N° D'ORDRE 3306

THESE

présentée devant

L'UNIVERSITE PAUL SABATIER DE TOULOUSE

(SCIENCES)

en vue de l'obtention du titre de

Docteur de l'Université Paul Sabatier

Spécialité : Informatique

en

TRAITEMENT AUTOMATIQUE DE LA PAROLE

présentée par

François PELLEGRINO

Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques

soutenue le Mardi 22 décembre 1998 devant le jury constitué de

M. Jean-Luc Gauvain	Directeur de Recherche CNRS, LIMSI, Orsay	Rapporteur
M. Jean-Marie Hombert	Professeur, Université Lumière - Lyon 2, Lyon	Rapporteur
M[me] Régine André-Obrecht	Chargée de Recherche CNRS, IRIT, Toulouse	Directeur de thèse
M. Louis-Jean Boë	Ingénieur de Recherche CNRS, ICP, Grenoble	Examinateur
M. Daniel Dours	Professeur, Université Paul Sabatier, Toulouse	Examinateur
M. Edouard Geoffrois	Ingénieur de l'Armement, DGA, ETCA, Arcueil	Examinateur
M. Jean-François Mari	Professeur, Université Nancy 2, Nancy	Examinateur
M. Guy Pérennou	Professeur, Université Paul Sabatier, Toulouse	Examinateur

UMR CNRS 5505 118, Route de Narbonne 31062 TOULOUSE CEDEX

à mes parents,

à mes ancêtres voyageurs, de toute langue, de tout pays.

Remerciements

Je n'ai pas l'impression qu'il soit particulièrement aisé de formuler des remerciements… La seule chose qui me rassure, en fait, est que c'est encore plus difficile de le faire de vive voix !

Je tiens à remercier les membres du jury qui ont fait l'effort de lire cette thèse et d'assister à la soutenance, en cette avant-veille de fête familiale. Je tiens tout particulièrement à remercier messieurs Jean-Luc Gauvain et Jean-Marie Hombert pour leur travail de rapporteurs ; ils ont tous deux apporté une expertise précieuse dans des domaines complémentaires de l'ingénierie linguistique. Je tiens par ailleurs à exprimer ma gratitude envers Monsieur Jean-Marie Hombert pour l'intérêt qu'il porte à ce travail depuis plusieurs années maintenant.

Les discussions que j'ai eues avec Monsieur Louis-Jean Boë sont, dans une large mesure, à l'origine de mon intérêt pour les aspects phonétiques et phonologiques du traitement automatique de la parole ; qu'il soit remercié pour ses analyses et sa disponibilité.

Je remercie grandement Monsieur Daniel Dours d'avoir accepté de faire partie de ce jury, replongeant ainsi dans le domaine de la parole et des voyelles…

Monsieur Edouard Geoffrois, en tant que responsable « parole » au sein de la DGA, a suivi de près les travaux présentés dans cette thèse ; il a ainsi été un interlocuteur privilégié, et je le remercie pour les remarques qu'il a formulées au cours de nos différentes réunions de travail.

J'ai eu plusieurs fois l'occasion d'apprécier les qualités humaines et scientifiques de Monsieur Jean-François Mari, et je lui suis reconnaissant d'avoir accepté le rôle d'examinateur pour cette thèse, démontrant encore une fois son attrait pour les domaines variés et pluridisciplinaires !.

Je tiens à remercier Monsieur Guy Pérennou, qui, d'une part m'a accueilli au sein de l'équipe IHM-PT de l'IRIT, et d'autre part, à accepter de faire partie de ce jury.

Il est difficile, voire impossible d'exprimer ma gratitude pour Madame Régine André-Obrecht en quelques mots… M'ayant accepté en stage de DEA un peu par hasard, elle m'a inoculé le virus de la recherche et de la « parole ». En se montrant chaque jour à l'écoute de ses étudiants et en instaurant avec eux une relation dépassant largement le cadre encadrant-encadré, elle est à l'origine d'une dynamique humaine et scientifique efficace : on sort forcément grandi de son contact, et j'espère ne pas déroger à la règle. Je souhaite grandement continuer à collaborer avec elle, et quoi qu'il advienne, je lui exprime encore toute ma gratitude et mon amitié.

Attribuer au hasard ma rencontre avec Régine est un (petit) mensonge, puisque c'est à Bernard Teston que je dois cette chance. Qu'il soit remercié pour ce coup de pouce donné au destin !

Je tiens également à remercier tous les membres de l'équipe IHM-PT. J'ai eu la chance pendant trois ans d'occuper un bureau ou la bonne humeur est la règle, et j'emporte avec moi les souvenirs des fous rires avec Nathalie et Isabelle, en espérant bien que de nombreux autres vont suivre ! Je tiens également à remercier Nathalie pour l'aide qu'elle m'a apporté dans la rédaction de ce manuscrit, m'aidant à clarifier ce qui était obscur (et à corriger ce que l'on appelle sobrement les fautes de frappe…).

Martine, Christine et les autres ont bien évidemment participé à la bonne humeur régnant dans l'équipe, et je les remercie pour les bons moments passés ensemble. Au fil du temps, des « collègues », comme on dit dans le sud-est, sont apparus et d'autres sont partis vers d'autres horizons, parmi les premiers, je tiens à remercier Jérôme et à lui souhaiter une bonne continuation, tandis que parmi les derniers, Laure Arnaud et Bruno gardent toute mon amitié et ne sont pas oubliés…

Je tiens également à remercier les potes, les amis fidèles qui ont fait de ces années passées à vivre à Toulouse et à bourlinguer dans les Pyrénées un plaisir qui place la barre bien haut pour la suite… Marion, Stéphane, Eric, Laure, Olivier, Beb's, Magda, et tous ceux que j'oublie à Toulouse et ailleurs.

A Corinne, enfin, c'est bien plus que de la gratitude que j'exprime…

Résumé de la thèse

Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques

Les systèmes actuels d'Identification Automatique des Langues (IAL) se composent d'un module de pré-traitement acoustico-phonétique et d'un module phonotactique pour chaque langue. Le pré-traitement, basé sur des modèles de Markov cachés, nécessite de grandes quantités de données étiquetées. Notre approche, basée sur une modélisation différenciée pour chaque classe phonétique, vise à obtenir une identification phonétique efficace sans recourir à des données étiquetées. Elle est appliquée à la modélisation des voyelles, par analogie avec les typologies linguistiques existantes pour les systèmes vocaliques :

- la détection des voyelles est indépendante de la langue et ne requiert aucun apprentissage,
- un MMG (Modèle de Mélange de Gaussiennes) est calculé pour chacune des langues. Sa topologie est déterminée en utilisant le critère d'information de Rissanen,
- en phase d'identification, la langue la plus vraisemblable est sélectionnée.

Les expériences sont menées sur les locuteurs masculins de cinq langues (coréen, espagnol, français, japonais et vietnamien) du corpus téléphonique OGI MLTS. Le taux de voyelles correctement détectées est de 93,5 %, avec un taux d'insertion de 10 %. Dans une tâche d'identification sans rejet, avec des énoncés d'une durée de 45 secondes, le taux d'identification correct obtenu avec des modèles de systèmes vocaliques atteint 80 %. Une étude similaire est menée sur les systèmes consonantiques. La prise en compte conjointe des modèles de systèmes vocaliques et consonantiques aboutit à un taux d'identification correcte de 85 %. Ces résultats sont similaires à ceux obtenus avec un modèle phonétique segmental global. Lorsqu'on fusionne les deux approches, différenciée et globale, le taux d'identification atteint 91 % : la modélisation des systèmes vocaliques est pertinente en IAL et l'approche différenciée permet d'obtenir une discrimination phonétique efficace sans utiliser de données étiquetées.

Mots-clefs : Identificati on automatique des langues, systèmes vocaliques, modèle de mélange de lois gaussiennes, critère d'information de Rissanen.

Abstract

A phonetic approach to automatic language identification: the acoustic modelling of the vowel systems

Current systems of Automatic Language Identification (ALI) consist of a module of acoustic-phonetic decoding which is considered as a pre-processing and of a phonotactic module for each language. The pre-processing, based on Hidden Markov Models, requires a big amount of labelled data. Our approach, based on a Differentiated Phonetic Modelling (DPM) for each phonetic class, aims at getting an efficient phonetic identification without depending on labelled data. DPM is applied to acoustic-phonetic modelling of the vowels, in analogy with existing linguistic typologies for the vowel systems:

- The detection of vowels is language-independent and it requires no training phase.
- A Gaussian Mixture Model is trained for each language. Its topology is decided using the Rissanen information criterion.
- During the phase of identification, the most likely language is selected.

The experiments are undertaken on male speakers of five languages (Korean, Spanish, French, Japanese and Vietnamese) of the multilingual telephone speech corpus OGI MLTS. The mean rate of correctly detected vowels reaches 93,5 %, with an insertion rate of 10 %. In the identification task without rejection, with 45 second duration utterances, the rate of correct identification reaches 80 %, whereas two thirds of the data (the non-vocalic segments) are ignored.

A similar consonant system model is also proposed. When the vowel system and the consonant system models are both considered, the correct identification rate reaches 85 %. These results are similar to those obtained with a global segmental model. Merging the two approaches (differentiated and global modelling) results in an overall identification rate of 91 % : Vocalic system modelling is relevant in ALI and differentiated modelling results in an efficient phonetic discrimination without using labelled data.

Keywords: Automatic language identification, vowel system, Gaussian mixture model, Rissanen criterion of information.

Table des matières

TABLE DES MATIÈRES **IX**

TABLE DES ILLUSTRATIONS **XVII**

INDEX DES FIGURES **XIX**
INDEX DES TABLEAUX **XXII**

INTRODUCTION GÉNÉRALE **1**

1ÈRE PARTIE DES HOMMES ET DES LANGUES **7**

INTRODUCTION **9**
CHAPITRE 1 DIVERSITÉ ET CARACTÉRISATION DES LANGUES **11**
1 LES LANGUES, INSTRUMENTS DU LANGAGE HUMAIN 12
1.1 Le langage : un média de communication humain 12
1.2 L'émergence du langage 13
1.3 L'acquisition du langage : un programme génétique ? 16
2 UNE BRÈVE HISTOIRE DES LANGUES 18
2.1 Vie et mort des langues : un état des lieux 18
2.2 Une histoire de la linguistique 19
2.2.1 Naissance de la linguistique 19
2.2.2 L'essor de la linguistique comparative 20
2.2.3 Les succès des comparaisons multilatérales 22
2.2.4 Les super-familles et les controverses actuelles 25
2.2.5 Linguistique, Archéologie et Génétique 27
3 DE LA TAXINOMIE A LA CARACTÉRISATION 29
4 LA DISCRIMINATION DES LANGUES PAR L'ÊTRE HUMAIN 31
4.1 Le nourrisson reconnaît-il les langues ? 31
4.2 La discrimination des langues chez l'adulte 33
4.3 Discussion 35
CHAPITRE 2 LA CARACTÉRISATION DES SYSTÈMES VOCALIQUES **37**
1 LA VOYELLE ET SON RÔLE DANS LA COMMUNICATION 37
1.1 La voyelle dans la communication écrite [Jean 87, Bottéro 93] 37
1.2 La voyelle dans la communication orale 39

1.2.1 Aspect acoustico-articulatoire 39
1.2.2 Le système vocalique : structure & substance 44
2 UNE TYPOLOGIE DES SYSTÈMES VOCALIQUES : [MADDIESON 84, VALLÉE 94] 49
2.1 La base de données UPSID 49
2.2 Une typologie des systèmes vocaliques 51
2.2.1 Travaux antérieurs 51
2.2.2 Choix méthodologiques 52
2.2.3 Résultats 54
2.2.4 Discussion 55
CONCLUSION **57**

2[ÈME] PARTIE DES ORDINATEURS ET DES LANGUES **59**

INTRODUCTION **61**
CHAPITRE 1 UN CADRE MULTILINGUE POUR LE TRAITEMENT AUTOMATIQUE DE LA PAROLE **63**
1 LE TRAITEMENT AUTOMATIQUE DE LA PAROLE DANS UN CADRE MULTILINGUE 63
2 UNE INTRODUCTION À L'IDENTIFICATION AUTOMATIQUE DES LANGUES 67
2.1 Les enjeux en IAL 67
2.1.1 Les enjeux applicatifs 67
2.1.2 Les enjeux scientifiques 69
2.1.3 Les enjeux militaires 69
2.2 Un survol historique (1973-1989) 70
2.2.1 La Génèse 70
2.2.2 Les Années 80 71
2.3 Vers les systèmes actuels 73
CHAPITRE 2 UN ETAT DE L'ART DE L'IAL **75**
1 LES CORPUS DE DONNÉES 75
1.1 EUROM_1 76
1.1 OGI MLTS [Muthusamy 92] 76
1.2 LDC CALLFRIEND 78
2 UN PANORAMA DES SYSTÈMES ACTUELS 78
2.1 Les approches statistiques 79
2.1.1 Rensselaer Polytechnic Institute, New York, Etats-Unis 79
2.1.2 LIMSI, France 79
2.1.3 Ensigma Ltd, Angleterre 80
2.1.4 ATT Bell Labs - Etats-Unis (Kadambe - Hieronymous) 80
2.1.5 ATT Bell Labs - Etats-Unis (Ramesh - Roe) 81
2.1.6 Université d'Aalborg, Danemark 81
2.1.7 BBN Systems and Technologies, Etats-Unis 82
2.1.8 Université de Tokyo, Japon 82
2.1.9 MIT, Etats-Unis, (Hazen - Zue) 83
2.1.10 MIT, Etats-Unis, (Zissman) 83
2.1.11 OGI, Etats-Unis, (Yan - Barnard) 85

2.1.12 Technical University of Ilmenau, Allemagne 85
2.2 Les approches neuro-mimétiques 86
2.2.1 OGI, Etats-Unis, (Berkling - Barnard) 86
2.2.2 OGI, Etats-Unis, (Muthusamy) 86
2.3 Les autres approches 86
2.3.1 Un système d'identification prosodique 86
2.3.2 Un système d'identification du locuteur 87
2.4 Discussion 87
2.4.1 Tableau récapitulatif 87
2.4.2 Tendances générales en IAL 88
CHAPITRE 3 LA MODÉLISATION PHONETIQUE DIFFERENTIEE 91
1 LES PERSPECTIVES D'UN DOMAINE EN EVOLUTION 91
1.1 Les limitations actuelles 91
1.2 Les axes de recherche 93
1.2.1 Discussion 93
1.2.2 Le projet DGA « Discrimination multilingue automatique » 96
2 L'APPROCHE PAR MODÉLISATION PHONETIQUE DIFFERENTIÉE 98
2.1 Les motivations 98
2.2 Le cadre statistique de l'étude 101
3 L'ÉTUDE RÉALISÉE : LA MODÉLISATION DES SYSTÈMES VOCALIQUES 103
3.1 Pourquoi une approche vocalique ? 103
3.2 Description de l'étude réalisée 104
CONCLUSION 105

3ÈME PARTIE LA MODÉLISATION DES SYSTÈMES VOCALIQUES APPLIQUÉE À L'IAL 107

INTRODUCTION 109
CHAPITRE 1 LA DÉTECTION AUTOMATIQUE DES SEGMENTS VOCALIQUES 111
1 LOCALISATION DES SEGMENTS VOCALIQUES 111
1.1 Un schéma synoptique du système 112
1.2 La segmentation du signal 113
1.3 La détection d'activité vocale 115
1.4 La détection des segments vocaliques 117
1.5 Discussion 124
2 EXPÉRIENCES EN DÉTECTION DES VOYELLES 124
2.1 Expériences sur un sous corpus issu de EUROM 1 125
2.2 Expériences sur le corpus OGI MLTS 128
2.2.1 Description des données employées 128
2.2.2 Résultats de la détection des segments vocaliques 130
2.3 Discussion 134
CHAPITRE 2 LA MODÉLISATION ACOUSTIQUE DES SYSTÈMES VOCALIQUES 135
1 PARAMETRISATION DES VOYELLES 135

1.1 Choix de l'espace de représentation 135
1.2 Algorithmes de paramétrisation cepstrale 136
2 MODÉLISATION DU SYSTÈME VOCALIQUE 138
2.1 Choix de la modélisation 138
2.2 Initialisation du modèle : la quantification vectorielle (QV) 139
2.3 Modèles par mélange de lois gaussiennes (MMG) 141
2.3.1 Méthode d'apprentissage : l'algorithme EM 141
2.3.2 Relation entre QV et MMG 142
2.4 L'algorithme LBG-Rissanen 143
2.5 Discussion 145
3 EXPÉRIENCES DE MODÉLISATION DES SYSTEMES VOCALIQUES 145
3.1 Expériences sur le corpus JLVoc 146
3.1.1 Description des données employées 146
3.1.2 Modélisation du système vocalique et classification des voyelles 149
3.2 Expériences sur le sous corpus issu de EUROM 152
3.3 Expériences sur le corpus OGI MLTS 156
3.3.1 Normalisation et représentation cepstrale 157
CHAPITRE 3 LA DISCRIMINATION DES SYSTÈMES VOCALIQUES APPLIQUÉE À L'IAL 161
1 DESCRIPTION DES ENSEMBLES D'APPRENTISSAGE ET DE TEST 161
2 LE SYSTEME DE REFERENCE 162
2.1 Synoptique du modèle MMG segmental global 162
2.2 Topologie du modèle et résultats expérimentaux 163
2.2.1 Choix de l'espace de paramètres 163
2.2.2 Influence du nombre de composantes du modèle 164
2.3 Résultats et discussion 165
3 LA DISCRIMINATION AUTOMATIQUE DES SYSTÈMES VOCALIQUES (SV) 166
3.1 Description du système 166
3.2 Expériences en discrimination des SV 167
3.2.1 Influence du type de décision (WTA ou multigaussienne) 167
3.2.2 Influence de la topologie du modèle et de l'espace d'observation 168
3.2.3 Modèle à nombre de gaussiennes variable : application de l'algorithme LBG-Rissanen 171
3.3 Expériences complémentaires 175
3.3.1 Procédures de normalisation 175
3.3.2 Fusion de décisions 180
4 DISCUSSION 182
CHAPITRE 4 VERS UN SYSTÈME D'IAL PHONÉTIQUE COMPLET 185
1 DISCRIMINATION DIFFÉRENCIÉE DES SYSTEMES VOCALIQUES ET CONSONANTIQUES 185
1.1 Description du système d'identification phonétique 185
1.2 Expériences en discrimination des systèmes consonantiques 186
1.2.1 Influence de la topologie du modèle et de l'espace d'observation 186
1.2.2 Modèles à nombre de gaussiennes variable 187
1.2.3 Procédure d'élagage 189
1.2.4 Discussion 190
1.3 Discrimination des systèmes vocaliques et consonantiques 190

2 UN APPORT DE LA MODÉLISATION DIFFÉRENCIÉE ? 191
2.1 Optimisation du modèle segmental global 192
2.2 Fusion des modèles global et différencié 193
3 RÉSULTATS DETAILLES DU SYSTÈME DERIVÉ DES MODELES PHONÉTIQUES GLOBAL ET VOCALIQUE 195
3.1 Matrices de confusion obtenues pour DEV_10 et DEV_ST 195
3.2 Prise en compte des hommes et des femmes 196
3.3 Expériences en discrimination FR-L 197
3.4 Expériences en identification de quatre langues 197
CONCLUSION 199

CONCLUSION GÉNÉRALE ET PERSPECTIVES 201

ANNEXES 207

ANNEXE 1 LISTE DES LANGUES DES CORPUS CITÉS 209
1 CALLFRIEND (12 LANGUES & 3 DIALECTES) 209
2 CALLHOME (6 LANGUES) 209
3 EUROM_1 (11 LANGUES) 209
4 GLOBALPHONE (9 LANGUES) 210
5 IDEAL (4 LANGUES) 210
6 OGI 22 LANGUAGES (22 LANGUES) 210
7 OGI MLTS (11 LANGUES) 210
ANNEXE 2 LES MMG ET L'ALGORITHME EM 211

RÉFÉRENCES BIBLIOGRAPHIQUES 213

Table des illustrations

INDEX DES FIGURES

Figure 1 : Chez le chimpanzé, l'épiglotte forme avec le voile du palais une cloison étanche ; les singes peuvent donc manger et respirer simultanément. Chez l'homme, l'abaissement du larynx rend cette faculté impossible (d'après [Ross 97]). 14

Figure 2 : La théorie du « goulot d'étranglement » (d'après [Victorri 97]). 15

Figure 3 – Expériences d'identification des langues par 7 auditeurs. La durée des stimuli est de 6 secondes (d'après [Muthusamy 93]) 34

Figure 4 – Représentation schématique de l'ensemble pharynx + conduits oral et nasal [Calliope 89] 40

Figure 5 - a) Le quadrilatère des voyelles cardinales de l'API (version 1996) - b) Projection de prototypes vocaliques dans F2/F1 (Hz) (d'après [Vallée 94]) 41

Figure 6 – Description des 37 symboles vocaliques d'UPSID [Vallée 94] 53

Figure 7 – Exemples de systèmes vocaliques a) Système à cinq voyelles le plus fréquent b) Système à 8 voyelles et 2 diphtongues Les cercles vides o correspondent aux qualités vocaliques absentes du système des monophtongues. 53

Figure 8 – Synoptique d'un système de dialogue multilingue basé sur des systèmes parallèles. 64

Figure 9 - Synoptique d'un système de dialogue multilingue basé sur des modules multilingues et des modules dépendants de la langue. 65

Figure 10 - Système de traduction automatique basé sur la traduction dans une langue « de passage ». Exemple de la traduction d'un énoncé en langue L_I en langue L_J 66

Figure 11 – Synoptique d'un système de dialogue intégrant une traduction automatique en langue « de passage ».............................. 66

Figure 12 – Schéma bloc du système de ATT Bell Labs (d'après [Kadambe 94]) AN = anglais, ES = espagnol et MA = mandarin 81

Figure 13 – Synopsis du Système MPR (Mixed Phoneme Recognition) d'après [Kwan 95]. 83

Figure 14 – Schéma du système Parallel PRLM (d'après [Zissman 96]) 3 des 6 décodeurs phonétiques sont représentés ; pour chacun d'entre eux, 3 des 11 modèles phonotactiques sont indiqués. 84

Figure 15 – Modèle basé sur des connaissances, avec une éventuelle composante statistique. 98

Figure 16 – Modèles statistiques avec utilisation amont des connaissances linguistiques. 98
Figure 17 – Schéma d'un modèle acoustico-phonétique basé sur la modélisation phonétique différenciée. 100
Figure 18 – Exemple de système d'IAL basé sur la modélisation phonétique différenciée. Cas de deux langues A et B. 101
Figure 19 – Exemple de système d'IAL basé sur la modélisation différenciée des systèmes vocaliques. Cas de deux langues A et B. 104
Figure 20 – Schéma bloc du système de détection des segments vocaliques. 112
Figure 21 – Exemple de signal de la base de données OGI. Le locuteur prononce la phrase « cent quatre-vingt jours par an, il pleut ».. 113
Figure 22 – Résultat de la segmentation automatique (traits verticaux) sur le phrase « cent quatre-vingt jours… »... 114
Figure 23 – Prise en compte des effets de bord dans la décision parole/silence. Les traits verticaux pointillés correspondent aux frontières originelles du segment *i*. Les traits verticaux pleins correspondent au segment tronqué sur lequel la décision est prise. 116
Figure 24 – Résultat de la détection d'activité vocale sur la phrase « cent quatre-vingt jours… ». 117
Figure 25 – Exemple de segmentation et de détection parole/silence en milieu bruité. . 117
Figure 26 – Analyse spectrale en canaux de Mel pour le /a/ de *quatre* et le /ʃ/ de *chambre.* a) Fenêtre d'analyse du /a/. b) Décomposition de l'énergie du /a/ en canaux de Mel. c) Fenêtre d'analyse du /ʃ/. d) Décomposition de l'énergie du /ʃ/ en canaux de Mel. 121
Figure 27 – Détection des voyelles pour la phrase « cent quatre-vingt jours… ». a) Signal acoustique segmenté automatiquement (traits verticaux fins). b) Fonction *Rec* et localisation des sons identifiés comme étant des voyelles (traits verticaux épais). . 122
Figure 28 – Détection des voyelles pour la phrase « euh à proximité d'un … ». a) Signal acoustique segmenté automatiquement. b) Fonction *Rec* et localisation des sons identifiés comme étant des voyelles (traits verticaux épais). 122
Figure 29 – Récapitulatif des algorithmes mis en œuvre au cours de la détection des segments vocaliques. 123
Figure 30 – Ensemble des segments détectés par l'algorithme de détection de voyelles. 126
Figure 31 – Taux de détection des voyelles et des semi-voyelles. 127
Figure 32 – Répartition des segments par catégorie. 127

Figure 33 – Description des systèmes vocaliques des cinq langues traitées. La schématisation employée est celle de la typologie établie à partir d'UPSID. (d'après [Vallée 94]). 128

Figure 34 – Exemple d'étiquetage en zones d'un signal OGI. Les labels phonémiques sont posés manuellement et ne sont pas fournis avec OGI. 130

Figure 35 – Taux de détection des zones vocaliques. 131

Figure 36 – Composition des ensembles de segments détectés (taux de pureté). 132

Figure 37 – Répartition des erreurs de détection sur le corpus OGI. 133

Figure 38 – Schéma synoptique de la chaîne de paramétrisation cepstrale. 137

Figure 39 – Variation de la distorsion d'un ensemble de données en fonction du nombre de classes du modèle. 144

Figure 40 – Variation du critère *I(Q)* en fonction du nombre de classes du modèle. 145

Figure 41 – Répartition des stimuli du corpus JLVoc dans le plan principal dérivé de la modélisation cepstrale. 147

Figure 42 – Exemple de représentation sous forme de *boxplot ;* la voyelle /a/. 148

Figure 43 – Répartition des données du corpus APP selon les 8 MFCC. Les axes sont identiques sur chaque *boxplot* 149

Figure 44 – Schéma synoptique de la modélisation (traits épais) et de la classification des voyelles (trais fins). 151

Figure 45 – Répartition des segments vocaliques détectés dans le premier plan factoriel. 152

Figure 46 – Répartition dans le premier plan factoriel des segments vocaliques extraits des locuteurs masculins APP pour les cinq langues. 158

Figure 47 – Description du fonctionnement du système de référence. Le processus d'apprentissage est spécifié en flèches pleines, et le processus d'identification en flèches pointillées. 163

Figure 48 – Description du fonctionnement du système de discrimination des SV. Le processus d'apprentissage est spécifié en flèches pleines, et le processus d'identification en flèches pointillées. 166

Figure 49 – Comparaison des taux d'identification correcte entre les modèles de taille constante pour les 5 langues (LBG) donnant le meilleur score et les modèles de taille variable (LBG-Rissanen). 173

Figure 50 – Schéma de la fusion statistique de décisions issues de deux modèles A et B. 180

Figure 51 – Comparaison des taux d'identification correcte entre les modèles de taille constante (algorithme LBG) donnant le meilleur score et les modèles de taille variable (algorithmes LBG-Rissanen ou LBG-log-Rissanen). 189

Figure 52 – Comparaison des taux d'identification correcte entre les modèles de taille constante (algorithme LBG) donnant le meilleur score et les modèles de taille variable (algorithmes LBG-Rissanen ou LBG-log-Rissanen). 193

Figure 53 – Bilan des expériences réalisées avec les modèles différencié et global sur le corpus DEV_ST. 194

INDEX DES TABLEAUX

Tableau 1 : Les 12 familles mondiales identifiées par les synthétiste s (d'après [Ruhlen 97]). 26

Tableau 2 – Synthèse des indices perceptifs cités par les auditeurs [Muthusamy 94a]. .. 34

Tableau 3 – Principales articulations supplémentaires (d'après [Vallée 94]) 43

Tableau 4 – Répartition des langues d'UPSID par famille, en nombre et en pourcentage (d'après [Vallée 94]). 50

Tableau 5 – Principaux corpus multilingues disponibles 76

Tableau 6 – Récapitulatif des études en IAL citées (*n. p.* indique *non précisé* , MMC Modèle de Markov Caché et RN Réseau Neuromimétique) 88

Tableau 7 – Description des différents corpus utilisés. * : les corp us comprennent plusieurs locuteurs mais nous n'en avons utilisé qu'un. 125

Tableau 8 – Répartition des phonèmes vocaliques et semi-vocaliques dans le corpus. .. 125

Tableau 9 – Description des classes majeures OGI. 129

Tableau 10 – Résultats de la détection de segments vocaliques. 131

Tableau 11 – Nombre de classes déterminé par LBG-Rissanen en fonction du nombre de qualités vocaliques du corpus. 150

Tableau 12 – Résultat de la classification des voyelles. 151

Tableau 13 – Répartition des voyelles détectées dans les 5 classes obtenues par LBG-Rissanen. 154

Tableau 14 – Répartition en pourcentage des voyelles détectées dans les cinq classes. 154

Tableau 15 – Répartition des voyelles détectées dans les 12 classes obtenues par LBG. ... 155

Tableau 16 – Répartition des locuteurs du corpus OGI MLTS. ... 156

Tableau 17 – Description des ensembles de test DEV_ST et DEV_10. Les durées sont indiquées en secondes. ... 162

Tableau 18 – Taux d'identification correcte obtenus sur le corpus DEV_ST avec le modèle global segmental de 20 lois gaussiennes. ... 164

Tableau 19 – Taux d'identification correcte obtenus sur le corpus DEV_ST avec le modèle global segmental en faisant varier la taille des modèles gaussiens. ... 164

Tableau 20 – Matrice de confusion obtenue sur le corpus DEV_ST avec 17 paramètres (8 MFCC + 8 ΔMFCC + D) et des modèles à 50 composantes. ... 165

Tableau 21 – Influence du type de décision (hypothèse W *inner-Take-All* WTA ou vraisemblance multigaussienne). Le vecteur de paramètres est 8 MFCC + D. ... 167

Tableau 22 – Influence du type de décision (hypothèse WTA ou vraisemblance multigaussienne). Le vecteur de paramètres est 8 MFCC E + 8 ΔMFCC + ΔE + D. 168

Tableau 23 – Matrice de confusion – (DEV_ST / 19 coefficients / 20 gaussiennes). ... 168

Tableau 24 – Pourcentage d'identification correcte sur le corpus DEV_ST dans l'hypothèse WTA. Le meilleur score obtenu pour chaque vecteur de paramètres est affiché en gras. ... 169

Tableau 25 – Nombre de gaussiennes déterminés par algorithme LBG-Rissanen. ... 172

Tableau 26 – Pourcentage d'identification correcte sur le corpus DEV_ST. Le nombre de composantes gaussiennes est fixé par LBG-Rissanen. ... 172

Tableau 27 – Matrice de confusion – (DEV_ST / 8 MFCC + D / LBG-Rissanen). ... 173

Tableau 28 – Nombre de gaussiennes déterminés par algorithme LBG-log-Rissanen. . 174

Tableau 29 – Pourcentage d'identification correcte sur le corpus DEV_ST. ... 174

Tableau 30 – Pourcentages d'identification correcte sur le corpus DEV_ST avec et sans normalisation par soustraction du biais calculé à l'apprentissage. ... 176

Tableau 31 – Pourcentages d'identification correcte sur le corpus DEV_ST avec et sans normalisation des scores par fonction discriminante. ... 177

Tableau 32 – Pourcentages d'identification correcte sur le corpus DEV_ST avec normalisation des scores par fonction discriminante et élagage des segments les moins discriminants. ... 178

Tableau 33 – Influence de la proportion d'élagage sur le score obtenu en identification avec le modèle #9 à 20 classes **sans** normalisation des scores. ... 178

Tableau 34 – Influence de la proportion d'élagage utilisé en conjonction avec la suppression de biais sur le modèle #9 à 20 classes **sans** normalisation des scores. 179

Tableau 35 – Scores d'identification obtenus en fusionnant les modèles LBG-Rissanen statiques et dynamiques. 181

Tableau 36 – Scores d'identification obtenus en fusionnant le modèle #9 amélioré avec différents modèles #2. 181

Tableau 37 – Scores d'identification obtenus en fusionnant le modèle #9 obtenu par LBG-log-Rissanen avec différents modèles #2. 181

Tableau 38 – Matrice de confusion obtenue sur le corpus DEV_ST par fusion des modèles #9 amélioré et #2 LBG-Rissanen. 182

Tableau 39 – Bilan des expérimentations en discrimination des SV sur les corpus masculins. 183

Tableau 40 – Pourcentage d'identification correcte sur le corpus DEV_ST dans l'hypothèse WTA. 187

Tableau 41 – Valeur moyenne (en secondes) des durées consonantiques et vocaliques sur les fichiers d'apprentissage de 45 secondes (APP_ST). 187

Tableau 42 – Nombre de segments consonantiques du corpus d'apprentissage et nombre de classes fixé par LBG-log-Rissanen dans l'espace d'observation #10. 188

Tableau 43 – Nombre de gaussiennes déterminé par algorithme LBG-Rissanen et LBG-log-Rissanen avec des corpus équilibrés. 188

Tableau 44 – Pourcentage d'identification correcte sur le corpus DEV_ST pour les modèles de taille variable. 189

Tableau 45 – Influence de la proportion d'élagage sur le score obtenu en identification avec les modèles #9 à 30 classes et #10 log-Rissanen. 189

Tableau 46 – Description des modèles employés en discrimination des systèmes consonantiques et vocaliques. 190

Tableau 47 – Taux d'identification correcte obtenu par fusion des scores consonantiques et vocaliques. 191

Tableau 48 – Nombre de composantes gaussiennes fixé par LBG-Rissanen (paramètres #2) et LBG-log-Rissanen (paramètres #9 et #10) pour la modélisation globale segmentale. 192

Tableau 49 – Influence de la proportion d'élagage sur le score obtenu en identification avec la modélisation globale. 193

Tableau 50 – Taux d'identification obtenus par fusion du modèle global et des modèles issus de la modélisation différenciée. 194

Tableau 51 – Matrice de confusion obtenue sur le corpus DEV_ST par fusion du modèle vocalique #9 initialisé par LBG-log-Rissanen avec le modèle global #9 appris par LBG (50 classes). 195

Tableau 52 – Matrice de confusion obtenue sur le corpus DEV_10 par fusion du modèle vocalique #9 initialisé par LBG-log-Rissanen avec le modèle global #9 appris par LBG (50 classes). 196

Tableau 53 – Taux d'identification correcte obtenus avec les locuteurs féminins et masculins et les modèles appris sur les locuteurs masculins seuls. 196

Tableau 54 – Taux d'identification correcte obtenus en discrimination FR-L sur les corpus DEV_10 et DEV_ST. 197

Tableau 55 – Taux d'identification correcte obtenus en identification de quatre langues en ensemble fermé sur les corpus DEV_10 et DEV_ST. 197

(1713) - © Heritage Map Museum

Introduction Générale

Le traitement automatique du langage parlé évoque, en tout premier lieu, la reconnaissance et la synthèse de la parole. Ces thèmes sont certainement les plus médiatisés même si la communauté des chercheurs en parole aborde une grande diversité d'activités allant de l'étude acoustique de la voix à la recherche des processus cognitifs d'acquisition du langage pour ne citer que ceux là. Bien évidemment, le grand public s'intéresse principalement au côté applicatif des recherches, ignorant généralement leur caractère pluridisciplinaire. Les sciences dites appliquées s'appuient à la fois sur la technologie et sur la recherche entreprise à un niveau plus fondamental ; que l'on retire l'un de ces deux fondements et l'édifice s'écroule… Les études menées par les linguistes et les ingénieurs durant des décennies ont permis, au cours des années 80, l'émergence de technologies vocales efficaces en laboratoire avant d'aboutir plus récemment au développement d'applications « grand public » grâce à l'explosion de la micro-informatique. De nos jours, les technologies vocales sont répandues, et déjà, la reconnaissance et la synthèse de la parole constituent une réalité quotidienne dépassant les frontières de notre simple communauté scientifique. Durant cette même période, d'autres thèmes de recherche ont quitté le domaine de l'utopie rêvée par quelques savants pour rejoindre celui de perspectives envisageables. L'Identification Automatique des Langues (IAL) est de ceux-là.

Depuis le début des années 90, reconnaître la langue parlée par un locuteur inconnu à partir de quelques secondes de parole n'est plus une chimère, et l'IAL s'impose comme un des enjeux majeurs du traitement de la parole au cours des décennies à venir. L'internationalisation des communications (Internet en est le plus parfait exemple) et l'accroissement de la demande pour des applications vocales interactives se conjuguant, la nécessité de développer des applications capables de traiter plusieurs langues devient manifeste. Les ingénieurs ont anticipé ce besoin en développant à partir des systèmes de reconnaissance de la parole et d'identification du locuteur les premiers systèmes opérationnels à l'aube des années 90. Ce mouvement, initié aux Etats-Unis s'est amplifié depuis, et, à ce jour, nombre de chercheurs se consacrent à ce sujet. En France, les recherches n'ont été que timidement amorcées, puisque à notre connaissance, les seuls travaux ayant précédé les nôtres en IAL sont ceux initiés au LIMSI en 1993 [Lamel 93].

C'est dans ce cadre un peu inexploré par les laboratoires français que notre étude a débuté. Fort heureusement, si la communauté ingénieuse [1] française se consacrait à d'autres thèmes, il n'en était pas de même pour la communauté des linguistes. De la phonologie à la linguistique comparative, nombre d'aspects complémentaires de la caractérisation des langues avaient déjà été longuement étudiés. En particulier, l'automne 1994 vit publier la thèse de Nathalie Vallée proposant une typologie des systèmes vocaliques de la base de données UPSID (UCLA Phonological Segmental Inventory Database) [Maddieson 84]. Cette thèse propose une typologie des langues originale puisque basée sur la substance phonétique des systèmes vocaliques et non sur des critères de proximité géographique ou morpho-syntaxique [Vallée 94]. L'idée de caractériser les langues et *a fortiori* de les identifier automatiquement à partir de la substance acoustico-phonétique de leur système vocalique a donc vu le jour dans l'esprit de Régine André-Obrecht et de Louis-Jean Boë. Cette approche originale offrait entre autres la particularité de ne pas nécessiter d'enregistrements étiquetés puisqu'elle s'appuyait sur une détection non supervisée des voyelles. Ainsi, il devenait possible de modéliser des langues pour lesquelles de tels corpus ne sont pas disponibles, soit l'écrasante majorité des quelques 5000 langues parlées dans le monde. Ce projet serait peut-être resté à l'étape d'esquisse surréaliste si la Délégation Générale pour l'Armement (DGA) n'avait pas lancé un appel d'offre au titre évocateur de « Discrimination Multilingue Automatique » et si l'IRIT, en même temps que l'ICP, le DDL et l'ILPGA n'avaient pas été sélectionnés. Par ailleurs, la même DGA assurait mon financement par le biais d'une allocation de recherche DGA/CNRS à partir du mois de septembre 95. Le présent manuscrit présente donc le bilan des recherches menées depuis trois ans dans ce cadre.

