533	182		L	3654	293		L					
aq2	117		599	40		L	1330	305		L	2436	158		L	3893	376							
aq3	150		581	70		L	1407	201		L	2479	215		L	3738	696		M					
aq4	135		687	76		L	1281	126		L	2468	191		L	3864	130		M					
bz1	108		517	42		M	1159	164		M	2250	71		M	3755	157		L					
bz2	103		518	137		L	<u>0</u>	0		M	2250	81		M	3661	148		L					
bz3	108		563	62		L	1150	137		M	2414	104		M	<b>3246</b>	920		L	<b>3805</b>	136			$F_2 = 1150$
bz4	103		586	106		L	1207	123		M	2272	102		M	<b>3247</b>	914		L	<b>3773</b>	102			$F_4 = 3805$
df1	170		672	160		L	<u>1362</u>	272		L	2529	95		L	3760	190		L					$F_4 = 3773$
df2	173		641	121		L	1287	99		M	2442	208		L	3658	396		L					
df3	170		600	177		L	1248	378		L	2561	134		L	3679	123		L					
gm1	186		632	73		L	1124	449		M	2330	228		L	3793	368		M					
gm2	186		618	51		L	1143	398		M	2215	185		M	<b>2984</b>	973			<u>4377</u>	387			pas de Fi vers 3000 Hz
gm3	177		670	48		L	1271	354		M	2284	91		L	3437	46		M					
gm4	186		606	33		L	1177	245		L	2377	138		M	3692	217							
jfm1	117		683	80		L	1289	93		M	2702	132		M	3735	232		L	4217	317		M	
jfm2	118		677	64		L	1321	197		M	2685	104		M	<b>3652</b>	626		L	<b>3920</b>	114			$F_4 = 3920$
jfm3	123		681	70		L	1346	180		M	2652	152		M	3854	194		L					
jg1	97		623	111		L	1279	228		M	2422	70		M	3367	222		M					
jg2	103		631	144		L	1366	300		L	2527	107		L	3360	232		M					
jg3	102		642	54		L	1148	172		M	2493	152		M	3372	164		M					
jg4	93		612	88		L	1158	261		M	2469	134		M	3422	245		M					
jlc1	116		597	89		L	1249	269		L	2480	100		M	<u>3388</u>	472		L	<u>3721</u>	241			
jlc2	121		627	75		L	1277	186		L	2507	92		M	<u>3312</u>	254		L	<u>3828</u>	253			
jlc3	112		585	113		L	1291	289		M	2513	104		M	<u>3362</u>	329		L	<b>3773</b>	246			
jlc4	117		611	57		L	1226	260		M	2505	161		L	<u>3321</u>	240		L	<u>3927</u>	295			
jmp1	98		570	66		L	1210	156		M	2260	131		M	3648	173		L	<u>4056</u>	757			pas de F5
jmp2	102		579	123		L	1188	266		M	2340	119		M	3722	249		L					
jmp3	104		556	55		M	1183	215		M	2297	140		M	3701	215		L					
jmp4	101		593	66		L	1205	149		M	2276	195		M	<u>3493</u>	403		L	<u>3894</u>	265			
jph1	125		596	39		L	1345	232		M	2397	236		M	<u>3346</u>	151		L	<u>4096</u>	546			pas de F5
jph2	129		613	42		L	1137	468		M	2385	297		L	<u>3275</u>	309		L					
jph3	133		638	43		L	1276	98		M	2355	139		M	<b>3262</b>	256		L	<u>3533</u>	603			un seul formant
jph4	137		636	46		L	1222	200		M	2369	94		M	<u>3285</u>	428		L	<b>3493</b>	440			un seul formant
ms1	106		600	75		L	1105	205		M	2073	127		M	3606	121		L					
ms2	115		629	88		L	1204	218		M	2121	86		M	3683	157		L					
ms3	114		611	80		L	<b>1230</b>	318		L	2205	143		M	3713	263		M					$F_2 = 1150$

Table 9. Formants intermédiaires de la voyelle /a/ de la phrase 16 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.

Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	128	117	140	598	0	0	1365	0	0	2518	0	0	ao	
aq2	216	190	117	593	0	0	1314	0	0	2419	0	0	aR	
aq3	152	135	150	584	0	0	1367	0	0	2476	0	0	ao	
aq4	288	222	135	626	0	0	1228	0	0	2524	0	0	aR	
bz1	160	156	108	524	0	0	1158	0	0	2258	0	0	aa	
bz2	216	161	103	515	0	0	1034	0	0	2242	0	0	aa	$F_2$ manuelle = 1150
bz3	192	154	108	563	0	0	1135	0	0	2393	0	0	aa	
bz4	280	205	103	553	0	0	1217	0	0	2255	0	0	aR	
df1	120	135	170	705	0	0	1343	0	0	2520	0	0	aR	
df2	112	143	173	652	0	0	1260	0	0	2432	1	0	aR	
df3	104	134	170	633	0	0	1229	0	3	2598	0	0	aR	
df4		124	160	675	0	0	1320	0	0	2559	0	0		
gm1	104	103	186	625	0	0	1176	0	2	2230	0	0	aR&	
gm2	128	125	186	617	0	0	1090	0	1	2138	0	0	a&	
gm3	144	138	177	660	0	0	1297	0	0	2266	0	0	aR	
gm4	232	206	186	614	0	0	1221	0	0	2312	0	0	aã	
jfm1	136	122	117	688	0	0	1278	0	0	2717	0	0	ao	
jfm2	112	111	118	691	0	0	1312	0	0	2689	0	0	ao	
jfm3	128	118	123	703	0	0	1306	0	0	2708	0	0	ao	
jfm4		98	142	713	0	0	1344	0	0	2801	0	0		
jg1	168	172	98	636	0	0	1323	2	2	2443	0	0	aR	
jg2	144	161	103	660	0	0	1319	0	1	2442	0	0	aR&	
jg3	200	205	102	638	0	0	1137	0	0	2467	0	0	aR&	
jg4	168	198	93	627	0	0	1136	0	0	2469	0	0	aR&	
jlc1	144	131	117	610	0	0	1211	0	0	2463	0	0	aR&	
jlc2	152	148	121	629	0	0	1221	0	0	2521	0	0	aR&	
jlc3	128	117	112	598	0	0	1299	0	0	2506	0	0	aR&	
jlc4	136	126	117	619	0	0	1219	0	0	2493	0	0	aR&	
jmp1	128	141	98	578	0	0	1186	0	0	2294	0	0	aR	
jmp2	136	147	103	576	0	0	1159	0	0	2379	0	0	aa	
jmp3	120	140	104	566	0	0	1152	0	0	2325	0	0	a	
jmp4	112	129	102	598	0	0	1170	0	0	2315	0	0	aa	
jph1	136	128	125	606	0	0	1263	0	0	2357	0	0	aR-ao	
jph2	136	141	129	627	0	0	1129	0	0	2275	0	0	ao	
jph3	136	142	133	652	0	0	1258	0	0	2391	0	0	ao	
jph4	120	128	137	637	0	0	1181	0	0	2376	0	0	ao	
ms1	176	151	106	605	0	0	1104	0	0	2080	0	0	aR&	
ms2	160	154	115	640	0	0	1174	0	0	2174	0	0	aR	
ms3	160	147	114	614	0	0	1167	0	0	2235	0	0	aR&	
ms4		91	106	609	0	0	1223	0	0	2128	0	0		

Table 9 bis. Formants finaux de la voyelle /a/ de la phrase 16.  
Analyse LPC à 50% ou à 80 ms du début de la voyelle, selon la répétition.