A notre connaissance, le rapport technique issu du contrat passé entre le CNET et le LIMSI en vue de la collecte d'un corpus multilingue [Gauvain 94] est le seul document édité en français sur l'IAL [2]. Le présent manuscrit a donc pour ambition de constituer l'introduction la plus complète possible au domaine de l'IAL et nous avons tenu à proposer un éclairage varié sur ce thème. Cela nous a parfois conduit à explorer des domaines connexes qui peuvent sembler relever de la culture plutôt que de la recherche en IAL, mais sans lesquels le panorama ne serait pas complet. De plus, il nous semble nécessaire de garder à l'esprit que le signal de parole est plus qu'une simple variation mécanique de pression acoustique et qu'il est avant tout le média privilégié de la communication humaine [3].

Ce manuscrit s'articule en trois parties elles-mêmes divisées en chapitres. Au cours de la première partie, intitulée « Des hommes et des langues », nous abordons plusieurs aspects complémentaires du langage parlé. Le premier chapitre est consacré à des aspects linguistiques et cognitifs des langues ; nous nous interrogeons sur

[1] Ce terme caractérise bien évidemment la communauté des ingénieurs et des informaticiens.

[2] En ne tenant pas compte des quelques communications récentes lors de congrès francophones.

[3] Et ce même si le mèl prend le pas sur le téléphone dans certains cas.

l'émergence du langage chez nos ancêtres primates et sur les relations que les langues entretiennent entre elles, évoluant dans le temps et dans l'espace comme la plupart des processus humains. Ces considérations nous permettent d'amorcer une réponse à la question : « qu'est-ce qu'une langue ? » en établissant un ensemble de traits qui, s'ils ne constituent pas une définition, permettent cependant d'appréhender la réalité des langages. Les expériences perceptives relatées en fin de chapitre confirment par ailleurs que les traits saillants diffèrent selon les langues et que l'homme est, dans une certaine mesure, capable de différencier une langue d'une autre. Le second et dernier chapitre de cette première partie introduit l'aspect phonologique de notre étude en présentant les systèmes vocaliques (SV). Après une description acoustique des voyelles, nous nous intéressons aux travaux entrepris par les linguistes sur les modélisations et les descriptions des SV des langues du monde, et particulièrement sur leur prédiction. Ce chapitre s'appuie notamment sur la thèse de Nathalie Vallée et nous reprenons plusieurs de ses conclusions sur les tendances universelles régissant l'organisation des SV des langues du monde.

La seconde partie de ce manuscrit, intitulée « Des ordinateurs et des langues » replonge le lecteur dans l'univers des technologies vocales et de l'IAL. Un premier chapitre précisera le cadre des applications multilingues avant de proposer une introduction à l'IAL, en particulier une description des enjeux de ce domaine. Le second chapitre constitue un état de l'art de l'IAL ; nous abordons le thème essentiel des corpus de données, sans lesquels aucun système automatique n'existerait, avant de décrire la plupart des systèmes d'IAL des années 90. Ce chapitre se conclut par une discussion sur les performances atteintes à l'heure actuelle. Cette discussion rebondit d'ailleurs au chapitre suivant, consacré à la présentation de l'approche que nous proposons, appelée la modélisation phonétique différenciée. Ce troisième chapitre débute par une réflexion sur les perspectives du domaine de l'IAL, où beaucoup reste à faire, même si certains semblent vouloir se satisfaire des performances atteintes à ce jour. Certaines considérations sur les méthodes appliquées actuellement nous permettent d'introduire la modélisation phonétique différenciée, qui, à notre avis, fournit un cadre théorique adéquat pour de nombreuses recherches mêlant linguistique et informatique. La première de ces études est bien évidemment présentée plus en détail puisqu'il s'agit de la modélisation différenciée des systèmes vocaliques.

Si les deux premières parties de ce manuscrit sont consacrées à des considérations plus ou moins théoriques, la troisième partie, intitulée « La modélisation des systèmes vocaliques appliquée à l'IAL », présente les expériences menées pour valider notre approche. Au cours du premier chapitre, nous étudions les algorithmes développés pour extraire automatiquement les voyelles présentes dans le signal, et cela sans apprentissage, supervisé ou non. Plusieurs expériences permettent d'évaluer la qualité du détecteur réalisé. Ces expériences sont réalisées en environnement varié, s'étendant de la parole lue en français à des enregistrements téléphoniques de parole spontanée multilingue (corpus OGI MLTS). Le second chapitre s'intéresse à la modélisation acoustique des SV. Dans un premier temps, nous étudions la représentation paramétrique adoptée pour les voyelles avant de développer les modèles statistiques employés pour modéliser les SV. Des expériences, là encore réalisées sur plusieurs corpus, permettent d'évaluer les méthodes et les algorithmes implantés. Le troisième chapitre présente les expériences réalisées en IAL par discrimination automatique des SV à partir des modèles présentés précédemment. Nous commençons par indiquer les résultats obtenus avec un système de référence basé sur une modélisation multigaussienne segmentale de l'ensemble des sons de chaque langue avant de nous consacrer aux expériences en discrimination des SV par le biais d'une tâche d'identification de cinq langues (coréen, espagnol, japonais, français et vietnamien) issues du corpus OGI MLTS. Au cours du quatrième et dernier chapitre, nous envisageons d'étendre le système d'IAL en intégrant une modélisation de l'ensemble des consonnes de chaque langue et en fusionnant les résultats obtenus en discriminations vocalique et consonantique. Ce système d'identification phonétique basé sur la modélisation différenciée des voyelles et des consonnes est alors évalué de manière approfondie par le biais de plusieurs tâches d'identification dérivées de l'identification des cinq langues déjà citées.

Le lecteur verra alors se profiler la conclusion de ce manuscrit tout en ayant, nous l'espérons, fait une agréable lecture.

(1714) - © Heritage Map Museum

1ère Partie

Des hommes et des langues

INTRODUCTION

A partir du signal acoustique de parole, on peut s'intéresser à la reconnaissance des sons produits, à la compréhension de l'énoncé ou encore à l'identité physique du locuteur. On peut bien évidemment chercher aussi à identifier la langue dans laquelle l'énoncé a été produit. On se heurte alors à une difficulté spécifique quant à la définition du problème : en effet, qu'est-ce que la langue ? Quels sont les facteurs qui permettent de considérer qu'une langue est différente d'une autre ? A l'évidence, il s'agit là d'une question qui est au cœur du débat linguistique, même si l'on peut considérer qu'elle est en marge des préoccupations des ingénieurs chargés de concevoir des systèmes d'identification automatique de la langue. A notre avis, il reste cependant essentiel d'appréhender ce thème de recherche en possédant une vision globale, allant de la diversité des langues du monde à leur caractérisation. Nous proposons dans cette partie plusieurs éclairages sur les langues du monde, au travers d'aspects linguistiques, cognitifs ou phonologiques.

Le premier chapitre présente une introduction au langage et aux langues humaines. Au travers d'une brève histoire de la linguistique (principalement de la linguistique comparative), nous espérons permettre au lecteur d'appréhender la dualité du langage, tenant à la fois d'un processus humain universel et d'un moyen d'expression où la diversité est reine. Nous introduisons aussi les différents niveaux d'analyse intervenant dans la caractérisation des langues par les experts avant de nous interroger finalement sur la capacité, pour chaque être humain, d'opérer des distinctions entre langues. Une brève présentation des expériences perceptives menées sur ce sujet nous paraît en effet constituer un bon préambule à l'identification *automatique* des langues.

Le second chapitre est consacré à un aspect plus restreint de la diversité du langage : la caractérisation des systèmes vocaliques. Dans notre optique d'identifier les langues à partir des voyelles détectées automatiquement, les travaux menés en phonologie sur les typologies vocaliques fournissent un cadre théorique incontournable. Chaque langue possède un système vocalique qui recouvre à la fois la notion de structure phonologique et de substance acoustico-phonétique, et les travaux typologiques récents permettent d'envisager leur application à l'identification automatique des langues.

Chapitre 1

DIVERSITE ET CARACTERISATION DES LANGUES

De nos jours, plusieurs milliers d'idiomes sont parlés à travers le monde. Les estimations vont de 4000 à 6000 langues distinctes, auxquelles il faut ajouter les variantes liées aux dialectes et aux différents parlers. La communauté linguistique ne reconnaît d'ailleurs pas unanimement les mêmes frontières entre dialecte et langue. En fait, il est inévitable de considérer les langues dans un contexte dynamique. En effet, si l'on accepte l'idée que le langage est propre à l'homme (nous reviendrons sur cette affirmation au paragraphe 1 de ce chapitre), il est régi, comme la plupart des traits humains par les règles de l'évolution. Ainsi, dialectes et langues évoluent, à la fois de manière endogène (apparition de nouveaux concepts, applications de règles phonétiques…), et exogène (influence des autres langues). Comme le remarquait non sans humour Antoine Meillet au début de ce siècle, l'empire romain est à l'origine d'une communauté d'individus géographiquement étendue sur des milliers de kilomètres et tous convaincus de parler le latin : au bilan, la diversité des dialectes et parlers a donné naissance à toute la famille de langues romanes.

Cette notion de famille est particulièrement importante en linguistique, principalement depuis les travaux des premiers philologues du XVI^{ème} siècle. Si, dès l'antiquité, plusieurs auteurs avaient relevé une affiliation entre certaines langues, nous verrons au paragraphe 2 que bien des siècles se sont écoulés avant que la recherche des familles de langues redevienne l'une des principales activités des linguistes. Cette démarche est à l'origine d'une meilleure compréhension de notre patrimoine linguistique, et, à l'heure actuelle, plusieurs théories dépassent le seul cadre de la linguistique pour s'intéresser plus globalement à l'histoire de l'Homme et à l'émergence du langage tout autant qu'à la diversité des langues.

Nous verrons enfin que l'étude de la parenté et de la diversité des langues a comme corollaire la recherche d'éléments les caractérisant (paragraphe 3) . A l'heure actuelle, les classifications reposent encore principalement sur des descriptions morpho-syntaxiques et lexicales, mais il est vraisemblable que les typologies des structures sonores établies ces dernières années offrent une alternative aux linguistes, en particulier grâce à l'émergence de tendances universelles qui peuvent, comme le supposait Nicolas S. Troubetzkoy dans les années 30, justifier à elles seules certaines ressemblances entre langues.

1 LES LANGUES, INSTRUMENTS DU LANGAGE HUMAIN

1.1 Le langage : un média de communication humain

Le langage est assurément un moyen de communication efficace. S'il n'est pas toujours évident d'énoncer clairement ce que l'on conçoit, il est encore moins facile de partager un concept sans l'énoncer ! Dès lors que l'on s'intéresse au langage, plusieurs aspects peuvent être abordés, de l'apparition historique (§ 1.2) du langage à l'acquisition par l'homme de sa langue maternelle (§ 1.3). S'il est admis que les animaux peuvent communiquer, parfois même de façon brillante, des expériences poussées menées sur des chimpanzés ont montré que, malgré leur intelligence, nos cousins éloignés n'arrivaient pas à élaborer un langage semblable à celui de l'homme… les œuvres complètes de Shakespeare ne seront sans doute jamais réécrites par un singe. Il semble donc que la faculté du langage ne soit pas parfaitement corrélée à la compréhension ou à l'intelligence, comme le note Jacques Mehler [Mehler 95]. On nomme généralement cette capacité humaine du langage *la double articulation* : « Les énoncés s'articulent en mots [et] les mots s'articulent en sons » comme le résume André Martinet [Martinet 68]. L'être humain peut produire une infinité de phrases à partir d'un nombre fini d'unités porteuses de sens (les monèmes [4]), et chacune de ces unités est elle même articulée à partir de quelques dizaines d'unités phonétiques distinctives, les phonèmes. C'est cette double articulation qui permet d'atteindre une richesse de communication unique. Le rôle du langage, vecteur de sens, semble donc être double car il permet à la fois de communiquer avec l'extérieur et de structurer une représentation interne du monde. L'un de ses aspects les plus frappants est la dualité entre universalité et diversité. En effet, il n'existe pas de prédisposition particulière à apprendre une langue plutôt qu'une autre, et par exemple, Derek Bickerton a montré que les langues créoles, bien que réparties à travers le monde, présentaient d'importantes similitudes grammaticales [Bickerton 95]. A cela s'ajoute le fait que chaque homme, possède à sa naissance les mêmes capacités de production de la parole, même si, pour les sourds, le langage se développera plutôt à partir de codes gestuels plutôt que sonores. A cet aspect universel, sur lequel nous aurons l'occasion de revenir, s'oppose une diversité de représentation considérable. Elle s'exprime à plusieurs niveaux, de la phonétique à la morpho-syntaxe. Cette capacité à développer des langages distincts, faisant références à des systèmes de concepts spécifiques, et à traduire ces concepts d'une langue à une autre sont fondamentalement humains. Pour communiquer par le langage, l'homme est capable de développer des stratégies adaptées à ses besoins. Nous avons évoqué les languages des signes qui représentent un cas extrême d'adaptation, mais on pourrait citer aussi l'écriture ou toutes les gammes de moyens de transmission acoustique, du langage

[4] Un monème est différent d'un mot ; en effet, un monème peut être réparti sur plusieurs mots (monème du pluriel dans une phrase), et inversement, à un mot peut correspondre plusieurs monèmes : le mot autoroute est constitué des deux monèmes auto- et –route [Martinet 68].

tambouriné au morse en passant bien entendu par la parole qui demeure le canal le plus utilisé.

1.2 L'émergence du langage

Il est difficile de s'intéresser à la manière dont le langage est apparu sans s'intéresser aux origines de l'homme. En effet il est clair, à moins d'intégrer des considérations métaphysiques, qu'à un moment dans la longue chaîne de l'évolution, un de nos ancêtres a acquis la faculté du langage. Il est quasiment certain que cette capacité est apparue en deux temps. L'apparition d'un code liant des gestes – visuels ou acoustiques – à des concepts sous la forme d'un "proto-langage" a sans doute précédé l'apparition de la double articulation. Langaney estime même que les primates vivant avant la séparation de l'homme et du chimpanzé, il y a plus de 4 millions d'années, disposaient de moyens de communication codés assez évolués [Langaney 97]. L'émergence du proto-langage semble nécessairement liée au développement du cerveau et à la notion de représentation abstraite du monde selon Bickerton. Tout comme Noam Chomsky [Chomsky 68], il penche pour une modification génétique rapide, du type macromutation accidentelle.

Ce scénario ne satisfait guère Steven Pinker, un élève de Chomsky pour qui l'apparition du langage ne peut qu'être le fait d'un processus fondamentalement évolutionniste qui relève donc de la sélection naturelle [Pinker 94]. Philip Lieberman souligne que l'appareil vocal de l'homme a évolué pour aboutir à un conduit vocal où le larynx est abaissé par rapport à celui de nos cousins chimpanzés (Figure 1). Cette configuration particulière a abouti selon lui à un système de communication qui minimise les erreurs de transmissions et qui maximise l'information véhiculée [Lieberman 91]. Par contre, cette modification morphologique fait de l'homme le seul mammifère ne pouvant pas respirer et manger en même temps. Cette limitation très forte (l'homme peut s'étouffer en mangeant) ne peut être en accord avec la sélection naturelle que si elle est compensée par un autre aspect qui pourrait bien être l'amélioration de la transmission des sons. Des études récentes réduisent pourtant l'importance de la morphologie du conduit vocal [Boë 98], montrant que celui de l'homme de Néandertal lui permettait déjà d'accéder à un espace vocalique maximal proche de celui de l'homme moderne. Cela ne signifie pas que nos ancêtres utilisaient cette capacité d'expression, mais tout au moins qu'il peut être nécessaire de rechercher des indices à un autre niveau, par exemple sur le développement cortical et sur l'innervation des articulateurs.

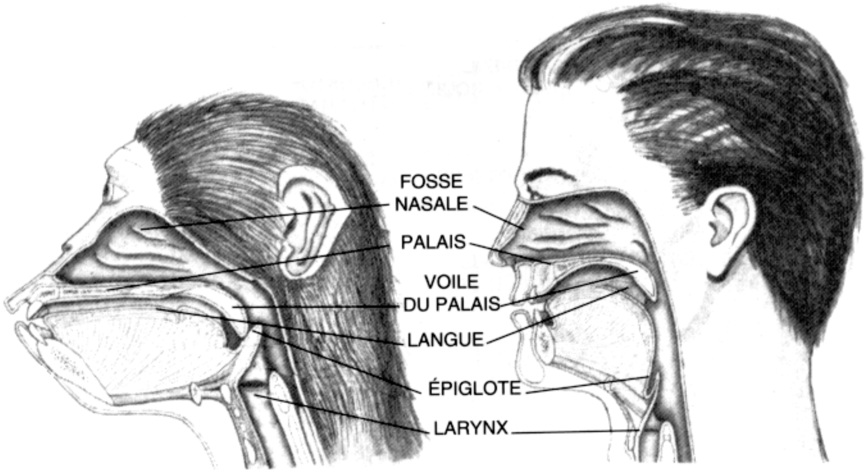

Figure 1 : Chez le chimpanzé, l'épiglotte forme avec le voile du palais une cloison étanche ; les singes peuvent donc manger et respirer simultanément. Chez l'homme, l'abaissement du larynx rend cette faculté impossible (d'après [Ross 97]).

Le passage du proto-langage au langage proprement dit pose d'autres problèmes de datations. Contrairement à ce que l'on pensait encore il y a à peine quinze ans, il est très peu probable que l'homme n'ait acquis la double articulation que très récemment (il y a moins de 10000 ans) et de manière simultanée en plusieurs endroits différents. Bien sûr, on peut trouver certaines personnes qui considèrent que le langage est apparu à Sumer [Halloran 97] il y a 5000 ans ou que le basque (ou plus exactement le saharan, nom qu'Edo Nyland donne à la langue mère du basque [Nyland 97]) est la première langue de l'humanité, parlée il y a 7000 ans…

Actuellement, les scientifiques penchent plutôt pour une origine antérieure du langage [Deacon 97] et la plupart des linguistes et des archéologues considèrent que son apparition est un événement si important qu'il a forcément était accompagné de changements importants à l'échelle humaine.

La génétique, la paléontologie et l'archéologie ont permis ces dernières années de retracer les grandes étapes de l'histoire des primates et des hominidés. La thèse actuelle est celle du goulot d'étranglement (Figure 2). Il y a plus d'un million d'années, l'un de nos ancêtres, *homo habilis* (ou *homo erectus* , son descendant), aurait colonisé la planète, de l'Europe à l'Asie. Il s'agit là d'une première dispersion géographique au cours de laquelle de petits groupes auraient évolué pour donner naissance à ceux que l'on nomme les *homo sapiens* archaïques (homme de Néandertal…). Parmi ces différentes populations, une branche a donné naissance à la lignée des *homo sapiens* modernes, à laquelle nous appartenons. Les faits tendent à prouver que, pendant plusieurs dizaines de milliers d'années, les communautés d'hominidés archaïques et modernes ont co-existés,avant que les lignées d' *homo sapiens* archaïques ne s'éteignent. La Terre n'aurait plus alors été peuplée que par des homo sapiens modernes vivant en Afrique. Leurs descendants auraient à leur tour colonisé la planète (seconde dispersion géographique) et donné naissance à toutes les communautés d' *homo sapiens* modernes.

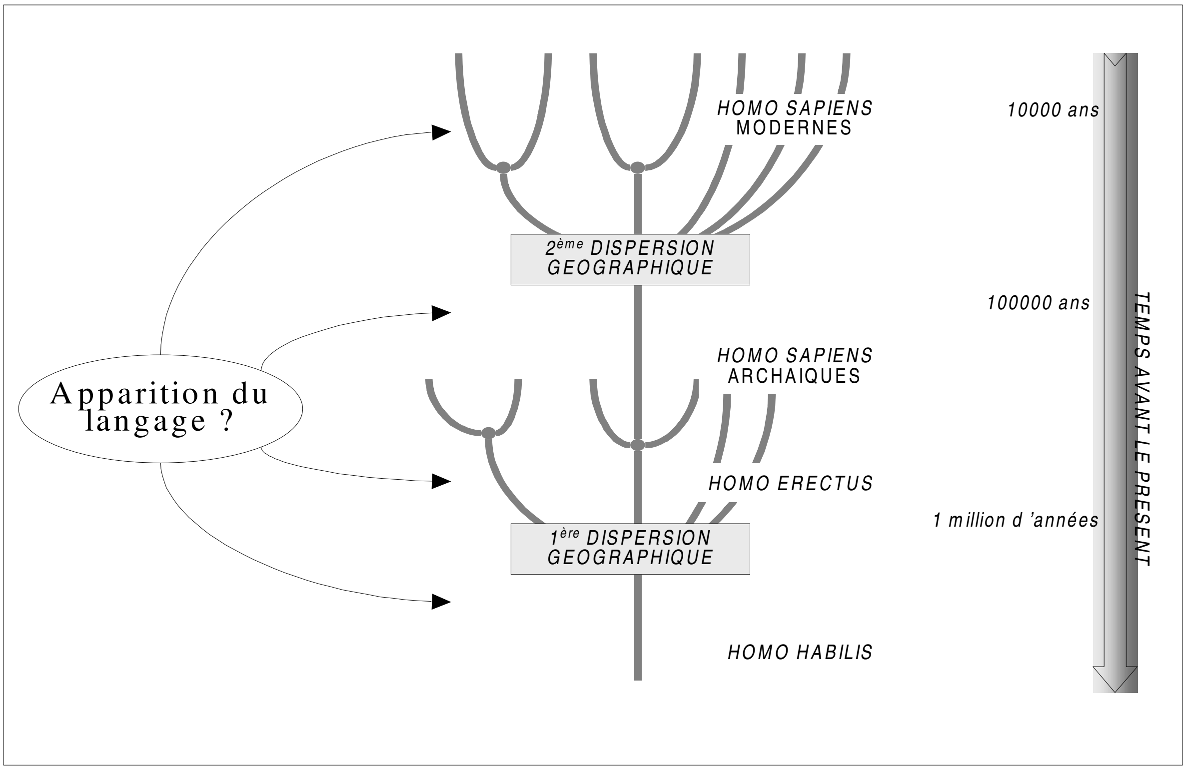

Figure 2 : La théorie du « goulot d'étranglement » (d'après [Victorri 97]).

Cette hypothèse que certains paléoanthropologues, partisans d'une thèse dite multirégionale réfutent, est confortée par le fait que l'ADN de l'homme de Néandertal est bien plus éloigné du notre, européen, que celui de n'importe quel homme vivant à l'heure actuelle sur Terre. Puisqu'il est peu probable que le langage soit apparu en plusieurs endroits simultanément, son émergence est sans doute antérieure à la seconde dispersion géographique ; il est soit apparu dans plusieurs groupes d' *homo sapiens* archaïques, soit dans une seule branche à la même époque. Une apparition encore antérieure est également envisageable si l'on minimise l'importance relative de la forme du conduit vocal dans le processus d'émergence du langage [Boë 98, Deacon 97]. Une autre hypothèse séduisante est actuellement défendue par le cogniticien Bernard Victorri : l'apparition de la double articulation dans une seule des branches de la population d' *homo sapiens* archaïques lui aurait donné un avantage décisif sur les autres populations disposant seulement d'un proto-langage. Si l'on suppose comme Victorri que les populations d' *homo sapiens* archaïques aient dû faire face à un dérèglement social, il est possible que l'apparition du langage doublement articulé ait permis à nos ancêtres d'évoluer et d'aboutir à une régulation sociale leur évitant le destin fatal des autres groupes d'hominidés. Une autre hypothèse, peut-être plus conforme à ce que l'on sait de la nature humaine est que les *homo* possédant la capacité du langage se soient imposés par la force à leurs cousins moins évolués.

Ruhlen propose quant à lui une autre hypothèse faisant coïncider l'apparition de la double articulation avec un événement historique majeur situé il y a quelques 40000 ans [Ruhlen 98]. En effet, après avoir utilisé durant plusieurs millions d'années des pierres taillées assez vulgairement comme outils, notre ancêtre se mit à produire à cette période des outils beaucoup plus évolués comme de fines aiguilles ou des harpons. Ce bond technologique pourrait, selon Ruhlen, être la conséquence de l'apparition d'une nouvelle forme de communication totalement nouvelle : le langage humain.

A l'heure actuelle, il n'est pas possible de trancher pour l'une ou l'autre des hypothèses présentées succinctement ici, et si certaines d'entre elles lèvent un coin du voile, le mystère reste entier sur l'apparition des capacités cognitives propices à l'émergence du langage.

1.3 L'acquisition du langage : un programme génétique ?

Si l'on sait depuis 1865, grâce à Paul Broca, que les centres cérébraux du langage sont principalement localisés dans l'hémisphère gauche du cerveau (chez quasiment tous les droitiers et chez deux tiers des gauchers), le processus physiologique d'acquisition du langage reste mystérieux. L'utilisation depuis 20 ans de techniques d'imagerie performantes (scanner, puis tomographie par émission de positons et imagerie par résonance magnétique) a affiné la localisation des structures neuronales impliquées dans la communication langagière ; selon Antonio Damasio, trois zones interviennent dans le processus [Damasio 97]. Une zone, répartie dans les deux hémisphères du cerveau, prend en charge la représentation du monde et sa conceptualisation. Une seconde zone, localisée dans les aires de Broca et de Wernicke, assure la représentation des règles rythmiques, syntaxiques et grammaticales, ainsi que la mémorisation des unités de base du langage, phonèmes ou gestes. Une troisième zone, localisée elle aussi dans l'hémisphère gauche, assurerait la médiation et la coordination entre concepts et mots. Plusieurs chercheurs pensent quant à eux que les structures du langage sont moins ponctuelles mais plutôt réparties en réseau traitant l'information en parallèle. Cette hypothèse est en partie basée sur le fait que les aphasies (troubles du langage consécutifs à un accident) ont tendance à s'atténuer avec le temps alors que les lésions cérébrales persistent. Cela laisse supposer que les réseaux traitant le langage peuvent au moins partiellement se réorganiser, même si une zone est détruite.

L'amélioration de notre connaissance neurologique ne permet cependant pas de comprendre le processus d'acquisition du langage. Chomsky a proposé dès les années 1950 une approche cognitive selon laquelle un module spécialisé du cerveau est codé génétiquement sous la forme d'une grammaire universelle, capable d'appréhender toute la diversité des langues humaines. Cette hypothèse, si elle a le mérite de justifier le fait que l'homme puisse apprendre à parler (il s'agit d'une capacité innée), est peut-être un peu exagérée. En effet, nos connaissances actuelles du développement du cerveau montrent plutôt que les tâches sont effectuées par des réseaux dynamiques, qui se sont développés par stimulation, principalement avant l'adolescence. Le cas des "enfants sauvages", ces enfants qui n'ont pas passé leurs premières années dans une société humaine est révélateur ; même lorsqu'ils retournent à la civilisation, ils n'arrivent pas à développer un langage plus sophistiqué que celui d'un enfant normal de deux ans. Cela semble indiquer que si le cerveau n'est pas stimulé de manière précoce en vue d'acquérir des structures neuronales prenant en charge la syntaxe, la deuxième articulation du langage humain, le cerveau perd cette capacité.

Mehler considère, à l'instar de Chomsky, que l'acquisition du système linguistique repose sur des structures génétiques indépendantes des capacités auditive et phonique de la personne, mais que ces capacités sont très rapidement affectées par l'environnement dans lequel baigne le bébé [Mehler 95]. L'acquisition du langage repose donc sur trois paramètres :

- ✓ l'information sensorielle, propre à chaque individu,
- ✓ des structures cognitives spécifiques, propres à l'être humain,
- ✓ et les contraintes formelles du langage, propres à l'être humain.

Dès lors, il ne s'agit pas d'un apprentissage par mise en corrélation de stimuli auditifs et visuels, mais d'un mécanisme plus complexe qui différencie l'homme de l'animal.

De nombreuses expériences cognitives ont porté sur les contrastes auxquels les bébés sont sensibles. Jakobson a montré que l'alternance des consonnes et des voyelles semblait primordiale. D'autres études, citées dans [Mehler 95], ont montré l'importance de l'intonation. La théorie développée par l'auteur est qualifiée d'apprentissage par l'oubli. En effet, les expériences cognitives ont montré que les bébés sont parfaitement sensibles à tous les contrastes acoustiques, que ce soit le lieu ou le mode d'articulation, ou encore les traits secondaires des voyelles (la nasalisation par exemple). Cette faculté disparaît vers l'âge de 12 mois, et le bébé n'est plus alors capable que de distinguer les contrastes pertinents dans sa langue (un bébé français ne fera pas de différence entre un /p/ et un /p ^{h}/, prononcé à l'anglaise et un bébé japonais deviendra insensible à la différence entre /l/ et /r/). Par contre, le bébé reste, tout comme l'adulte, capable de percevoir des contrastes portant sur des sons ne partageant aucune dimension articulatoire avec les sons de sa langue maternelle, comme les clicks par exemple pour des enfants français ou anglais. Cela suggère que le phénomène n'est pas lié à une perte sensorielle de la perception, mais plutôt à une réorganisation de la structure cognitive qui représente les sons en fonction de la langue apprise. Il est aussi intéressant de noter que le bébé plongé dans un environnement multilingue arrive à ne pas mélanger les deux langues et à "apprendre" plusieurs systèmes phonétiques. Il semble donc que dès son plus jeune âge l'enfant est capable, grâce à des indices sans doute phonétiques et prosodiques, d'identifier si deux énoncés sont prononcés dans la même langue ou non. Nous reviendrons sur cet aspect de l'identification des langues au paragraphe 4 du présent chapitre.

2 UNE BREVE HISTOIRE DES LANGUES

2.1 Vie et mort des langues : un état des lieux

A ce jour, plus de 5000 langues sont parlées à travers le monde. La diversité linguistique n'est pas uniformément répartie puisque à peine 4 % d'entre elles sont parlées en Europe et au Moyen-Orient, alors que 15 % viennent des tribus autochtones d'Amérique et que l'Afrique et l'Asie représentent les 84 autres pour cent [Krauss 92]. 200 à 250 langues sont parlées par plus d'un million de personnes, tandis que la moitié d'entre elles recensent moins de 6000 locuteurs. A notre époque de mondialisation de la communication, la pression est forte d'adopter des langages partagés par des communautés importantes. Cela se fait généralement aux dépends d'idiomes moins répandus et le constat est sans appel : d'ici un siècle, 70 à 90 pour cent des langues auront disparu. 90 % des langues aborigènes australiennes périssent déjà alors que dans le Nord de la Russie 27 langues sur 30 ne sont plus transmises aux jeunes générations, et que dire de certaines langues de l'Alaska pour lesquelles les locuteurs se comptent sur les doigts d'une main ?

Cette évolution historique semble inéluctable, que l'on juge ce processus naturel (évolutionniste) ou non (catastrophique). Il est évident que plusieurs paramètres interviennent dans la survie ou l'extinction d'une langue et, à la pression de l'internationalisation des communications (où l'anglais prédomine à l'heure actuelle, comme l'ont fait le latin et le français précédemment) s'ajoutent les orientations prises en politique linguistique. Le Catalan Miguel Siguan présente une analyse poussée de la situation des langues en Europe [Siguan 96] ; il propose certains axes pour promouvoir la diversité des langues et pour assurer non seulement la subsistance mais aussi le développement des langues faibles. « Par langues faibles, il faut comprendre aussi bien les langues minoritaires par rapport aux langues étatiques, que les langues étatiques peu utilisées par rapport aux langues plus employées, ou encore les grandes langues comme l'espagnol [5] ou le français par rapport à la pression qu'elles subissent de l'anglais ». L'auteur note encore que selon les pays, la politique linguistique est très variée, allant du monolinguisme (cas de la France) au plurilinguisme d'état (cas du Luxembourg) en passant par des solutions fédérales (exemple de la Suisse) ou régionales (autonomie linguistique des régions en Espagne). Il fait appel à la notion de niveau de conscience linguistique que doivent atteindre les locuteurs des langues minoritaires pour assurer leur pérennité et à la volonté politique qui doit l'accompagner. Le philosophe Charles Taylor cite dans un entretien accordé à la revue en ligne *DiversCité Langues* [Taylor 97], l'exemple de l'islandais, parlé par seulement 225 000 personnes et qui se maintient parfaitement, adoptant constamment des néologismes pour s'adapter à l'évolution des techniques et des concepts. A la lecture du livre de Siguan, il apparaît que le soutien étatique semble indispensable à la survie d'une langue. Taylor est quant à lui

[5] L'auteur de précise pas s'il fait référence au castillan ou au catalan…

plus nuancé ; en Inde, le bengali est parlé par 200 millions de personnes, et même si la volonté politique a été, au moment de l'indépendance, d'instaurer une langue officielle unique (l'hindi), la réalité sociale et historique en a décidé autrement : il y a aujourd'hui 15 langues officielles en Inde. Par contre, il considère que la situation dans certains pays africains va mener à une évolution différente. Dans ces pays, où coexistent des sociétés la plupart du temps multilingues et où le nombre de langues parlées est très important, il semble peu probable qu'elles se maintiennent toutes, et l'on risque d'arriver à une marginalisation de certaines langues au profit d'autres, parlées par des communautés plus étendues. Cette situation rappelle celle qu'a connue l'Amérique latine où les deux langues indigènes les plus répandues sont actuellement le quéchua au Pérou et le guaraní au Paraguay.

La mort des langues est un processus naturel, auquel la mondialisation de la communication et du commerce donne une ampleur jamais atteinte. Il ne faut pourtant pas oublier que des langues sont nées jusqu'à une date récente. Les circonstances de ces apparitions nous ramènent à une époque peu glorieuse de la société occidentale où des esclaves parlant des langues souvent différentes étaient déracinés de leur communauté d'origine pour travailler dans les plantations. Pour communiquer entre eux, ils durent développer des langages rudimentaires communs, les pidgins, reposant sur une syntaxe et un vocabulaire appauvri [Bickerton 95]. Les enfants de ces esclaves, nés dans les colonies, ont développé en quelques générations (quelques décennies) ces pidgins qui se sont transformés en langues proprement dites, les créoles. Les études ont montré que ces langues possédaient une similarité de structure étonnante, par delà la distance géographique ou l'éloignement linguistique des pidgins qui leur ont donné naissances. Le créole hawaïen datant d'à peine un siècle, on a pu retracer la manière dont il avait émergé, et il est fort possible que la structure qu'il partage avec les autres créoles soit caractéristique de l'acquisition du langage par les enfants.

A ces langues apparues récemment, on peut associer l'hébreu, qui, après avoir pratiquement disparu durant des siècles, a vécu une renaissance sous la forme d'une langue étatique, ou encore l'espéranto, inventé par le docteur Zamenhof il y a un siècle. Cette langue internationale, créée *a posteriori* à partir des langues indo-européennes demeure pratiquée par près de deux millions de personnes, et elle rassemble des communautés réparties sur toute la planète.

2.2 Quelques repères linguistiques

2.2.1 Naissance de la linguistique

L'histoire a retenu peu de choses des études menées sur les langues avant la fin du XVI ème siècle pour la simple raison qu'elles sont rares et souvent anecdotiques. Au XII ème siècle avant notre ère, le pharaon Psamétik 1 er réalise une expérience portant sur l'origine des langues : il fait élever ensemble deux nourrissons à qui personne ne parle jusqu'à ce qu'ils échangent leur premier mot. Ce mot, "bekos", existant en phrygien (il signifie "pain"), le pharaon déduit que cette langue est la première langue de l'humanité…

Platon, au VI ème siècle avant notre ère, relève dans ses dialogues socratiques des ressemblances entre le phrygien et le grec ; ce travail étymologique peut sans conteste être qualifié de précurseur. Virgile le poursuivra plusieurs siècles plus tard en affiliant le latin au grec, mais un long laps de temps s'écoulera à nouveau avant que l'on reparle de classification ou d'origine des langues. Si dans le monde arabe, les grammairiens perçoivent assez tôt la parenté entre l'arabe et l'hébreu (langues qui seront appelées sémitiques à partir de 1781), les savants européens ne s'intéresseront à ce sujet que bien plus tard. Entre le XVI ème et le XVIII ème siècles les grammairiens et les philologues font émerger des parentés entre langues, principalement pour des groupes de langues de la famille qui sera nommée ultérieurement indo-européenne (romanes, celtes…). L'approche employée est étymologique et basée sur l'écrit, ne prenant en compte aucune possible correspondance phonétique entre mots. Des recherches entreprises par les théologiens chrétiens [6] donnant à l'hébreu ancien le rôle de langue originelle jusqu'aux travaux de Gottfried Wilhelm Leibniz, le grand philosophe et savant allemand, les essais foisonnent. Leibniz regroupe, dans l'optique de son grand projet d'unification des représentations entre logique mathématique et philosophie, plusieurs langues que l'on sait aujourd'hui appartenir à la famille indo-européenne, ainsi que l'égyptien (ce que ne renieraient pas les partisans du nostratique comme nous le verrons plus tard). Il est aussi à l'origine d'une tentative de construction d'une langue universelle, la *Caracteristica Universalis* .

La tradition veut que l'on retienne comme date de naissance de la linguistique le jour où, en 1786, Sir William Jones, juge en poste en Inde et érudit, postule dans un discours à la société asiatique de Calcutta que le latin, le grec, le sanskrit, le gotique et l'avestique (ancien perse) sont issues d'une même langue, aujourd'hui disparue [Jones 1786].

A cette époque, les affirmations de parenté entre ces langues ne sont guère révolutionnaires, puisqu'elles s'appuient en partie sur des travaux entrepris dès le XVIème siècle ; par contre, il est certain que le discours de Sir Jones fait apparaître des notions d'évolution des langues et d'affiliation absolument visionnaires, presque darwiniennes en fait, alors que *On the Origin of Species* ne paraît qu'en 1859 !

2.2.2 L'essor de la linguistique comparative

Au cours du XIX ème siècle, la phonétique et la méthodologie comparative empruntée à la biologie feront de la linguistique une véritable discipline scientifique.

[6] Le but était d'étayer le récit biblique de la confusion des langues survenue à Babel (Babylone).

Après le discours visionnaire de W. Jones, son homonyme Thomas Jones propose en 1813 la dénomination de famille indo-européenne pour synthétiser les travaux de ses prédécesseurs.

Entre 1815 et 1825, Rasmus Rask, Franz Bopp et Jacob Grimm posent les fondements de la grammaire comparée et de la linguistique comparative grâce à la découverte de lois phonétiques. Ces lois établissent les règles d'évolution des sons au cours de la vie des langues. En effet, certains changements sont probables et couramment observés, alors que d'autres sont très rares. L'un des exemples les plus célèbres concerne l'évolution du son /t/ dans les langues indo-européennes : Ce son, déjà présent dans les langues indo-européennes anciennes, a évolué vers le son /θ/ dans les langues germaniques (*fa<u>th</u>er*) alors qu'il est resté inchangé dans les langues romanes (*pa<u>t</u>er*). De telles équivalences permettent d'envisager une affiliation entre langues par évolution phonétique, fournissant là un outil extrêmement précieux pour les linguistes. En 1861, August Schleicher, un botaniste très influencé par la théorie de l'évolution, jugera même que la méthode peut-être appliquée de manière plus vaste ; il s'intéresse aux similitudes des formes phonétiques non pas seulement pour justifier des parentés, mais aussi et surtout pour les rechercher. La phonétique se transforme alors en un outil d'investigation plus qu'en une simple méthode de justification. Schleicher, fort de ces comparaisons, est le premier à proposer une classification arborescente des familles et invente la notion de proto-mot.

Au cours de la deuxième moitié du XIX^ème^ siècle les recherches ne se limitent plus à l'indo-européen (par exemple vers 1858, Wilhelm Bleek unifie un groupe de langues africaines très étendu sous le nom de "Bantou") et plusieurs centaines de familles sont découvertes. L'idée de rapprocher l'indo-européen d'autres familles de langues commence à voir le jour, sous le regard courroucé des plus conservateurs. Des discussions s'engagent sur l'origine des langues et atteignent une rare violence pour une communauté aussi sage, à tel point qu'en 1866 la société de linguistique de Paris interdit d'aborder le sujet ! Fort heureusement pour la science, cette proscription n'empêche pas les savants de développer leurs propres idées, et le danois Holger Pedersen avance au début du XX^ème^ siècle que les langues indo-européennes et d'autres familles (altaïques, ouraliques…) sont issues d'une langue commune plus ancienne qu'il appelle le "nostratique" (du latin *noster* notre).

Parallèlement, l'italien Alfredo Trombetti, après avoir étudié des langues assez diverses géographiquement parlant, considère qu'elles ne peuvent être qu'apparentées. Cette réflexion audacieuse rejoint les considérations formulées au paragraphe 1.2 sur l'universalité du langage, mais elle dépasse le seul cadre biologique puisque Trombetti sera le premier à envisager dès 1926 une relation entre le basque et plusieurs langues du Caucase et à supposer une parenté entre certaines langues d'Amérique que l'on regroupera plus tard sous le nom de famille na-déné.

Aux Etats-Unis, Edward Sapir conjecture une relation entre la famille nord-américaine na-déné et la famille sino-tibétaine. Devant les réactions violentes que cette hypothèse osée provoque dans la communauté linguiste, il ne persévère pas dans cette voie… que les soviétiques Sergueï Nikolaïev et Sergueï Starostine exhumeront en 1984.

Pendant des décennies, le débat se poursuit entre partisans d'une grande prudence méthodologique — ceux que l'on appelle les "orthodoxes" — et défenseurs d'une méthode de comparaison multilatérale plus audacieuse. En 1950, cette controverse est loin d'être close, mais un jeune linguiste américain du nom de Joseph Greenberg, partisan de l'approche la plus radicale, va obtenir un succès retentissant.

2.2.3 Les succès des comparaisons multilatérales

Au cours des années 50, Greenberg travaille sur l'émergence de grandes familles linguistiques en Afrique, et en particulier sur des langues dont les liens de parenté sont encore mal connus. Ces travaux vont l'amener à modifier assez sérieusement certaines hypothèses que l'on pensait acquises. Il suggère en effet, contre l'avis des autorités en la matière, que le foyer originel des langues bantoues, parlées dans une bonne partie de l'Afrique subéquatoriale, n'est pas situé au centre de la zone, mais plutôt en bordure Nord Ouest, à la frontière du Nigeria et du Cameroun. Cette conclusion s'appuie sur deux principes très importants :

- ✓ l'hypothèse de l'aire ancestrale (énoncée par Sapir) : cette hypo thèse veut que l'implantation la plus ancienne d'une famille de langues soit l'emplacement où la langue parlée présente la plus grande diversité linguistique. Cela s'explique par le fait que la langue ancestrale est celle qui a évoluée le plus longtemps.
- ✓ Le principe des moindres mouvements : si une famille de langues est parlée sur une aire très étendue et que l'on trouve une famille apparentée sur une aire beaucoup plus réduite contiguë, il est très probable que la famille étendue ait occupé une zone plus réduite proche de la zone de la famille parente éloignée.

Greenberg montre que les langues les plus proches des langues bantoues sont des langues alors classées dans la famille soudanique occidentale et parlées dans l'est du Nigéria. L'hypothèse de l'aire ancestrale le conforte dans son opinion, puisque les parlers bantous du Nord Ouest sont, linguistiquement parlant, les plus éloignés des autres parlers, alors même que, du zoulou (parlé en Afrique du sud) au swahili (parlé en Afrique de l'est), l'homogénéité linguistique est grande malgré la distance. Il avance dès lors que la famille bantoue, même si elle est très étendue, ne représente qu'un sous-groupe partageant avec les langues soudaniques occidentales un même groupe qu'il nomme le Bénoué-Congo (lui-même sous groupe du nigéro-congolais).

Parmi les détracteurs de Greenberg, le plus violent est l'éminent bantouiste londonien Malcolm Guthrie, directeur de la School of Oriental and African Studies pour la bonne raison qu'il défend une thèse totalement opposée sur l'origine géographique des langues bantou : il situe la zone originelle au Congo, et il considère les parlers bantou du Nord Ouest comme dégénérés pour expliquer leurs différences avec le bantou congolais. Il s'appuie entre autres sur le fait qu'il est très peu vraisemblable que des tribus du Nord

Ouest aient colonisé le reste de l'Afrique méridionale en dépit de la présence d'un rempart naturel fort impressionnant : la forêt équatoriale. La bataille entre les deux linguistes dure plus de dix ans, mais peu à peu, de nouveaux éléments donnent raison à Greenberg. En particulier, l'apparition au Nigeria de la métallurgie, associée à l'agriculture avant que ne débute l'expansion proprement dite (il y a environ 2300 ans), fournissait un avantage décisif aux tribus du sud-est du Nigeria sur les populations de chasseurs-cueilleurs peuplant l'Afrique subéquatoriale. Ces faits historiques abondaient dans le sens de Greenberg, alors que l'hypothèse de Guthrie avait du mal à s'en accommoder. Le lecteur désireux d'étudier plus précisément l'expansion des langues bantoues peut se reporter à [Philippson 1997].

Les langues bantou ne constituant qu'une facette du travaille de Greenberg, celui-ci ne s'arrête pas là, et il travaille à une unification des langues africaines en un nombre réduit de familles en suivant son approche taxinomique de "comparaison multilatérale". En 1963, il publie sa classification en 4 familles : khoisan, nigero-khordofanien (dont la famille bantoue), nilo-saharien et afro-asiatique (avec comme branche le sémitique et le tchadique). Cette taxinomie, qui, après avoir été combattue et débattue est finalement acceptée, aura demandé plus de dix ans, et elle représente une "diminution de l'entropie" considérable, comparée aux quelques 6000 langues et dialectes parlés en Afrique. L'émergence de super-familles en Afrique va s'accompagner d'autres travaux synthétistes du même type, entraînant la communauté linguiste dans de nouveaux débats comme nous allons le voir maintenant.

Au cours des décennies qui suivent, les linguistes débattent de plus en plus de l'isolement des familles de langues ou leur rapprochement en super-familles. De manière un peu simpliste, on peut dire que deux écoles s'affrontent. La première, représentant le courant "orthodoxe", émet de sérieux doutes sur la possibilité d'apparenter des langues anciennes – ou des proto-langues reconstruites – parlées il y a plusieurs milliers d'années et d'en déduire l'existence de familles plus anciennes. Les tenants de cette école jugent que la méthode de comparaison employée (et inspirée de Schleicher) ne peut pas être appliquée de manière rigoureuse à des langues ou des familles de langues anciennes reconstruites. Cette limitation est due au manque de données, aussi bien du point de vue quantitatif que qualitatif et elle implique que si on leur applique la méthode de comparaison multilatérale, on aboutit à des résultats au mieux non vérifiables et au pire erronés. L'école la plus radicale, que l'on pourrait nommer "synthétiste", estime au contraire que les méthodes de comparaison appliquées traditionnellement sont trop rigides et qu'elles ne permettent pas de mettre en évidence des similitudes qu'une comparaison opérant dans un cadre plus vaste (en tolérant que pour certains mots, on ne retrouve pas de ressemblance évidente) révèle.

Les deux écoles s'affrontent en particulier sur la définition d'une limite temporelle au-delà de laquelle les similitudes entre langues ne peuvent plus être étudiées. Antoine Meillet et bien d'autres considéraient que rechercher des parentés entre des langues parlées il y a plus de 6 000 ans est une hérésie. Johanna Nichols [Nichols 92] considère quant à elle que la limite de fiabilité est plutôt située entre 8 000 ans (âge probable de la langue indo-européenne) et 10 000 ans (âge estimé de la famille afro-asiatique). Parmi les arguments avancés par les orthodoxes, on relèvera les suivants : tout d'abord, il est très vraisemblable qu'après 8 000 ans, les changements phonétiques et linguistiques survenus entraînent la disparition des traces de la supposée proto-langue dans un bruit de fond important. Ensuite, les linguistes cherchant à rassembler les langues ne prennent pas assez en compte les phénomènes tels que l'emprunt ou tout simplement le hasard, et privilégient systématiquement l'origine commune lorsque des similitudes entre langues sont détectées. Enfin, et c'est là un point de désaccord méthodologique important, les synthétistes ont tendance à conclure de manière hâtive, prenant des indices de présomption de parenté pour des preuves. Ce point de discorde sera nuancé par la mise en corrélation des résultats obtenus par plusieurs disciplines (en particulier la génétique) comme nous le verrons au paragraphe 2.2.5.

A ces arguments développés par les orthodoxes, les synthétistes – ou monogénétistes – opposeront une défense véhémente, certains laissant même entendre qu'un sentiment "euro-centriste" déplacé pousse certains linguistes à accorder un statut supérieur à l'indo-européen par rapport aux autres langues. Le fait que les orthodoxes fixent comme limite 8 000 ans, soit l'âge supposé de la langue proto-indo-européenne, n'est pas étranger à ce jugement.

Les synthétistes développent également d'autres arguments, répondant plus précisément aux attaques formulées contre leur méthodologie. Ils réfutent le facteur du bruit noyant les traces des langues anciennes, considérant que "tout ne change pas" en 6 000 ans, et produisant des exemples de mots ayant peu évolué depuis : le proto-mot **nepōt* (l'étoile signifie qu'il s'agit d'une forme reconstruite) qui signifie neveu en proto-indo-européen est toujours présent en roumain sous la forme *nepot* et en français, népotisme reste un terme étymologiquement lié à la notion de neveu. Ensuite, le choix de mots dans un lexique de base (parties du corps, chiffres, pronoms de la première et de la deuxième personne du singulier…) minimise les risques d'emprunts. Enfin, il est difficile d'expliquer par le hasard le fait que des combinaisons de formes précises se retrouvent dans plusieurs familles : par exemple, les familles où l'opposition entre les pronoms de la première et la deuxième personne du singulier porte sur les sons /m/ et /t/ sont nombreuses en Europe et en Asie. Si l'on retrouvait cette opposition dans toutes les familles de langues du monde, on pourrait arguer qu'il s'agit d'une opposition universelle, mais ce n'est pas le cas puisque pour la plupart des langues parlées en Amérique, l'opposition porte sur les sons /n/ et /m/. Pour les synthétistes, l'existence de ces similarités est imputable, dans la très grande majorité des cas, à un lien de parenté entre les langues – ou les familles – concernées, et une étude multilatérale des lexiques de base permet de le révéler.

Nous allons voir à présent certaines des discussions où les différentes écoles, des plus conservatrices aux plus hardies, s'affrontent.

2.2.4 Les super-familles et les controverses actuelles

Dans les années 1960, plusieurs linguistes russes de Moscou reprennent l'hypothèse du nostratique de Pedersen et les travaux de Trombetti sur des parentés lointaines entre langues. Aaron Dolgopolsky et Vladislav Illich-Svitych s'attellent, d'abord séparément, puis ensemble, à construire la macro-famille à laquelle l'indo-européen est apparenté. Ils reconstruisent donc la famille nostratique à partir des familles indo-européenne, ouralienne, altaïque, afro-asiatique et incluent le dravidien et le kartvélien (géorgien). La méthode employée est empruntée à l'école traditionnelle, puisque les correspondances phonétiques sont appliquées aux proto-langues reconstruites pour chacune des familles. Bien que la méthodologie soit commune avec celle ayant fait émerger l'indo-européen, le fait que le nostratique ait dû être parlé il y a quelques 12 000 ans, amènent les linguistes orthodoxes, tels Eric Hamp et C. Watkins à la critiquer vivement (emploi de données mal classées, non prise en compte d'emprunts…).

Si le différend entre nostraticistes et orthodoxes porte plus sur la limite fatidique des 8 000 ans fixée dès le début du siècle par les spécialistes de l'indo-européen, il est clair que la méthode de comparaison multilatérale sans reconstruction phonétique préalable que Greenberg a depuis appliquée à l'indo-européen est profondément contestée par les mêmes Hamp et Watkins. La super-famille eurasiatique reconstruite par Greenberg présente bien évidemment des traits communs avec le nostratique, même si, la méthode n'étant pas la même (Greenberg considère certaines langues ignorées par les nostraticistes car leurs reconstructions ne sont pas assez sûres comme les langues eskimo-aléoutes ou tchouktchi-kamtchatkiennes) des différences subsistent. En particulier, Greenberg considère le dravidien, le kartvélien et la famille afro-asiatique comme distincte de l'eurasiatique ou tout au plus apparentées à un niveau plus lointain. Le nostraticiste Starostine rejoint d'ailleurs Greenberg sur le fait que le dravidien et l'afro-asiatique sont moins apparentés au nostratique que les autres langues. Comme le dit Merritt Ruhlen, « […] il est clair que les nostraticistes et Greenberg sont alliés dans leur tentative de dépasser les limites artificielles arbitrairement imposées par les indo-européanistes ».

A l'heure actuelle, plusieurs autres controverses demeurent entre linguistes. Outre les discussions portant sur un nombre restreint de langues ou sur la reconstruction des proto-langues, un nombre réduit de problématiques globales se dessinent et Greenberg et son ancien disciple Ruhlen sont toujours au cœur des débats :

- ✓ la première controverse porte sur une unification des nombreuses langues américaines en familles. Là encore, Greenberg s'oppose très vivement depuis 1987 à un nombre important de ses confrères. En effet, là où les linguistes orthodoxes voient une mosaïque de plusieurs centaines de langues, Greenberg suggère une unique famille qu'il nomme amérinde. Les langues parlées en Amérique se regrouperaient donc en trois familles : l'eskimo-aléoute (issu de la famille eurasiatique) à l'extrême nord, le na-déné, parlé principalement au Canada et l'amérinde partout ailleurs. Ces trois familles étant plus proches des langues asiatiques qu'elles ne le sont entre elles, elles seraient issues de trois phases de migration asiatique survenues à des époques distinctes. D'après les études menées sur la diversité des langues et l'hypothèse d'aire ancestrale, on peut dire que la première vague a colonisé les deux continents, puis qu'une seconde vague s'est installée en Amérique du nord et enfin que les locuteurs des langues eskimo-aléoutes ont traversé le détroit de Béring.

- ✓ le deuxième débat porte sur la synthèse des super familles (qui ne sont déjà pas admises par les orthodoxes) en familles plus anciennes encore, comme l'hypothétique super-famille déné-caucasienne rassemblant entre autres le basque et les langues na-déné. A partir de ces familles très anciennes, les monogénétistes recherchent des racines communes, pour faire émerger une langue plus ancienne, et Ruhlen va même jusqu'à envisager de trouver ainsi des racines provenant peut-être de la langue mère de l'humanité.

Famille	**Localisations principales**	**Exemples de langues**
Khoisan	Afrique	!kung, sandawe
Nigéro-kordofanien	Afrique	swahili, douala
Nilo-saharien	Afrique	massaï
Afro-Asiatique	Afrique	haoussa, arabe, hébreu
Dravidien	Péninsule indienne	brahoui, tamil
Kartvélien	Géorgie	géorgien
Eurasiatique	Europe, Asie, Amérique du Nord et Groenland	finnois, mongol, japonais, coréen, eskimo, français
Dene-caucasien	Euskadi, chaîne himalayenne, Amérique du Nord…	basque, sino-tibétain, caucasien, navajo
Austrique	Asie du Sud-Est, Madagascar	vietnamien, miao, thaï, maori
Indo-pacifique	Nouvelle Guinée	tasmanien, yareba
Australien	Australie	pama-nyunga
Amérinde	Amériques du Nord et du Sud	Algonquin, quechua, miwok

Tableau 1 : Les 12 familles mondiales identifiées par les synthétistes (d'après [Ruhlen 97]).

Cette monogénèse est défendue par Merritt Ruhlen et John Bengston, contre la plupart des autres linguistes. Ces deux hommes se sont attachés à la recherche de racines mondiales dont ils ont reconstruit 27 étymologies. Le débat est loin d'être clos, et la plupart des linguistes tels Claude Hagège multiplient les critiques, les arguments développés étant principalement méthodologiques. Pour Ruhlen, toute tentative d'étude de parenté entre langues débute par une phase de taxinomie, suivie d'une tentative de reconstruction de la langue mère (approche multilatérale empruntée à Greenberg). Pour les opposants, il faut d'abord reconstruire la langue ancêtre, puis rechercher les langues affiliées. Cette démarche étant extrêmement rigoriste, il est quasiment impossible de rassembler des langues reconstruites. Les synthétistes se contentent de critères de parenté moins stricts et Ruhlen arrive ainsi à dégager 12 super-familles linguistiques (Tableau 1).

D'autres arguments sont développés par les opposants à Ruhlen et Bengston, en particulier par Johanna Nichols. Elle a établi une typologie basée sur la morphologie, le lexique et la syntaxe pour quelques 200 langues. Elle est arrivée à la conclusion que l'on peut distinguer deux types de zones linguistiques, suivant qu'il s'agit d'une aire d'expansion d'une famille de langues ou d'une zone résiduelle (Caucase, Balkans). Dans les zones d'extension, elle admet que les principes généraux pris en compte par Ruhlen et Greenberg s'appliquent (à savoir que la parenté est le principal facteur de similarité entre langues) mais elle juge qu'une modélisation plus fine et moins arborescente est nécessaire dans les zones résiduelles. En particulier, dans ces zones où le multilinguisme est courant, Nichols avance l'hypothèse que des modifications linguistiques soient apparues et aient diffusé dans toute la zone, donnant au final des caractères communs à des langues qui pouvaient ne pas être apparentées précédemment.

Le débat entre monogénétistes et orthodoxes évolue actuellement avec la prise en compte d'études menées dans d'autres disciplines qui permettent elles aussi d'établir des relations entre histoire et populations, comme nous allons le voir maintenant.

2.2.5 Linguistique, Archéologie et Génétique

Depuis 10 ou 15 ans, en effet, les recherches sur les origines des langues se rapprochent des recherches menées sur les origines de l'homme. En effet, la linguistique ne permet pas de dater les événements dès que l'on s'éloigne d'un passé récent, alors que génétique et archéologie sont plus aptes, dans une certaine mesure, à conclure. En effet, l'histoire des migrations humaines est riche en enseignement, et la prise en compte de plusieurs sources d'information permet d'affiner les hypothèses. Prenons par exemple la localisation du foyer de peuplement indo-européen originel. Parmi plusieurs hypothèses (Russie du Sud-Est, Eurasie septentrionale, foyer multiple…), nous retiendrons celles qui situent le foyer de peuplement au cœur du croissant fertile en Anatolie (Turquie – hypothèse défendue par Colin Renfrew [Renfrew 90]), ou en Ukraine (hypothèse défendue par J. P. Mallory [Mallory 89]). Ces études se basent en partie sur l'apparition de l'agriculture, à la fois dans les vestiges archéologiques et dans les proto-langues (en nostratique, il n'y a pas de lexique caractéristique de l'agriculture alors que dans les langues filles, un tel lexique existe). Les deux hypothèses aboutissent à des scénarios d'expansion de l'indo-européen différents : Pour Renfrew, soutenu par Ruhlen, c'est l'apparition de l'agriculture qui a permis à nos ancêtres de se répandre dans toute l'Europe et l'Asie continentale. Pour Mallory, il s'agit des conquêtes des guerriers à cheval. Ces deux hypothèses sont en fait conciliables, et l'hypothèse d'une migration en deux vagues est envisageable, en particulier si l'on se réfère aux travaux entrepris dans une autre discipline : la génétique.

En effet, l'équipe de Luigi Cavalli-Sforza, qui est à l'origine d'une classification génétique très souvent en accord avec la classification linguistique de Ruhlen, penche pour une double expansion des langues indo-européennes. A partir d'un premier foyer situé en Anatolie, les émigrants auraient colonisé plusieurs endroits, vers l'ouest en passant par la Grèce, et en Asie continentale en contournant la mer Caspienne. Ensuite, une deuxième vague de migration aurait eu lieu à partir de ce foyer ukrainien, vers l'Ouest en particulier, apportant la technologie de ces peuples de guerriers cavaliers aux agriculteurs déjà installés.

Comme nous venons de l'écrire, Cavalli-Sforza est à l'origine de l'un des plus importants arguments en faveur des synthétistes : l'histoire génétique des populations, évaluée à partir de l'ADN mitochondrial et d'autres facteurs génétiques, aboutit à un arbre de classification aux ramifications assez proches de l'arbre linguistique de Ruhlen.

Plusieurs différences subsistent, mais une bonne partie d'entre elles s'expliquent assez bien avec des considérations simples : par exemple, le fait que les hongrois soient très proches des autres indo-européens génétiquement et très différents linguistiquement se justifie par une substitution de langues : en effet, ils ont adopté (de force) la langue (ouralienne et non indo-européenne) des conquérants Magyar qui les ont dominés. Un autre exemple est donné par le basque. En effet, la langue est indubitablement issue d'une famille non indo-européenne (la famille déné-caucasienne selon Ruhlen) et effectivement, génétiquement parlant, les basques présentent encore aujourd'hui le plus fort taux de Rhésus négatif dans la population au monde. Cette particularité génétique témoigne de la différence biologique ayant toujours existée entre la population basque et ses voisines. Bien évidemment, ces différences génétiques tendent à disparaître alors que l'identité linguistique de la langue basque demeure. Comme le dit Ruhlen, cela s'explique par le fait que « les langues ne font pas l'amour ».

D'autres équipes travaillent sur le parallèle entre classification génétique et linguistique, et principalement l'équipe de Langaney à Genève. Pour sa part, il relève quelques remarques méthodologiques appelant à plus de prudence pour certains des résultats de Cavalli-Sforza, arguant – entre autres – de la disparité des données utilisées pour ses tests. En 1992, Pellegrini conclut d'ailleurs que l'approche pluridisciplinaire (archéologie, paléontologie et génétique) ne permet pas de lever l'incertitude sur les datations de manière sûre pour l'histoire ancienne [Pellegrini 92]. Par contre, l'équipe de Genève, après avoir recommencé certaines des expériences sur des populations subsahariennes avec des marqueurs génétiques "normalisés" reproduit certains des résultats de l'équipe de Cavalli-Sforza pour des périodes allant de – 8 000 ans à nos jours [Sanchez-Mazas 91-92].

Il est intéressant de noter que la génétique fait appel à des hypothèses proches du principe de l'aire ancestrale et du principe de moindre mouvement : les travaux entrepris en génétique sont basés sur l'étude statistique des taux de mutations survenus dans certains gènes puisque, plus une espèce a évolué, plus ses gènes ont muté entre son ancêtre et sa forme actuelle. De plus, la séquence de gènes ancestrale peut être trouvée en comparant cette espèce à une espèce proche dont elle s'est séparée. Dans le cas de l'homme, les recherches portent depuis une dizaine d'années sur des séquences de gènes portées par l'ADN mitochondrial, presque exclusivement apporté par le patrimoine génétique de la mère ; cette précaution permet de mieux tracer l'évolution des populations féminines plutôt que mixtes et elle permet de mieux rendre compte des mouvements migratoires. Par contre, les vitesses de mutation étant mal connues (on les considère comme constantes au cours du temps), la datation précise des migrations est ardue. La séquence génétique ancestrale a été étudiée en comparant l'ADN humain et l'ADN du chimpanzé, l'un des derniers animaux à s'être séparé de notre branche d' *homo erectus* . Ces études ont abouti à l'hypothèse de l'Eve africaine, raccourci sémantique signifiant que le plus ancien groupe de femmes avait vécu en Afrique. Cette hypothèse est toujours controversée, mais elle vient de trouver un soutien éclatant avec la publication récente de résultats convergents. Ces travaux, entrepris par les équipes de Peter Underhill et Michael Hammer (cités dans [Magnan 98]), ont porté sur la population masculine (séquence d'ADN portée par le chromosome sexuel Y) et ils convergent vers l'Afrique comme foyer originel. De manière plus précise, il semble même que la population la plus proche de cet Adam génétique, comme se sont empressés de l'appeler les journalistes, soit la population khoisane. Une hypothèse riche en spéculation quant à l'origine des langues…

En résumé, nous pouvons dire que l'existence d'une population ancestrale parlant un langage unique est une hypothèse séduisante mais non encore avérée, même si des éléments provenant de plusieurs disciplines convergent vers cette genèse scientifique.

3 DE LA TAXINOMIE A LA CARACTERISATION

Le paragraphe précédent était principalement consacré aux résultats établis par les linguistes en taxinomie des langues. Nous allons maintenant aborder la description des approches développées, des classifications linguistiques à l'émergence de typologies phonologiques.

Greenberg citait en 1958 trois approches majeures en classification des langues : la **classification génétique** , terme qui était alors employé sans ambiguïté possible pour l'approche généalogique, la **classification aréale** , opérant des regroupements sur des bases de proximité géographique et la **classification typologique** , basée sur les propriétés structurelles des langues.

Historiquement, les deux premières approches sont basées sur la recherche d'un vocabulaire commun entre les langues. Le développement de la grammaire comparée a permis par la suite d'élaborer une description morpho-syntaxique des langues plus apte à faire émerger des structures linguistiques communes. Ainsi, l'approche typologique est apparue après les premières classifications, avec l'émergence de règles structurelles. Elle s'est développée depuis deux siècles pour aboutir à des critères que l'on retrouve aujourd'hui dans la littérature classique [Hagège 82] et qui portent sur différents niveaux grammaticaux.

On peut distinguer avec Hagège les langues agglutinantes (concaténation de suffixes ajoutés à une racine), flexionnelles (verbes conjugués, noms et pronoms déclinés) ou isolantes (utilisation de mots courts invariables). Le fait que les langues ne distinguent pas les verbes des noms (cas du santali en Inde) ou au contraire marquent une distinction verbo-nominale nette (comme le français par exemple), ou encore que la construction de la phrase marque préférentiellement l'agent (construction ergative) ou le patient (construction accusative), sont autant de critères de distinction qui permettent d'aboutir à une description précise de la structure morpho-syntaxique de chaque langue.

Si, dès 1861, Schleicher a proposé de baser les grammaires utilisées sur la nature des sons de chaque langue, les premières typologies des structures sonores sont apparues bien plus tard. L'une des principales raisons tient en un mot : hétérogénéité. A une époque où enregistrer la voix était extrêmement difficile, voire impossible, les observations reposaient sur les personnes qui étaient en contact direct avec les langues étudiées, sans possibilité de vérification *a posteriori*. Les premiers travaux significatifs entrepris reviennent sans doute à Troubetzkoy en 1928. Son activité au sein du cercle de linguistes de Prague aboutit à la notion de « lois de formation des systèmes [vocaliques] » qui préfigure la recherche d'universaux du langage. Durant les décennies 1930 et 1940, les travaux du cercle de Prague seront poursuivis par Troubetzkoy et Jakobson. Après la guerre, les procédures de description phonologique des langues deviennent plus strictes et les données ainsi rassemblées fournissent une matière conséquente aux phonologues. A cette époque, il est clair dans l'esprit des linguistes, que si linguistique et phonologie ont en commun l'étude des structures des langues, la phonétique est une science relevant de la biologie et non pas de l'étude linguistique. Cette dichotomie qui est encore aujourd'hui la règle pour certains linguistes est toutefois ébranlée dès la fin des années 40 lorsque Jakobson propose que les unités des systèmes *phonologiques* des langues soient soumises à des règles d'invariances *substantielles*. Hockett poursuivra sur cette voie, en étudiant les tendances universelles phonologiques. Le *Symposium on Universals in Linguistic Theory* de 1967 est le théâtre du débat autour de ces disciplines, et il préfigure les travaux de Lindblom qui, en 1972 bâtit avec Liljencrants la linguistique "*substance-based*". Les modèles qu'ils établissent permettent de rendre compte au niveau de la réalisation phonétique des sons de l'émergence de certaines caractéristiques des systèmes vocaliques phonologiques. Depuis cette date, de nombreux autres travaux ont été menés en ce sens [Stevens 72, Lindblom 75, Maddieson 84, Schwartz 89] et les relations entre phonétique, phonologie et linguistique ont abouti à des approches unificatrices (telle la phonologie "intégrative" de Ohala [Ohala 91]) et à des typologies très poussées des systèmes phonologiques et principalement des systèmes vocaliques [Vallée 94]. Plus qu'à une simple taxinomie, ces études visent à atteindre la substance universelle du langage humain et à approfondir notre connaissance sur son acquisition autant que sur son évolution. Le chapitre suivant de ce manuscrit nous permettra d'appréhender cet aspect de la linguistique par le biais de la caractérisation des systèmes vocaliques.

4 LA DISCRIMINATION DES LANGUES PAR L'ETRE HUMAIN

Avant de nous intéresser à l'identification *automatique* des langues il nous paraît nécessaire de nous interroger sur la manière dont chacun d'entre nous distingue – avec plus ou moins de bonheur et d'efficacité – une langue d'une autre. En effet, relativement peu de personnes vivent dans un environnement rigoureusement monolingue. Que ce soit dès la naissance (on estime qu'au moins la moitié des enfants du monde vivent dans un environnement multilingue) ou plus tard dans le cadre de l'apprentissage plus tardif de langues étrangères – là encore avec plus ou moins de bonheur – la plupart des habitants de cette planète ont, au pire entendu des locuteurs parler, au mieux, appris à s'exprimer, dans plusieurs langues. Ainsi, chacun d'entre nous pourra dire si un énoncé est prononcé dans une langue qui fait probablement partie des idiomes qu'il connaît ou dire au contraire que cette langue lui est totalement inconnue. Cette faculté dépend bien évidemment du "bagage linguistique" de chacun, et les recherches entreprises dans ce domaine cognitif se heurtent souvent à la disparité dudit bagage. Quoi qu'il en soit, on peut aussi s'intéresser à un problème connexe : comment les enfants plongés dans un environnement multilingue apprennent-ils de manière cohérente plusieurs langues ? En effet, le constat est là, un bébé à qui l'on parle régulièrement deux langues a toutes les chances d'apprendre correctement ces deux langues. Nous n'avons qu'effleuré cet aspect particulier de l'acquisition du langage au paragraphe 1.3 et il nous semble particulièrement intéressant de s'y attarder ici.

4.1 Le nourrisson reconnaît-il les langues ?

Quelle que soit la théorie de l'acquisition du langage que l'on adopte, il est nécessaire d'envisager chez le nourrisson le cas où il est plongé dans un environnement multilingue. Il est généralement admis que le nourrisson développe au cours de sa première année une bonne partie de ses compétences phonologiques, alors que les structures morpho-syntaxiques du langage n'atteindront leur maturité que plus tard. Durant ces premières années essentielles, comment un enfant génère-t-il plusieurs systèmes phonologiques, puis syntaxiques distincts alors qu'il s'agit là d'un apprentissage en quelque sorte non supervisé : lorsque une même personne de l'entourage de l'enfant s'expriment dans plusieurs langues, l'identité de la langue parlée est contenue implicitement dans le message et non explicitée par ailleurs [7]. Pourtant, les erreurs de "codage" entre les langues sont plutôt rares, tout au moins au niveau phonologique et prosodique. Si l'on adopte l'approche par codage phonologique proposée par Mehler, l'enfant effectue un filtrage du signal acoustique de manière à adopter la

[7] Si les deux parents s'expriment chacun dans une langue et non plus en utilisant plusieurs langues, l'identité de la langue parlée peut alors être rattachée à l'identité du locuteur.

représentation phonologique correspondant à l'énoncé [Mehler 95]. Etant donné que le "bagage linguistique" des nourrissons est faible, voire inexistant, il est probable que ce filtrage s'opère à partir d'informations de bas niveau, par opposition au contenu syntaxico-sémantique dont l'adulte peut disposer pour identifier les langues (cf. paragraphe suivant). Parmi ces informations de bas niveau, on peut relever les traits potentiels suivants :

- ✓ traits acoustico-phonétiques (présence ou absence de certains sons, probabilités de rencontrer ces sons dans les langues),
- ✓ traits phonotactiques (probabilités d'occurrence de séquences de sons),
- ✓ traits prosodiques (traits relatifs à l'intonation, au rythme et à l'accentuation).

Des expériences assez nombreuses ont été menées pour d'une part, tester la faculté des nourrissons à discriminer plusieurs langues entre elles, et d'autre part faire émerger les traits utilisés. On peut citer par exemple [Mehler 86] pour les travaux menés sur le français et le russe, [Moon 93] pour la discrimination entre anglais et espagnol. Ces expériences ont été menées avec des bébés extrêmement jeunes (moins d'une semaine) et elles ont montré que dès cet âge, le nouveau né est capable de distinguer sa langue maternelle d'une autre et même de faire des distinctions entre certaines langues étrangères [Nazzi 98]. Dès lors qu'il s'agit d'identifier les processus mis en jeu par le nourrisson pour opérer une discrimination entre langues, les hypothèses invoquent généralement la prosodie en se basant sur les faits suivants.

La plupart des expériences sont menées à partir de parole naturelle puis à partir de parole artificiellement modifiée, soit en opérant un filtrage passe-bas coupant à 400 Hz de manière à ne conserver que la fréquence fondamentale[8], soit en opérant une déstructuration de l'énoncé, par exemple en réorganisant les phonèmes de manière à supprimer la prosodie. Le fait que dans le premier cas (parole filtrée) le bébé soit toujours capable de distinguer les langues proposées et pas dans le second (parole déstructurée) incite à penser que la prosodie et la structure de l'énoncé sont utilisés par le nourrisson. A l'heure actuelle, bien des points demeurent obscurs, mais il semble effectivement que l'on puisse écarter les traits acoustico-phonétiques des informations utilisées : la durée relativement courte des énoncés utilisés invalide au moins partiellement l'hypothèse selon laquelle le bébé ferait "des statistiques" sur les fréquences d'occurrences des sons qu'il entend. Par contre il reste possible que les traits phonotactiques soient utilisés, même s'ils disparaissent lorsque l'on filtre les énoncés. Ce point rejoint une hypothèse formulée par Mehler selon laquelle le trait primordial utilisé par le bébé est lié au rythme, et plus particulièrement aux séquences consonne-voyelle CV. En effet, cette information ne disparaît pas complètement des énoncés filtrés, et des expériences complémentaires ont montré que des bébés français[9] étaient très sensibles à cette structure CV ou plus exactement à la structure syllabique. Certaines expériences

[8] on peut tout de même arguer que le premier formant peut être conservé avec cette méthode.

[9] c'est-à-dire élevé dans un environnement monolingue français.

suggèrent par ailleurs que les nourrissons seraient sensibles au *type prosodique* de chaque langue (syllabique, moraïque, à accentuation…) et que la discrimination se baserait là-dessus. Dans cet ordre d'idées, on peut citer [Jusczyk 93] qui indique que des bébés américains de six mois distinguent l'anglais du norvégien mais pas du hollandais, langue ayant une structure prosodique proche de l'anglais.

A l'heure actuelle, il est difficile de conclure sur les processus mis en œuvre par le nourrisson lorsqu'il distingue une langue d'une autre. L'hypothèse d'une représentation du rythme à l'aide d'unités rythmiques propres au langage et effectuées à partir des séquences CV est cependant au cœur des débats. Il est possible que les expériences menées par ailleurs sur l'identification des langues ainsi que sur les représentations rythmiques chez les adultes apportent certains éléments de réponse.

4.2 La discrimination des langues chez l'adulte

Comme nous l'avons précisé en introduction de ce chapitre, il est difficile de quantifier l'influence des différents facteurs interagissant lorsque l'on étudie les capacités des adultes en identification des langues. En effet, les expériences perceptives sont généralement effectuées avec des auditeurs candides afin que le bagage linguistique d'experts ne biaise pas leur décision. Cette précaution a cependant comme conséquence de complexifier le dépouillement des résultats puisque il est alors difficile d'analyser la démarche menant à la décision. De nombreuses expériences ont porté sur l'apport des différentes composantes du langage pour la discrimination perceptive. Dès lors que l'on s'intéresse à la prosodie, on a généralement recours à plusieurs artifices pour en isoler les composantes rythmique, intonative et accentuelle. Franck Ramus [Ramus 98] cite plusieurs expériences réalisées au début des années 80 sur l'intonation et le rythme, au moyen de stimuli naturels filtrés ou déstructurés :

- ✓ des locuteurs adultes discriminent entre l'anglais américain, le japonais et le cantonais en se basant sur le rythme et la prosodie [Ohala 79],
- ✓ des locuteurs anglais distinguent le français de l'anglais grâce à l'intonation uniquement [Maidment 83],
- ✓ des locuteurs néerlandais distinguent le néerlandais de l'anglais en se basant sur l'intonation [Willems 82].