Répé- tition	F <sub>0</sub> (Hz)	A	F <sub>1</sub> (Hz)	B <sub>1</sub> (Hz)	A	T	F <sub>2</sub> (Hz)	B <sub>2</sub> (Hz)	A	T	F <sub>3</sub> (Hz)	B <sub>3</sub> (Hz)	A	T	F <sub>4</sub> (Hz)	B <sub>4</sub> (Hz)	A	T	F <sub>5</sub> (Hz)	B <sub>5</sub> (Hz)	A	T	Remarques	
aq1	148		283	50		M	1896	128		F	2950	153		M	4458	209		M						
aq2	129		334	154		M	1903	84		F	3036	130		M	3717	979			4395	335				
aq3	163		276	92		M	1934	33		F	3006	82		M	3622	385								
aq4	156		330	119		M	1945	19		M	2942	119		M	4260	361								
bz1	112		292	54		F	1926	81		F	2765	148		M	4422	492		M						F <sub>4</sub> = 3500 F <sub>5</sub> = 4422
bz2	112		291	45		F	1882	26		F	2701	91		M	3454	273		M	4206	302				
bz3	109		320	62		F	1930	66		F	2724	154		M	3448	198	L		4396	943				
bz4	109		291	50		F	1849	42		F	2780	128		M	3569	346		L	4027	698				
df1	177		381	63		L	2176	99		M	3111	262		M	3792	98		M						
df2	177		411	37		L	2049	80		F	2908	186		M	3849	177		M						
df3	166		254	96			2159	109			3017	364			3871	519								
gm1	200		241	43		F	2215	66		M	2977	239		M	3918	115		M						
gm2	195		292	174		F	2304	75		M	2967	524			4274	580								
gm3	200		279	118		F	2212	21		M	2909	134		M	3729	259		M						
gm4	205		271	79		F	2215	83		M	2850	94		M	3792	192		M						
jfm1	122		249	53		M	2348	66		M	0	0		L	3458	411			3796	154				F <sub>3</sub> = 3500 F <sub>4</sub> = 3900
jfm2	123		293	25		M	2413	107		M	0	0		L	3472	248			3868	220				F <sub>3</sub> = 3500 F <sub>4</sub> = 3900
jfm3	119		260	45		M	2323	53		M	0	0		L	3567	129			3908	159				F <sub>3</sub> = 3500 F <sub>4</sub> = 3900
jg1	102		270	63		M	2056	110			2847	87			3571	327								
jg2	100		251	74		M	2014	74		M	2908	122		M	3853	419		M						
jg3	114		283	42		M	2023	80		M	2757	129		M	4460	348								F <sub>4</sub> = 3200 F <sub>5</sub> = 4460
jg4	123		307	35		M	2033	117		M	2748	142		M				M						F <sub>4</sub> = 3400 F <sub>5</sub> = 4500
jlc1	123		274	81		M	2137	45		M	2889	181			3405	178			3790	431				
jlc2	121		315	44		M	2221	36		M	2937	137			3779	310								F <sub>4</sub> = 3400
jlc3	121		274	54		M	2232	102	M		3068	148			3447	150			3959	279				
jlc4	123		311	100		L	2148	66		M	2880	110			3785	130								F <sub>4</sub> = 3500
jmp1	109		318	49		M	1880	193		M	2804	199		M	3612	263		M						
jmp2	104		309	37		M	1874	156		M	2884	146		M	3579	346		M						
jmp3	107		316	62		M	1905	149		M	2875	114		M	3541	242		L	4358	781				
jmp4	105		308	57		M	1864	133		M	2916	113			3616	232			4345	579				
jph1	137		321	31		M	1977	350		F	2683	83		M	3447	343		M						
jph2	140		317	84		L	1935	105		M	2741	128		M	3480	148		L						
jph3	137		306	29		M	1939	342		F	2749	108		M	3492	275		M						
jph4	145		323	59		M	2032	481		F	2843	116		F	3406	118		M						
ms1	123		247	19		F	1962	67		F	3031	119		M	3625	105		M						
ms2	115		259	43		F	1964	111		F	3050	98		M	3624	123		M						
ms3	112		266	58		M	1962	162		F	2971	143		M	3594	152		M	4259	442				

Table 10. Formants intermédiaires de la voyelle /i/ de la phrase 11 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	192	170	148	287	0	0	1891	0	0	2938	0	0	i	
aq2	192	155	129	331	0	0	1914	0	0	3008	0	0	i	
aq3	152	117	163	260	0	0	2024	0	0	2992	0	0	i	
aq4	192	151	156	305	0	0	1944	0	0	2925	0	0	i	
bz1	232	146	112	293	0	0	1916	0	0	2770	0	0	i	
bz2	256	172	112	297	0	0	1867	0	0	2718	0	0	i	
bz3	152	134	109	331	0	0	1958	0	0	2751	0	0	i	
bz4	208	142	109	291	0	0	1851	0	0	2776	0	0	i	
df1	48	59	177	409	0	0	2193	0	0	2885	2	0	i	
df2	64	80	177	416	0	0	1992	0	0	2915	0	0	pje	
df3	24	37	166	380	2	0	2119	0	0	2965	0	0	pje	
df4		57	186	433	0	0	2108	1	0	2922	2	0		
gm1	128	143	200	242	0	0	2246	0	0	2994	0	0	i	
gm2	112	122	195	285	0	0	2327	0	0	3138	2	0	i	
gm3	104	108	200	288	0	0	2245	0	0	2954	0	0	i	
gm4	88	108	205	293	0	0	2247	0	0	2888	0	0	i	
jfm1	112	103	123	249	0	0	2356	0	0	3373	1	1	i	
jfm2	120	115	123	294	0	0	2420	0	0	3432	1	0	i	
jfm3	96	100	120	268	0	0	2335	0	0	3547	0	0	i	
jfm4		103	153	404	0	0	2370	0	0	<del>2908</del>	2	1		(3310, 2922, 2895)
jg1	64	82	102	272	0	0	2057	0	0	2828	0	0	i	
jg2	128	127	100	258	0	0	2038	0	0	2911	0	0	i	
jg3	80	97	115	287	0	0	2043	0	0	2743	0	0	i	
jg4	88	95	125	317	0	0	2003	0	0	2661	1	0	i	
jlc1	136	120	123	292	0	0	2146	0	0	<u>0</u>	3	0	i	$F_3$ manuelle = 2944
jlc2	120	130	121	301	0	0	2194	0	0	2989	2	0	i	
jlc3	136	128	121	285	0	0	2327	1	0	3117	0	0	i	
jlc4	104	99	123	296	0	0	2142	0	0	2951	0	0	i	
jmp1	72	96	109	326	0	0	1891	0	0	2788	0	0	i	
jmp2	56	67	105	313	0	0	1882	0	0	2869	0	0	i	
jmp3	80	96	107	328	0	0	1918	0	0	2832	0	0	i	
jmp4	80	102	105	316	0	0	1866	0	0	2962	2	0	i	
jph1	80	98	137	327	0	0	1988	0	0	2682	0	0	i	
jph2	96	118	140	322	0	0	1969	0	0	2778	0	0	i	
jph3	96	111	137	314	0	0	1961	2	0	2780	0	0	i	
jph4	96	111	145	333	0	0	2060	0	1	2827	0	0	i	
ms1	192	166	123	247	0	0	1964	0	0	3036	0	0	piR&	
ms2	128	121	115	262	0	0	1989	0	0	3053	0	0	i	
ms3	72	73	112	269	0	0	1962	0	0	2987	0	0	i	
ms4		138	112	272	0	0	1941	0	0	3010	0	0		

Table 10 bis. Formants finaux de la voyelle /i/ de la phrase 11.  
Analyse LPC à 50% ou à 80 ms du début de la voyelle, selon la répétition.

Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	142		284	156		M	1956	79		M	2918	153		M	<del>3679</del>	343			<del>4484</del>	190			
aq2	123		326	196		M	1925	153		M	2937	168		M	<del>3679</del>	323			<del>4252</del>	440			$F_1 = 280$
aq3	135		261	85		M	1908	15		M	3008	336		M	<del>3867</del>	689			<del>4326</del>	267			
aq4	135		281	70		M	1866	30		M	2778	146		M	<del>3492</del>	370			<del>4184</del>	433			
bz1	108		294	33		F	1830	85		F	0	0	F		3430	104		M					$F_3 = 2503 F_5 = 4200$
bz2	103		315	67		F	2551	112		F	0	0	F		3431	451		M	<del>3594</del>	209			$F_2 = 1900 F_3 = 2551 F_5 = 4130$
bz3	111		317	82		F	1945	64		F	2639	155		F	3419	139		M					
bz4	106		308	57		F	1881	95		F	2644	177		F	3522	123		M					
df1	173		<del>343</del>	133			<del>2271</del>	192			<del>2966</del>	259			<del>3729</del>	95							trop de friction
df2	170		337	77		M	2236	54		M	<del>3126</del>	555			<del>3656</del>	376							trop de friction
df3	0		0	0			0	0			0	0			0	0			0	0			pas de i
gm1	195		259	78		F	2223	56		M	2856	313		F	3940	219		M					
gm2	190		261	65		F	2213	24		M	2968	157		M	3909	83		M					
gm3	205		259	65		F	2226	89		M	2832	93		M	<del>3880</del>	118		L					
gm4	200		323	153		F	<del>2118</del>	227			<del>2939</del>	286			<del>3818</del>	125		L					
jfm1	119		274	42		M	2281	67			0	0			<del>3439</del>	316		L	<del>3837</del>	275			$F_3 = 3400 F_4 = 3800$
jfm2	117		280	33		M	2354	69			<del>3304</del>	292			<del>3736</del>	171							
jfm3	115		281	30		M	2251	57		L	<del>3155</del>	382		L	<del>3752</del>	176		L					
jg1	106		246	54		F	2113	75		M	2867	84		M	3721	577		M					
jg2	105		269	58		F	2116	49		M	2838	74		M	3831	168		M					
jg3	140		310	51		M	2015	71		M	2820	147		M	<del>4286</del>	779							un Fi entre 3000 et 4000 Hz
jg4	135		292	95		M	2012	108		M	2800	141		M	3405	232		M	<del>4479</del>	176			
jlc1	123		231	63		F	2121	88		M	<del>3221</del>	110		M	<del>4254</del>	385							$F_3 = 2940 F_4 = 3220$
jlc2	117		251	54		F	2170	103		L	2847	197		L	<del>3829</del>	583							$F_4 = 3275$
jlc3	115		213	57		F	2112	106			<del>3280</del>	84			<del>4003</del>	383							$F_3 = 2847 F_4 = 3280$
jlc4	114		244	92		F	<del>2186</del>	111			<del>2890</del>	115			<del>3737</del>	339							$F_4 = 3338$
jmp1	100		305	71		M	1921	86		M	2712	85		M	3623	293		M					
jmp2	102		310	26		M	1917	116		M	2777	126		M	3542	366		M					
jmp3	100		306	62		M	1915	79		M	<del>2850</del>	255			<del>3492</del>	185			<del>4483</del>	493			
jmp4	103		321	60		L	1869	127		M	2672	104		M	<del>3552</del>	246		L	<del>4443</del>	251			
jph1	126		286	21		M	1897	59		M	2686	138			3543	427							
jph2	131		321	48		M	1944	53		M	2628	259		M	3430	174		M					
jph3	140		301	76		M	1984	125		M	2784	122		M	3416	63		M					
jph4	153		372	151		L	1930	94		M	<del>3383</del>	82	L		<del>3920</del>	581							$F_3 = 2618 F_4 = 3383$
ms1	114		253	24		F	1933	59		F	2951	153		F	3557	112		L					
ms2	111		252	44		F	1954	98		F	<del>3031</del>	116			<del>3686</del>	240							
ms3	117		249	37		F	1980	142		F	<del>3106</del>	197			<del>3764</del>	238							

Table 11. Formants intermédiaires de la voyelle /i/ de la phrase 15 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	F <sub>0</sub> (Hz)	F <sub>1</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>2</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>3</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	Ecoute	Remarques
aq1	104	99	142	298	0	0	1989	0	0	2879	1	0	i	
aq2	176	147	125	286	1	0	1880	0	0	2939	0	0	i	
aq3	136	116	135	255	0	0	1917	0	0	2950	1	0	i	
aq4	144	123	135	287	0	0	1933	0	0	2807	0	0	i	
bz1	168	177	109	301	0	0	1849	0	0	2524	2	0	i	
bz2	176	166	103	313	0	0	1936	0	0	2546	0	0	i	
bz3	192	180	111	316	0	0	1948	0	0	2630	0	0	i	
bz4	152	154	106	306	0	0	1902	0	0	2598	0	0	i	
df1	48	70	173	368	0	0	2234	0	0	2985	2	0	pj&	
df2	88	117	170	349	0	0	2260	0	0	3083	1	1	i	
df3	0	0	181	0	5	0	0	5	0	0	5	0	supjāsõ	pas de iR
df4		98	156	315	0	0	2307	0	0	3185	0	0		
gm1	112	129	195	276	2	0	2248	0	0	2850	2	0	i	
gm2	112	136	190	257	0	0	2221	0	0	2974	0	0	i	
gm3	112	132	205	265	0	0	2230	0	0	2859	0	0	i	
gm4	88	115	200	275	0	0	2201	0	0	2862	0	2	i	
jfm1	128	137	119	270	0	0	2270	0	0	3428	0	0	i	
jfm2	72	90	117	286	0	0	2370	0	0	3299	0	0	i	
jfm3	104	125	115	278	0	0	2284	0	0	3457	2	1	i	
jfm4		136	131	310	0	0	2326	0	0	3401	2	0		
jg1	96	116	106	249	0	0	2113	0	0	2861	0	0	i	
jg2	120	122	105	283	0	0	2126	0	0	2803	0	0	i	
jg3	128	172	140	312	0	0	2063	0	0	2852	0	0	i	
jg4	120	149	137	294	0	0	2012	0	0	2894	0	0	i	
jlc1	112	159	123	238	0	0	2133	0	0	2917	0	0	i	
jlc2	128	172	117	249	0	0	2148	0	0	2828	0	0	i	
jlc3	104	143	115	222	0	0	2125	0	0	2862	0	0	i	
jlc4	144	185	114	261	0	0	2198	0	0	2918	0	0	i	
jmp1	72	107	100	314	0	0	1970	0	0	2698	0	0	i	
jmp2	72	110	102	318	0	0	1927	0	0	2751	0	0	i	
jmp3	56	98	100	321	0	0	1944	0	0	2712	2	0	i	
jmp4	48	79	103	323	0	0	1879	0	0	2654	0	0	i	
jph1	80	95	126	292	0	0	1970	1	0	2676	0	0	i	
jph2	72	98	131	323	0	0	1964	0	0	2629	0	0	i	
jph3	80	100	140	316	0	0	1997	0	0	2836	2	0	i	
jph4	64	89	153	385	0	0	1900	0	0	2572	2	0	i	
ms1	192	191	114	253	0	0	1934	0	0	2966	0	0	i	
ms2	136	177	111	266	0	0	1961	0	0	3035	0	0	i	
ms3	200	173	117	251	0	0	1976	0	0	3095	0	0	i	
ms4		234	114	279	0	0	1883	0	0	2984	1	0		