A l'opposé de ces expériences menées avec des stimuli dédiés aux expériences perceptives sur un nombre restreint de langues (généralement deux ou trois), il est intéressant de citer celles conduites par Yeshwant Muthusamy à OGI (Oregon Graduate Institute) lors de l'enregistrement du corpus multilingue MLTS (cf. 2[ème] partie). Ce corpus, dédié à l'identification automatique des langues, a permis de réaliser plusieurs séries d'expériences perceptives avec des stimuli réels non altérés (pas de filtrage) et sur un nombre relativement important de langues (10 langues) [Muthusamy 93, Muthusamy 94a]. L'auteur de cette étude a demandé à 7 auditeurs de langue anglaise américaine d'identifier des séquences de 1 à 6 secondes de parole après une phase d'apprentissage.

Les taux d'identification sont bien évidemment meilleurs pour 6 secondes (54,6 % d'identification correcte) que pour 1 seconde (37,0 %).

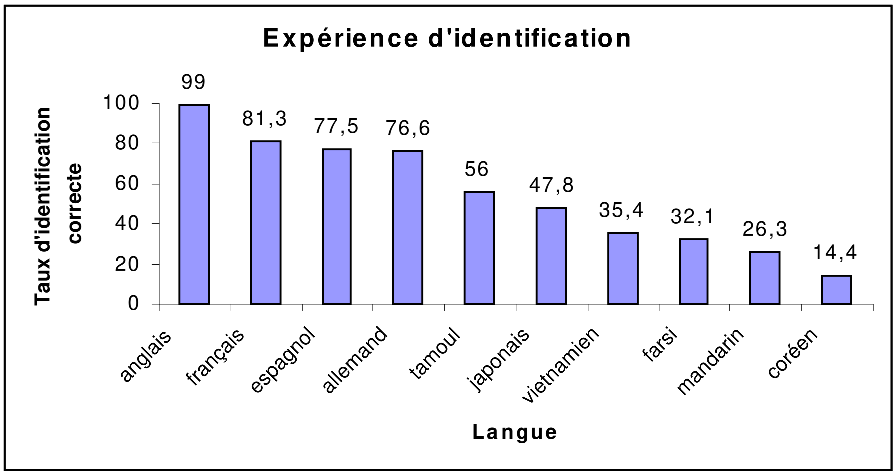

Figure 3 – Expériences d'identification des langues par 7 auditeurs. La durée des stimuli est de 6 secondes (d'après [Muthusamy 93])

Si l'on étudie les résultats pour les stimuli de 6 secondes (Figure 3), on constate que la langue la mieux reconnue est bien évidemment l'anglais (99 %) et que les langues française, espagnole et allemande sont plutôt bien reconnues (plus de 75 %). Comme le dit Muthusamy, ces langues sont « languages that the listeners were most often exposed to ». On remarquera par ailleurs que le coréen semble défier les auditeurs du test (score à peine supérieur au hasard).

Langue	Indices caractéristiques
allemand	Présence du mot *ich* – fréquence des sons vélaires aspirés
anglais	-
coréen	Aucune caractéristique marquante
espagnol	Débit de parole rapide – présence de la séquence /εs/
farsi	Présence de sons aspirés – fréquence du son /ʃ/
français	Présence de sons nasalisés – intonation particulière
Japonais	Présence de crisp occlusives– Présence de mots tels que *watashiwa* ou *mashita*
mandarin	Présence de tons caractéristiques
tamil	Présence des phonèmes /r/ et /l/
vietnamien	Intonation particulière – présence de sons nasalisés (particulièrement la vélaire /ŋ/)

Tableau 2 – Synthèse des indices perceptifs cités par les auditeurs [Muthusamy 94a].

Cette expérience, si elle permet d'estimer si un auditeur se familiarise avec certaines langues étrangères ne répond pour autant pas à des questions précises sur la stratégie qu'ils adoptent. D'autres études plus poussées ont été menées à partir des mêmes enregistrements, et les auditeurs ont été interrogés sur les indices leur ayant permis de prendre leurs décisions [Muthusamy 94a].

Les locuteurs comprenant tous l'anglais américain, ils n'ont pas relevé de traits distinctifs de niveau infra-sémantique alors que c'est le cas pour les autres langues (Tableau 2). L'analyse des réponses montre que les niveaux phonétique, phonotactique, morphologique et prosodique peuvent intervenir dans le processus de décision. Les informations prosodiques sont d'ailleurs les plus complexes à analyser car il est peu évident de déterminer quelle composante (rythmique, mélodique…) est réellement prise en compte.Pour aller plus loin dans cette étude, il est nécessaire de mettre en place des protocoles plus stricts, comme ceux que Ramus a exploité à partir de voix resynthétisée [Ramus 98]. Il a ainsi pu mener des expériences perceptives en jouant sur différents paramètres, en particulier en supprimant les informations phonotactiques (en remplaçant les consonnes par /s/ et les voyelles par /a/), en supprimant le rythme (en remplaçant tous les sons par /a/) et en modifiant la prosodie [10]. Les résultats montrent que des auditeurs adultes français discrimine le japonais de l'anglais à partir du rythme seul, mais pas de l'intonation. Ces deux langues étant de type prosodique différent (à accentuation pour l'anglais, moraïque pour le japonais), Ramus a étudié si une telle discrimination demeurait avec des langues moins différentes au niveau prosodique. Il a donc renouvelé l'expérience avec le polonais vs. l'anglais et l'espagnol (langue syllabique) et là encore les résultats sont semblables (le rythme est nécessaire et suffisant pour identifier une langue de l'autre).

4.3 Discussion

Les diverses expériences rapportées dans ce paragraphe confirment l'importance de la prosodie et des variations d'ordre suprasegmental dans le processus humain de discrimination entre langues. A notre avis, le fait que le rythme soit prédominant par rapport à la mélodie intonative suppose une relation assez forte entre prosodie (énergie, F_0 et durée des segments) et structure phonotactique (enchaînement des sons) au niveau des catégories phonétiques (le signal filtré masquant l'identité du son mais pas vraiment sa classe phonétique). Les interrogations soulevées par ces hypothèses rejoignent les préoccupations des linguistes et des cogniticiens sur le codage du langage chez l'Homme, en particulier sur les unités – syllabiques ou autres – privilégiées. Les expériences cognitives menées sur ce thème ne doivent pour autant par faire oublier l'importance des autres niveaux de représentation du langage, en particulier l'inventaire phonétique et la présence de séquences de phonèmes ou de mots caractéristiques dans l'énoncé.

[10] Le lecteur curieux d'entendre le résultat gagnera à aller voir le site de F. Ramus à l'adresse : http://www.ehess.fr/centres/lscp/persons/ramus/resynth/ecoute.htm

Chapitre 2

LA CARACTERISATION DES SYSTEMES VOCALIQUES

Le terme de *voyelle* revêt à la fois une signification orale – il fait référence à une catégorie de sons – et une signification écrite puisqu'il renvoie aussi aux symboles utilisés dans le langage écrit pour transcrire ces sons. L'histoire du langage a longtemps dissocié ces deux aspects, et aujourd'hui, si les voyelles sont présentes à l'oral dans toutes les langues du monde, certaines écritures ne les représentent pas. La voyelle joue en effet un rôle double dans la communication, bien plus fort à l'oral qu'à l'écrit comme nous allons le voir dans le paragraphe 1. Nous nous interrogerons aussi au cours de ce paragraphe sur la manière dont chaque langue construit ses propres contrastes entre voyelles à partir d'un espace vocalique commun. Cette question a « hanté » l'esprit des linguistes dès que la linguistique a porté un intérêt à la phonologie. Il se dégage des études menées par Nathalie Vallée [Vallée 94] certaines tendances universelles qui permettent de mieux appréhender la diversité des langues du monde telle qu'elle apparaît dans la base de données UPSID [Maddieson 84].

1 LA VOYELLE ET SON ROLE DANS LA COMMUNICATION

1.1 La voyelle dans la communication écrite [Jean 87, Bottéro 93]

Comme nous l'avons vu au chapitre précédent, l'homme a acquis la faculté langagière il y a déjà assez longtemps (40 000 ans selon Ruhlen) ; les premières peintures rupestres datent approximativement de cette époque, voire d'un passé plus ancien. Il est quasiment certain que les premières langues furent uniquement de tradition orale, et que l'écriture a émergé lors de la sédentarisation des populations (soit par nécessité, lorsque les communautés se sont agrandies, soit pour des raisons pratiques, la plupart des supports de l'époque étant difficilement transportables). Dès 30 000 ans avant notre ère, des codes arbitraires (cordelettes à nœuds…) avaient déjà été utilisés par certaines peuplades pour compter ou pour marquer les objets, mais ces systèmes ne permettaient pas de véhiculer des concepts évolués. Les écrits les plus anciens exhumés par les fouilles archéologiques sont des documents administratifs : ils ont été découverts bien plus récemment (4500 ans avant JC) au royaume de Sumer, au cœur de l'actuel Irak, entre Tigre et Euphrate. Si les écritures cunéiformes de Mésopotamie ont gardé pendant longtemps leurs secrets, on sait actuellement que des civilisations pluriséculaires se sont côtoyées, à partir du V^{ème} millénaire avant notre ère, du pays d'Elam (sud-ouest de l'actuel Iran) aux contrées bordant la méditerranée (Canaan…). La lente évolution d'une écriture idéographique vers une écriture alphabétique a souvent été marquée par une étape syllabique au cours de laquelle les symboles utilisés correspondaient soit à des consonnes seules, soit à ces mêmes consonnes en contexte vocalique (vieux-perse en Iran, méroïtique en Nubie…). Les premiers alphabets proprement dits sont apparus au cours du II ème millénaire avant notre ère, en écriture cunéiforme (ville d'Ugarit) ou cursives (phénicien). Là encore, les voyelles n'étaient pas notées, tout comme dans l'écriture démotique égyptienne par exemple. Comment, dans ces conditions, en est-on arrivé à noter les voyelles par écrit ? Plusieurs facteurs ont certainement interagi : le fait que plusieurs familles linguistiques [11] aient été en relation dans une zone d'échange permanente et que les systèmes d'écriture se soient développés à partir de deux axes principaux (hiéroglyphique en Egypte et cunéiforme à Sumer) sont sans aucun doute des facteurs propices à l'évolution, à la fois des langues orales et écrites. Il semble acquis que les langues parlées à cette époque disposaient de peu de voyelles ou du moins de peu de contrastes vocaliques. De manière générale, la connaissance de la suite de consonnes d'un mot (et antérieurement de la suite de syllabes) aboutissait sans ambiguïté à l'oralisation correcte. Lorsque le doute subsistait, on notait différemment la consonne. Par exemple, en vieux perse, la plupart des symboles correspondaient aux consonnes en contexte /a/. Si le contexte était différent (/i/ ou /u/), la consonne était notée en utilisant un autre signe. Cet "alphabet" comportait deux consonnes /g/, 3 consonnes /d/… qui correspondaient en fait à des syllabes. On peut y voir l'origine des notations diacritiques des voyelles employées dans certaines langues (éthiopien) et qui ont abouti plus tard à la transcription explicite des voyelles. Une autre origine plus probable de l'apparition des voyelles dans nos alphabets actuels est le phénicien. Cette langue, parlée dans le Croissant Fertile vers 1200 avant notre ère était issue d'une langue sémitique plus ancienne encore (parfois nommée proto-cananéen). Elle utilisait un système alphabétique composé de 22 signes consonantiques. Rapidement, cette langue essaima et donna naissance à la plupart des systèmes d'écritures employés aujourd'hui, des alphabets latin et grec au brahmi [12]. Il semble bien que la langue parlée en Grèce au début du premier millénaire avant notre ère ait multiplié les contrastes vocaliques, et qu'une réorganisation de l'alphabet phénicien se soit opérée vers 700 avant JC : certains symboles consonantiques peu usités en grec ont alors été utilisés pour noter des voyelles. Par exemple, le symbole Aleph, signifiant à l'origine un bœuf et transcrivant en phénicien le stop glottal fut utilisé par les grecs sous le nom de Alpha pour transcrire le son 'a'. Parallèlement, certaines langues issues de l'araméen se mirent aussi à employer des symboles consonantiques ou des symboles diacritiques pour noter leurs propres voyelles.

Ce rapide résumé de l'apparition des voyelles dans la langue écrite est bien incomplet, et bien d'autres systèmes d'écritures mériteraient d'être mentionnés (du

[11] l'ancien perse est une langue indo-européenne tout comme le sont les langues anatoliennes. L'accadien et le cananéen sont des langues hamito-sémitiques, tout comme l'égyptien antique, de parenté plus lointaine. Le Sumérien est à l'heure actuelle considéré comme un isolat linguistique.

linéaire A employé en Crête au début du deuxième millénaire avant notre ère au chinois ou aux écritures méso-américaines). Nous pouvons cependant noter que deux facteurs se sont conjugués pour donner naissances aux alphabets modernes : la **multiplication** des contrastes vocaliques dans les langues antiques et une tendance généralisée à **simplifier** l'écriture. En effet, la fréquence croissante des contrastes vocaliques a généralement provoqué l'apparition d'un système de notation des voyelles et l'extension des champs d'utilisation des documents écrits se satisfaisait plus d'une écriture simple où le nombre de symboles employés est réduit (le sumérien utilisait près de cinq cents signes, alors que le vieux perse en conservera uniquement quarante-deux). Ces deux tendances ont donc abouti à l'émergence d'alphabets utilisant des consonnes et des voyelles (sous formes diacritiques ou non) plutôt qu'à la multiplication des symboles dans les écritures syllabiques. Dès lors, l'influence des voyelles à l'écrit n'a pas été démentie et l'histoire de l'écriture nous révèle sans ambages que les évolutions de la langue écrite ont généralement été précédées de changements phonétiques dans la langue parlée. Comme nous l'avons vu au premier chapitre, l'évolution orale des langues est un processus naturel qui accompagne l'évolution des cultures et nous ne chercherons donc pas à justifier l'évolution diachronique des contrastes vocaliques dans les langues du monde. Nous allons cependant aborder l'aspect plus spécifiquement oral de la voyelle dans la communication tant au niveau acoustique qu'au niveau de la structure des systèmes vocaliques.

1.2 La voyelle dans la communication orale

1.2.1 Aspect acoustico-articulatoire

Plusieurs critères de nature articulatoire existent pour classer les sons produits par l'être humain en différentes catégories. Si l'on s'appuie sur la présence ou l'absence de vibration des cordes vocales, on obtient deux catégories de sons, respectivement voisés et non voisés. Si, de plus, on sépare les sons voisés en fonction de la présence ou de l'absence d'une obstruction du conduit vocal, on parvient à séparer les consonnes voisées des voyelles. Une voyelle est donc caractérisée par le passage libre de l'air depuis la glotte lorsque les cordes vocales sont en vibration. Tout comme la séparation entre sons voisés et non voisés peut-être difficile à établir (cas de la voix craquelée), il existe certains sons qui empruntent à la fois aux voyelles et aux consonnes, comme les spirantes et les semi-voyelles. Intuitivement, on s'attend à ce que la structure acoustique des voyelles soit relativement simple, puisque le conduit vocal se comporte comme un résonateur stimulé par l'air passant au niveau des cordes vocales et ouvert à l'extrémité buccale. Bien évidemment, cette vision simpliste est erronée. En effet, le conduit vocal ne se réduit pas à un unique résonateur mais à un système complexe (Figure 4) où des constrictions peuvent se produire et où plusieurs paramètres de contrôle entrent en interaction. Dès lors, on peut considérer qu'une voyelle va être caractérisée par un

[12] l'écriture hébraïque et arabe sont aussi issues du proto-cananéen par l'intermédiaire de l'araméen.

nombre variable de paramètres liés à son timbre et à la présence (facultative) d'articulations supplémentaires.

Dans le cadre de la phonétique traditionnelle, le timbre de la voyelle est fonction :

- ✓ du nombre de résonateurs oraux (présence ou absence [13] du résonateur labial),
- ✓ de la forme du résonateur buccal (position antérieure, centrale ou postérieure de la masse de la langue),
- ✓ du volume du résonateur buccal (aperture faible, moyenne ou grande de la bouche).

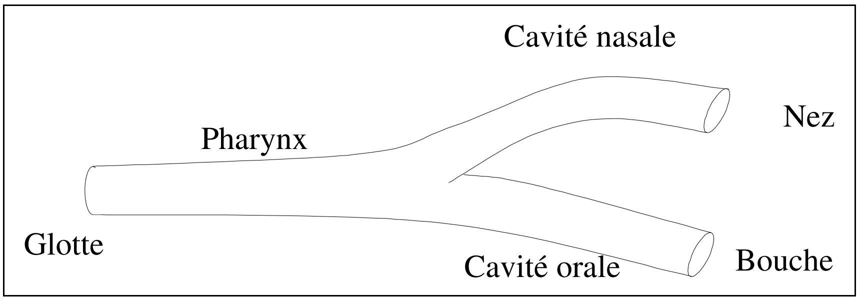

Figure 4 – Représentation schématique de l'ensemble pharynx + conduits oral et nasal [Calliope 89]

Chacun de ces paramètres résulte d'une configuration adoptée par les articulateurs de la parole. Il existe donc des limitations physiologiques aux conformations possibles tant pour le nombre de configurations intermédiaires (pour le résonateur labial, on s'accorde à ne considérer que deux positions, lèvres arrondies ou non) que pour les limites absolues : le son /a/ est le plus ouvert et les sons plus fermés que /i/ par exemple, ne sont plus des voyelles. De même, le son /u/ est le plus postérieur possible (au regard de la position de la masse de la langue) tandis que la voyelle /i/ est la plus antérieure que l'on puisse réaliser.

La représentation classique de cet espace est héritée d'une longue tradition phonéticienne [Hellwag 1781, Bell 1867, Jones 1918] qui a abouti à une représentation articulatoire sous forme d'un quadrilatère (Figure 5.a). La représentation acoustique issue des travaux de Delattre en 1948 a confirmé dans une certaine mesure sa pertinence. En effet, la représentation des voyelles dans un plan défini par les axes F_2 et F_1 (respectivement deuxième et première résonances du conduit vocal) fait émerger un triangle relativement isomorphe avec le quadrilatère de l'API (Figure 5.b). En particulier, les trois voyelles /i/ /a/ et /u/ occupent dans les deux représentations des positions extrêmes.

[13] « l'absence » de résonateur labial consiste à ne pas arrondir les lèvres.

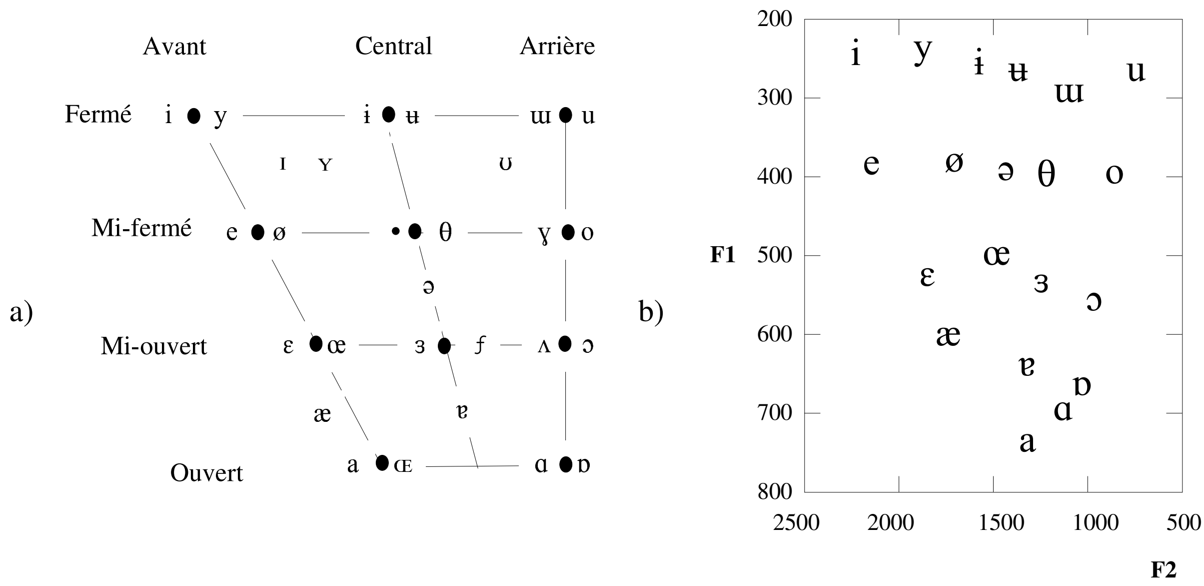

Figure 5 - a) Le quadrilatère des voyelles cardinales de l'API (version 1996)
- b) Projection de prototypes vocaliques dans F2/F1 (Hz) (d'après [Vallée 94])

Dans la représentation de l'API, lorsque deux symboles occupent le même point (par exemple /i/ et /y/), le symbole de gauche est une voyelle non arrondie alors que le symbole de droite correspond à la voyelle arrondie correspondante. Les voyelles les plus à gauche du diagramme sont donc qualifiées de voyelles antérieures non arrondies et à l'inverse, les symboles les plus à droite sont donc des voyelles postérieures arrondies. Ces voyelles, qui délimitent également le triangle vocalique acoustique sont appelées dans leur ensemble les voyelles périphériques, par opposition à toutes les autres voyelles qui se projettent à l'intérieur du quadrilatère /i, a, ɒ, u/ ou du triangle /i, a, u/.

Les travaux entrepris dès la fin des années 50 sur la théorie acoustique de la production de la parole [Stevens 55, Fant 60] et sur les possibilités d'inversion entre espaces acoustique et articulatoire ont cependant fait surgir de nouvelles interrogations. En effet, la paramétrisation du conduit vocal à partir de la fonction d'aire [Fant 60, Maeda 85] amène à une description tridimensionnelle de la totalité du conduit vocal (position X_c et aire A_c de la constriction ainsi que l'aire aux lèvres A_l) et non plus seulement à une description bidimensionnelle au niveau de la bouche (hauteur et avancement de la langue). Ainsi, pour les voyelles /ɒ/, /ɑ/, /a/ et /æ/, le lieu de constriction est pharyngal et non plus buccal. Cette représentation est donc plus fine puisqu'elle modélise mieux les relations acoustico-articulatoires et elle a, depuis, été largement attestée [Wood 79, Ladefoged 82, Maeda 90, Boë 92]. Par contre, elle présente des caractéristiques qui paradoxalement, l'éloignent de l'espace de représentation acoustique F_2/F_1 [Boë 95]. On peut noter entre autres que :

- les représentations dérivées des espaces X_c/A_c ou X_c/A_l ne sont pas isomorphes avec l'espace F2/F1 (/i/, /a/ et /u/ n'occupent pas les points extrêmes de l'espace, /u/ occupe une position intermédiaire entre /i/ et /a/…),
- lors d'un abaissement monotone de la mâchoire à partir de la voyelle /i/, l'aire de la constriction se comporte de manière non monotone, voire discontinue.

Des études récentes se sont attaquées à ce paradoxe : comment une modélisation bi-dimensionnelle peu précise et ne prenant pas toujours en compte le lieu réel de constriction du conduit vocal, peut-elle s'avérer plus proche de la réalité acoustique qu'un modèle plus fin comme celui issu des paramètres de la fonction d'aire ? Certains éléments de réponse ont été apportés en intégrant des contraintes morphologiques pour le mouvement de la langue [Boë 95]. Ces études portant sur le déplacement à volume constant de la langue ont montré que les paramètres traditionnels relatifs à la langue (haut/bas et avant/arrière) contrôlent le lieu et l'aire de la constriction, qu'elle soit palatale, vélaire ou même pharyngale.

Cette unification des espaces de représentation vocaliques soulève de nouvelles questions, en particulier du point de vue du contrôle moteur. En effet, si la corrélation entre les représentations bidimensionnelles s'explique à la lumière de ces résultats, l'intégration de contraintes articulatoires de coût et de continuité de mouvement dans le cadre de la modélisation tridimensionnelle du conduit vocal risque d'amener à prendre en compte de nouveaux paramètres.

Quelle que soit la représentation adoptée, la visualisation du système nécessite la définition d'unités vocaliques de référence. Ces prototypes, même s'ils ne correspondent pas réellement à des voyelles enregistrées, ont une réalité à la fois au niveau de la production et de la perception de la parole. En effet, le nombre de timbres vocaliques de référence est limité par une discrétisation de l'espace qui est le fait de contraintes acoustico-articulatoires [Boë 89] et de contraintes perceptives. Cet espace des *voyelles possibles* restreint donc de manière stricte le nombre de timbres vocaliques sur lesquels des contrastes perceptifs peuvent s'appuyer. L'être humain a dépassé cette restriction en faisant porter des contrastes sur des articulations supplémentaires : en présence d'une limitation physiologique sur les axes de contrôle principaux, l'homme a augmenté le nombre de dimensions utilisées. Ces traits secondaires (Tableau 3) peuvent revêtir différents aspects temporels ou articulatoires, et ils augmentent de manière considérable la taille de l'espace des voyelles possibles.

Il est intéressant de noter que cette augmentation s'accompagne d'une complexification globale de la production des voyelles puisqu'elle nécessite la mise en œuvre d'articulateurs supplémentaires. Le fait que l'être humain, au cours de l'histoire du langage, ait développé une stratégie « coûteuse » sur le plan articulatoire s'inscrit dans une logique évolutionniste où un autre aspect de la communication a bénéficié de cette stratégie. Comme nous venons de le signaler, la contrepartie positive a été l'augmentation du nombre de contrastes vocaliques possibles, et donc l'augmentation de la quantité d'information portée par les voyelles. Ainsi, il apparaît que chaque langue développe sa propre stratégie vis à vis des voyelles, en s'appuyant à la fois sur des principes articulatoires de production et sur des principes contrastifs liés à la perception de la parole.

Nom du trait	Description du trait
allongement	augmentation de la durée par tenue de la position de la langue
ultra-brièveté	diminution de la durée
nasalité	activation du résonateur nasal par abaissement du voile du palais
aspiration	superposition d'un souffle au son voisé
vélarisation	superposition à la constriction principale d'une constriction au niveau du vélum.
laryngalité	constriction laryngale effective caractérisée par des cordes vocales rapprochées et raides
pharyngalité	constriction pharyngale, élévation du larynx
dévoisement	affaiblissement de la vibration des cordes vocales
rétroflexion	rétraction de la langue, apex relevé

Tableau 3 – Principales articulations supplémentaires (d'après [Vallée 94])

Dès lors, on peut se demander pour quelles raisons chaque langue privilégie certaines distinctions ou au contraire en délaisse d'autres. Cette question phonologique sur la structure des systèmes vocaliques est-elle liée à la substance même des voyelles au sens phonétique (la manière dont elles se réalisent) ? Le clivage a longtemps été total entre ces deux aspects, mais les travaux menés par Johan Liljencrants et Björn Lindblom au début des années 70 ont montré que la prise en compte des réalisations acoustiques des voyelles est essentielle si l'on souhaite justifier l'émergence de systèmes vocaliques spécifiques.

Au cours de ce paragraphe, nous avons vu apparaître succinctement la notion de coût articulatoire, liée aux mouvements des différents agents intervenant au cours de la phonation, et la notion de contraste perceptif, liée au système auditif humain. De manière synthétique, on peut dire que pour communiquer, l'homme fait appel au principe d'efficacité fonctionnelle, qui est une optimisation de la transmission de l'information par un compromis coût/efficacité adéquat. Nous allons voir maintenant que cette notion, empruntée à la théorie de l'information, est à la base des travaux menés sur les structures des systèmes vocaliques.

1.2.2 Le système vocalique : structure & substance

Le principe d'efficacité fonctionnelle implique que l'on optimise *globalement* le processus de communication. Cette transmission de l'information se réalise successivement dans les espaces articulatoire, acoustique et perceptif.

L'être humain ne peut guère influencer le rendement de son système perceptif : lorsque nous écoutons quelqu'un dans un milieu bruyant, tout au plus pouvons nous orienter une oreille (si possible celle qui fonctionne le mieux) dans la direction de la source de la voix, et placer notre main en cornet autour du pavillon. Même si notre optimisation auditive s'arrête là, une deuxième source d'amélioration du rendement perceptif demeure possible : elle repose sur l'aspect audiovisuel de la communication parlée et nous renvoie donc à l'optimisation de l'utilisation de l'information visuelle (déplacement du regard vers la bouche du locuteur). L'espace acoustique est le canal de transmission de la voix ; généralement, si l'auditeur peut diminuer le bruit environnant (éteindre la radio…) il aura tendance à le faire, toujours dans le but d'améliorer la transmission. Si les bruits ne relèvent pas de son autorité, aucun contrôle ne peut être exercé. Il apparaît donc que l'optimisation globale de la communication se réduit généralement à une optimisation de la production de la parole. Cette optimisation doit aller dans l'intérêt de l'auditeur (minimiser les erreurs de perception auditive et visuelle) et dans celui du locuteur (minimiser le coût de production). Il est donc assez logique de considérer que les structures des systèmes vocaliques sont largement conditionnées par le principe d'efficacité fonctionnelle au niveau articulatoire.

André Martinet indiquait en 1970 que « [l'] on peut attendre des éléments distinctifs d'une langue (…) qu'ils ne se confondent pas les uns avec les autres. On peut donc supposer qu'ils tendront à être aussi différents les uns des autres que le permettent les organes qui contribuent à leur production ». Cette hypothèse que l'auteur nomme la *différenciation maxima*, l'amène à formuler une remarque de nature diachronique. Il avance en effet qu'un système phonologique évoluera vers une « équidistance entre les phonèmes ». Il tempère ensuite cette opinion en considérant les variations de prononciation d'un phonème par rapport à ce qui est la norme à une époque donnée : « Ces variations seront freinées et stoppées si elles se rapprochent dangereusement de ce qui est la norme d'un autre phonème. Elles seront tolérées si elles n'exposent jamais l'usager à ne pas être compris » [Martinet 70]. Ces critères diachroniques alimentent la notion de contraste perceptif maximal tout en autorisant une certaine paresse articulatoire. A la même époque, Stevens d'une part, et Liljencrants associé à Lindblom d'autre part vont intégrer des critères semblables dans des approches synchroniques visant à expliciter les systèmes vocaliques à partir de leur substance phonétique (respectivement la *théorie quantique* et la *théorie de la dispersion vocalique maximale*). Depuis, d'autres études « substance-based » ont été réalisées, et les modèles se sont affinés. La plupart des experts s'accordent cependant sur la nécessité de prendre aussi en compte des critères linguistiques, comme ce que Louis ten Bosch appelle la « charge fonctionnelle » de la voyelle [ten Bosch 95] et que Martinet appelait le « rendement fonctionnel ».

- **la théorie quantique**

La théorie quantique de la parole [Stevens 72] est basée sur :

- l'existence de zones articulatoires stables, où des changements articulatoires relativement importants ne mènent pas à un changement d'unité perçue par l'auditeur,
- l'existence de zones instables où un changement articulatoire rapide entraîne le passage d'une unité perceptive à une autre (existence d'un seuil perceptif).

Elle s'appuie sur l'organisation des systèmes phonologiques des langues du monde et sur des considérations plus générales que nous avons déjà formulées (les espaces articulatoires et acoustiques sont continus alors que l'espace perceptif est discrétisé). En constatant que la plupart des systèmes phonologiques basent leurs unités phonétiques sur des oppositions de traits distinctifs plutôt que sur l'augmentation systématique du nombre de dimensions [Chomsky 68], Stevens avance que ce choix est directement lié à la nature quantique de la parole et donc à la dualité stabilité/contraste des relations entre production et perception. Ses travaux ont donc porté sur l'établissement de règles sur le choix des traits discriminants à partir de considérations sur le contraste et la stabilité des sons produits. Pour les systèmes vocaliques, Stevens a établi plusieurs lois ayant trait aux positions relatives des formants F_1, F_2 et F_3 assurant le caractère quantique des voyelles. Plusieurs imperfections assez nettes demeuraient cependant sur les systèmes prédits par ces règles, et des critiques ont été formulées, à la fois sur l'aspect nécessaire du caractère quantique des voyelles et sur les difficultés à définir des zones stables vis à vis de plusieurs paramètres articulatoires. Stevens a été amené en 1989 à réévaluer à la baisse le rôle tenu par les zones de stabilités perceptives et à privilégier l'existence de zones de contraste pour expliquer la formation des systèmes vocaliques. A l'heure actuelle, la théorie quantique ne permet pas de prédire de manière efficace certaines tendances des langues du monde. En effet, les raisons amenant un système à développer une voyelle quantique plutôt qu'une autre restent obscures. Par exemple, le fait que le son /y/ soit beaucoup moins fréquent que le son /i/ ne s'explique pas dans le cadre de cette théorie. De plus, il semble que certaines voyelles non quantiques soient préférées à d'autres qui, elles, le sont. Il semble donc nécessaire d'intégrer d'autres considérations que la présence de discontinuités perceptives correspondant aux traits articulatoires distinctifs pour expliquer et prédire les différents systèmes vocaliques des langues du monde.

- **la théorie de la dispersion maximale**

A l'époque où Stevens ébauche la théorie quantique au MIT, Liljencrants et Lindblom développent en Europe la théorie de la dispersion maximale basée sur le postulat suivant : un système phonologique est optimal s'il maximise la distance perceptive entre ses unités de référence. Dans cette approche, le choix des unités se fait donc par une optimisation globale du système résultant. La première difficulté a donc été d'obtenir une mesure de distance dans l'espace perceptif. Cette étape a été franchie en exploitant la notion de *voyelle possible* dans l'espace articulatoire, puis en transposant cet espace au niveau acoustique grâce à un modèle de production. Le passage à l'espace perceptif a consisté en un changement d'échelle, de manière à passer d'une échelle formantique spectrale linéaire à une échelle perceptive de Mel.

En intégrant les contraintes de maximisation des distances intra-systémiques, la théorie de la dispersion maximale prédit de manière correcte les systèmes vocaliques les plus courants possédant jusqu'à 6 voyelles. Au delà de ce nombre, le modèle a tendance à prévoir des voyelles intérieures hautes (entre /i/ et /u/) plutôt que des voyelles intérieures centrales comme cela est le cas en réalité. Différentes études perceptives sur le rôle prépondérant du premier formant et sur la perception des intensités (plus proche d'une échelle logarithmique que linéaire) ont mené à des améliorations sensibles du modèle [Lindblom 75, Bladon 81, Lindblom 86]. La prise en compte par Crothers en 1978 de cercles vocaliques (intégrant une notion de dispersion pour chaque voyelle) plutôt que d'unités vocaliques ponctuelles comme références permet de simuler efficacement les systèmes réels en optimisant les distances entre les sous-espaces vocaliques correspondant à chaque timbre. Ce modèle se révèle efficace, puisque l'optimisation des systèmes possédant plus de 4 timbres vocaliques aboutit à l'émergence de voyelles intérieures.

- **la théorie de la dispersion adaptative**

Au milieu des années 80, Lindblom va apporter une modification majeure à son modèle en concédant que les systèmes vocaliques ne recherchent pas forcément le contraste maximal, mais plutôt à maintenir un contraste suffisant. Cette nouvelle théorie, dite du *contraste perceptif suffisant*, implique qu'il existe d'autres contraintes qui régissent la formation des systèmes phonologiques. Ces contraintes tendent à privilégier des contrastes basés sur des traits distinctifs portés par un nombre réduit d'axes plutôt que sur des dimensions articulatoires supplémentaires (nasalité, pharyngalité) comme le montrent les études de Ian Maddieson en 1986. Ce principe, tout comme dans le cadre quantique, tend vers une économie articulatoire que Lindblom modélise en intégrant un coût articulatoire à ses équations dans le cadre d'une nouvelle théorie, la théorie de la *dispersion adaptative*. Ce coût articulatoire est calculé à la fois sur des critères statiques (minimisation de l'effort pour atteindre la cible articulatoire à partir d'une position neutre) et dynamiques (minimiser les distances parcourues lors du passage d'un phonème à un autre).

Si la théorie quantique avait tendance à optimiser l'efficacité fonctionnelle des systèmes vocaliques en prenant en compte uniquement des critères locaux (discontinuités perceptives entre voyelles), les théories issues des travaux de Liljencrants et Lindblom sur le contraste maximal privilégient plutôt une optimisation globale du système. Ces deux approches se rejoignent pour considérer la notion de contraste perceptif comme étant au centre de la prédiction des systèmes vocaliques. Elles atteignent cependant leurs limites dès lors que le nombre de timbres vocaliques des systèmes dépassent 6 ou 7. La fin des années 80 verra apparaître une nouvelle théorie qui intégrera à la fois des critères d'optimisation globale du système et des contraintes acoustico-articulatoires locales sur les voyelles qui le composent : la *théorie de la dispersion-focalisation*.

- **la théorie de la dispersion-focalisation (TDF)**

Cette théorie est apparue en 1989 à l'Institut de la Communication Parlée de Grenoble [Schwartz 89, Vallée 94]. Elle se base sur les différentes évolutions de la théorie de la dispersion maximale, tout en cherchant à améliorer les prédictions obtenues.

La première source d'amélioration est la prise en compte de l'importance spécifique du premier formant F_1 par rapport aux formants suivants. Considérant que, même s'il est avéré au niveau perceptif que F_1 joue un rôle prépondérant par rapport à F_2, la pondération proposée par Lindblom est inefficace, Jean-Luc Schwartz et al. proposent de pondérer F_1 dans l'espace tridimensionnel des trois premiers formants plutôt que dans le simple espace F_1/F_2. La prise en compte des distances entre les différents formants dans une échelle perceptive de Bark a ensuite permis de se ramener à une optimisation du système par dispersion dans un espace à 2 dimensions F_1/F'_2. F'_2 est appelé second formant effectif et il est obtenu par pondération de F_2 par F_3 et éventuellement par F_4. Cette pondération des différents formants a pour but, comme le suggère Lindblom, d'éviter la prolifération de voyelles prédites entre /i/ et /u/.