Table 11 bis. Formants finaux de la voyelle /i/ de la phrase 15.  
Analyse LPC à 50% ou à 80 ms du début de la voyelle, selon la répétition.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	135		359	74		M	1953	86		M	2530	109		M	<del>3326</del>	661		L					
aq2	140		382	66		M	1901	89		F	2492	81		F	<del>3578</del>	868		L					
aq3	137		351	39		M	1891	95		F	2541	171		M									
aq4	142		352	50		M	1904	30		F	2556	128		L	<del>3618</del>	655		L					
bz1	111		385	52		M	1681	156		F	2380	136		M	3454	99		L					
bz2	108		368	41		M	1731	118		F	2408	49		M	3637	167		L					
bz3	112		435	91		M	1732	179		F	2463	127		M	<del>3447</del>	261		M					
bz4	111		407	40		M	1745	84		M	2440	78		M	3718	172		L					
df1	173		435	159		L	1811	189		M	2912	141		M	3815	170		F					
df2	173		432	156		L	1825	118		M	2829	199		L	3774	115		M					
df3	177		426	91		L	1775	160		M	3032	191		M	3966	185		F					
gm1	177		415	16		L	1896	94		M	2611	256		L	<del>3966</del>	355		L					
gm2	177		385	26		L	1969	205		M	2688	220		M	<del>4456</del>	152		L					$F_4 = 3300 \ F_5 = 4456$
gm3	190		407	28		L	2012	130		M	2732	66		M	<del>3906</del>	235		L					
gm4	195		404	16		L	1960	152		M	2661	83		L	<del>3914</del>	463		L					
jfm1	121		381	104		M	<u>0</u>	0		M	<u>2379</u>	116			3852	246		L					$F_2 = 2379 \ F_3 = 2900$
jfm2	115		399	43		M	2160	60		M	<u>2537</u>	343		M	3905	86		M					$F_3 = 2800$
jfm3	117		410	30		M	2112	88		M	<u>2594</u>	442		M	3788	157		L					$F_3 = 2750$
jg1	95		408	60		M	1897	67		M	2578	77		M	<del>3276</del>	379		L					
jg2	102		405	82		M	1924	71		M	2574	84		M	3304	395		M					
jg3	114		468	130		L	1890	51		M	2538	88		M				M					$F_4 = 3179$
jg4	112		438	71		M	1869	64		M	2466	73		M				M					$F_4 = 3157$
jlc1	129		372	100		L	2033	128		M	2646	126		M	<del>3310</del>	538		L	<del>3647</del>	265		L	
jlc2	123		363	76		L	2017	71		M	2600	155		M	<del>3354</del>	164		L	<del>3851</del>	244			
jlc3	114		341	66		L	2033	166		M	2596	180		M	<del>3413</del>	364		L	<del>3970</del>	515			
jlc4	108		361	60		M	2004	118		M	2594	153		M	3414	262		L	<del>4198</del>	683			pas de F5
jmp1	95		412	59		M	1805	123		F	2559	90		M	3579	144		L					
jmp2	101		425	95		M	1852	173		F	2495	92		M	3490	169		L					
jmp3	97		431	110		M	1902	275		F	2562	113		M	<del>3467</del>	344		L					
jmp4	98		443	128		L	1830	227		F	2518	141		M	3521	178		L					
jph1	133		450	110		L	1894	41		M	2533	113		M	<del>3388</del>	685		L	<del>3622</del>	144		L	
jph2	137		423	120		M	1856	40		F	2498	124		M	<del>3373</del>	210		L	<del>3657</del>	215		L	
jph3	140		377	105		M	1829	66		F	2407	49		M	<del>3266</del>	416		L	<del>3515</del>	103		L	un seul Fi entre 3000 et 4000 Hz
jph4	140		444	94		M	1833	161		M	2443	61		M	<del>3344</del>	99		L	<del>3663</del>	196		L	
ms1	111		360	50		M	2049	96		M	2580	94		M	3733	214		M					
ms2	112		348	43		M	1936	62		M	2569	147		M	<del>3797</del>	226		L					
ms3	117		383	77		M	1970	64		M	<del>2555</del>	217		M	3721	193		M					

Table 12. Formants intermédiaires de la voyelle /e/ de la phrase 1 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.

Répé- tition	Durée (ms)	Durée (%)	F <sub>0</sub> (Hz)	F <sub>1</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>2</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	F <sub>3</sub> (Hz)	df <sub>2</sub>	df <sub>1</sub>	Ecoute	Remarques
aq1	96	105	135	364	0	0	1934	0	0	2513	0	0	e	
aq2	104	110	142	380	0	0	1887	0	0	2477	0	0	e	
aq3	120	111	137	351	0	0	1926	0	0	2526	0	0	e	
aq4	104	96	142	383	0	0	1898	0	0	2521	0	0	e	
bz1	88	100	111	401	0	0	1707	0	0	2369	0	0	e	
bz2	104	110	108	380	0	0	1728	0	0	2412	0	0	e	
bz3	80	83	112	438	0	0	1741	0	0	2478	0	0	e	
bz4	88	87	111	430	0	0	1747	0	0	2431	0	0	e	
df1	72	83	173	451	0	0	1848	0	0	2901	0	0	e	
df2	88	107	173	462	0	0	1843	0	0	2814	0	0	e	
df3	80	101	177	449	0	0	1796	0	0	3017	0	0	e	
df4		101	163	420	0	0	1884	0	0	2759	0	0		
gm1	72	82	177	416	0	0	1855	0	0	2555	1	0	e	
gm2	88	97	177	394	0	0	1968	0	0	2665	0	0	e	
gm3	88	98	190	421	0	0	1987	0	0	2655	0	0	e	
gm4	88	97	195	420	0	0	1973	0	0	2756	2	0	e	
jfm1	72	70	121	386	0	0	2336	0	0	<u>0</u>	3	0	i-e	F <sub>3</sub> manuelle = 2900
jfm2	104	111	115	417	0	0	2157	0	0	<u>2521</u>	0	0	e	F <sub>3</sub> manuelle = 2800
jfm3	112	110	117	415	0	0	2112	0	0	<u>2532</u>	1	0	e	F <sub>3</sub> manuelle = 2750
jfm4		102	137	444	0	0	2192	0	0	2751	0	0		
jg1	96	111	95	409	0	0	1887	0	0	2564	0	0	e	
jg2	88	92	102	422	0	0	1938	0	0	2541	0	0	e	
jg3	96	102	115	478	0	0	1884	0	0	2535	0	0	e	
jg4	88	101	112	452	0	0	1846	0	0	2488	0	0	e	
jlc1	96	100	129	391	0	0	2015	0	0	2625	0	0	e-i	
jlc2	80	90	123	383	0	0	2021	0	0	2580	0	0	i	
jlc3	88	91	114	385	0	0	1985	0	0	2576	0	0	i	
jlc4	88	91	109	393	0	0	1946	0	0	2554	0	0	i	
jmp1	80	94	95	422	0	0	1804	0	0	2525	0	0	e-i	
jmp2	80	93	101	426	0	0	1830	0	0	2493	0	0	e	
jmp3	64	79	97	434	0	0	1876	0	0	2524	0	0	e	
jmp4	72	85	98	452	0	0	1827	0	0	2482	0	0	e	
jph1	96	100	133	436	0	0	1896	0	0	2529	0	0	e	
jph2	88	100	137	422	0	0	1879	0	0	2487	0	0	e-i	
jph3	88	96	140	421	0	0	1843	0	0	2406	0	0	e-i	
jph4	88	92	140	443	0	0	1826	0	0	2428	0	0	e-i	
ms1	104	109	111	376	0	0	2011	0	0	2546	0	0	e-i	
ms2	104	95	112	364	0	0	1895	0	0	2581	0	0	e-i	
ms3	104	107	117	393	0	0	1908	1	0	2582	0	0	e-i	
ms4		77	108	398	0	0	1812	0	0	2443	0	0		

Table 12 bis. Formants finaux de la voyelle /e/ de la phrase 1.  
Analyse LPC à 50% du début de la voyelle.