La seconde source d'amélioration porte sur un critère de stabilité intrinsèque de chaque voyelle. Ce critère, appelé *focalisation*, est basé sur diverses études à la fois acoustico-articulatoires et perceptives. Les conclusions sont d'une part que les voyelles présentant une zone fréquentielle où plusieurs formants sont proches sont préférées aux autres [Boë 86] et d'autre part que ces mêmes voyelles sont préférées au niveau perceptif [Schwartz 89b]. Le critère de focalisation retenu est calculé dans l'espace tridimensionnel $F_1/F_2/F_3$ et il privilégie l'émergence de voyelles focales dans les systèmes vocaliques. Il permet de prédire la présence du /y/ dans des systèmes où /i/ et /u/ sont présents, bien que les distances entre /i/ et /y/ soient faibles dans le plan F_1/F'_2.

La TDF est donc basée sur une minimisation globale du coût du système, estimé à la fois par des contraintes contrastives globales intervocaliques et des contraintes de stabilité intra-vocaliques. Il s'agit là d'une synthèse de critères perceptifs et acoustico-articulatoires, à la fois globaux et locaux, pondérés par deux paramètres, l'un réglant la pondération de F_1 par rapport à F'_2 et le second réglant l'influence du facteur de focalisation par rapport au facteur de dispersion. Le lecteur trouvera une description plus complète de cette théorie ainsi que les résultats des expériences en prédiction dans [Vallée 94, Schwartz 97].

Les différents systèmes qui ont été présentés jusqu'à présent ont tendance à donner – ou à redonner – au critère de stabilité des voyelles un rôle fondamental dans la prédiction des systèmes vocaliques. Une autre approche déductive, développée par René Carré et Mohamad Mrayati tend à montrer qu'un modèle contrastif, à la fois sur le plan acoustique et articulatoire, peut être suffisant pour justifier l'émergence de certains systèmes.

- **la prédiction des systèmes vocaliques à partir du modèle DRM**

La théorie proposée dans [Carré 95] est basée sur le modèle de production DRM (Distinctive Region Model). Ce modèle est un modèle de tubes possédant deux configurations (fermé-fermé dans le cas d'une constriction centrale, fermé-ouvert dans le cas d'une constriction antérieure ou postérieure). Il est commandé par trois paramètres (position de la constriction, degré de constriction et aire aux lèvres). Le constat de Carré et Mrayati est que pour passer d'un système de n voyelles à un système de $n+1$ voyelles, on assiste dans les théories antérieures à une réorganisation totale du système vocalique, de manière à rétablir le contraste maximal (ou suffisant). En s'appuyant sur des simulations de transitions vocaliques selon deux modes de commande (longitudinal si la constriction est déplacée dans le tube, transversal si les aires des différentes régions du tube sont modifiées), et en considérant que la règle de contraste suffisant s'applique dans l'espace de commande plutôt que dans l'espace perceptif, les auteurs de cette théorie construisent les systèmes vocaliques les plus probables de manière incrémentale en respectant à la fois le critère de contraste acoustique et le critère de simplicité articulatoire.

- **Discussion**

Les différents modèles présentés dans ce paragraphe ont tous apporté des connaissances supplémentaires sur la structure des systèmes vocaliques. L'intégration de contraintes locales, liées à la réalisation de chaque voyelle, et de contraintes globales sur la distinctivité du système global semble être l'approche la plus prometteuse. A l'heure actuelle, il paraît certain que la prédiction des systèmes vocaliques s'inscrit dans un cadre synchronique propice à expliquer les lois régissant leur émergence mais peut-être aussi dans un cadre diachronique plus apte à révéler – peut-être en intégrant des contraintes linguistiques – des règles d'apparition ou de disparition de contrastes.

Quelle que soit l'approche envisagée, ces modèles de prédiction s'appuient sur d'importantes données de manière à valider les systèmes produits. A l'heure actuelle, la base de données de référence est UPSID [Maddieson 84]. Si les théories de prédiction en ont largement tiré profit, d'autres axes de recherches ont aussi exploité cette manne phonologique, comme en atteste l'établissement par Nathalie Vallée d'une typologie des systèmes vocaliques en vue de la recherche d'universaux du langage.

2 UNE TYPOLOGIE DES SYSTEMES VOCALIQUES : [MADDIESON 84, VALLEE 94]

2.1 La base de données UPSID

La base de données UPSID (UCLA Phonological Segment Inventory Database) consiste, dans sa version initiale, en une description phonologique de 317 langues du monde [Maddieson 84]. Elle a depuis été étendue à 534 langues dans la version la plus récente (SUPERB UPSID [Lindblom et al. 92]). Le résultat est, pour reprendre la formule de Ian Maddieson traduite par Vallée, « une fenêtre par laquelle il est possible d'entrevoir un état actuel des langues du monde ». Si la première version d'UPSID a demandé environ six ans pour être publiée, elle reposait en partie sur les archives phonologiques du *Stanford Project on Language Universals* que l'on doit à Greenberg. L'objectif de la base UPSID est donc de fournir un matériau de référence dans la recherche d'universaux du langage. Une telle ambition passe nécessairement par un choix particulièrement réfléchi des langues étudiées, de manière à fournir un échantillon représentatif de la diversité des langues. Dans la version originale de la base, vingt familles de langues sont représentées [Maddieson 86] ou plus exactement 17 familles et 3 isolats linguistiques. La classification proposée par Maddieson s'accorde en bonne partie avec celle développée par d'autres auteurs et les divergences observées par rapport aux classifications de Greenberg, Ruhlen ou Ross ne biaisent pas véritablement la représentation des langues du monde obtenue. Pour chaque famille, au moins une langue par sous-famille ou par branche a été prise en compte avec un critère de distance génétique retenu pour la distinction des branches fixé à 1500 ans. Cette valeur élevée nous ramène à la séparation entre les sous-familles germaniques nordique (islandais, danois…) et occidentale (allemand, anglais…). La prudence méthodologique adoptée a pour but de réduire les similarités liées à la parenté génétique de manière à ne pas biaiser les statistiques sur les universaux du langage. Le lecteur trouvera dans le Tableau 4 un résumé des familles présentes ainsi que leur nombre de représentants dans UPSID$_{317}$.

Pour chaque langue de la base de données, une description phonologique du système vocalique et du système consonantique est fournie. Pour le système vocalique, cette description retient sept degrés d'aperture (fermé à ouvert), trois positions d'articulation (avant, central, arrière) et un indice de protrusion des lèvres (arrondies vs. non arrondies) pour qualifier le timbre vocalique. A cela s'ajoutent si nécessaire des symboles diacritiques caractérisant les contrastes supplémentaires (Tableau 3). Plusieurs reproches ont été adressés aux concepteurs d'UPSID en particulier sur la validité des unités phonétiques choisies et sur l'homogénéité des descriptions. Il est bien entendu que cette base de données a été réalisée à partir de transcriptions auditives réalisées par des experts différents sur un intervalle de temps assez important. Cela introduit bien évidemment des variations au niveau des descriptions obtenues pour des raisons à la fois perceptives (chaque être humain, fût-il un linguiste expert est plus sensible à certains contrastes qu'à d'autres [Mehler 95]) et linguistiques (les connaissances phonétiques *a priori* de l'expert agissent comme une projection de la réalité acoustique dans un espace personnalisé [Kingston 91]). Il en résulte une relative imprécision des descriptions des sons lorsqu'ils ne coïncident pas avec l'inventaire phonétique personnel du transcripteur. Une autre critique a été formulée sur le choix de l'allophone retenu comme unité phonologique. L'approche retenue dans UPSID donne la priorité à l'allophone le plus courant, qui est donc statistiquement le plus valide. Lorsque par contre, il n'y a pas de forme réellement dominante, l'unité retenue est la position médiane du nuage d'observations. On peut alors reprocher à cette méthode d'introduire artificiellement des unités n'ayant pas de réalité acoustique mais là encore plutôt une réalité statistique. Ces critiques n'enlèvent aucunement à la base UPSID sa valeur, et il nous semble que dans toute entreprise humaine de collecte de données, ces mêmes limitations apparaissent, liées tant à la variabilité des phénomènes humains qu'au biais introduit par l'expert. A notre avis, UPSID est une base de données considérable, et ses limitations intrinsèques doivent être intégrées par les utilisateurs lors de son exploitation.

Famille	Nombre de représentants	%
Khoisan	2	0,63
Nigéro-kordofanien	31	9,78
Nilo-saharien	21	6,62
Afro-asiatique	21	6,62
Dravidien	5	1,58
Burushasky (isolat)	1	0,32
Caucasien	3	0,95
Indo-européen	21	6,62
Basque (isolat)	1	0,32
Ouralo-altaïque	22	6,94
Aïnou (isolat)	1	0,32
Paléo-sibérien	4	1,26
Eskimo-aléoute	2	0,63
Sino-tibétain	18	5,68
Austro-thaï	25	7,89
Austro-asiatique	6	1,89
Indo-pacifique	26	8,20
Australien	19	5,99
Nord-amérindien	51	16,09
Sud-amérindien	37	11,67
TOTAL	317	100

Tableau 4 – Répartition des langues d'UPSID par famille, en nombre et en pourcentage (d'après [Vallée 94]).

2.2 Une typologie des systèmes vocaliques

L'approche typologique en linguistique est issue d'une longue tradition taxinomique héritée des sciences expérimentales. Nous avons vu (Chapitre 1, paragraphe 3) que c'est d'abord au niveau de la morpho-syntaxe que les propriétés des langues ont été étudiées, et que les typologies phonologiques sont bien plus récentes. Bien évidemment, établir une typologie n'est pas une fin en soi, mais cette étape se révèle indispensable pour atteindre les tendances universelles du langage et ainsi mieux connaître la faculté langagière humaine. Ce paragraphe est consacré à une typologie des systèmes vocaliques établie à Grenoble par Vallée dans le cadre de sa thèse : *« Systèmes vocaliques : de la typologie aux prédictions »* et il doit donc beaucoup à son manuscrit.

2.2.1 Travaux antérieurs

Ce vingtième siècle aura vu naître l'idée de typologie des systèmes vocaliques, et plusieurs études le jalonneront. On peut bien évidemment citer comme première pierre à l'édifice la note de Troubetzkoy : « J'ai mis au net tous les systèmes vocaliques que je connaissais par cœur (34 en tout) et j'ai essayé de les comparer les uns aux autres […]. Les résultats sont extrêmement curieux [….]. Tous les systèmes se réduisent à un petit nombre de types et peuvent être représentés par des schémas symétriques […]. Plusieurs lois "de la formation des systèmes" se laissent dégager sans peine […]. Je crois que les lois empiriques acquises ainsi seront d'une grande importance […]. Elles devront être applicables à toutes les langues, aussi bien aux langues mères (Ursprachen) reconstruites théoriquement qu'aux divers stades de développement des langues historiquement attestées. »

Les différentes études réalisées depuis cette époque ont fait émerger des tendances différentes bien que portant généralement sur deux points :

- la fréquence d'occurrence de certaines voyelles ou de certains sous-systèmes vocaliques dans les langues du monde (fréquence du système /i a u/…),
- l'existence de « règles implicationnelles » sur la formation des systèmes (la présence d'une voyelle V_1 implique la présence d'une voyelle V_2 dans le système).

Si l'on précise le cadre de chacune des études menées, il n'y a aucun paradoxe à la non concordance absolue entre les conclusions : d'une part le nombre de langues considérées n'a fait que croître pour passer des 34 langues citées par Troubetzkoy à 203 il y a vingt ans [Crothers 78] et plus de 300 dans les années 80 et 90 [Maddieson 84, Vallée 94] ; d'autre part, des différences méthodologiques assez nettes sont observées, en particulier sur les critères de définition des types retenus (sous-systèmes organisés selon l'aperture pour Hockett ou sous-systèmes périphériques et intérieurs pour Crothers).

Un autre point méthodologique ajoutant à la difficulté à recouper les différents résultats est la prise en compte de règles de regroupement des timbres vocaliques de manière à diminuer le nombre de types de systèmes ou plus exactement, à augmenter le nombre de représentants de chaque type. Crothers considère par exemple que le système /i a o ɯ/ est du type / i a o u/. Hockett était parfois amené à considérer pour le / e/ un degré d'aperture à part entière ou à le regrouper avec le / a/. Il est clair qu'un tel filtrage – il s'agit là encore d'un filtrage par expertise – conditionne le degré de précision et même la nature des universaux obtenus. Dans le même ordre d'idée, les articulations supplémentaires superposées aux timbres de base ont rarement été prises en compte (Hockett sépare les sous-systèmes de voyelles longues ou nasales ; Crothers identifie toute voyelle au timbre de base correspondant, faisant souvent disparaître des oppositions essentielles…). Dans le cadre de sa thèse, Vallée disposait d'un matériau plus conséquent que ses prédécesseurs, et nous allons maintenant voir quels choix méthodologiques ont été les siens.

2.2.2 Choix méthodologiques

En se basant sur les observations formulées par Lass sur les typologies des systèmes vocaliques existantes [Lass 84], Vallée a mis a profit la richesse d'UPSID pour prendre en compte le maximum d'oppositions vocaliques. Toutes les voyelles ont donc été prises en compte (qu'elles possèdent zéro, une ou plusieurs articulations supplémentaires) sans qu'aucune équivalence typologique ne soit réalisée. Par exemple, les systèmes à trois voyelles /i a ɯ/, /i æ u/ et /i a u/ constituent trois types différents. Lorsque des diphtongues sont présentes, elles sont aussi prises en compte, sous la forme des deux segments vocaliques reliés par une flèche.

La désignation de chaque type de système vocalique se fait par deux nombres séparés d'un point. Le premier nombre est la taille du système en prenant en compte toutes les voyelles, et le second nombre est en fait un ordre lié à la fréquence du système dans UPSID. Par exemple, le système référencé 5.1 est le système le plus fréquent comportant 5 voyelles.

La représentation des systèmes se fait au moyen d'une grille (Figure 6) issue de la représentation phonétique traditionnelle. Comme l'explique Vallée : « notre "espace phonologique" est en quelque sorte une *idéalisation* de la coupe sagittale d'un conduit vocal, la face orientée vers la gauche, de telle sorte que les voyelles d'aperture fermée soient situées en haut de la grille, les voyelles d'aperture ouverte en bas, les voyelles articulées à l'avant de la cavité buccale (voyelles antérieures) à gauche et les voyelles articulées plus en arrière (postérieures) à droite ; les voyelles arrondies étant placées à coté des non arrondies partageant la même aperture et la même localisation sur l'axe antéro-postérieur. »

Sur cette grille et pour chaque système, un point noir situe les symboles présents dans le système, augmenté si nécessaire de symboles diacritiques. Lorsqu'elles existent dans le système, les diphtongues sont aussi représentées sur le même schéma lorsque cela ne rend pas la lecture impossible.

Prenons comme exemple le système le plus représenté dans UPSID : /i 'e' a 'o' u/. Il sera désigné 5.1 (système le plus fréquent à 5 voyelles) et représenté par la grille Figure 7.a alors qu'un système plus complexe à huit voyelles avec opposition de longueur et deux diphtongues comme celui du dagomba, parlé au Ghana sera représenté par la grille Figure 7.b.

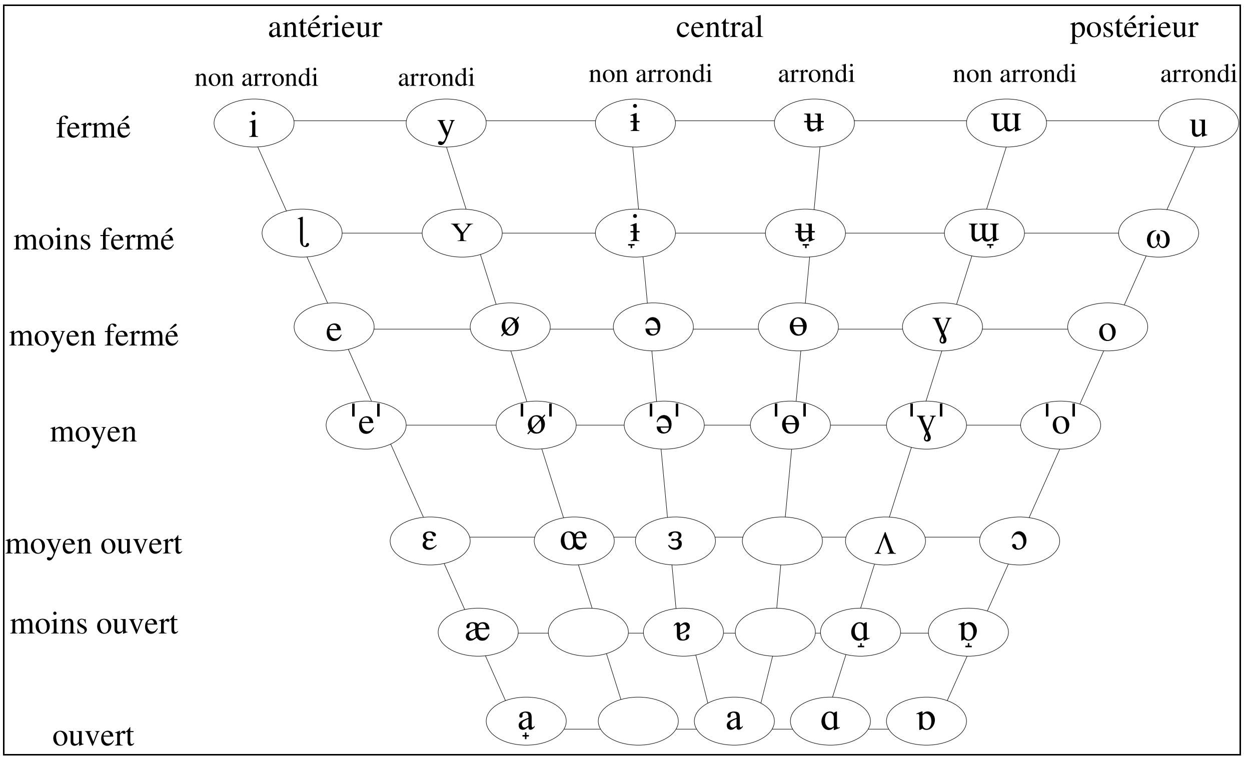

Figure 6 – Description des 37 symboles vocaliques d'UPSID [Vallée 94]

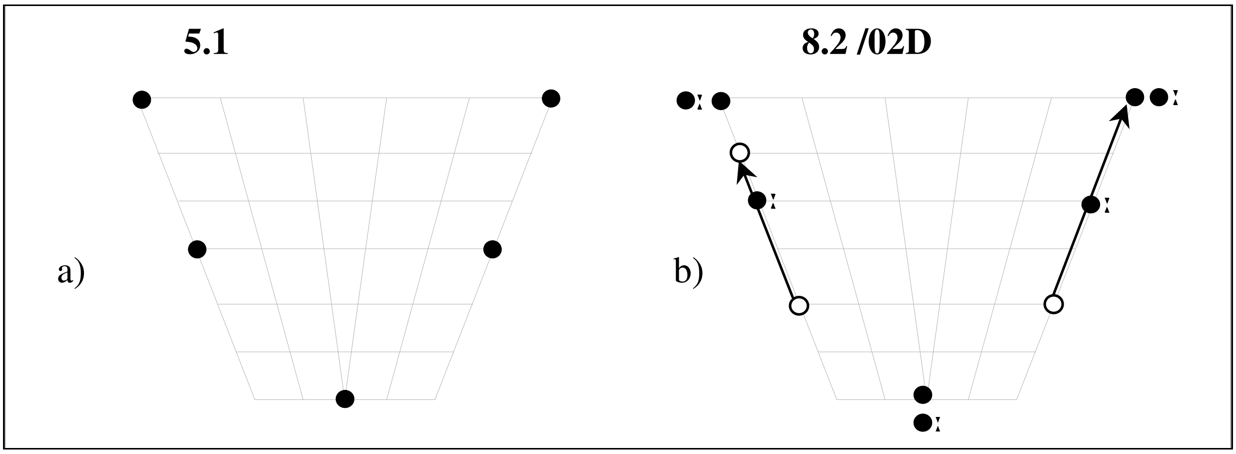

Figure 7 – Exemples de systèmes vocaliques a) Système à cinq voyelles le plus fréquent
b) Système à 8 voyelles et 2 diphtongues
Les cercles vides o correspondent aux qualités vocaliques absentes du système des monophtongues.

2.2.3 Résultats

Le propos de ce paragraphe n'est pas de reprendre toutes les conclusions formulées par Vallée sur les systèmes vocaliques issus de sa typologie. Nous noterons cependant quelques statistiques intéressantes ainsi que quelques tendances universelles qui s'en dégagent.

Les 317 langues présentes dans UPSID se répartissent en 219 systèmes vocaliques distincts possédant entre 3 et 20 voyelles (une langue khoisane fait figure d'exception avec 24 voyelles et 22 diphtongues…). 23,3 % des langues d'UPSID possèdent 5 voyelles, et plus de 80 % en possèdent au plus 10. Si l'on s'intéresse de plus près à la substance des systèmes, il apparaît que tous les systèmes reposent sur au moins 3 voyelles périphériques alors que seulement 44,4 % d'entre eux possèdent au moins une voyelle intérieure. Le sous-système /i a u/ est présent dans plus de 90 % des langues, et si l'on prenait en compte les regroupements des systèmes proches comme le proposait Crothers, ce triangle serait quasiment universel. Les voyelles suivantes les plus fréquentes sont /'e'/ et /'o'/, voyelles moyennes antérieure non arrondie et postérieure arrondie ; la voyelle intérieure la plus fréquente est quant à elle le /'ə'/.

De manière générale, pour les systèmes possédant jusqu'à 5 segments vocaliques, aucune articulation supplémentaire n'est utilisée dans le cadre d'oppositions vocaliques. Cette tendance décroît lorsque l'on s'approche de la limite de 9 voyelles, et à partir de 10 segments vocaliques, il y a au moins une dimension supplémentaire qui apparaît (principalement la nasalité, présente dans 22,4 % des langues), voire plusieurs, pouvant porter sur les mêmes unités (apparition de voyelles nasales longues par exemple). L'exploitation de cet inventaire statistique à la Prévert permet de faire émerger plusieurs tendances universelles intéressantes.

D'une part, et cela confirme dans une certaine mesure les travaux antérieurs, on retrouve le système le plus fréquent à *n* voyelles dans le système le plus fréquent à *n+1* voyelles. Cette tendance, facile à observer pour un nombre faible de voyelles est mise en défaut pour les passages de 6 à 7 et de 8 à 9 voyelles. En effet, dans ces cas précis, on assiste à une réorganisation du système vocalique « afin de préserver la loi de symétrie et la disposition des voyelles sur la périphérie du triangle vocalique ».

D'autre part, la préférence générale des langues va vers des systèmes simples, employant moins de 10 segments vocaliques. Si cette tendance est respectée, les systèmes privilégient les voyelles périphériques, en particulier pour des systèmes possédant moins de 5 voyelles. Si par contre elle n'est pas vérifiée, la stratégie adoptée consiste à ajouter une dimension supplémentaire plutôt qu'à surcharger l'espace vocalique primaire (le nombre de timbres vocaliques n'augmente pas). La dimension généralement choisie est la nasalité, même si les travaux de Vallée montrent que dans ce cas, peu de timbres nasalisés sont utilisés (généralement les 3 voyelles nasales issues de /i a u/).

La seconde articulation la plus utilisée est l'allongement (présent dans 16,6 % des langues), et la tendance, lorsque le nombre de voyelles devient très important est à la superposition de plusieurs traits distinctifs sur les timbres primaires.

2.2.4 Discussion

Les conclusions tirées de la typologie des systèmes vocaliques élaborée par Vallée sont bien entendu provisoires et l'exploitation de ladite typologie reste plus que jamais à l'ordre du jour [Schwartz 97]. Néammoins, le vaste travail entrepris nous inspire déjà quelques réflexions.

La méthodologie employée permet d'obtenir une typologie d'une précision sans doute jamais atteinte auparavant. Au vu des résultats, il semble que l'on puisse dégager des tendances universelles absolues d'une part (prédominance des sons « extrêmes » comme le /i/ ou le /u/, tendance générale à augmenter le nombre de dimensions articulatoires lorsque le nombre de contrastes vocaliques augmente…) et des tendances peut-être moins « impérieuses » puisqu'elles sont rendues universelles par leur prédominance statistique (prédominance des systèmes à cinq voyelles…) d'autre part. Cela signifie-t-il alors que les systèmes vocaliques des langues du monde sont issus de stratégies différentes ou que d'autres paramètres sont à prendre en compte ? La prise en compte d'autres phénomènes est bien évidemment nécessaire (paramètres lexicaux, sociologiques, historiques…), mais elle ne suffira peut-être pas à déterminer si plusieurs stratégies d'optimisation de la communication parlée ont été développées par l'homme ou non. On peut aussi penser que les systèmes vocaliques obtenus peuvent être exploités en opérant des regroupements de types *a posteriori*, par opposition aux regroupements opérés par Crothers par exemple de manière *a priori*. En effet, il est tout à fait plausible que les travaux à venir sur ce thème fassent émerger encore de nouveaux universaux.

Une autre remarque sur les résultats établis par Vallée repose sur les 219 systèmes vocaliques recensés. Ils ont été établis pour les 317 langues concernées grâce à des descriptions collectées par des linguistes. Il est évident qu'il s'agit là de spécialistes capables de percevoir et de quantifier des contrastes qui échapperaient au commun des mortels. Il serait à notre avis intéressant de réaliser des expériences perceptives d'identification de langues ayant des systèmes vocaliques proches (au sens typologique), de manière à étudier si le contraste perceptif persiste chez des auditeurs « candides » ou non. Ce type d'expérience est par nature difficile à réaliser, car le nombre de paramètres perceptifs mis en jeu est difficile à régler, comme nous allons le voir au cours du chapitre suivant, consacré à l'identification des langues par l'homme.

CONCLUSION

Ici s'achève cette première partie, placée sous le signe de la variété ! En effet, nous avons vu que, tout en possédant les mêmes capacités physiologiques et cognitives, les différentes communautés humaines ont développé des langages d'une grande diversité. Cette diversité a pour corollaire la difficulté à définir ce qu'est une langue par opposition à un parler ou un dialecte. Deux idiomes sont considérés comme des langues distinctes si elles sont suffisamment contrastées. Cette notion de contraste est donc essentielle en linguistique, et elle se base sur la caractérisation des langues.

Chaque langue s'est construite et a évolué sous l'influence de nombreux paramètres, allant des facteurs liés à l'optimisation de la transmission de l'information (règles d'évolution phonétiques) jusqu'à des pressions d'ordre social (accents, emprunts aux langues étrangères). La notion majeure de parenté reste d'actualité, et il semble qu'une vingtaine de grandes familles linguistiques rendent compte des relations assez récentes entre langues (jusqu'à 6000 ans avant J.C. environ). Nous avons cependant vu au cours du second chapitre que ces familles ne s'appuient pas sur la diversité phonologique du langage, et que les typologies basées sur la structure des langues font apparaître des classes linguistiques bien différentes. Même si des langues sont apparentées, les traits qui les distinguent peuvent être divers, et bien souvent ils relèvent de plusieurs niveaux (de la phonologie à la syntaxe). A l'heure actuelle cette diversité peut être étudiée à partir de bases de données conséquentes (UPSID pour la phonologie) et des tendances universelles se dégagent. Les expériences cognitives révèlent pour leur part que l'être humain est capable d'extraire du signal des traits discriminants qui peuvent relever de plusieurs niveaux comme le montrent les commentaires des auditeurs des expériences rapportées par Muthusamy.

Le travail présenté dans ce manuscrit s'articulant autour des systèmes vocaliques des langues, nous avons choisi de développer cet aspect phonologique de la caractérisation des langues. Les timbres vocaliques offrent un espace de représentation homogène, que ce soit dans l'espace acoustique ou articulatoire. La représentation discrète des voyelles sous forme de quadrilatère est basée sur une réalité perceptive, et la typologie des systèmes vocaliques qui en découle prend en compte des segments phonologiques enregistrés à cette fin. Les voyelles correspondantes n'ont donc généralement pas subi de phénomène de coarticulation, comme c'est le cas en parole continue spontanée. Les travaux que nous avons mené sur la détection des voyelles (1er chapitre de la 3ème partie) apportent un éclairage nouveau sur ce sujet.

(1606-1616) - © Heritage Map Museum

2ème Partie

Des ordinateurs et des langues

INTRODUCTION

Dès l'avènement des grands systèmes informatiques au cours des années 50, l'homme a caressé le rêve de parler aux machines. Si pendant de longues années, les progrès intervenus dans le domaine des Interfaces Homme-Machine (IHM) ont été confinés au cœur des laboratoires, l'explosion de la micro-informatique a permis à ce rêve de se réaliser, au moins de manière partielle. Cette réalité est née des perfectionnements considérables établis en informatique et en Traitement Automatique de la Parole (TAP).

Le TAP est un enjeu majeur du domaine des Interfaces Homme-Machine et du dialogue assisté par ordinateur. Quel que soit le type d'application envisagé, la finalité d'un système de TAP est de faciliter une tâche de l'utilisateur. Il peut s'agir de pallier à une déficience physique de cette personne, de lui permettre de travailler plus efficacement par une utilisation conjointe de plusieurs modalités, d'extraire rapidement des informations à sa place, de minimiser son effort lorsqu'il doit utiliser une machine ou encore tout simplement d'augmenter son confort. Même si l'on peut parfois s'interroger sur la notion de progrès sous-tendue par l'utilisation massive de technologies vocales[14], il reste évident que, dans la majorité des cas, il ne s'agit pas de superflu mais bien de progrès essentiels. Cela fait maintenant près d'un demi-siècle que la réalité du TAP rejoint petit à petit la science-fiction, et à l'aube du XXIème siècle, de nouveaux enjeux voient le jour : les domaines d'application du TAP s'étendent et s'orientent globalement vers une *diversification* : on souhaite traiter des énoncés plus variés, prononcés par des locuteurs inconnus, dans des conditions non standard, et de plus en plus, on désire permettre au système de fonctionner dans plusieurs langues : les systèmes de TAP de demain devront fonctionner dans un environnement multilingue.

[14] Serait-ce réellement un progrès de demander à un robot qu'il vous apporte une bière belge ou une boisson gazeuse américaine plutôt que d'aller la chercher au réfrigérateur ?

Nous nous intéressons, au cours du premier chapitre, aux enjeux de la multilingualité en traitement de la parole. Il nous semble que parmi les défis envisagés, l'Identification Automatique des Langues (IAL) joue un rôle particulièrement important : il s'agit dans bien des cas du premier pas vers la multilingualité. Le second chapitre présente un état de l'art du domaine de l'IAL sur lequel nous nous appuyons pour analyser les tendances qui se sont dégagées à l'issue de vingt-cinq années de recherche. Le troisième et dernier chapitre propose – à partir de ces tendances et des limitations actuellement rencontrées – une réflexion sur les perspectives qui s'offrent aux chercheurs en IAL. Nous introduisons alors une approche originale et prometteuse, la Modélisation Phonétique Différenciée, qui constitue le cadre d'étude de cette thèse et qui a pour but une meilleure exploitation des caractéristiques acoustico-phonétiques des langues. Nous concluons en proposant une évaluation de cette approche dans le cadre de la modélisation des systèmes vocaliques.

Chapitre 1

UN CADRE MULTILINGUE POUR LE TRAITEMENT AUTOMATIQUE DE LA PAROLE

Le traitement automatique de la parole est un domaine particulièrement complexe, où les performances atteintes dépendent grandement des conditions d'utilisation. Les recherches menées durant de longues années ont permis d'obtenir des conditions de fonctionnement étendues, tout en améliorant le niveau de performance. Les applications disponibles actuellement dans le commerce sont considérablement plus souples que les systèmes de la génération précédente. Cette vulgarisation implique d'augmenter encore l'adaptabilité des applications, la demande pour des systèmes moins spécifiques (indépendants du vocabulaire, indépendants du locuteur et indépendants de la langue) étant très forte. Si l'on sait que l'augmentation de la taille des vocabulaires traités passe par une modélisation plus exhaustive de celui-ci, ou encore que pour obtenir un système indépendant du locuteur il peut-être nécessaire d'adapter les modèles appris à chaque nouveau locuteur (ou de rapprocher l'énoncé du locuteur inconnu des modèles appris), qu'en est-il pour adapter un système à un contexte multilingue ?

Pour répondre à cette question, nous allons examiner les contraintes qui s'ajoutent aux difficultés habituelles du TAP dans un cadre multilingue, que ce soit avec des applications intrinsèquement multilingues ou lorsque l'on veut dépasser la dépendance vis à vis de la langue dans des applications classiques (paragraphe 1). Ces constatations nous amèneront naturellement à introduire la notion d'Identification Automatique des Langues (paragraphe 2) avant de la développer dans le chapitre suivant.

1 LE TRAITEMENT AUTOMATIQUE DE LA PAROLE DANS UN CADRE MULTILINGUE

Un premier aspect que l'on peut considérer comme intrinsèquement multilingue est lié à l'utilisation croissante de l'informatique dans le cadre de l'enseignement de langues étrangères. Les méthodes consistant à quantifier le degré d'apprentissage d'une langue étrangère L2 par un locuteur ayant pour langue maternelle L1 sont nombreuses [Botha 97, Goddijn 97, Kumpf 97]. A l'heure actuelle, ces systèmes s'intéressent surtout aux caractéristiques acoustiques et prosodiques des énoncés produits par l'élève, même s'ils abordent parfois l'aspect morpho-syntaxique. Ce domaine encore marginal est pourtant amené à se développer, en partie par la volonté politique de promouvoir l'enseignement assisté par ordinateur.

Une autre application multilingue amenée à se développer est la traduction automatique. Il s'agit là d'un domaine qui dépasse largement le simple cadre du TAP et qui mobilise un nombre considérable de chercheurs et d'industriels dans le monde. On peut citer par exemple le consortium C -STAR II (Consortium for Speech Translation Advanced Research) qui rassemble plus de vingt partenaires de dix pays [15]. Chaque partenaire a pour tâche de concevoir un système acceptant la langue de son pays en entrée et au moins une des langues des autres partenaires en sortie [Waibel 96] dans une tâche de planification de voyage. Par rapport aux systèmes de traduction à partir du texte, la prise en compte d'énoncés oraux augmente nettement la difficulté de la tâche. Au bilan, l'énoncé à traduire est altéré voire dégradé par les différents niveaux de traitement, de la reconnaissance acoustico-phonétique à la compréhension. A l'heure actuelle, il est vain d'espérer un système de traduction efficace quel que soit le contexte, mais les applications pour des systèmes de traduction intégrant reconnaissance et synthèse vocale dans un cadre limité seraient déjà nombreuses : on peut envisager ainsi de petits systèmes portables d'aide à la conversation pour les personnes se déplaçant en pays étranger ou encore des systèmes de réservation d'avion couplés à des systèmes de traduction [Rayner 93, Morimoto 93]. Cette dernière application est un exemple type de ce que l'on peut attendre d'un système de dialogue multilingue. Elle va d'ailleurs nous servir à illustrer notre propos.

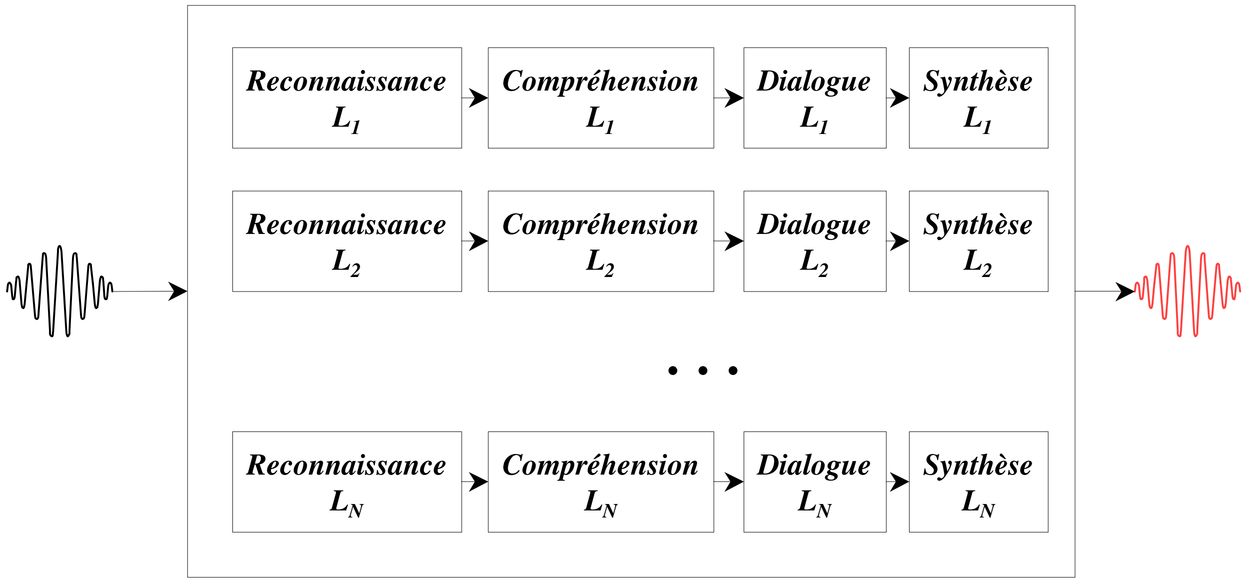

Figure 8 – Synoptique d'un système de dialogue multilingue basé sur des systèmes parallèles.

---

[15] Les principaux partenaires Français sont le LIMSI et le CLIPS

A l'heure actuelle, on est capable de concevoir des systèmes de dialogues multilingues [Glass 93] en se reposant sur la mise en parallèle de systèmes dépendants du langage (Figure 8). Ce type d'architecture présente des avantages (chaque système peut-être amélioré individuellement, on peut facilement traiter une nouvelle langue) mais il est aussi particulièrement coûteux puisque chaque module (compréhension, dialogue…) doit être réalisé en *N* exemplaires.