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques		
aq1	145		351	167		F	756	119		F	2291	153		M	3101	338		M							
aq2	160		370	267		M	835	204		M	2233	260		M	3357	173		M							
aq3	153		369	195		M	847	268		M	2299	236		M	3350	372		M							
aq4	137		355	381		M	764	102		M	2365	170		M	3122	114		M							
bz1	111		358	57		M	825	96		M	2292	117		F	3393	157		F							
bz2	106		362	50		M	830	58		M	2370	48		F	3342	185		F	4182	477					
bz3	109		349	150		M	641	316			2402	544			3237	703								F1, F2 peu visibles ; ni F3 ni F4	
bz4	111		361	129		M	782	365			2492	757			3040	959			4349	566				F1, F2 peu visibles ; ni F3 ni F4	
df1	170		392	129		M	779	171		M	2120	441			2826	596			3615	289				friction	
df2	170		377	160			774	201			2214	431			3242	332								friction	
df3	170		386	96		M	1001	274		M	2011	496			3120	363									
gm1	195		0	0			734	135			2293	143		M	3235	161		M	4078	416					$F_1$ brute = 421
gm2	195		361	63			717	47			2354	172		F	3376	203		M							
gm3	228		0	0			719	76			2290	139		F	3234	151		M							$F_1$ brute = 425
gm4	200		364	115			754	33			2383	75		M	3116	251		M	3736	246					
jfm1	121		338	80		M	830	48		M	2641	122		F	3661	200		F	3912	532					
jfm2	114		357	94		M	820	239		F	2180	634			2737	452			3745	243					$F_3$ brute = 2737 ; F3, F4 peu visibles
jfm3	112		370	304			837	95			0	0			2729	280			3508	136					$F_3$ brute = 2729 ; ni F3 ni F4
jg1	106		0	0			767	630			2407	635			3151	348		M							$F_1$ brute = 422 ; pas de F3
jg2	108		332	196			824	345			2202	804			2784	452		M	3793	985					pas de F3 ; F4 peu visible
jg3	112		0	0			935	810			2590	307		M	3668	631		M							$F_1$ brute = 454 ; F3, F4 peu visibles
jg4	121		0	0			638	216			2334	920		M	3346	394		L							F4 peu visible
jlc1	115		345	138		M	801	242		M	2394	278		M	3039	478			3356	220		M			$F_4$ = 3356
jlc2	121		0	0			607	692			2566	234		M	3471	399		M							
jlc3	111		334	149		M	732	128			2410	259		M	2701	590			3326	188					$F_4$ = 3326
jlc4	115		404	308			614	416			2645	431			3322	209									F3, F4 peu visibles
jmp1	108		361	124		L	908	105		F	2325	225		F	3340	297		F	4295	290					
jmp2	109		358	105			796	83			2279	257			3246	96		F	4390	209					pas de F3
jmp3	97		354	40		M	804	54		M	2318	88		F	3296	124		F	4152	525					
jmp4	106		354	40		M	837	38		M	2229	136		F	3249	151		F	4176	649					
jph1	142		379	86		L	841	118		M	2577	186		M	3156	76		M							
jph2	133		356	58		M	792	103		M	2427	203		F	3225	195		F							
jph3	148		362	45			716	75		M	2475	196		F	3234	152		M							
jph4	156		372	109			833	361		M	2283	626		M	2843	802			3249	327					$F_3$ = 2400
ms1	111		386	196			826	135			2139	157			3096	550			3720	298					F3, F4 peu visibles
ms2	106		0	0			0	0			2191	421		M	3316	804			4062	903					pas de F4
ms3	112		415	321		M	909	311		M	2185	232		M	3572	292									

Table 13. Formants intermédiaires de la voyelle /u/ de la phrase 8 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	56	52	145	336	0	1	736	0	0	2365	1	0	puR <sup>u</sup>	
aq2	64	67	160	398	0	0	812	1	1	2235	2	0	puR <sup>u</sup>	
aq3	40	43	153	381	0	1	789	0	0	2367	1	0	puR <sup>u</sup>	
aq4	56	50	137	357	0	1	728	0	1	2417	0	0	puR <sup>u</sup>	
bz1	40	40	111	363	0	0	810	0	0	2300	0	0	p <sup>R</sup> u	
bz2	56	68	106	364	0	0	830	0	0	2383	0	0	p <sup>R</sup> u	
bz3	56	73	109	341	0	0	666	0	2	2409	2	1	puR <sup>u</sup>	
bz4	56	68	111	360	0	0	739	0	2	2322	2	0	puR <sup>u</sup>	
df1	56	62	170	400	0	0	780	0	0	2083	2	1	phhu	
df2	56	67	170	449	2	1	753	0	2	2038	3	2	phhu	(2214, 2540, 2767)
df3	24	28	170	378	0	0	1006	0	0	2038	0	3	phhu	
df4		55	160	446	0	2	817	0	0	2038	3	2		(2072, 2684, 2237)
gm1	72	78	195	421	0	0	744	0	0	2359	1	0	puR <sup>u</sup>	
gm2	72	72	195	346	0	0	705	0	0	2401	0	0	poR <sup>u</sup>	
gm3	64	78	228	410	0	0	714	0	0	2346	0	0	puR	
gm4	72	65	200	358	0	0	748	0	0	2421	0	0	puR <sup>u</sup>	
jfm1	56	57	121	336	0	0	819	0	0	2664	0	0	puR <sup>&amp;</sup>	
jfm2	40	43	114	335	0	0	860	0	0	2177	2	2	puR <sup>&amp;</sup>	$F_3 = 2737$
jfm3	48	41	112	380	0	1	809	0	0	2705	0	0	puR <sup>&amp;</sup>	
jfm4		60	137	390	0	0	841	0	1	0	4	0		$F_3$ hors limite = 2872
jg1	80	80	109	0	4	0	539	2	0	2614	2	0	puR <sup>u</sup>	$F_1 = 422$ $F_2 = 767$
jg2	48	55	108	341	0	2	713	1	0	2500	2	0	puR <sup>u</sup>	
jg3	48	52	114	429	0	1	784	2	0	2538	0	0	puR <sup>u</sup> ou pR <sup>o</sup>	
jg4	40	45	121	0	4	0	665	0	0	2503	2	0	pR <sup>o</sup>	(0, 0, 360)
jlc1	56	55	115	340	0	1	734	1	0	2469	1	0	puR <sup>&amp;</sup>	
jlc2	48	53	121	338	2	0	653	0	1	2592	0	0	puR <sup>&amp;</sup>	
jlc3	48	46	111	335	0	1	718	0	0	2466	1	0	puR <sup>&amp;</sup>	
jlc4	48	56	115	393	2	2	573	0	1	2563	1	0	pu&	$F_2$ manuelle = 660
jmp1	48	57	108	365	0	0	877	0	0	2332	0	0	puR <sup>u</sup>	
jmp2	32	44	109	359	0	0	806	0	0	2482	2	1	puR <sup>&amp;</sup>	
jmp3	48	59	97	356	0	0	801	0	0	2372	1	0	puR <sup>&amp;</sup>	
jmp4	48	52	106	362	0	0	781	0	0	2302	1	0	pu <sup>R</sup> &	
jph1	40	46	142	375	0	0	759	1	0	2571	0	0	puR <sup>u</sup>	
jph2	64	71	133	355	0	0	756	0	0	2463	0	0	puR <sup>u</sup>	
jph3	56	62	148	361	0	0	673	0	0	2478	0	0	puR <sup>u</sup>	
jph4	48	50	156	371	0	0	781	0	1	2329	2	1	puR <sup>u</sup>	
ms1	64	61	111	384	2	0	794	1	0	2142	1	0	puR <sup>&amp;</sup>	
ms2	64	64	106	420	2	0	818	2	1	2252	0	0	puR <sup>u</sup>	
ms3	64	58	112	394	0	1	833	1	0	2225	0	0	puR <sup>u</sup>	
ms4		57	109	419	0	0	824	0	1	1976	2	0		

Table 13 bis. Formants finaux de la voyelle /u/ de la phrase 8.  
Analyse LPC à 50% du début de la voyelle.

Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	142		301	137		M	783	235		F	2156	128		F	<del>3536</del>	266			4296	318			pas de F4
aq2	137		266	110		M	745	219		F	<del>2098</del>	179			<del>3467</del>	250			4198	639			ni F3 ni F4
aq3	133		230	104		M	699	114		F	<del>2110</del>	271			<del>2944</del>	465			4037	267			ni F3 ni F4
aq4	135		305	270			<del>973</del>	977			<del>2228</del>	195			<del>3410</del>	367			<del>4365</del>	386			$F_2 = 795$ ; $ni F4 ni F5$
bz1	109		304	76		F	<del>696</del>	140		F	<del>2084</del>	290			<del>3225</del>	419			<del>3726</del>	458			$F_2 = 760$ ; ni F3 ni F4
bz2	108		311	63		M	733	60		M	2097	159		F	<del>3212</del>	508		L	<del>3750</del>	785			F3, F4 peu visibles
bz3	108		299	57		M	730	62		M	2027	203		F	3376	133		M					
bz4	106		306	51		M	701	97		M	2162	171		F	<del>3350</del>	344		L	<del>3647</del>	296			$F_4 = 3500$
df1	173		<del>328</del>	157			<del>678</del>	211			<del>2000</del>	275			<del>2857</del>	511			<del>3967</del>	482			rien au-dessus de 1000 Hz
df2	168		332	162		M	747	108		M	2632	360		F	3329	239		F					F3, F4 peu visibles
df3	173		<del>257</del>	204			<del>647</del>	449			<del>2003</del>	605			<del>3201</del>	618			<del>3725</del>	664			rien au-dessus de 1000 Hz
gm1	177		273	97		F	660	91		M	<del>2100</del>	209			<del>2938</del>	498			<del>4044</del>	265			ni F3 ni F4
gm2	181		303	115		M	705	113		M	2123	209		F	<del>3074</del>	501			4195	120			pas de F4
gm3	181		296	123		M	717	192		M	2109	77		F	<del>3986</del>	245							pas de F4
gm4	181		264	63		M	665	60		M	<del>1985</del>	525		F	<del>3046</del>	388			<del>4208</del>	69			pas de F4
jfm1	115		290	79			700	392			<del>2427</del>	354			3803	87							rien au-dessus de 1000 Hz
jfm2	126		278	198			666	104			<del>2622</del>	506			<del>3958</del>	179							rien au-dessus de 1000 Hz
jfm3	111		280	174		M	715	157		M	<del>2451</del>	446			<del>3497</del>	790			4171	315			rien au-dessus de 1000 Hz
lg1	108		<del>263</del>	63			<del>720</del>	214			<del>2395</del>	283			<del>3136</del>	916			<del>3908</del>	704			rien au-dessus de 1000 Hz
lg2	102		297	118		M	715	223		M	2317	242		F	<del>3834</del>	199							pas de F4
lg3	114		286	80		M	724	46		M	2343	208		F	<del>3919</del>	329							
lg4	97		277	38		M	735	104		M	<del>2289</del>	346			<del>3013</del>	292			<del>4027</del>	185			rien au-dessus de 1000 Hz
jlc1	111		<del>318</del>	177		M	<del>702</del>	190		M	2643	258		F	3483	185		M	<del>4333</del>	744			
jlc2	117		<del>309</del>	98			<del>708</del>	181			<del>2406</del>	581			<del>3354</del>	828			<del>3613</del>	215			$F_3 = 2585$ $F_4 = 3500$
jlc3	115		288	134			687	104			<del>2325</del>	440			<del>2936</del>	746			<del>3639</del>	136			rien au-dessus de 1000 Hz
jlc4	119		<del>281</del>	158			<del>738</del>	121			<del>2372</del>	665			<del>3097</del>	480			<del>3780</del>	223			rien au-dessus de 1000 Hz
jmp1	101		294	128		M	689	71		M	2187	104		F	<del>2999</del>	266			<del>3919</del>	205			pas de F4
jmp2	96		305	74		M	748	28		M	2239	76		F	<del>3167</del>	214			<del>4400</del>	769			F4 peu visible
jmp3	96		277	135		M	739	50		M	<del>2262</del>	324			3293	423			<del>4278</del>	532			rien au-dessus de 1000 Hz
jmp4	101		307	88		M	716	38		M	<del>2163</del>	224			<del>3043</del>	297			<del>3802</del>	430			rien au-dessus de 1000 Hz
jph1	131		316	42		M	773	40		M	2424	212		F	3138	67		M	<del>3719</del>	437			pas de F5
jph2	137		340	25		M	739	73		M	2422	48		F	3255	324		M					
jph3	135		326	47		M	719	82		F	2421	166		F	3145	86		F					
jph4	142		306	32		M	697	106		F	2401	180		F	3207	191		M					
ms1	112		322	125		M	724	476		M	2024	24		M	3699	240		M					
ms2	112		321	85		M	754	198		F	2035	168		F	<del>3151</del>	726			<del>3959</del>	274			F4 peu visible
ms3	115		315	95		M	742	212		F	2077	227		F	<del>2969</del>	446			<del>3867</del>	219			pas de F4

Table 14. Formants intermédiaires de la voyelle /u/ de la phrase 12 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.



Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	280	262	142	277	0	0	791	0	0	2126	0	0	buR	
aq2	264	243	137	268	0	0	793	0	2	2128	1	0	buR	
aq3	280	248	133	245	0	0	692	0	0	2118	0	0	buR	
aq4	216	165	135	315	0	0	<b>966</b>	2	2	2201	0	0	buR	$F_2$ manuelle = 795
bz1	336	293	109	302	0	0	<b>709</b>	0	0	2134	0	0	buR	$F_2$ manuelle = 760
bz2	320	302	108	322	0	0	739	0	0	2124	0	0	buR	
bz3	360	350	108	308	0	0	692	0	0	2051	0	0	buR	
bz4	296	288	106	309	0	0	735	0	0	2157	0	0	buR	
df1	184	196	173	341	0	0	688	0	0	2051	2	1	buR	
df2	192	194	168	334	0	0	757	0	0	2596	0	0	buR	
df3	176	205	173	234	0	2	607	0	2	1980	2	2	buR	
df4		235	156	356	0	0	671	0	1	2658	0	0		
gm1	288	262	177	272	0	0	661	0	0	2043	1	0	buohh	
gm2	248	267	181	317	0	0	689	0	0	2069	0	0	buohh	
gm3	256	262	181	282	0	0	654	0	0	2089	0	0	buō	
gm4	248	234	181	275	0	0	675	0	0	2063	2	1	buōhh	
jfm1	208	158	117	290	0	0	711	0	1	2450	0	0	buR	
jfm2	240	183	126	282	0	1	665	0	0	□	3	2	buR	(1975, 2154, 0)
jfm3	208	175	111	303	0	1	677	0	0	□	3	1	buR	(2598, 2451, 2144)
jfm4		194	135	□	3	1	□	3	2	□	3	1		(362, 250, 0); (0, 653, 506); (2679, 1998, 0)
ig1	184	181	108	277	0	0	716	0	1	2343	1	1	buR	
ig2	184	169	103	297	0	0	742	0	2	2288	1	0	buR	
ig3	208	215	114	279	0	0	731	0	0	2311	0	0	buR	
ig4	184	181	97	283	0	0	750	0	0	2301	0	0	buR	
jlc1	216	221	111	329	0	0	722	0	0	2619	0	0	buR	
jlc2	240	233	117	303	0	0	703	0	0	2659	0	0	buR	
jlc3	200	220	115	290	0	1	674	0	0	2351	2	1	buR	
jlc4	160	147	119	284	0	0	725	0	0	□	3	1	puR	(2228, 0, 2561)
jmp1	200	220	101	283	0	0	682	0	0	2191	0	0	buR-puR	
jmp2	192	212	96	317	0	0	722	0	0	2227	0	0	puR	
jmp3	200	236	96	292	0	0	715	0	0	2095	2	0	puR	
jmp4	144	135	101	321	0	1	704	0	0	2213	1	0	puR	
jph1	208	191	131	312	0	0	763	0	0	2430	0	0	buR <sup>4c</sup>	
jph2	240	211	137	326	0	0	704	0	0	2424	0	0	buR	
jph3	240	204	135	325	0	0	735	0	0	2417	0	0	buR <sup>4c</sup>	
jph4	208	183	142	306	0	0	702	0	0	2425	0	0	buo	
ms1	256	196	112	316	0	0	761	0	2	2025	0	0	buR	
ms2	232	219	112	318	0	0	767	0	1	2052	0	0	buR	
ms3	240	202	115	331	0	0	783	0	1	2087	0	0	buR	
ms4		230	108	323	0	0	795	0	0	2012	0	0		

Table 14 bis. Formants finaux de la voyelle /u/ de la phrase 12.  
Analyse LPC à 80 ms du début de la voyelle, sauf pour jmp4 (50%).