Dans un but de réduction de la complexité des systèmes multilingues, les recherches portent – entre autres – sur la recherche d'espaces de modélisation indépendants des langues ou plus exactement adaptés à plusieurs langues. Ces études traitent généralement soit de la reconnaissance de la parole (recherche d'une unification des représentations acoustico-phonétiques [Corredor-Ardoy 97, Anderson 97]), soit de la synthèse vocale (recherche d'unités phonétiques adéquates et de représentations prosodiques internationales [Hirst 98]). L'objectif est alors de concevoir un système de dialogue reposant sur des modules indépendants de la langue et sur d'autres, spécifiques à chaque langue (Figure 9).

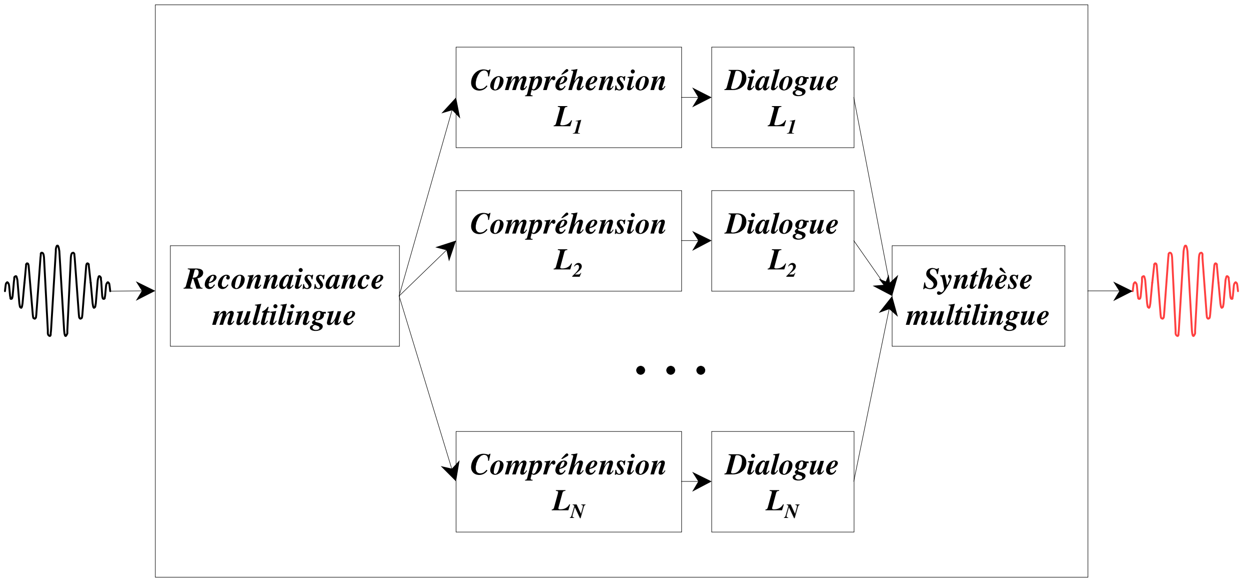

Figure 9 - Synoptique d'un système de dialogue multilingue basé sur des modules multilingues et des modules dépendants de la langue.

Dans un avenir plus lointain, il est tentant d'envisager l'utilisation conjointe de systèmes de traduction automatique et de dialogue multilingue [Dorr 98] pour augmenter encore l'adaptabilité de l'application à une nouvelle langue, par exemple en appliquant une approche de traduction en deux étapes, reposant sur l'emploi d'une langue intermédiaire dans le processus de traduction (Figure 10).

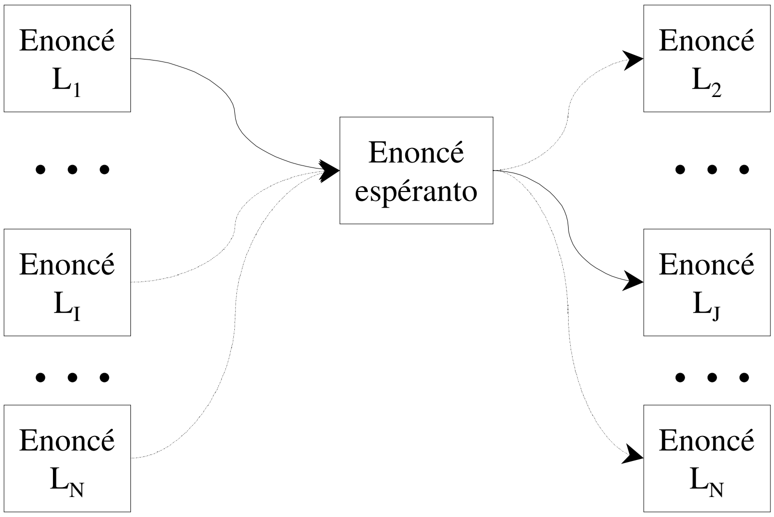

Figure 10 - Système de traduction automatique basé sur la traduction dans une langue « de passage ». Exemple de la traduction d'un énoncé en langue L_I en langue L_J

L'utilisation d'un tel système ne nécessiterait qu'un unique système de dialogue (dans notre exemple en espéranto), au prix de la conception du système de traduction correspondant (Figure 11). On peut cependant noter que, pour une application de réservation par exemple, il s'agit de traduire l'intention de l'utilisateur, et que le système de dialogue en langue intermédiaire peut donc être assez rudimentaire si le système d'analyse extrait correctement les informations pertinentes de l'énoncé.

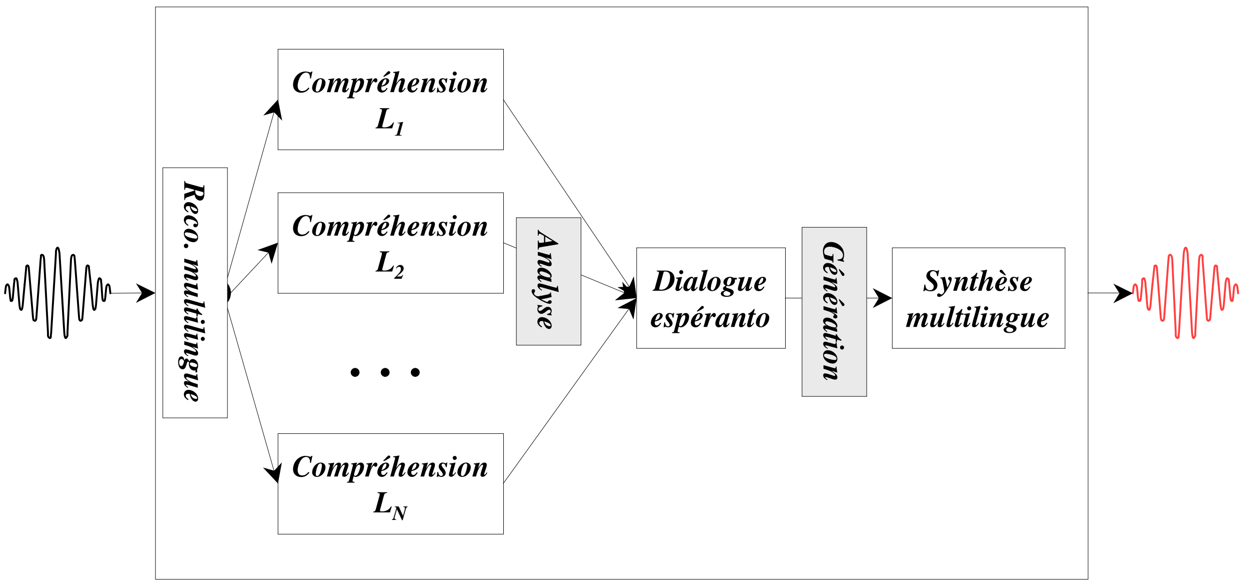

Figure 11 – Synoptique d'un système de dialogue intégrant une traduction automatique en langue « de passage ».

Quoi qu'il en soit, l'intégration de ces technologies implique encore de mener des recherches dans bien des domaines (traduction, compréhension, dialogue, synthèse…), et, à l'heure actuelle, seules les applications basées sur des systèmes parallèles (Figure

8) sont opérationnelles, aussi bien en reconnaissance de la parole qu'en synthèse (système Jupiter II de Dolphin Computer Access Ltd, ProVerbe de Elan Informatique…). Cette approche mène à des systèmes lourds dès lors que le nombre de langues augmente, puisqu'il est nécessaire d'optimiser une représentation spécifique à chacune des langues au moyen d'une quantité de données généralement importante.

La recherche d'une unification des représentations est un défi considérable pour dépasser ces limitations. Un autre axe de développement rendu nécessaire par les applications multilingues résulte de la nécessité de connaître *a priori* la langue traitée. La première possibilité est de demander à l'utilisateur d'identifier sa langue par un média non vocal (en cliquant sur une icône de son drapeau…), mais si l'on souhaite axer l'interaction sur la modalité parole, il est nécessaire d'identifier automatiquement la langue parlée par l'utilisateur. Il s'agit là d'une des nombreuses applications de l'Identification Automatique des Langues, comme nous allons le voir maintenant.

2 UNE INTRODUCTION A L'IDENTIFICATION AUTOMATIQUE DES LANGUES

L'Identification Automatique des Langues (IAL) est un domaine du traitement de la parole plutôt récent. L'objectif poursuivi est de concevoir un système qui, à partir d'un énoncé prononcé par un locuteur, détermine la langue qu'il a employée. A cette formulation élémentaire correspond une réalité bien plus complexe, puisque l'IAL peut se décliner et s'envisager dans un nombre assez important d'applications, où les conditions sont très variables, du nombre de locuteurs au nombre de langues, et de la longueur de l'énoncé à ses conditions d'enregistrements.

2.1 Les enjeux en IAL

2.1.1 Les enjeux applicatifs

L'ère actuelle est une ère de communication multilingue, que ce soit entre humains (au sein des grandes mégalopoles ou par téléphone interposé), ou entre humains et machines (IHM). Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres. Ces systèmes d'IAL peuvent être envisagés dans une tâche d'assistance au dialogue humain (DH) ou au sein d'IHM.

L'exemple le plus célèbre de situation de DH multilingue, cité par Muthusamy [Muthusamy 94b] est assez significatif des besoins à venir : aux Etats-Unis, les numéros des services d'urgence sont centralisés et accessibles en appelant le 911. La nécessité d'avoir un standard téléphonique disposant d'un service d'interprètes efficace est depuis longtemps une réalité dans ce pays multi-ethnique, à tel point que ATT a mis en place un service, (nommé *ATT Language Line*) chargé de diriger chaque appel vers le correspondant qui pourra comprendre la langue employée. Ce service d'interprètes gère

140 langues, et l'aiguillage des appels est à l'heure actuelle réalisé de manière entièrement manuelle : lorsqu'un appel arrive à un standardiste et qu'il ne s'agit pas d'une langue qu'il connaît, il le renvoie vers un autre standardiste en fonction de la langue – ou du type de langue – qu'il croit avoir reconnu. Ainsi l'appel peut transiter par plusieurs standardistes pendant un temps assez long, comme le rapporte Muthusamy [16]. Cet exemple montre à quel point il peut être important, dans le cadre d'une intervention d'urgence, de pouvoir identifier une langue rapidement. Un système d'IAL permettrait de confronter l'avis du standardiste avec la décision de la machine, et si cela ne se révèle pas plus efficace, on peut imaginer un système automatique qui donne à l'interprète une liste des langues potentielles correspondant à l'appel.

Si l'exemple du 911 est le plus marquant, il n'est pas pour autant unique, et dans bien des cas, l'IAL permettrait d'apporter une assistance au DH. Que ce soit dans un aéroport international, dans un grand hôtel ou pour un standard de réservation de billets pour une coupe de monde de football, le dialogue serait facilité si les gens pouvaient s'exprimer dans leur propre langue.

L'IAL a aussi un rôle croissant à jouer au sein des IHM, que ce soit pour permettre leur utilisation dans des pays plurilingues ou dans un cadre international. Si l'on reprend l'exemple de l'Espagne, un système de réservation de billets de train par téléphone doit être capable de recevoir un appel en castillan, en catalan, en basque ou en aranais [17], de même que tout système de dictée vocale par exemple. Ces contraintes linguistiques sont courantes à travers le monde, même si elles nous sont peu familières dans un pays où une seule langue officielle subsiste. Une autre application des IHM en plein développement consiste à mettre en des lieux publics (aéroports, gares, mais aussi offices du tourisme) des bornes d'information à disposition des voyageurs. Si, à l'heure actuelle, ces bornes peuvent être activées en deux ou trois langues en appuyant sur un bouton, l'extension du nombre de langues traitées se satisfera mieux d'un système d'identification de la langue. On retrouve aussi ce type d'utilisation dans le cadre de systèmes de synthèse multilingue et de traduction automatique lorsque plusieurs langues sont acceptées en entrée.

En résumé, on peut dire que les applications commerciales de l'IAL recouvrent surtout le besoin de routage initial du système vers la langue adaptée à l'utilisateur. Nous allons maintenant voir que du point de vue scientifique, ce n'est pas forcément ce type d'utilisation qui prédomine.

---

[16] En s'exprimant en Tamoul, Muthusamy a attendu plusieurs minutes avant d'obtenir un correspondant le comprenant, après avoir été dirigé – sans succès – vers trois interprètes d'Asie du Sud-Est. La situation a été débloquée uniquement lorsqu'il a prononcé le nom anglais ‘Tamil’.

[17] L'aranais est la langue officielle la plus récente en Espagne

2.1.2 Les enjeux scientifiques

Un corollaire implicite de l'IAL est la notion de distance entre langues, et de distance entre locuteurs et langues. De manière générale, on considère que les locuteurs utilisant un système automatique multilingue s'expriment dans leur langue naturelle. En effet, dès lors que l'on s'exprime dans une langue étrangère L2, le système phonologique employé est intermédiaire entre ceux de L1 et L2, et les structures morpho-syntaxiques sont aussi teintées d'un « accent » plus ou moins fort. Un des premiers thèmes scientifiques que l'on peut envisager pour des systèmes d'IAL est donc l'enseignement des langues étrangères. Dans un laboratoire de langue, il n'est pas toujours possible à l'enseignant d'écouter chaque étudiant et de le corriger ; cette tâche peut être partiellement prise en compte par un système automatique : le locuteur, en se perfectionnant, doit s'éloigner du modèle de sa langue maternelle et se rapprocher de celui de la langue L2. Pour qu'un tel système soit efficace, il est nécessaire que la distance qu'il calcule ait une réalité perceptive, c'est-à-dire que la distance perceptive soit corrélée à la distance automatique. Ce paradigme intervient aussi pour la validation croisée de modèles phonologiques et de modèles automatiques. Nous avons vu au cours de la première partie que des modèles phonologiques de systèmes vocaliques assez nombreux existent, et l'une des validations expérimentales possibles consisterait à vérifier que les distances entre les systèmes prédits sont conformes à ce qui est réellement observé. Une telle mise en correspondance de systèmes automatiques et phonologiques peut se révéler très fructueuse à la fois pour les informaticiens et les linguistes. Un autre thème de recherche intervenant aussi en IAL consiste à étudier la caractérisation et la modélisation des langues. En particulier, la recherche d'unités inter-langues [Corredor Ardoy 97] rejoint les préoccupations des linguistes et des phonologues tout autant que des cogniticiens s'intéressant à l'universalité du langage.

D'autres motivations scientifiques existent pour l'IAL, mais ce qui apparaît actuellement est plutôt une sous-exploitation manifeste des systèmes automatiques puisqu'ils sont orientés vers la prise de décision en faveur d'une langue ou d'une autre et non vers une exploitation scientifique du processus de décision.

2.1.3 Les enjeux militaires

Un autre domaine d'application essentiel est le domaine militaire. Le US Department of Defense est à l'origine des premières recherches menées en IAL au sein des laboratoires Texas Instruments au début des années 70. Depuis, de nombreux autres projets ont vu le jour, aux Etats-Unis comme en Europe (Projet DGA 95/118 : discrimination automatique multilingue). Cet intérêt est pleinement motivé par l'éventail des applications entrevues par les militaires, que ce soit dans le cadre de la communication ou de ce qu'il est commun d'appeler « l'intelligence » ou le renseignement militaire.

Pour ce qui est de la communication militaire, les enjeux sont proches des enjeux commerciaux : la communauté multilingue concernée peut être réduite (exemple de la

Force d'Action Rapide franco-germanique) ou à l'inverse plus étendue (pays de l'OTAN, intervention de casques bleus de l'ONU) ; les applications peuvent être des IHM (systèmes en service dans plusieurs pays) ou des assistances au dialogue humain (mise en présence de militaires de différents pays).

Le renseignement militaire est, pour sa part, demandeur de systèmes plus fins à la fois capables d'identifier la langue parlée par un locuteur *a priori* non coopératif et de fournir des renseignements sur ce locuteur : il existe des langues parlées sur des espaces couvrant des millions de kilomètres carrés, et une localisation plus précise de l'origine de la personne peut-être souhaitée. Dans ce cas, il est nécessaire d'opérer une modélisation plus fine de chaque langue, et la prise en compte de caractéristiques dialectales requiert alors des connaissances linguistiques poussées. Une des différences principales avec les applications civiles demeure que, dans le cas le plus général, on n'est pas assuré que la langue parlée par le locuteur fasse partie des langues apprises par le système : cela implique que le système intègre une décision de rejet. Il est bien évident qu'un système ne peut pas reconnaître toutes les langues du monde mais il peut déjà être extrêmement intéressant d'identifier la famille à laquelle une langue appartient. Un système automatique effectuant cette présélection est parfaitement envisageable, mais là aussi, cela implique d'intégrer de manière efficace des connaissances linguistiques fondamentales (quelles familles employer, quels aspects modéliser…). Une dernière application repose là encore sur la notion de distance entre langues puisqu'il s'agit de confirmer si un locuteur donné s'exprime réellement dans la langue qu'il annonce ou plutôt dans une autre. Toutes ces applications nécessitent une exploitation scientifique poussée des résultats fournis par les systèmes en IAL ou en identification des familles linguistiques, et les préoccupations des militaires rejoignent donc à la fois les enjeux des linguistes et ceux des informaticiens.

Comme cela a été précisé au début de ce paragraphe, les premières recherches en IAL ont été entreprises dans les années 70 sous l'impulsion du DoD. Nous allons maintenant étudier les approches envisagées durant les années 70 et les années 80 de manière à dégager les termes actuels de la problématique de l'IAL.

2.2 Un survol historique (1973-1989)

Dans [Muthusamy 93], l'auteur signale qu'au cours de ces deux décennies, seulement 14 articles sur le sujet ont été publiés. Ce chiffre est à comparer avec les dizaines d'articles qui, depuis le début des années 90, ont été édités dans les différents congrès et revues sur la communication parlée.

2.2.1 La Génèse

- **Texas Instruments**

Dès 1973, Texas Instruments consacre un effort de recherche soutenu à l'Identification Automatique des Langues. Ces recherches, menées jusqu'en 1980, visent à identifier une langue parmi sept langues candidates (non précisées) à partir d'enregistrements de parole lue [Leonard 80]. L'approche choisie est d'établir des modèles statistiques de motifs phonétiques caractéristiques de chaque langue ainsi que de leur fréquence d'occurrence. L'obtention des motifs pertinents repose sur une sélection manuelle par les experts de TI au cours d'une première phase d'analyse du signal. La phase suivante consiste à implanter des algorithmes d'extraction semi-automatique de ces motifs. Cette approche « experte » implique une bonne connaissance des langues étudiées, et effectivement, lorsque cette connaissance fait défaut, les taux d'identification chutent. Les améliorations des différentes versions ont surtout porté sur la détection des motifs caractéristiques à partir du signal (seuil sur les fréquences d'occurrence, utilisation de critères d'entropie).

- **House et Neuberg**

En 1977, une autre étude en IAL est publiée par House et Neuberg. Il s'agit de modéliser par un réseau de Markov les séquences de macro-classes phonétiques (plosives, fricatives, voyelles...) rencontrées dans chaque langue [House 77]. Ces travaux, menés à partir de transcriptions phonétiques de textes (et non de parole) écrits en huit langues, ont inspiré par la suite la plupart des études menées dans les années 90 et intégrant des considérations phonotactiques. Le fait que le système ne traite pas de parole réelle ne permet évidemment pas à cette époque de savoir si cette approche est efficace ou non.

- **Discussion**

Les deux voies expérimentées au cours des années 70 ont donc été la modélisation acoustique d'une part, et la modélisation phonotactique d'autre part. Les expériences montrent que l'IAL peut se baser sur ces deux approches, mais aucune tentative intégrant les deux aspects n'a été publiée. D'autre part, les conditions expérimentales (étiquetage manuel ou semi-automatique, transcription phonétique et non signal réel) ne permettent pas encore de préciser les performances que l'on peut espérer de tels systèmes.

2.2.2 Les Années 80

- **Li et Edwards**

Les principes de modélisation stochastique de séquences de sons vont être repris par Li et Edwards sur des enregistrements de parole lue par des locuteurs masculins de cinq langues [Li 80]. A partir de six classes acoustico-phonétiques (noyaux syllabiques, consonnes fricatives voisées…) trois modèles statistiques (à base de chaînes de Markov) seront développés. Le premier modélise la succession des segments dans chaque langue alors que les deux autres modélisent, dans un cas, la succession de syllabes, et dans l'autre les relations intra-syllabiques. La méthode employée permet en fait de modéliser les séquences de consonnes de chaque langue. Il est intéressant de noter que le système obtenu discrimine efficacement les langues à structure syllabique simple (deux langues asiatiques tonales) des langues européennes caractérisées par des séquences consonantiques plus longues.

Au cours des années 80, les travaux de Li et Edwards ont été les seuls à exploiter une modélisation markovienne. Les autres études menées utiliseront des approches à base de fonctions de décision polynomiales, de règles ou de quantification.

- **Cimarusti et Ives**

Cimarusti et Ives s'intéressent en 1982 à l'identification de neuf langues à partir d'une analyse LPC (Linear Predictive Coefficients) du signal acoustique [Cimarusti 82]. Une centaine de paramètres sont extraits (coefficients d'autocorrélation, coefficients cepstraux, fréquences des formants...). Un classificateur polynomial est optimisé de manière itérative pour les données d'apprentissage puis appliqué sur les données de test. Les résultats obtenus sont bons, mais le faible nombre de locuteurs (tous masculins) ne garantit pas que le système soit indépendant du locuteur.

A partir de ces travaux, Ives développe en 1986 un autre système à base de règles issues du seuillage de 50 des paramètres précédemment utilisés. Les règles qui émergent sont au nombre de 9 et elles sont en partie de nature prosodique (basées sur la fréquence fondamentale F_0 du signal, la variance du second formant F_2, le nombre de voyelles, la densité d'énergie spectrale dans des filtres...).

Les résultats obtenus [Ives 86] sont là encore bons, mais difficiles à évaluer car l'indépendance des corpus de test et d'apprentissage n'est pas clairement précisée.

- **Foil et Goodman**

Les travaux que nous venons de citer ont tous été réalisés sur des corpus de données enregistrés en laboratoire, dans des conditions « propres ». Foil s'intéresse pour sa part à de la parole enregistrée à la radio en trois langues, dans des conditions de bruit bien réelles.

Le système conçu se base à l'origine sur l'identification prosodique (intonation et rythme) des langues [Foil 86]. Le système obtenu par classification bayésienne à partir de F_0 et de l'enveloppe du signal se révèle décevant, et une autre approche est alors développée : par quantification vectorielle, Foil obtient 10 vecteurs formantiques caractéristiques de chaque langue, puis il opère une classification sur les données de test, calculant ainsi une distorsion pour chaque langue.

Ces travaux sont poursuivis par Goodman en 1989, qui améliore la robustesse de chaque élément en développant un nouvel algorithme de suivi de formants, en ajoutant un module de décision voisé/non voisé plus performant et en utilisant une distance euclidienne pondérée pour la classification [Goodman 89]. Il faut ajouter à cela qu'il conserve 60 vecteurs caractéristiques par langue au lieu des 10 retenus par Foil.

- **Discussion**

Si les travaux entrepris dans les années 70 avaient privilégié les approches phonétiques et phonotactiques, on assiste dans les années 80 à une diversification des paramètres discriminants testés. Chaque approche, qu'elle soit basée sur des contraintes phonotactiques, sur des paramètres spectraux ou sur la prosodie, atteint des résultats intéressants bien qu'il soit difficile de juger des conditions expérimentales. On peut noter que les paramètres prosodiques ne sont pas efficacement utilisés dans l'approche bayésienne de Foil mais qu'une approche à base de règles [Ives 86] peut se révéler efficace.

2.3 Vers les systèmes actuels

Le bilan général des années 70 et 80 est contrasté. Des recherches avec des données extrêmement disparates et des méthodes variées ont été menées, mais le domaine de l'IAL n'a pas atteint sa maturité : les systèmes ne sont pas complètement détaillés, les protocoles expérimentaux ne sont pas uniformisés, et aucune approche ne se dégage comme étant une référence.

C'est sur ce constat en demi-teinte que débutent les années 90. En fait, la situation va changer de manière radicale en deux ou trois ans sous l'action conjuguée de plusieurs facteurs, principalement liés à la Reconnaissance Automatique de la Parole (RAP). En effet, au cours des années 80, l'amélioration des systèmes de RAP a été prodigieuse ; on a assisté à la mise sur le marché des premiers systèmes efficaces, les serveurs vocaux opérationnels ont fait leur apparition et, vers 1990, les IHM sont sorties des laboratoires et ont pénétré le monde réel… Cette situation aura plusieurs effets sur l'IAL. Tout d'abord, les enjeux applicatifs deviennent plus pressants puisque dorénavant les systèmes d'IHM sont opérationnels ; il devient urgent d'envisager de les doter d'une capacité multilingue. Ensuite, la plupart des équipes de recherche souhaitent appliquer le savoir-faire acquis au cours des années 80 en RAP à d'autres domaines. On assiste alors à un regain d'intérêt pour des domaines considérés jusqu'alors comme « secondaires » comme l'identification du locuteur [18] ou de la langue. Dès lors, la dynamique se met en marche : sous l'impulsion conjuguée de la demande applicative et de l'offre scientifique, des corpus de données sont enregistrés [Muthusamy 92]. Il n'en fallait pas plus pour que le domaine de l'IAL émerge comme un thème majeur du TAP, et que les modélisations markoviennes ou neuromimétiques, massivement employées en RAP, s'imposent en IAL.

[18] Dans le cas de l'identification du locuteur, le mouvement a été plus précoce que pour la langue.

Chapitre 2

UN ETAT DE L'ART DE L'IAL

Les années 90 ont vu s'intensifier les recherches menées à travers le monde en IAL. Depuis quelques années, la communauté scientifique dispose de corpus multilingues conséquents (paragraphe 1), et chacun a pu expérimenter les méthodes qu'il envisageait et comparer les résultats obtenus avec ceux des autres laboratoires. Cette situation d'émulation a mené à la publication d'un nombre important d'articles sur l'IAL, et les systèmes conçus atteignent aujourd'hui des performances intéressantes (paragraphe 2).

1 LES CORPUS DE DONNEES

Si les systèmes expérimentaux développés au cours des décennies 70 et 80 ont été validés sur des corpus spécifiques et parfois insuffisamment décrits dans la littérature, les systèmes récents reposent pour la plupart sur des corpus internationalement reconnus. Ces corpus ont été enregistrés à l'initiative d'organismes publics comme le NIST (National Institute of Standards and Technology) aux Etats-Unis ou à la demande des laboratoires eux-mêmes, et la communauté scientifique mondiale dispose actuellement de bases de données et de corpus d'évaluation conséquents. On peut regrouper ces corpus en deux grands types, selon qu'ils sont constitués de données enregistrées en haute qualité (studio anéchoïque, microphone Hi-Fi) ou d'appels enregistrés via le canal téléphonique (bruit ambiant, coupure à 3,5 kHz). Le Tableau 5 présente un aperçu des corpus multilingues couramment utilisés en IAL. La taille du corpus (en nombre de locuteurs ou en durée) ne figure pas dans ce tableau puisque sa signification intrinsèque est faible et qu'il serait nécessaire de prendre en compte d'autres facteurs (équilibrage du corpus entre les langues, équilibrage du nombre de locuteurs masculins et féminins…). Les langues présentes dans chacun de ses corpus sont précisées en Annexe 1.

Nous n'allons pas décrire ici la totalité des corpus cités (nous renvoyons le lecteur désireux d'obtenir de plus amples renseignements sur les différentes données aux références données en notes) mais uniquement les corpus EUROM_1, OGI MLTS et CALLFRIEND. Le premier a longtemps été le seul corpus enregistré en studio pour un nombre important de langues, tandis que le second a été le standard utilisé par les systèmes d'IAL lors des campagnes d'évaluation du NIST et que le troisième est en passe de devenir un corpus de référence pour les campagnes futures.

Nom du corpus	Nombre de langues	Conditions d'enregistrement	Type de parole	Transcriptions (type – quantité)
CALLFRIEND[19]	12 (15)	Téléphone	Conversation	-
CALLHOME[20]	6	Téléphone	Conversation	Orthographique – partielle
EUROM_1[21]	11	Studio	Lue	Phonétique – totalité
GlobalPhone[22]	9	Studio	Lue	Orthographique - totalité
IDEAL[23]	4	Téléphone	Mixte (spontanée / lue)	-
OGI 22 languages[24]	22	Téléphone	Principalement spontanée	-
OGI MLTS[25]	11	Téléphone	Principalement spontanée	Phonétique – partielle

Tableau 5 – Principaux corpus multilingues disponibles

1.1 EUROM_1

Cette base de données a été développée dans le cadre du contrat européen ESPRIT SAM. 11 langues européennes ont ainsi été enregistrées. Il s'agit d'enregistrements de parole dite « de laboratoire », c'est-à-dire de données lues, recueillies dans un studio anéchoïque et échantillonnées à 20 kHz. La base est composée pour chaque langue de phrases, de mots isolés et de logatomes (successions de séquences Voyelle-Consonne-Voyelle) prononcés par 30 locuteurs masculins et 30 locuteurs féminins.

La totalité du corpus a été phonétiquement étiquetée par des experts. EUROM_1 constitue donc un matériau de choix pour les études phonétiques et phonologiques nécessaires en IAL.

1.1 OGI MLTS [Muthusamy 92]

Cette base de données est la référence dans le monde de l'IAL. Elle a été utilisée successivement avec 6, 9, 10 puis 11 langues pour les campagnes de test du NIST jusqu'en 1995. Contrairement à EUROM_1, il s'agit ici de parole téléphonique échantillonnée à 8 kHz dans une ambiance le plus souvent bruitée. Voici un extrait du protocole d'enregistrement proposé par OGI :

1. Quelle est votre langue natale ?

---

[19] cf. http://www.ldc.upenn.edu/ldc/catalog/html/package/cf.html

[20] cf. http://www.ldc.upenn.edu/ldc/catalog/html/package/ch.html

[21] cf. http://www.icp.grenet.fr/Relator/multiling/eurom1.html

[22] [Schultz 97]

[23] [Lamel 98]

[24] cf. http://www.cse.ogi.edu/CSLU/corpora/22lang/

[25] cf. http://www.cse.ogi.edu/CSLU/corpora/mlts.html

2. Quelle langue parlez-vous la plupart du temps ?
3. Enumérez les chiffres de zéro à dix, SVP.
4. Récitez les sept jours de la semaine, SVP.
5. Parlez-nous du climat de la ville où vous habitez.
6. Décrivez la pièce d'où vous nous appelez.
7. …

Chaque réponse est enregistrée avec un temps de réponse limité (10 secondes pour la question 5 par exemple). A cela s'ajoute l'enregistrement d'une minute de parole spontanée pour chaque locuteur, enregistrée sous le nom de « story ». On obtient finalement pour chaque locuteur environ deux minutes de parole.

Cette base de données est particulièrement intéressante car elle permet de se rapprocher des conditions d'utilisation réelles de systèmes d'IAL : milieu bruité (parfois même très bruité), canal téléphonique, présence de pauses et d'hésitations dans les énoncés, parole spontanée. Par contre, certains aspects peuvent se révéler plutôt gênants ; en particulier, pour le corpus dit « français », plusieurs locuteurs ont un fort accent de français québécois. Dans ce cas précis, il s'agit plus d'un corpus francophone que français. Il est vraisemblable que cet aspect se retrouve pour d'autres langues (anglophone, hispanophone…).

D'autre part, nous n'avons pas encore évoqué les transcriptions phonétiques du corpus. Elles sont de deux types, soit sous forme de classes phonétiques majeures (voyelle, fricative, silence ou closion, explosion, ou autres consonnes) soit sous forme phonétique classique.

L'étiquetage phonétique classique est disponible pour six langues[26] pour un nombre de « story » variable allant de 64 (japonais) à 210 (anglais). Il a été réalisé manuellement par des experts. L'étiquetage en classes majeures est quant à lui disponible pour toutes les langues pour une partie du corpus (environ 8 minutes par langue) mais il est réalisé de manière semi-automatique : la procédure d'étiquetage automatique est corrigée manuellement par un expert. L'étiquetage obtenu se révèle parfois inexact ou ambigu (cas de bruits non produits par le locuteur et étiquetés comme des occlusives).

Ces quelques remarques étant formulées, on peut considérer que OGI MLTS est la base de données sur laquelle les performances de la plupart des systèmes sont évaluées à l'heure actuelle, et qu'elle représente à ce titre une contribution majeure au domaine de l'IAL. La plupart des tests sont réalisés (sur recommandation du NIST) avec le signal nommé « story », limité aux 45 premières secondes.

[26] Allemand, anglais, espagnol, hindi, japonais et mandarin.

1.2 LDC CALLFRIEND

Le dernier corpus que nous allons décrire ici est aussi l'un des plus récents puisque le projet CALLFRIEND a débuté en 1996. Alors qu'OGI proposait un protocole permettant d'obtenir de la parole quasi-spontanée au moyen de questions, l'approche employée par le LDC (Linguistic Data Consortium) pour collecter de la parole continue « naturelle » est encore plus simple : pour chacune des langues du corpus, 60 conversations totalement spontanées ont été enregistrées, avec une durée allant de 5 à 30 minutes.

Le corpus comprend aussi des informations sur chaque interlocuteur (sexe, age, niveau d'éducation, numéro de téléphone) ainsi que sur chaque appel (qualité de la transmission, nombre d'interlocuteurs). Il s'agit dans tous les cas d'appels locaux et les locuteurs s'expriment toujours dans leur langue natale.

Ce corpus bénéficie de l'expérience dont on dispose aujourd'hui sur l'enregistrement des corpus multilingues, et on peut s'attendre à ce qu'un meilleur contrôle des enregistrements ait été réalisé. Par contre, aucune transcription n'est disponible, et le fait que plusieurs locuteurs soient enregistrés ensemble ajoute des difficultés supplémentaires dans une tâche d'IAL. En 1996, la campagne de test du NIST a principalement portée sur ce corpus, et il est donc en passe de devenir incontournable.

De ces trois corpus, celui qui a été le plus employé ces dernières années en IAL est OGI MLTS, comme le montrera le tableau récapitulatif des principaux travaux entrepris au cours des années 90 (Tableau 6).

2 UN PANORAMA DES SYSTEMES ACTUELS

Si l'on s'interrogeait au début des années 70 sur les paramètres discriminants pertinents pour distinguer les langues entre elles, les expériences menées jusqu'en 1989 ont apporté quelques éléments de réponse. Depuis cette date en effet, la majeure partie des systèmes s'appuient comme nous allons le voir sur une modélisation phonotactique du langage. De manière à peine exagérée, on peut dire que la modélisation des sons du langage sert uniquement à transformer l'énoncé d'un espace continu acoustique en une suite de symboles discrets. Certains systèmes tirent cependant avantage du décodage acoustico-phonétique, soit en en exploitant explicitement un score d'identification de la langue, soit en optimisant globalement le décodage phonétique et le modèle phonotactique qui le suit.

On peut considérer que les systèmes actuels relèvent principalement de deux approches méthodologiques : la modélisation statistique (2.1), basée sur la recherche de la langue la plus vraisemblable par rapport à des modèles et la modélisation neuro-mimétique (2.2) qui généralise la notion de règle de manière très performante. D'autres approches, moins influencées par la RAP que les précédentes, sont encore – heureusement – développées. On peut citer les expériences de S. Itahashi [Itahashi 95] basées sur une modélisation purement prosodique et celles de K. P. Li [Li 94], basées quant à elle sur une méthode d'identification du locuteur (2.3).

2.1 Les approches statistiques

2.1.1 Rensselaer Polytechnic Institute, New York, Etats-Unis

Les travaux présentés en 1991 dans [Savic 91] sont caractéristiques du passage des années 80 aux années 90 : alors que le système présenté – basé sur une modélisation markovienne de chaque langue et sur des paramètres issus de la détection de la fréquence fondamentale F_0 – est novateur, les informations sur le corpus utilisé demeurent incomplètes.

Pour chaque langue (elles sont au moins 4 dans le corpus), un réseau de Markov ergodique à 5 états est appris (premier module) et, à partir d'un algorithme de détection de F_0, la distribution fréquentielle et les variations de la fréquence fondamentale sont évaluées (second module). Un classificateur (non décrit) est utilisé pour prendre en compte les résultats issus des deux modules. Il est difficile d'évaluer l'apport de chacun d'entre eux puisque les résultats ne sont pas communiqués dans l'article.

2.1.2 LIMSI, France

Les travaux menés au LIMSI par L. Lamel et J.L. Gauvain s'appuient en grande partie sur la connaissance acquise en modélisation phonétique sur les réseaux de Markov en RAP. Les expériences ont été menées sur les 10 langues du corpus initial OGI MLTS [Lamel 94].

A partir des données d'apprentissage, un large modèle ergodique est appris pour chaque langue ; chaque modèle élémentaire représente une unité au niveau phonétique et non une macro-classe phonétique comme dans les travaux précédents [House 77, Savic 91]. L'originalité de l'algorithme est qu'il effectue une optimisation des modèles en prenant en compte la vraisemblance conjointe du signal et de la suite d'unités décodées alors que dans la plupart des systèmes, cette modélisation des séquences décodées (modèle phonotactique) ne modifie pas les modèles acoustico-phonétiques (il s'agit alors d'un post-traitement). Cette approche sera reprise entre autres par Marc Zissman (cf. 2.1.10). Les résultats obtenus au LIMSI, portant sur des données de 10 secondes de durée, sont proches de 60 % d'identification correcte [Lamel 94].

Actuellement, deux axes de recherche sont privilégiés au LIMSI en IAL. Le premier traite de l'intégration de modèles de langage basés sur des mots ou des groupes de mots [Jardino 96] au système et le second cherche à établir une représentation phonétique unifiée par regroupement de phonèmes dépendants des langues en un modèle de décodage acoustico-phonétique global [Corredor-Ardoy 97].

2.1.3 Ensigma Ltd, Angleterre

Les travaux rapportés dans [Tucker 94] ont été réalisés sur 3 langues (Anglais, Hollandais et Norvégien) issues du corpus EUROM_1.

L'approche adoptée repose sur deux ensembles de modèles. Le premier ensemble, indépendant du langage (IL) est obtenu à partir de modèles estimés sur TIMIT (donc en anglais américain). Le second modèle, dépendant du langage repose sur une ré-estimation des modèles IL avec les données de chaque langue, après un alignement de chaque phrase sur les modèles IL. La modélisation permet de constituer ainsi un ensemble d'unités spécifiques pour chaque langue.

En phase de test, la décision finale est prise en tenant compte des scores générés lors des décodages effectués avec les différents ensembles de phonèmes, ainsi que d'une statistique calculée sur la fréquence d'occurrence de ces phonèmes. A partir de données de test de 10 secondes de durée, le taux d'identification correcte obtenu est de 90 %.

2.1.4 ATT Bell Labs - Etats-Unis (Kadambe - Hieronymous)

Les travaux menés au laboratoire ATT Bell par S. Kadambe et J. L. Hieronymous ont eux aussi porté sur 3 langues, extraites du corpus OGI MLTS.

L'approche présentée (Figure 12) est classique : il s'agit de calculer un score acoustico-phonétique à partir de Modèles de Markov Cachés (MMC) puis d'utiliser la suite de phonèmes ainsi générée en entrée d'un système à base de grammaire N-gramme pour générer un score phonotactique.

La reconnaissance phonétique est effectuée par un système développé au laboratoire Bell, basé sur des MMC Continus à Durée Variable, tandis que le score phonotactique est généré par un modèle trigramme. Les résultats obtenus atteignent 91 % d'identification correcte avec des phrases de test de 50 secondes [Kadambe 94].

Il est intéressant de noter que si la reconnaissance phonétique se passe mal, il n'est pas possible de « récupérer » son erreur ; elle se répercute au niveau du modèle phonotactique en pénalisant la langue parlée. C'est pour cette raison que la plupart des autres auteurs ont choisi de connecter un modèle phonotactique pour chacune des langues à identifier en sortie de chacun des décodeurs phonétiques (cf. Figure 14).

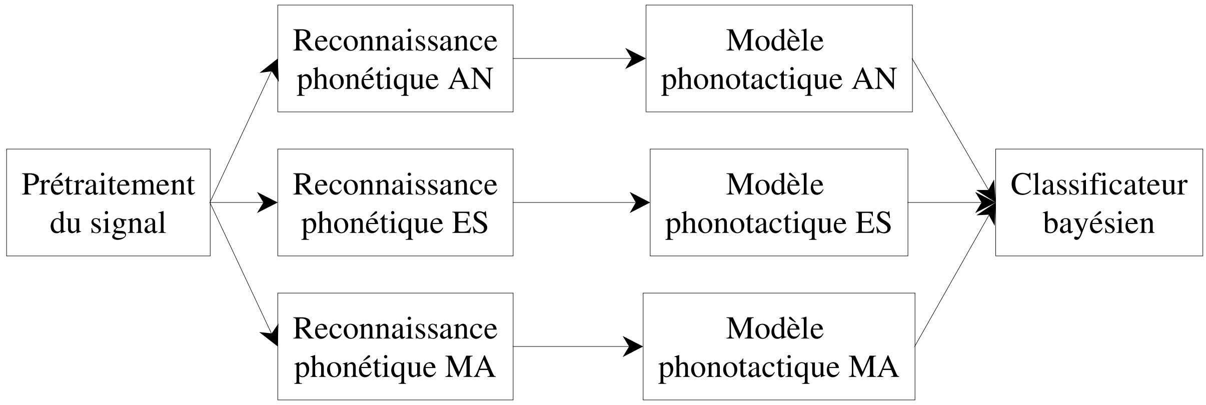

Figure 12 – Schéma bloc du système de ATT Bell Labs (d'après [Kadambe 94])
AN = anglais, ES = espagnol et MA = mandarin

2.1.5 ATT Bell Labs - Etats-Unis (Ramesh - Roe)

Nous trouvons dans [Ramesh 94] un autre travail développé au sein des laboratoires Bell par R. Ramesh et D. B. Roe.

L'étude réalisée vise à démontrer l'apport d'une modélisation par mots-clefs dans une tâche d'Identification des Langues dans un cadre restreint puisqu'il s'agit d'enregistrements ayant trait à une tâche d'opérateur téléphonique dans 4 langues. Les données utilisées durant la phase de développement ont été enregistrées chez ATT sur une ligne téléphonique numérique. Les tests finals sont faits avec des données OGI MLTS dans les mêmes langues.

Un réseau de Markov est utilisé, avec une modélisation explicite des 30 à 40 mots-clefs retenus par langue, et une modélisation du reste de la parole en unités (de nature non précisée) représentées par des réseaux gauche-droite de 5 à 10 états.

Pour ne pas détériorer les résultats en passant du canal numérique (apprentissage) au canal analogique (test), une normalisation cepstrale et une modélisation du canal (par un modèle MMC adapté sur une phase de « silence ») sont étudiés. Les résultats mentionnés font état de 96 % d'identification correcte avec des séquences de test de durée de 5 à 10 secondes (il s'agit des énumérations de chiffres).

2.1.6 Université d'Aalborg, Danemark

En introduisant la notion de polyphonèmes (communs à plusieurs langues) et de monophonèmes (spécifiques à chaque langue), P. Dalsgaard et O. Andersen [Dalsgaard 94, Andersen 97] aboutissent à un réseau décrivant les 4 langues de leur corpus (issues de EUROM 0) avec 114 phonèmes modélisés par des MMC à trois états.

Une première phase d'identification, menée par algorithme de Viterbi, aboutit classiquement à la reconnaissance de la suite de phonèmes la plus probable (au sens de Viterbi). Une deuxième phase est alors appliquée, par modification des poids de chaque phonème (initialement équi-pondérés) en tenant compte des matrices de confusion inter-phonèmes établies durant l'apprentissage pour chaque langue : on augmente ainsi l'importance relative des phonèmes les plus discriminants. Avec des phrases de test de 2 minutes, le score d'identification obtenu est de 88,1 %.

2.1.7 BBN Systems and Technologies, Etats-Unis

Les travaux menés par M. A. Lund et H. Gish [Lund 95] s'articulent eux aussi autour d'un système de reconnaissance phonétique complété par un modèle de langage. Les tests ont été effectués avec 9 langues issues du corpus OGI MLTS.

L'originalité de l'approche réside dans le fait que l'effort de recherche porte entièrement sur le modèle phonotactique : le décodage phonétique utilise en effet un réseau de Markov développé uniquement sur la langue anglaise, sous forme de MMC à 3 états représentant les 50 phonèmes retenus. Cela revient à projeter les données acoustiques de toutes les langues dans l'espace phonétique de la langue anglaise.

Deux approches sont étudiées pour les modèles de langage, toutes deux basées sur des grammaires bigrammes :

- ✓ La première consiste à générer des grammaires, non seulement sur les phonèmes mais aussi sur des unités plus longues, appelées pseudo-mots (extraites par un algorithme dynamique basé sur la similarité des séquences de phonèmes),
- ✓ La seconde approche, itérative, est une méthode d'application de l'algorithme EM aux grammaires bigrammes sur les phonèmes.

Les résultats fournis ne portent que sur des expériences de discrimination entre deux langues (Langue 1 vs Langue 2), et en moyenne le score obtenu est de 92,4 %.

2.1.8 Université de Tokyo, Japon

Nous voici encore en présence d'un système conjuguant une étape de reconnaissance phonétique et un modèle de grammaire N-gramme [Kwan 95]. 5 langues ont été retenues dans le corpus OGI MLTS, et plusieurs approches ont été testées.

La première étude porte sur la structure à utiliser en phase de reconnaissance phonétique, à savoir s'il est plus efficace d'employer plusieurs systèmes de reconnaissance en parallèle, chacun modélisant une langue, (comme le système présenté Figure 12) ou un système unique modélisant l'ensemble des langues, et construit à partir de tous les phonèmes, communs à toutes les langues ou spécifiques (comme le système présenté Figure 13).

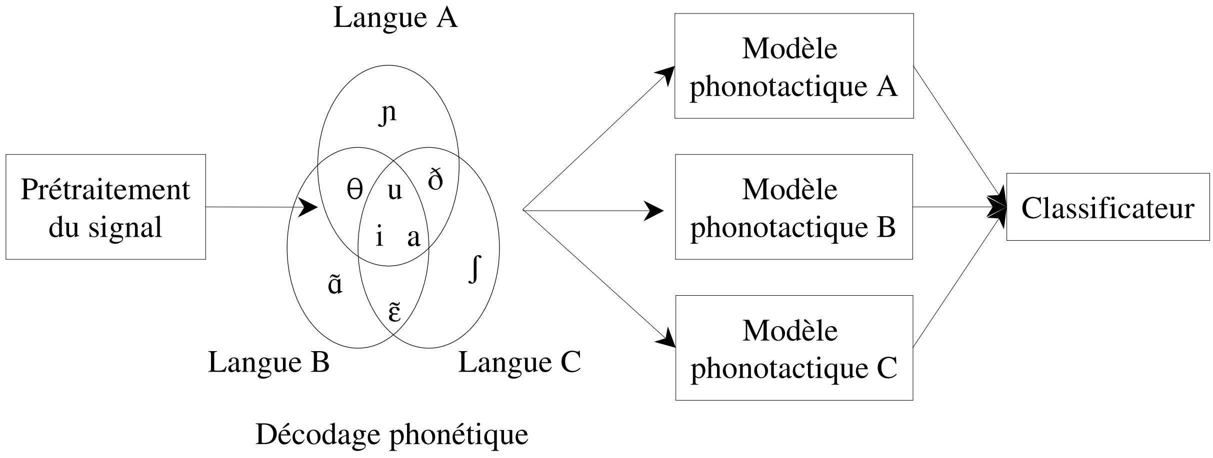

Figure 13 – Synopsis du Système MPR (Mixed Phoneme Recognition) d'après [Kwan 95].

La seconde étude porte sur le modèle de langage, Kwan et Hirose étudiant l'efficacité de modèles unigrammes et bigrammes pour chaque paire de langues.

Les meilleurs résultats en discrimination (Langue 1 vs Langue 2) sont obtenus avec le système de décodage MPR et les modèles unigrammes (78 % d'identification correcte en moyenne). Selon les auteurs, la dégradation – inattendue – des résultats en utilisant des grammaires bigrammes s'expliquent par le manque de données pour certaines langues.

2.1.9 MIT, Etats-Unis, (Hazen - Zue)

Il est proposé dans [Hazen 97] un système d'IAL basé sur le système de reconnaissance phonétique SUMMIT [Zue 90]. A ce module de base (indépendant de la langue) s'ajoutent un module phonotactique (fondé sur une grammaire trigramme) et un module prosodique. Ce dernier prend en compte, d'une part F_0 au niveau segmental (modèle statistique gaussien) et d'autre part la durée des segments issus de SUMMIT (là encore un modèle statistique gaussien par classe phonétique). Les expériences rapportées ont tendance à montrer que le module prosodique est inefficace, et que le modèle trigramme est performant puisque le taux d'identification correcte atteint 78,1 % avec les phrases de test de 45 secondes pour 11 langues du corpus OGI MLTS.

2.1.10 MIT, Etats-Unis, (Zissman)

Un autre système développé au MIT est décrit dans [Zissman 96]. Parmi plusieurs approches, celle offrant les meilleurs résultats est basée sur une architecture hybride entre décodage acoustico-phonétique spécifique à chaque langue et décodage commun. Le système de base (PRLM : Phone Recognition followed by Language Modeling) repose non pas sur le seul système de décodage SUMMIT mais sur des modules acoustico-phonétiques développés avec HTK [27]. En sortie de ce seul décodeur, un modèle n-gramme est appliqué pour chacune des langues à identifier. L'identification est donnée par la vraisemblance phonotactique la plus élevée.

Cette approche effectuant une projection aveugle dans un espace phonétique *a priori* peut se révéler inefficace si les caractéristiques phonétiques des langues à identifier sont trop éloignées du décodeur employé. Pour éviter cet effet, Zissman utilise plusieurs décodeurs acoustico-phonétiques de manière à optimiser la couverture de l'espace acoustico-phonétique et à permettre de rattraper les éventuelles erreurs de décodage d'une langue. Cette approche (Parallel PRLM) est proposée dans le cas où l'on ne dispose pas d'un décodeur acoustico-phonétique pour chacune des langues traitées. Ce système, implanté avec 6 décodeurs (allemand, anglais, espagnol, hindi, japonais et mandarin) obtient 80 % d'identification correcte avec les 11 langues du corpus OGI MLTS (Figure 14).

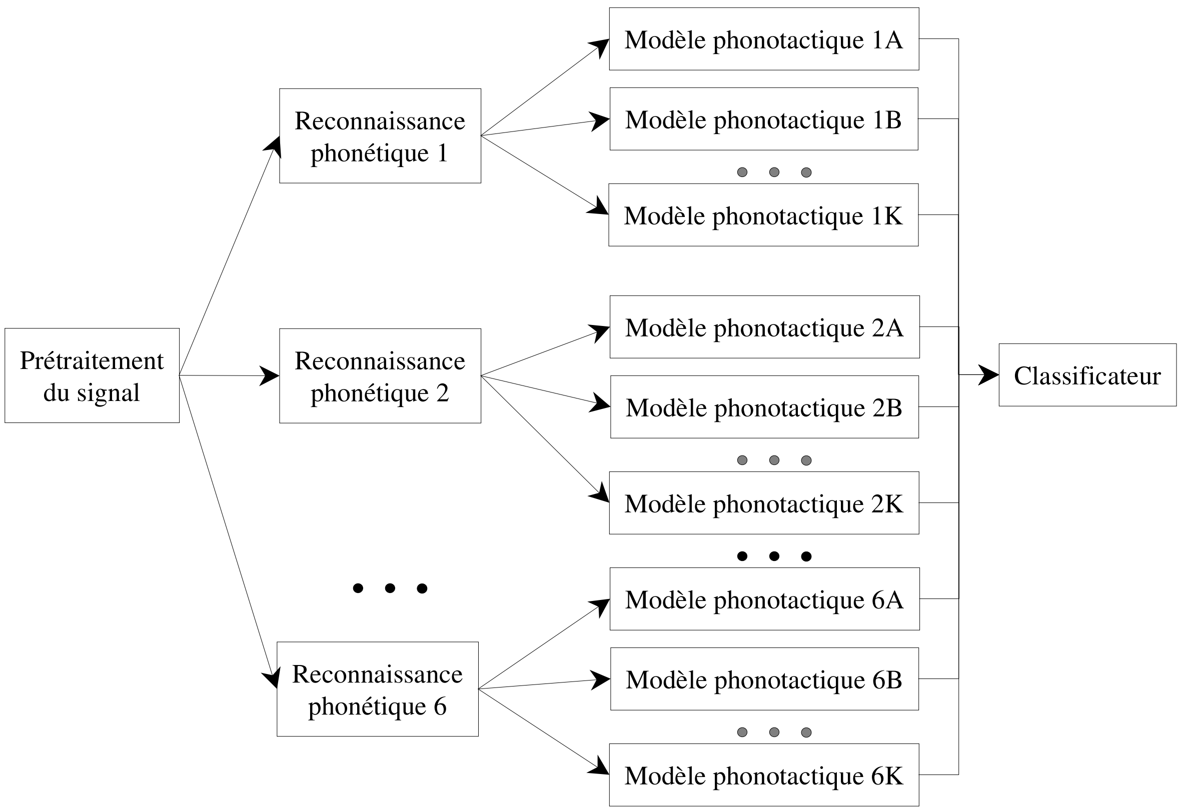

Figure 14 – Schéma du système Parallel PRLM (d'après [Zissman 96])
3 des 6 décodeurs phonétiques sont représentés ; pour chacun d'entre eux, 3 des 11 modèles phonotactiques sont indiqués.

Sur la figure, les décodeurs phonétiques sont numérotés de 1 à 6 et les langues à reconnaître de A à K. Il y a un modèle phonotactique par langue à reconnaître pour

[27] pour plus d'information, consulter le serveur http://www.entropic.com/htk/html

chacun des décodeurs phonétiques, soit un total de 66 modèles phonotactiques dans le cas présent. En intégrant plusieurs améliorations (modélisation acoustico-phonétique dépendante du sexe, modélisation de la durée…), le score d'identification correcte atteint 89 % avec l es phrases de test de 45 secondes.

Ne disposant pas d'un décodeur acoustico-phonétique pour chacune des 11 langues, Zissman n'a pas pu tester l'approche consistant à optimiser conjointement les unités du décodeur et les modèles de langage (approche proposée dans [Lamel 94]).

2.1.11 OGI, Etats-Unis, (Yan - Barnard)

Le système développé à OGI [Yan 96] repose sur la même architecture que le Parallel PRLM de Zissman. 6 décodeurs à base de MMC sont utilisés pour fournir les séquences phonétiques aux 66 modèles de langage. L'effort de recherche a principalement porté sur l'optimisation de ces derniers et sur le classificateur final (réseau de neurones). Le modèle phonotactique s'articule en trois modules :

- ✓ une grammaire bigramme classique prenant en compte le contexte passé du phonème (modèle backward),
- ✓ une grammaire bigramme prenant en compte le contexte futur du phonème (modèle forward),
- ✓ un modèle de durée.

L'objectif est bien évidemment d'approcher les performances que l'on peut attendre d'un modèle trigramme sans en avoir la complexité et surtout le besoin considérable en données. Avec cette méthode et une classification par réseau de neurones, Yan atteint 86,7 % d'identification correcte sur les 11 langues avec les énoncés de 45 secondes de durée.

2.1.12 Technical University of Ilmenau, Allemagne

Les travaux entrepris par J. Navrátil et W. Zühlke [Navrátil 97] visent aussi à optimiser les modèles de langage utilisés en IAL. A partir d'un unique décodeur phonétique implanté sous HTK avec les phonèmes anglais, des modèles de langage optimisés par deux approches sont appris :

- ✓ La première approche est basée sur une grammaire bigramme tenant compte des deux contextes précédant chaque phonème par l'intermédiaire du calcul de classes d'équivalences à partir de ces deux contextes,
- ✓ la seconde approche est basée sur des arbres de décision binaires.

Les expériences montrent que les deux modèles ainsi établis améliorent les performances par rapport à un modèles bigramme classique et que l'utilisation conjointe des deux modèles est encore plus efficace : le taux d'identification obtenu avec les phrases de 45 secondes de 9 langues du corpus OGI MLTS s'élève à 90,6 %.

2.2 Les approches neuro-mimétiques

2.2.1 OGI, Etats-Unis, (Berkling - Barnard)

E. Barnard mène avec K. Berkling des travaux basés sur une modélisation acoustico-phonétique par réseaux de neurones [Berkling 95]. Ces recherches visent à réduire la quantité de données nécessaires lors de l'apprentissage en regroupant les phonèmes.

Un premier algorithme est développé pour extraire des unités acoustiques « mixtes » (à mi-chemin entre phonèmes et classes phonétiques majeures) pertinentes en IAL. Le système s'articule ensuite autour de deux modules :

- ✓ un réseau de neurones est chargé de reconnaître la suite d'unités prononcées parmi les 60 unités issues de l'algorithme ,
- ✓ un classificateur (linéaire ou neuronal) calcule des statistiques sur les fréquences d'occurrence des différentes unités et suites d'unités pour chacune des langues.

Les meilleurs résultats sont de 59 % d'identification correcte avec 6 langues du corpus OGI MLTS.

2.2.2 OGI, Etats-Unis, (Muthusamy)

Y. K. Muthusamy, auteur d'un fameux article sur l'IAL [Muthusamy 94] a développé un système neuronal à deux niveaux [Muthusamy 93] :

- ✓ le premier effectue la segmentation du signal en 7 classes majeures. Le chemin trouvé par le réseau neuronal est optimisé par un algorithme de Viterbi,
- ✓ le second prend en charge l'identification de la langue à partir du chemin trouvé et d'un grand nombre de critères statistiques (fréquences d'occurrence de motifs bigrammes et trigrammes…). Il s'agit là encore d'un réseau neuronal.

Les meilleurs résultats obtenus par Muthusamy sont de 62,4 % d'identification correcte sur les phrases de 45 secondes de 10 langues du corpus OGI MLTS.

2.3 Les autres approches

2.3.1 Un système d'identification prosodique

S. Itahashi propose un système d'identification des langues basé uniquement sur la prosodie [Itahashi 95]. Il ne s'agit pas là de la première tentative ([Ives 86], [Foil 86], [Hazen 97]) mais assurément d'une des plus réussies. Après avoir opéré une détection voisé/non voisé, Itahashi effectue l'approximation de la courbe de F_0 par des fonctions linéaires par morceaux afin d'extraire 21 paramètres statistiques (liés aux variations de F_0 et d'énergie) de chaque zone voisée.

A partir d'une analyse discriminante utilisant une distance de Mahalanobis, les résultats obtenus avec les locuteurs masculins de 6 langues du corpus OGI MLTS sont remarquables puisqu'ils atteignent 63,3 % d'identification correcte avec 20 secondes de parole.

2.3.2 Un système d'identification du locuteur

L'approche proposée par K. P. Li en 1994 est bien différente des autres approches contemporaines et de ses travaux des années 80 [Li 80, Li 94]. Il s'agit en effet d'appliquer des techniques de reconnaissance du locuteur à l'IAL. Le processus, constitué de deux étapes, suppose que la langue à identifier est celle du locuteur le plus proche dans la base d'apprentissage.

La première phase consiste à extraire les noyaux syllabiques du signal acoustique avec un réseau de neurones, puis à calculer 75 paramètres acoustiques et spectraux. La seconde étape consiste alors à déterminer le locuteur le plus proche.

Cette approche se révèle pertinente puisque les résultats obtenus sur 10 langues du corpus OGI MLTS sont de l'ordre de 78 % d' identification correcte pour les phrases de 45 secondes de durée.

2.4 Discussion

Nous allons dans ce paragraphe récapituler les résultats obtenus par les différents systèmes et étudier les tendances générales qui s'en dégagent.

2.4.1 Tableau récapitulatif

Le Tableau 6 donne, pour chaque système cité dans ces pages, plusieurs informations sur son architecture et sur les résultats qu'il obtient :

Référence : il s'agit de l'article d'où sont tirées les informations suivantes.

Reconnaissance phonétique : lorsque le système repose sur un décodage acoustico-phonétique, son type (markovien ou neuronal) est mentionné. Lorsqu'il y a un chiffre entre parenthèses, il s'agit du nombre de décodeurs mis en parallèle dans l'expérience ; si ce chiffre n'est pas mentionné, il y a un décodeur par langue.

Modèle phonotactique : ce champ indique le type du modèle phonotactique utilisé (généralement N-gramme).

Prosodie : ce champ indique si un modèle prosodique est employé dans le système.

Corpus : on trouve ici le nom du corpus sur lequel les résultats indiqués sont obtenus.

Nombre de langues : il s'agit bien évidemment du nombre de langues sur lesquelles le test a porté.

Résultats : on trouve ici le résultat en pourcentage d'identification correcte dans une tâche d'identification des langues n'intégrant pas de décision de rejet. La durée des stimuli utilisés pour le test est également précisée.

Référence	Reconnaissance phonétique	Modèle phonotactique	Prosodie	Corpus	Nombre de langues	Résultats	
						Durée	%
Savic 91	MMC	-	oui	*n. p.*	*n. p.*	*n. p.*	*n. p.*
Berkling 94	RN	-	-	OGI MLTS	3	45 s	74,2
Dalsgaard 94	MMC (1)	-	-	EUROM 0	4	2 min.	88,1
Kadambe 94	MMC	N-gramme	-	OGI MLTS	3	45 s	91
Lamel 94	MMC	N-gramme	-	OGI MLTS	10	10 s	59,7
Li 94	RN	-	-	OGI MLTS	10	45 s	78
Muthusamy 93	RN (1)	RN	-	OGI MLTS	10	45 s	62,4
Ramesh 94	MMC	Modèles de mots	-	OGI MLTS[28]	4	5 à 10 s	96
Tucker 94	MMC	N-gramme	-	EUROM 1	3	10 s	90
Itahashi 95	-	-	oui	OGI MLTS[29]	6	20 s	63,3
Kwan 95	MMC (1)	N-gramme	-	OGI MLTS	5	45 s	78
Lund 95	MMC (1)	N-gramme	-	OGI MLTS	9	*n. p.*	*n. p.*
Hazen 97	MMC (1)	N-gramme	oui	OGI MLTS	11	45 s	78,1
Yan 96	MMC (6)	N-gramme	-	OGI MLTS	11	45 s	86,7
Zissman 96	MMC (6)	N-gramme	-	OGI MLTS	11	45 s	89
Andersen 97	MMC	N-gramme	-	OGI MLTS	3	45 s	83,7
Corredor-Ardoy 97	MMC (1)	N-gramme	-	IDEAL	4	10 s	91
Navrátil 97	MMC (1)	N-gramme + Arbre binaire	-	OGI MLTS	9	45 s	90,6

Tableau 6 – Récapitulatif des études en IAL citées (*n. p.* indique *non précisé*, MMC Modèle de Markov Caché et RN Réseau Neuromimétique)

2.4.2 Tendances générales en IAL

Sur les 18 études succinctement décrites dans ce chapitre, 14 utilisent une modélisation statistique markovienne pour effectuer un décodage acoustico-phonétique, 2 effectuent ce décodage via des réseaux neuromimétiques et 2 autres systèmes reposent sur des critères différents (modélisation statistique de la prosodie dans [Itahashi 95] et modélisation neuronale du locuteur dans [Li 94]).

[28] Il s'agit d'une modélisation par mots-clefs, seules les énumérations de chiffres sont prises en compte

[29] Seuls les locuteurs masculins sont pris en compte

La plupart des systèmes s'articulent en deux modules, le premier effectuant un décodage acoustico-phonétique de manière à fournir une ou plusieurs séquences d'unités phonétiques discrètes en entrée d'un second module, généralement basé sur une grammaire statistique, qui modélise alors les contraintes phonotactiques de la langue (12 systèmes sur 18). Les systèmes qui n'exploitent pas ce type d'information obtiennent généralement de moins bons résultats.

Si l'on étudie plus précisément les systèmes de décodage acoustico-phonétique, la tendance actuelle (95-97) privilégie l'usage d'un unique décodeur, commun à toutes les langues. Il peut être construit à partir des unités phonétiques d'une seule langue [Lund 95, Hazen 97, Navrátil 97] ou en faisant émerger un ensemble d'unités couvrant l'espace phonétique de toutes les langues [Berkling 95, Kwan 95, Corredor-Ardoy 97]. Une seconde approche consiste à utiliser plusieurs décodeurs dépendants d'une langue en parallèle, même s'ils ne correspondent pas aux langues à identifier [Yan 96, Zissman 96]. L'objectif est alors d'augmenter la robustesse de l'ensemble en recombinant plusieurs scores phonotactiques pour chaque langue plutôt qu'en en calculant un seul. Cette tendance fait reposer l'identification quasiment intégralement sur les modèles de langage, puisqu'en sortie du (ou des) décodeur(s) acoustico-phonétique, aucun score d'identification n'est généré.

Le principal avantage de ces méthodes est de réduire considérablement la quantité de données étiquetées nécessaires pour certaines langues, voire de supprimer ce besoin : lorsqu'on ne dispose pas de données étiquetées pour une langue, on la décode avec d'autres systèmes phonétiques. Cet avantage est essentiel dès lors que l'on augmente le nombre de langues traitées, et il justifie à lui seul l'usage qui est fait de cette méthode. Cela dit, il nous semble qu'un tel décodage du signal est particulièrement réducteur, et qu'il sous-exploite « l'identité phonétique » de chaque langue : en opérant une *projection* de données acoustiques d'une langue X dans l'espace phonétique de la langue Y, on perd une partie de l'information qui peut se révéler capitale. Le fait d'utiliser plusieurs décodeurs Y_1 à Y_M permet certes de diminuer ces pertes (par analogie avec du traitement d'antennes), mais, comme le souligne Zissman, le choix des décodeurs est alors crucial puisqu'il conditionne les performances du système.

Les recherches qui ont été menées au cours de cette thèse visent à réaliser une meilleure exploitation de « l'identité phonétique » des langues à identifier sans recourir pour autant à des données étiquetées pour chacune d'entre elles. L'approche choisie, appelée Modélisation Phonétique Différenciée (MPD) fait l'objet du chapitre suivant.

Chapitre 3

LA MODELISATION PHONETIQUE DIFFERENTIEE

A l'heure actuelle, les systèmes d'IAL sont capables d'identifier à 90 % une langue parmi 11 en exploitant 45 secondes de parole. Ces résultats, obtenus avec des corpus de parole de qualité médiocre (signal bruité, parole téléphonique…) sont déjà impressionnants, mais tous les problèmes sont loin d'être résolus !

Les axes majeurs du développement de l'IAL consistent à augmenter le nombre de langues traitées, à traiter des énoncés où plusieurs locuteurs interviennent (dialogue…) et à diminuer la durée de parole nécessaire à l'identification correcte (paragraphe 1.1). Les corpus OGI *22 languages* et CALLFRIEND permettent d'évaluer la robustesse des systèmes dans un cadre plus étendu, et déjà, la plupart des systèmes sont testés sur des énoncés de 10 secondes [30], en plus de ceux de 45 secondes.

L'adaptation des systèmes à ces conditions plus exigeantes passera-t-elle par des optimisations ponctuelles, requérant de plus en plus de puissance de calcul et de données d'apprentissage (une surenchère qui est déjà quotidienne en RAP), ou les solutions d'avenir viendront-elles d'une intégration de connaissances phonologiques plus poussées ? La réponse n'est probablement pas catégorique, mais elle viendra certainement d'une collaboration entre ingénieurs et linguistes (paragraphe 1.2). En France, la DGA (Délégation Générale pour l'Armement) est à l'origine du projet « Discrimination multilingue automatique » qui a rassemblé durant deux ans quatre laboratoires aux compétences diverses dans le but d'étudier la pertinence de certains indices discriminants en IAL. Après avoir brièvement présenté ce projet (paragraphe 1.2.2), nous nous attacherons à détailler la Modélisation Phonétique Différenciée (MPD) qui constitue le cadre des travaux entrepris au cours de cette thèse (paragraphe 2).

1 LES PERSPECTIVES D'UN DOMAINE EN EVOLUTION

1.1 Les limitations actuelles

Si les performances des systèmes d'IAL dont on dispose actuellement sont bonnes (obtenir près de 100 % d'identification correcte pour 4 ou 5 langues peut se révéler suffisant dans certaines applications), on se heurte cependant à des limitations qui peuvent s'avérer gênantes dès que l'on complexifie le problème.

[30] sur de tels énoncés, le système *parallel PRLM* de Zissman obtient encore 79 % d'identification correcte sur 11 langues.

La première des limitations est classique puisqu'il s'agit de l'absence de décision de rejet : il es t préférable que le système ne rende pas de décision plutôt qu'il en retourne une erronée. Cette constatation bien connue en TAP est parfaitement transposable à l'IAL, et les chercheurs en ont pleinement conscience. En effet, plusieurs travaux récents [Kwan 97, Parris 97] portent sur ce point précis [31], et il est prévisible que cette tendance s'accentue car l'intégration d'une décision de rejet efficace à un système est un préalable à son utilisation hors d'un laboratoire.

Une autre restriction porte sur le décodage acoustico-phonétique. En effet, il est certain que les décodeurs les plus performants sont entraînés avec une grande quantité de données étiquetées manuellement. Il s'agit là de l'aspect le plus coûteux de l'IAL car de tels corpus nécessitent de mobiliser un expert de la langue étudiée pendant un temps assez long. Si, pour certaines langues (anglais américain, anglais britannique, français, allemand…) on dispose de ce type de corpus (quoi que rarement enregistrés au travers du canal téléphonique), il est évident que dans la très grande majorité des langues, la quantité de données étiquetées disponibles n'est pas suffisante pour obtenir un décodeur robuste. La principale stratégie employée pour éluder ce problème consiste à ne pas utiliser de décodeur spécifique à chacune des langues (cf. paragraphe 2.4 du chapitre précédent).

Les différentes approches alors mises en œuvre (décodeur « trans-langue » ou mise en parallèle de plusieurs décodeurs déjà construits) présentent chacune des avantages en matière de réduction de coût par rapport à l'utilisation d'un décodeur par langue, mais généralement elles exploitent uniquement la suite de symboles phonétiques et non les scores de vraisemblance de la séquence. Ces approches utilisent donc les caractéristiques phonétiques de chaque langue de manière implicite dans la procédure d'identification (pas de score phonétique), et il nous semble que cela est bien moins efficace que d'envisager l'exploitation explicite de ces paramètres (en calculant en sortie du décodeur phonétique un score de vraisemblance pour chaque langue).

Etant données les performances atteintes par les systèmes actuels, on pourrait considérer que cette non-optimalité est sans importance, d'autant plus qu'elle permet de concentrer l'effort de recherche sur les modèles phonotactiques sur lesquels repose intégralement la décision. En fait, il n'en est rien, et les travaux de Zissman ont bien montré que la qualité de la modélisation acoustico-phonétique restait cruciale [Zissman 96]. De plus, lorsque l'on réduit la durée des stimuli utilisés pour l'identification (dans le cadre d'une application commerciale, 10 secondes est déjà un temps relativement long) ou lorsque l'on augmente le nombre de langues, il paraît essentiel d'exploiter le maximum d'informations contenues dans le signal.

[31] Depuis 1995 les campagnes d'évaluation NIST comprennent des tests avec rejets (test dits « en ensemble ouvert »).

Cette optimisation peut passer par une diversification des paramètres modélisés (phonétiques, phonotactiques et prosodiques), une optimisation de chaque module et/ou la prise en compte d'informations linguistiques.

1.2 Les axes de recherche

Dans un premier temps, nous allons analyser les modules qui composent classiquement un système d'IAL, et nous nous intéresserons ensuite aux travaux menés dans le cadre du projet DGA « Discrimination Multilingue Automatique » qui visait à préciser dans quel cadre les expertises linguistique et informatique pouvaient interagir.

1.2.1 Discussion

Si l'on se réfère aux approches classiques en IAL [Zissman 96], on peut décomposer le processus d'identification de la langue en une séquence de trois traitements :

- ✓ la génération d'une représentation acoustico-phonétique du signal,
- ✓ un test d'adéquation entre cette représentation et un ou plusieurs modèle(s) appris pour chacune des langues possibles,
- ✓ l'exploitation de ces scores pour prendre une décision.

L'IAL étant une discipline en évolution, de nombreuses améliorations peuvent être apportées à chacun des processus mis en jeu. Les différentes techniques d'optimisation issues de la RAP n'ont pas toutes été exploitées et elles seront pour la plupart efficaces en IAL. D'autre part, des méthodes originales de représentation multilingue [Andersen 97, Corredor-Ardoy 97] sont à la base d'approches prometteuses.

- **La représentation acoustico-phonétique**

Les méthodes employées actuellement relèvent pour la plupart de la RAP, que ce soit pour le prétraitement (filtrage RASTA, détection d'activité vocale) ou pour la modélisation proprement dite (modélisation de la durée, séparation des modèles féminins et masculins, choix des unités…). En IAL il est également nécessaire d'optimiser la couverture phonétique des langues à identifier. Que ce soit par la sélection d'unités communes à plusieurs langues [Corredor-Ardoy 97, Andersen 97] ou par le choix de plusieurs décodeurs spécifiques à une langue [Zissman 96, Yan 96], l'optimisation des espaces de décodage n'en est qu'à ses balbutiements. En particulier, dès lors que le *sens* de l'énoncé n'est pas l'information que l'on cherche à modéliser, la nature des unités phonétiques à utiliser est peut-être à revoir : il est possible que les unités les plus pertinentes en IAL ne soient pas les mêmes qu'en RAP ou qu'en identification du locuteur. La pertinence du niveau syllabique a déjà été évoquée du point de vue cognitif (Chapitre 3 de la première partie), et il est possible qu'il constitue une représentation adéquate et originale. D'autre part les unités utilisées sont généralement indépendantes du contexte (phonèmes plutôt qu'allophones), de manière à réaliser le compromis classique entre apprentissage efficace d'un petit nombre de modèles et description plus précise de l'énoncé. Nous pensons qu'une approche exploitant une segmentation *a priori* du signal, de manière à faire émerger des sous-unités transitoires et stables se révélerait payante [André-Obrecht 88, André-Obrecht 93].

Cette recherche d'un niveau de représentation indépendant de la langue, outre qu'elle permettrait de diminuer efficacement la quantité de données étiquetées nécessaires à l'apprentissage, présente une perspective séduisante du point de vue scientifique et dépassant largement le simple thème de l'IAL. Par contre, il reste particulièrement ardu d'exploiter un unique décodage acoustico-phonétique pour établir la vraisemblance d'une séquence symbolique par rapport aux différentes langues, comme le montrent les expériences menées en ce sens [Kwan 95].

- **La caractérisation de la langue**

Si à l'heure actuelle, les modèles utilisés pour identifier les langues cherchent principalement à modéliser des contraintes phonotactiques, il ne faut pas négliger pour autant d'autres aspects, comme la caractérisation acoustico-phonétique (abordée ci-dessus) et la modélisation prosodique.