Répé- tition	$F_0$ (Hz)	A	$F_1$ (Hz)	$B_1$ (Hz)	A	T	$F_2$ (Hz)	$B_2$ (Hz)	A	T	$F_3$ (Hz)	$B_3$ (Hz)	A	T	$F_4$ (Hz)	$B_4$ (Hz)	A	T	$F_5$ (Hz)	$B_5$ (Hz)	A	T	Remarques
aq1	140		287	53		F	794	85		F	2155	58		F	<del>3046</del>	198			<del>4150</del>	206			pas de F5
aq2	117		321	95		F	822	104		F	2264	60		M	3162	80		M	4211	242		F	
aq3	150		276	82		F	807	87		M	2198	31		M	<del>3204</del>	182			4056	143		F	pas de F4
aq4	135		287	189		M	768	121		M	2241	139		M	3120	69		M	4302	329		F	
bz1	108		297	38		F	796	61		F	2223	56		F	3389	159		M	<del>4475</del>	520			F4 peu visible ; pas de F5
bz2	103		304	93		F	744	119		M	2257	87		F	3184	85		M	<del>4235</del>	539		L	$F_5 = 4500$
bz3	108		300	50		F	787	103		M	2229	90		F	3488	171		M					
bz4	103		300	66		F	805	110		M	2238	105		F	3322	141		F					
df1	170		<del>331</del>	156			<del>660</del>	91			<del>2599</del>	555			<del>3221</del>	348							rien au-dessus de 1000 Hz
df2	173		<del>348</del>	203			<del>725</del>	222			<del>2309</del>	772			<del>3117</del>	603							rien au-dessus de 1000 Hz
df3	170		<del>366</del>	168			<del>682</del>	155			<del>0</del>	0			<del>2992</del>	387			<del>3802</del>	901			rien au-dessus de 1000 Hz
gm1	186		374	27		M	785	47		M	2330	140		M	3155	435		M	3857	607		F	
gm2	186		390	25		M	793	55		M	2390	22		M	3245	160		F	4359	344			rien entre 4000 et 5000 Hz
gm3	177		374	33		M	792	32		M	2371	138		M	<del>3626</del>	144		L					
gm4	186		354	57		M	768	77		M	2380	45		M	3283	176		L	4230	576		M	F4, F5 peu visibles
jfm1	117		313	45		F	807	52		F	2523	58		F	3861	102		F					F3 peu visible
jfm2	118		329	67		M	873	263		M	2615	84		M	3865	78		M					F3 peu visible
jfm3	123		342	77		M	828	98		M	2593	40		M	3730	132		M	<del>4001</del>	725			
lg1	97		302	151		M	796	278		M	2392	198		M	<del>2941</del>	690			4490	394			rien entre 3000 et 5000 Hz
lg2	103		308	158		M	791	104			<del>2535</del>	398			<del>3783</del>	850							rien au-dessus de 1000 Hz
lg3	102		289	77		M	806	135		M	2284	214		F	<del>2945</del>	594							rien au-dessus de 3000 Hz
lg4	93		<del>0</del>	0			<del>852</del>	553			<del>2396</del>	365			<del>3314</del>	344							u trop bref
jlcl	116		303	143		M	695	111		M	2488	178		M	<del>3106</del>	500			3457	189		M	
jlcl2	121		301	118		M	755	27		M	2522	147		M	<del>3088</del>	287		L	<del>3633</del>	240		L	F4, F5 peu visibles
jlcl3	112		309	86		M	747	85		M	2443	176		F	<del>3244</del>	325			<del>3594</del>	335		M	F3 peu visible ; $F_4 = 3400$
jlcl4	117		312	77		M	747	126		M	2510	125		M	<del>3170</del>	508		L	<del>3795</del>	284		L	F4, F5 peu visibles
jmp1	98		338	59		M	788	100		M	2249	104		M	3283	163		M	3912	650		M	
jmp2	102		359	50		M	860	160		M	2268	202		M	3145	442		M	3895	260		M	
jmp3	104		358	55		M	830	115		M	2320	583		M	3233	485		M	4061	598		M	F3 peu visible
jmp4	101		346	56		M	895	91		M	2266	133		M	3319	108		M	4000	454		M	
jph1	125		349	58		L	802	136		M	2340	118		M	2929	94		M					
jph2	129		360	46		M	800	113		M	2258	301		M	<del>2914</del>	347		M	<del>3232</del>	189			$F_4 = 3232$
jph3	133		337	63		M	780	86		M	2423	279		M	3083	103		M					
jph4	137		337	45		L	896	75		M	<del>2254</del>	467		L	<del>2885</del>	446		L	<del>3203</del>	80		M	$F_3 = 2400$ $F_4 = 3203$
ms1	106		289	64		M	654	75		M	2008	117		F	3114	255			3921	569			rien au-dessus de 2000 Hz
ms2	115		309	70		M	739	46		M	2032	147		F	3175	502			3700	291		F	F3 peu visible
ms3	114		315	141		M	700	117		M	2070	59		M	<del>3177</del>	910			3670	450			F4 peu visible

Table 15. Formants intermédiaires de la voyelle /u/ de la phrase 16 dans la première étude.  
Analyse LPC à 40% du début de la voyelle.

Répé- tition	Durée (ms)	Durée (%)	$F_0$ (Hz)	$F_1$ (Hz)	$df_2$	$df_1$	$F_2$ (Hz)	$df_2$	$df_1$	$F_3$ (Hz)	$df_2$	$df_1$	Ecoute	Remarques
aq1	136	124	140	291	0	0	795	0	0	2173	0	0	buR	
aq2	80	71	117	329	0	0	822	0	0	2296	0	0	buR	
aq3	168	149	150	278	0	0	808	0	0	2190	0	0	buR	
aq4	152	117	135	293	0	0	758	0	0	2263	0	0	buR	
bz1	288	280	108	292	0	0	816	0	0	2230	0	0	buR	
bz2	176	132	103	313	0	0	755	0	0	2256	0	0	buR	
bz3	160	128	108	309	0	0	784	0	0	2240	0	0	buR	
bz4	200	146	103	298	0	0	808	0	0	2237	0	0	buR	
df1	40	46	170	356	2	1	658	2	0	1941	4	0	buR	(2599, 0, 0)
df2	56	72	173	332	1	1	723	0	1	1941	2	2	buR	
df3	48	62	170	363	2	0	660	2	0	1941	4	0	buR	(0, 0, 2661)
df4		50	160	392	0	0	762	0	0	2650	1	0		
gm1	88	88	186	377	0	0	775	0	0	2364	0	0	buR	
gm2	96	94	186	387	0	0	787	0	0	2404	0	0	buR	
gm3	120	115	177	368	0	0	767	0	0	2329	0	0	buR	
gm4	136	121	186	351	0	0	757	0	0	2372	0	0	buR	
jfm1	104	93	117	317	0	0	806	0	0	2541	0	0	buR	
jfm2	80	80	118	329	0	0	857	0	0	2597	0	0	buR	
jfm3	56	52	123	342	0	0	811	0	0	2617	0	0	buR	
jfm4		66	142	335	0	0	707	0	0	2705	0	0		
jg1	64	66	98	315	0	0	782	0	0	2403	0	0	buR	
jg2	48	54	103	313	2	0	776	2	0	2490	2	1	buR	
jg3	80	82	102	303	0	0	774	0	0	2323	0	0	buR	
jg4	8	10	93	303	4	0	774	3	2	2388	2	1	buR	(478, 0, 0); (852, 505, 651)
jlc1	136	124	117	313	0	0	706	0	0	2509	0	0	buR	
jlc2	184	179	121	300	0	0	743	0	0	2514	0	0	buR	
jlc3	176	160	112	302	0	0	735	0	0	2432	0	0	buR	
jlc4	160	149	117	322	0	0	727	0	0	2515	0	0	buR	
jmp1	56	62	98	347	0	0	786	0	0	2284	0	0	buR	
jmp2	96	104	103	364	0	0	840	0	0	2294	0	0	b <sup>R</sup> <sub>u-o</sub>	
jmp3	88	103	104	365	0	0	818	0	0	2303	0	0	buR	
jmp4	96	111	102	351	0	0	887	0	0	2266	0	0	b <sup>R</sup> <sub>u-o</sub>	
jph1	80	75	125	346	0	0	776	0	0	2367	0	0	bu <sup>R</sup> <sub>u</sub>	
jph2	72	75	129	364	0	0	770	0	0	2308	0	0	bo	
jph3	88	92	133	347	0	0	<u>720</u>	0	0	2402	0	0	bo <sup>R</sup>	$F_2$ manuelle = 760
jph4	88	94	137	351	0	0	826	0	0	2338	0	0	bo	
ms1	240	205	106	289	0	0	663	0	0	2068	0	0	buR	
ms2	128	124	115	315	0	0	730	0	0	2069	0	0	buR	
ms3	152	140	114	326	0	0	707	0	0	2104	0	0	buR	
ms4		64	106	359	0	0	800	0	0	2377	0	0		

Table 15 bis. Formants finaux de la voyelle /u/ de la phrase 16.  
Analyse à 50% ou à 80 ms du début de la voyelle, selon la répétition.



# Abstract

Methods for achieving automatic speaker recognition may be classified into two categories : pattern recognition based approaches that implicitly use interspeaker and intraspeaker variability of speech and approaches which explicitly take into account the sources of interspeaker and intraspeaker differences. The latter examine linguistic units in order to extract features which are relevant for speaker characterization. The aim of the present study is precisely to analyse the relative effectiveness of the three first formants of French vowels in the context /p-vowel-R/.

These selected trigrams are a part of a larger set of preselected acoustic and phonetic parameters which lead us to record and digitalize a set of seventeen sentences, uttered four times by ten male speakers coming from the same region.

We begin to expound the required knowlegdes about speech production process and intraspeaker and interspeaker variability for understanding issues in speaker characterization.

We then present a review of the studies in automatic speaker recognition according to the above classification. With regards of methods adapted from the domain of speech recognition, we restrict our review to the more recent works. In the other hand, we try to be more exhaustive for studies in speaker characterization.

Following this, we develop the different stages of our study. After describing and proving the preselected acoustic and phonetic parameters, we deal with the the hand-labeling of our corpus according to strict rules. We examine furthermore the effectiveness of the three first formants of French vowels in the context /p-vowel-R/. For that purpose, we firstly propose an automatic method to determine reliable values of the three frequencies of the first formants of selected vowels. We besides discuss the reliability of the results. For every vowel, we then analyse the relative effectiveness of each combination of formants frequencies and differences between these frequencies. This analyse is based on the computation of three "relevance indicators" from the results of speaker identification experiments.

## Keywords

Speech processing, speaker recognition, speaker characterization, formant computation, speech variability, speech labeling,

# Résumé

Les recherches en reconnaissance automatique du locuteur peuvent être réparties en deux grandes classes, d'une part les recherches qui sont fondées sur les techniques de reconnaissance de formes issues de la reconnaissance automatique de la parole, et d'autre part, celles qui ont pour objectif d'exploiter explicitement la variabilité interlocuteur et la variabilité intralocuteur de la parole. Pour cela, ces dernières tentent d'extraire du signal de parole des paramètres acoustiques et phonétiques qui caractérisent au mieux le locuteur. Notre étude se situe dans cette seconde classe.

Dans un premier temps, nous présentons les connaissances des domaines de la production de la parole et de la variabilité de la parole qui sont nécessaires à la compréhension de la problématique de la caractérisation du locuteur.

Puis, nous plaçons notre travail dans le cadre plus général de la reconnaissance automatique du locuteur en présentant une synthèse des travaux les plus récents dans les deux classes de recherche que nous avons définies au début de ce résumé.

Enfin, nous développons les différentes étapes de notre étude.

Après une présentation des paramètres acoustiques et phonétiques sélectionnés comme susceptibles de caractériser au mieux le locuteur, nous décrivons les phases d'élaboration et d'étiquetage manuel de notre corpus, ce qui nous conduit à exposer notre vision de la problématique de l'étiquetage.

Puis, nous détaillons l'étude de la pertinence des trois premiers formants de sept voyelles orales dans un contexte /p-voyelle-R/. Pour cela, nous exposons tout d'abord une méthodologie de détermination automatique de valeurs robustes des trois premiers formants des voyelles et nous commentons les résultats obtenus.

Ensuite, nous présentons, pour chacun des triplets sélectionnés, l'analyse de la pertinence des combinaisons de formants et des écarts entre les formants pour l'identification du locuteur. Cette analyse est fondée sur l'établissement de trois indicateurs issus d'expériences d'identification d'un locuteur parmi dix.

Enfin, nous terminons par une interprétation de ces résultats et leur comparaison avec les conclusions d'autres études comme celles qui concernent la normalisation des fréquences formantiques en reconnaissance automatique de la parole.

## Mots clés

Traitement automatique de la parole, caractérisation du locuteur, reconnaissance du locuteur, identification du locuteur, vérification du locuteur, variabilité intralocuteur, variabilité interlocuteur, formants, étiquetage, segmentation.



# UNIVERSITE DE NANCY I

NOM DE L'ETUDIANT : Mademoiselle MELLA Odile

NATURE DE LA THESE : DOCTORAT DE L'UNIVERSITE DE NANCY I  
en INFORMATIQUE

VU, APPROUVE ET PERMIS D'IMPRIMER

NANCY, le 10 FEV. 1993 n°66

LE PRESIDENT DE L'UNIVERSITE DE NANCY I



# Résumé

Les recherches en reconnaissance automatique du locuteur peuvent être réparties en deux grandes classes, d'une part les recherches qui sont fondées sur les techniques de reconnaissance de formes issues de la reconnaissance automatique de la parole, et d'autre part, celles qui ont pour objectif d'exploiter explicitement la variabilité interlocuteur et la variabilité intralocuteur de la parole. Pour cela, ces dernières tentent d'extraire du signal de parole des paramètres acoustiques et phonétiques qui caractérisent au mieux le locuteur. Notre étude se situe dans cette seconde classe.

Dans un premier temps, nous présentons les connaissances des domaines de la production de la parole et de la variabilité de la parole qui sont nécessaires à la compréhension de la problématique de la caractérisation du locuteur.

Puis, nous plaçons notre travail dans le cadre plus général de la reconnaissance automatique du locuteur en présentant une synthèse des travaux les plus récents dans les deux classes de recherche que nous avons définies au début de ce résumé.

Enfin, nous développons les différentes étapes de notre étude.

Après une présentation des paramètres acoustiques et phonétiques sélectionnés comme susceptibles de caractériser au mieux le locuteur, nous décrivons les phases d'élaboration et d'étiquetage manuel de notre corpus, ce qui nous conduit à exposer notre vision de la problématique de l'étiquetage.

Puis, nous détaillons l'étude de la pertinence des trois premiers formants de sept voyelles orales dans un contexte /p-voyelle-R/. Pour cela, nous exposons tout d'abord une méthodologie de détermination automatique de valeurs robustes des trois premiers formants des voyelles et nous commentons les résultats obtenus.

Ensuite, nous présentons, pour chacun des triplets sélectionnés, l'analyse de la pertinence des combinaisons de formants et des écarts entre les formants pour l'identification du locuteur. Cette analyse est fondée sur l'établissement de trois indicateurs issus d'expériences d'identification d'un locuteur parmi dix.

Enfin, nous terminons par une interprétation de ces résultats et leur comparaison avec les conclusions d'autres études comme celles qui concernent la normalisation des fréquences formantiques en reconnaissance automatique de la parole.

## Mots clés

Traitement automatique de la parole, caractérisation du locuteur, reconnaissance du locuteur, identification du locuteur, vérification du locuteur, variabilité intralocuteur, variabilité interlocuteur, formants, étiquetage, segmentation.