Les modèles phonotactiques les plus courants sont issus d'une modélisation statistique par grammaire N-gramme : pour chaque énoncé à identifier, le système calcule la vraisemblance de la suite d'unités phonétiques décodées $\{\phi_1,\phi_2,...,\phi_T\}$ par rapport au modèle établi pour la langue L en considérant que la vraisemblance de chaque unité la constituant est fonction de son contexte passé limité à N unités :

$$\Pr(\phi_1,\phi_2,...,\phi_T|L) = \prod_{i\in[1..T]} \Pr(\phi_i|\phi_{i-N+1},...,\phi_{i-1},L)$$

Généralement, on se limite à $N = 2$ (modèle bigramme), là encore pour réaliser le compromis entre apprentissage efficace d'un nombre raisonnable de modèles[32] et prise en compte d'un contexte plus important, nécessitant plus de données d'apprentissage. De nombreuses heuristiques existent pour améliorer de tels systèmes : modèles trigrammes interpolés [Yan 96], modèles trigrammes utilisant une matrice d'unités de passage [Navrátil 97], modèles multigrammes à séquences de longueur variable [Deligne 96]. D'autres études sont menées également sur des formulations moins courantes de contraintes phonotactiques, comme l'usage d'arbres binaires [Navrátil 97].

Un autre axe de recherche, étudié avec plus ou moins de succès par le passé [Foil 86, Itahashi 95, Yan 96] repose sur la modélisation de la prosodie des langues. Que ce soit par l'intonation, l'accentuation ou le rythme, il est acquis qu'un certain niveau de discrimination existe entre les langues, comme le confirment les expériences perceptives (cf. le premier chapitre de la première partie). La modélisation de la prosodie n'est

[32] Si l'on dispose de P unités phonétiques, le nombre de modèles à apprendre est de l'ordre de P^2 pour les modèles bigrammes et de l'ordre de P^3 pour des modèles trigrammes.

cependant pas un problème trivial, et si les expériences en synthèse de la parole tendent à se généraliser, il n'en est pas de même pour l'IAL et la question demeure : comment obtenir une modélisation pertinente à partir de connaissances linguistiques qualitatives ?

A l'heure actuelle, il semble qu'une fusion précoce d'informations phonétiques et prosodiques (par exemple en entraînant un modèle de Markov avec des vecteurs de paramètres cepstraux et prosodiques) ne soit guère efficace [33] [Hazen 97]. Cela implique de développer une approche spécifique à ce type d'information et de réaliser une fusion tardive des décisions des différents modules.

Outre les difficultés liées à la mise au point d'une nouvelle modélisation (on peut citer [Farinas 98] sur ce sujet), on voit alors apparaître le problème de l'exploitation de scores issues d'approches hétérogènes.

- **La prise de décision**

La dernière étape du processus d'IAL repose sur la détermination de la langue parlée à partir de scores générés par les différents modèles mis à contribution.

Généralement, ces scores sont de nature quantitative (vraisemblance ou distance) et lorsqu'ils sont issus d'un unique modèle par langue, la prise de décision se réduit alors à sélectionner la vraisemblance maximale ou la distance minimale. Dès lors que l'on met à contribution plusieurs modèles (modèles acoustico-phonétiques, phonotactiques ou prosodiques), on se heurte à des difficultés supplémentaires que l'on regroupe généralement sous le terme de fusion de données. Il est en effet nécessaire de normaliser les différents flux de manière à ne pas biaiser les résultats, et d'autre part il peut être judicieux de pondérer différemment chaque source d'information, de manière à réaliser une fusion optimale. Des réseaux de neurones à la logique floue, les outils pour atteindre cet objectif ne manquent pas, mais ils nécessitent généralement une bonne connaissance des modèles employés, et dans la plupart des cas, les classificateurs se limitent à une approche bayésienne.

Un autre type de difficultés peut survenir si l'on souhaite intégrer des informations de nature qualitative (et non plus quantitative) dans le module de décision. La prise en compte de critères linguistiques entre la plupart du temps dans ce cadre. L'expérience a montré en RAP que l'intégration de ce type de connaissances était souvent décevante ; cela laisse supposer qu'il peut-être plus efficace de développer des approches spécifiques, comme nous allons le voir maintenant.

[33] Le décalage entre caractéristiques prosodiques supra-segmentales et unités phonétiques segmentales justifie en grande part cette inefficacité.

1.2.2 Le projet DGA « Discrimination multilingue automatique »

Le projet « Discrimination multilingue automatique » est basé sur une collaboration entre quatre laboratoires français, réunis dans le cadre d'une convention (n° 95/118) avec la DGA entre 1996 et 1998. Les partenaires sont :

- l'Institut de la Communication Parlée (ICP) de Grenoble,
- l'Institut de Linguistique et Phonétique Générales et Appliquées (ILPGA) de Paris,
- le laboratoire de Dynamique Du Langage (DDL) de Lyon,
- l'Institut de Recherche en Informatique de Toulouse (IRIT).

Outre la mise en place d'une synergie entre linguistes et informaticiens, l'objectif principal du projet était double puisqu'il s'agissait de :

- rechercher des critères discriminants pour l'IAL,
- définir une typologie en vue de la détermination robuste d'identifiants linguistiques (système vocalique, prosodie).

Le constat de départ était que, la modélisation classique en IAL reposant sur des méthodes empruntées à la RAP, il pouvait être fructueux de rechercher d'autres paramètres, ignorés dans ces systèmes. Le travail des différents laboratoires a donc consisté à déterminer quelles informations pouvaient être d'une part porteuses de l'identité de la langue et d'autre part extraites de manière robuste du signal. A cette fin les laboratoires de linguistique (DDL, ICP et ILPGA) ont enregistré des données multilingues selon un protocole commun mis au point par le DDL et l'ILPGA. La collecte de ces corpus visait à fournir un matériau où plusieurs paramètres étaient contrôlés. Ces paramètres étaient principalement :

- le débit (rapide, normal ou lent),
- le mode d'articulation (hyperarticulé, normal ou hypoarticulé),
- le style de texte (littéraire, scientifique…).

Les travaux menés à l'ICP ont principalement porté sur l'établissement à partir d'UPSID de typologies (vocaliques [Vallée 94] et consonantiques [Vallée 98]) et sur la recherche d'universaux du langage [Schwartz 97] tandis que l'ILPGA et le DDL s'intéressaient plus globalement à la recherche d'indices discriminants. On peut classiquement regrouper les traits étudiés en traits segmentaux et en traits suprasegmentaux, selon qu'ils portent sur les sons constitutifs d'une langue (nature, fréquence d'occurrence) ou sur les caractéristiques prosodiques des énoncés [Hombert 98, Vaissière 98]. Cette classification est en fait légèrement restrictive puisque les études suprasegmentales ont également porté sur les enchaînements de sons – ou de séquences – et sur l'émergence d'unités plus étendues que les segments phonétiques, telles que les syllabes [Hombert 98]. Les études menées au DDL ont d'autre part porté sur les variations dialectales observées dans les différentes zones géographiques de langue arabe [Barkat 97], et sur l'extraction d'indices discriminants entre ces parlers. Ces travaux ayant donné lieu à une collaboration poussée avec les recherches présentées dans ce manuscrit, nous y reviendrons au Chapitre 2 de la troisième partie.

Les résultats obtenus à l'heure actuelle confirment qu'il existe une multitude d'indices pertinents permettant de réduire le champ d'investigation lorsque l'on se trouve en présence d'une langue inconnue (présence de certaines consonnes, présence de voyelles nasales [34], enchaînement des syllabes, présence de syllabes récurrentes…). Plusieurs questions demeurent sur l'apport de tels critères dans un système automatique. On peut en effet se demander a propos de ces connaissances d'une part si elles ne sont pas *implicitement* modélisées dans les modèles phonétiques et phonotactiques, et d'autre part sous quelle forme les modéliser de manière *explicite*.

On peut par exemple s'intéresser à la modélisation des informations phonotactiques. Même si les modèles phonotactiques sont particulièrement pertinents en IAL, il nous semble qu'ils ne prennent pas pleinement en compte la réalité multiniveau de la parole : dans chaque langue, des règles régissent l'enchaînement des unités acoustico-phonétiques mais aussi les enchaînements d'unités plus longues (typiquement les syllabes). Il est probable qu'un modèle de type bi-multigramme [Deligne 98] puisse mieux intégrer ce type de contraintes multiniveaux qu'un modèle n-gramme classique mais la mise au point de tels modèles requiert une expertise assez importante pour faire apparaître les niveaux pertinents. Il est évident qu'une collaboration entre linguistes et informaticiens est nécessaire, tout comme elle peut l'être pour la modélisation acoustico-phonétique, un domaine où beaucoup reste à faire.

Au vu des expériences menées en RAP, les approches basées sur la modélisation des connaissances linguistiques conduisent à un relatif échec (Figure 15). L'utilisation d'approches inductives en IAL mène là encore à des résultats mitigés [Samouelian 96]. Nous pensons qu'une intégration *en amont* des connaissances par rapport aux modèles statistiques pourrait se révéler efficace. L'idée n'est plus de faire reposer la décision sur un modèle de connaissances, mais plutôt d'utiliser ces connaissances au sein d'un pré-traitement linguistique (Figure 16). L'objectif est de prendre en compte cette expertise dès la conception des modèles statistiques, qui sont par la suite seuls à intervenir dans le processus de décision. Ces modèles statistiques peuvent être de différents types, adaptés aux informations modélisées (mélodie, rythme, système vocalique…) et ils peuvent être optimisés séparément, alors que l'approche statistique (ou neuronale) classique n'autorise pas cette liberté (cas d'un modèle acoustico-phonétique standard). Le paragraphe suivant décrit l'approche mise en œuvre au cours de cette thèse. Le pré-traitement linguistique consiste à segmenter l'énoncé en classes phonétiques, comme nous allons maintenant le voir.

[34] lorsqu'elles n'assimilent pas le trait de nasalité d'une consonne voisine.

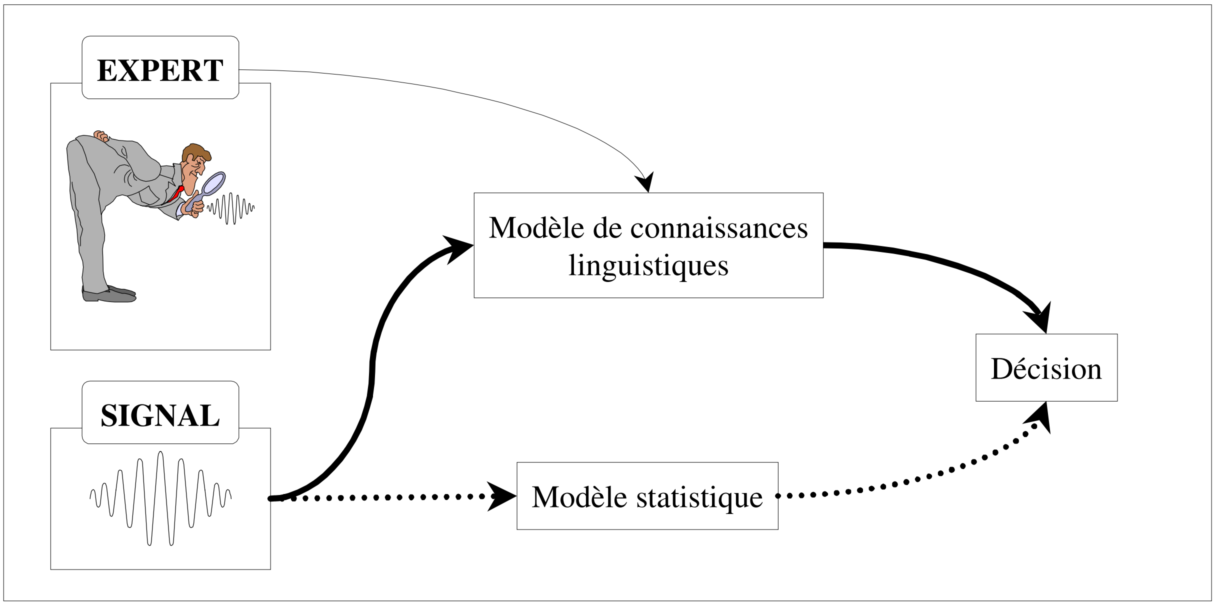

Figure 15 – Modèle basé sur des connaissances, avec une éventuelle composante statistique.

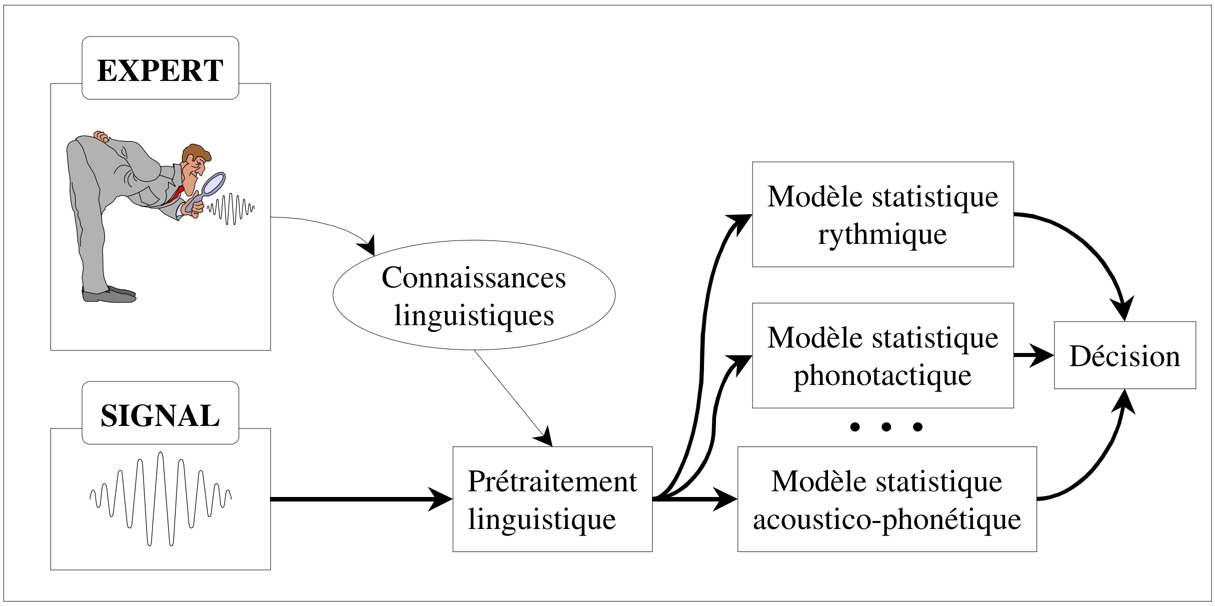

Figure 16 – Modèles statistiques avec utilisation amont des connaissances linguistiques.

2 L'APPROCHE PAR MODELISATION PHONETIQUE DIFFERENTIEE

2.1 Les motivations

Les meilleures performances en IAL sont obtenues à l'heure actuelle grâce à la discrimination par modèles phonotactiques. L'un des avantages principaux de ces modèles est qu'ils peuvent être appris automatiquement sans données étiquetées. Pour être efficaces, de tels systèmes nécessitent cependant de disposer d'un ou plusieurs décodage(s) acoustico-phonétique(s) performant(s) en entrée. Il est évident que la conception de ces décodeurs est particulièrement coûteuse sur le plan de la collecte des données et de leur étiquetage. Les systèmes standards reposent donc sur un paradoxe : l'étape la plus coûteuse du processus (le décodage acoustico-phonétique) n'est pas *explicitement* exploitée pour discriminer les langues, mais *implicitement* employée comme pré-traitement à la modélisation phonotactique. Cette réalité est à notre avis sous-optimale pour deux raisons :

- elle ne résulte pas d'une optimisation de la représentation acoustico-phonétique dans un but de discrimination,
- elle n'exploite pas de manière quantitative la vraisemblance des séquences phonétiques produites.

Nous préconisons donc une utilisation conjointe des modèles acoustico-phonétiques et phonotactiques pour calculer le score d'identification, de manière à exploiter le maximum d'informations. De plus, il nous semble opportun d'envisager la prise en compte d'informations *a priori* dans les modèles acoustico-phonétiques. Nous proposons donc une approche, la Modélisation Phonétique Différenciée, basée sur les remarques suivantes :

- les typologies des langues [Vallée 94, Vallée 98] distinguent les systèmes vocaliques des systèmes consonantiques, et à l'intérieur de ces systèmes, certaines consonnes sont encore représentées séparément (fricatives, plosives…),
- ces typologies peuvent être utilisées pour identifier une langue – ou un groupe de langues – à partir de sa description phonologique,
- les paramètres les plus pertinents pour caractériser un son peuvent dépendre de sa classe phonétique : décrire une plosive en terme de formants est inadapté alors qu'il s'agit d'une représentation adéquate pour d'autres sons,
- lorsque l'on modélise ensemble des sons de nature homogène (par exemple les voyelles), on peut plus facilement prendre en compte certaines contraintes spécifiques (limites acoustiques de l'espace vocalique dans cet exemple).

Ces remarques nous amènent à envisager une modélisation différenciée de chaque système phonologique (système vocalique, système des consonnes fricatives…) redéfini en tenant compte de contraintes liées à la représentation acoustico-phonétique de la parole spontanée (coarticulation…). Un tel système est présenté sur la Figure 17. Il s'articule en trois phases, abordant chacune un aspect du problème.

Le pré-traitement doit réaliser un étiquetage du signal en segments des différentes classes phonétiques retenues. Les algorithmes mis en œuvre peuvent être de différents types (analyse spectrale, modèles statistiques…), mais dans le but de réaliser un système ne requérant aucun étiquetage manuel, nous avons écarté les algorithmes basés sur un apprentissage supervisé. D'autre part, nous avons opté pour des algorithmes totalement indépendants de la langue traitée, de manière à appliquer un pré-traitement unique lors de la phase d'identification de la langue.

Un modèle acoustico-phonétique différencié est mis en œuvre pour chacune des classes phonétiques retenues. La paramétrisation adoptée peut-être la même pour chaque classe ou adaptée à sa structure acoustique. Les modèles sont bien entendu dépendants de chaque langue, de manière à pouvoir effectuer une discrimination basée sur leurs vraisemblances.

Au cours d'une dernière étape, les scores de vraisemblance des différents modèles relatifs à une langue sont recombinés. Cette opération peut nécessiter des techniques de normalisation, en particulier si les espaces paramétriques sont distincts pour chaque modèle.

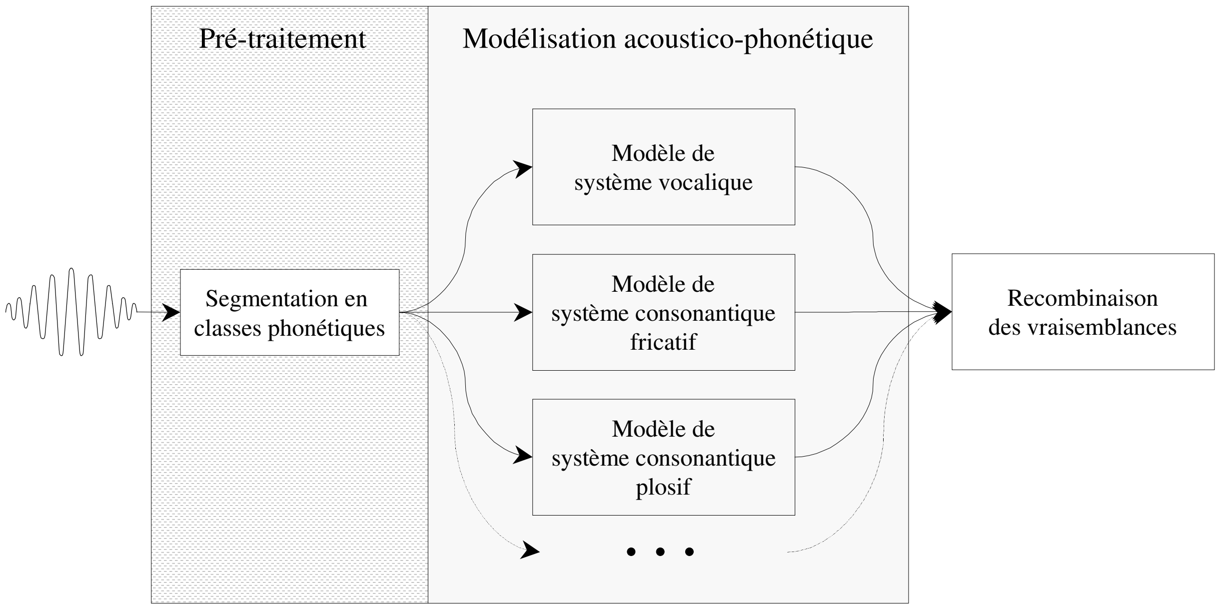

Figure 17 – Schéma d'un modèle acoustico-phonétique basé sur la modélisation phonétique différenciée.

Lorsque les modèles adoptés pour chaque classe phonétique le permettent (cas de modèles de Markov par exemple), on peut ensuite exploiter les contraintes phonotactiques de chaque langue dans un modèle de type grammaire n-gramme et on obtient alors un système d'IAL comparable aux systèmes de référence (Figure 18), mais exploitant explicitement les caractéristiques acoustico-phonétiques, sans nécessiter de données étiquetées.

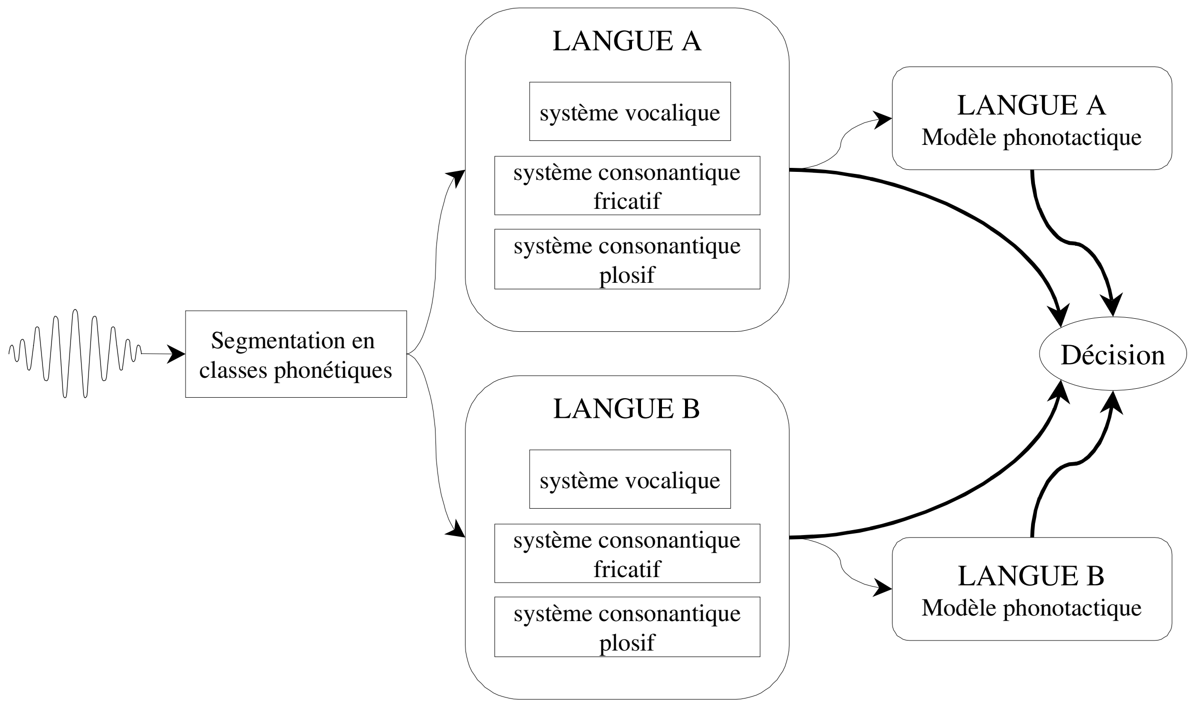

Figure 18 – Exemple de système d'IAL basé sur la modélisation phonétique différenciée. Cas de deux langues A et B.

2.2 Le cadre statistique de l'étude

Nous allons maintenant proposer une formulation statistique du problème de l'IAL tel qu'il se pose dans l'approche phonétique différenciée. Nous nous plaçons dans le cadre de l'identification d'une langue en ensemble fermé, c'est-à-dire que lorsque un énoncé est prononcé dans une langue inconnue, celle-ci fait partie des N_L langues apprises par le système.

Soit $L = \{L_1, L_2, \ldots, L_{N_L}\}$ l'ensemble des N_L langues modélisées. Le problème est de déterminer la langue L^* dans laquelle un énoncé inconnu a été prononcé. Dans notre approche basée sur le maximum de vraisemblance, on va déterminer la langue la plus vraisemblable qui est donc une estimation $\hat{L}^*$ de L^*. Dans la suite, nous allons utiliser par commodité la notation L^* plutôt que $\hat{L}^*$.

Soit $\Omega = \{\Omega_1, \Omega_2, \ldots, \Omega_M\}$ l'ensemble des M classes phonétiques du modèle ; elles sont indépendantes de la langue. A l'issue du pré-traitement acoustique, l'énoncé est paramétré par une suite de T segments phonétiques $\Phi = \{\phi_1, \phi_2, \ldots, \phi_T\}$ à laquelle correspond une suite de T vecteurs d'observation $O = \{o_1, o_2, \ldots, o_T\}$. Nous avons alors par définition :

$$L^* = \arg\max_{1 \leq i \leq N_L} \left(\Pr(L_i | O) \right) \qquad (1.)$$

Le théorème de Bayes nous permet d'écrire :

$$L^* = \arg\max_{1\le i\le NL}\left(\frac{\Pr(O|L_i)\Pr(L_i)}{\Pr(O)}\right) = \arg\max_{1\le i\le NL}(\Pr(O|L_i)\Pr(L_i)) = \arg\max_{1\le i\le NL}(\Pr(O|L_i)) \qquad (2.)$$

en considérant que la probabilité d'apparition *a priori* de chaque langue est équiprobable. Si l'on introduit la séquence de symboles phonétiques Φ dans la formule précédente, on a :

$$L^* = \arg\max_{1\le i\le NL}(\Pr(O|L_i)) = \arg\max_{1\le i\le NL}\left(\sum_{\Phi\in\Omega^T}\Pr(O,\Phi|L_i)\right) \qquad (3.)$$

qui peut encore se reformuler :

$$L^* = \arg\max_{1\le i\le NL}\left(\sum_{\Phi\in\Omega^T}\left\{\Pr(O|\Phi,L_i).\Pr(\Phi|L_i)\right\}\right) \qquad (4.)$$

Dans le cadre de notre étude, la suite Φ est obtenue par un traitement déterministe, et elle est donc unique pour un énoncé donné ; la langue la plus vraisemblable est alors donnée par l'équation :

$$L^* = \arg\max_{1\le i\le NL}\left(\Pr(O|\Phi,L_i).\Pr(\Phi|L_i)\right) \qquad (5.)$$

On retrouve dans cette équation les deux termes liés à la caractérisation acoustico-phonétique de la langue $\left(\Pr(O|\Phi,L_i)\right)$ et à la modélisation des contraintes phonotactiques $\left(\Pr(\Phi|L_i)\right)$. Par contre, et à l'inverse d'un modèle classique où le cardinal de l'ensemble des unités phonétiques est relativement important (de l'ordre de 40 pour des unités indépendantes du contexte en français), la vraisemblance phonotactique est ici calculée à partir d'une suite Φ de classes phonétiques en nombre plus réduit (de l'ordre de la dizaine au maximum) et indépendantes de la langue. On peut donc s'attendre à ce que le pouvoir de discrimination entre langues du modèle phonotactique soit plus faible que dans l'approche classique, même si des contraintes différentes existent dans chaque langue au niveau des enchaînements de classes phonétiques. Pour pallier à ce possible manque d'efficacité, nous proposons donc d'introduire une suite Ψ^i issue du décodage de O dans les modèles différenciés de la langue i : le symbole ψ_k^i correspond à l'état acoustico-phonétique du modèle de la classe ϕ_k dans laquelle le k[ième] segment a été décodé. La vraisemblance phonotactique est donc donnée par :

$$\Pr(\Phi|L_i) = \sum_{\Psi^i}\Pr\left(\Phi,\Psi^i|L_i\right) = \sum_{\Psi^i}\left\{\Pr(\Phi|\Psi^i,L_i).\Pr(\Psi^i|L_i)\right\} \qquad (6.)$$

Si l'on suppose que les segments sont indépendants les uns des autres conditionnellement aux états, on a :

$$\Pr(\Phi|L_i) = \sum_{\Psi^i}\left(\Pr(\Psi^i|L_i).\prod_{k=1}^{T}\Pr(\phi_k|\psi_k^i,L_i)\right) = \sum_{\Psi^i}\Pr(\Psi^i|L_i) \qquad (7.)$$

car $\Pr(\phi_k | \psi^i_k, L_i) = 1 \quad \forall k \in [1,...,T]$ puisque ψ^i_k est un état du modèle différencié correspondant à la classe ϕ_k dans la langue i.

D'après (4.) et (7.), la langue la plus vraisemblable est donnée par :

$$L^* = \arg\max_{1 \leq i \leq N_L} \left(\Pr(O | \Phi, L_i) . \sum_{\Psi^i} \Pr(\Psi^i | L_i) \right) \qquad (8.)$$

Si l'on s'intéresse maintenant à la vraisemblance acoustico-phonétique, en supposant là encore que les segments sont indépendants les uns des autres conditionnellement aux états, on a

$$\Pr(O | \Phi, L_i) = \prod_{k=1}^{T} \Pr(o_k | \phi_k, L_i) = \prod_{\phi_k = \Omega_1} \Pr_{\Omega_1}(o_k | L_i) . \prod_{\phi_k = \Omega_2} \Pr_{\Omega_2}(o_k | L_i) ... \prod_{\phi_k = \Omega_M} \Pr_{\Omega_M}(o_k | L_i) \qquad (9.)$$

La vraisemblance est donc le produit des vraisemblances dans chacun des modèles phonétiques différenciés. Cette formulation permet en fait d'optimiser chacun des modèles différenciés indépendamment des autres. C'est cette approche que nous avons développé par la suite, de manière à étudier le pouvoir discriminant porté par les systèmes vocaliques.

3 L'ETUDE REALISEE : LA MODELISATION DES SYSTEMES VOCALIQUES

3.1 Pourquoi une approche vocalique ?

Le préambule nécessaire à la mise en place d'un système d'IAL basé sur la modélisation différenciée réside dans l'implantation du module réalisant l'étiquetage en classes majeures. Nous avons souhaité développer des algorithmes indépendants de la langue et, dans la mesure du possible, ne requérant pas d'apprentissage. Nous nous sommes donc orientés vers des méthodes d'analyse spectrale du signal acoustique. La mise au point de tels algorithmes ne se fait pas instantanément, et nous avons dirigé notre effort initial vers une seule classe phonétique. Notre choix s'est porté sur les voyelles pour plusieurs raisons, empruntant à la fois à la phonologie (cf. première partie, Chapitre 2) et au traitement de signal.

L'existence d'une typologie des systèmes vocaliques récente [Vallée 94] fournit un cadre phonologique idéal en vue de l'étude acoustico-phonétique des systèmes vocaliques. De plus, il se dégage des études réalisées que ces systèmes présentent un fort caractère discriminant tout en offrant les avantages d'une représentation homogène. La représentation acoustique des voyelles présente en effet une isomorphie remarquable avec leur représentation articulatoire (avant/arrière, ouvert/fermé) alors que ce n'est pas le cas des consonnes, pour lesquelles les variations de lieu et de mode d'articulation entraînent des changements de structure acoustique.

Sur le plan du traitement du signal acoustique, nous avons principalement travaillé avec des données enregistrées au travers du canal téléphonique (corpus OGI MLTS) et la bande passante résultante altère moins les voyelles que certaines consonnes comme les fricatives. De manière plus prosaïque, la détection automatique des voyelles nous a semblé moins complexe à mettre en œuvre que la détection de certaines consonnes.

3.2 Description de l'étude réalisée

Nous avons vu (équations 8 et 9) que dans le cadre de la modélisation phonétique différenciée, la langue la plus vraisemblable est donnée par :

$$L^* = \underset{1 \le i \le NL}{\arg\max}\left(\left[\prod_{\phi k=\Omega 1}\Pr_{\Omega 1}(o_k|L_i).\prod_{\phi k=\Omega 2}\Pr_{\Omega 2}(o_k|L_i)...\prod_{\phi k=\Omega M}\Pr_{\Omega M}(o_k|L_i)\right].\left[\sum_{\Psi^i}\Pr(\Psi^i|L_i)\right]\right) \quad (10.)$$

Nous avons été amenés à simplifier cette expression, car notre étude a porté uniquement sur les possibilités d'identification des langues par discrimination des systèmes vocaliques acoustiques. La description en classes phonétiques se réduit donc à la classe vocalique et à la classe consonantique : $\Omega = \{C, V\}$. D'autre part, nous n'avons pas pris en compte la modélisation des contraintes phonotactiques dans notre étude ; dans ce cadre restreint, L* est donné par l'expression :

$$L^* = \underset{1 \le i \le NL}{\arg\max}\left(\prod_{\phi k=V}\Pr_V(o_k|L_i).\prod_{\phi k=C}\Pr_C(o_k|L_i)\right) \quad (11.)$$

Le système résultant (Figure 19) permet d'étudier spécifiquement l'apport de la modélisation acoustico-phonétique des systèmes vocaliques dans une tâche d'IAL en comparant les résultats avec une approche par modélisation globale. L'ensemble de ces travaux, de la détection automatique à la prise de décision, fait l'objet de la troisième partie de ce manuscrit.

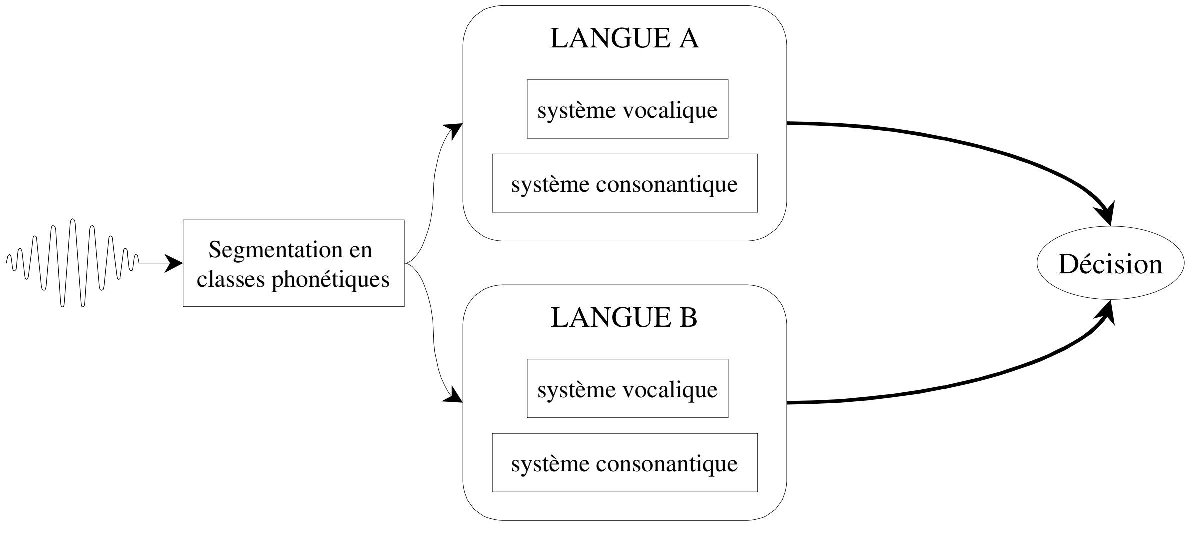

Figure 19 – Exemple de système d'IAL basé sur la modélisation différenciée des systèmes vocaliques. Cas de deux langues A et B.

CONCLUSION

Nous avons vu au cours de cette seconde partie que l'IAL est devenu un domaine majeur du traitement automatique de la parole. Les enjeux sont nombreux, et l'internationalisation des sociétés et des communications est à l'origine d'une demande pressente pour des systèmes opérationnels. L'état de l'art proposé met en évidence le fait que l'IAL vit une époque charnière. Les systèmes actuels obtiennent de bons résultats lorsque l'on dispose d'une quantité de données étiquetées convenable et lorsque le locuteur parle suffisamment longtemps. Bon nombre de langues, parlées par des millions de personnes, n'offrent pas de tels corpus. Alors que les méthodes se basent principalement sur des techniques de RAP, nous pensons qu'il est nécessaire de profiter de l'expertise des linguistes pour établir de nouveaux indices discriminants extractibles automatiquement. C'est l'objectif des travaux entrepris dans le cadre du projet DGA 95/118, et plus particulièrement de la modélisation phonétique différenciée introduite dans ce manuscrit. Notre approche vise à modéliser les sons du langage dans des espaces homogènes plutôt que dans un espace global et commun. L'ambition poursuivie est double puisqu'il s'agit d'utiliser plus efficacement les caractéristiques acoustico-phonétiques des langues (trop souvent délaissées au profit des modèles phonotactiques) et de s'affranchir de la nécessité de disposer de données étiquetées en effectuant la différentiation des sons par une approche non supervisée. Durant ces trois années, nous nous sommes focalisés sur l'étude de la faisabilité de cette approche, en étudiant tout particulièrement le pouvoir discriminant des systèmes vocaliques générés automatiquement à partir du signal.

(1598) - © Heritage Map Museum

3ème Partie

La modélisation des systèmes vocaliques appliquée à l'IAL

INTRODUCTION

La réflexion menée au cours de la seconde partie à partir d'un état de l'art de l'IAL nous amène à envisager une approche en discrimination multilingue basée sur la caractérisation phonétique des langues. Cette approche vise à tirer profit des connaissances linguistiques disponibles en élaborant des modèles acoustico-phonétiques différenciés inspirés de typologies phonologiques. Nous avons choisi d'étudier en priorité la capacité discriminante de modèles vocaliques pour plusieurs raisons détaillées précédemment. Ces travaux ont pour but principal d'évaluer la faisabilité d'un système d'IAL intégralement non supervisé, c'est-à-dire ne requérant à aucun moment d'étiquetage du signal par un expert humain, de manière à disposer d'un système totalement portable d'une langue à une autre. L'autre objectif poursuivi est de vérifier si la modélisation différenciée est aussi, voire plus efficace qu'une modélisation globale en IAL.

Le système conçu durant cette thèse est basé sur la localisation automatique des voyelles dans le signal, suivie d'une modélisation statistique multigaussienne du système vocalique acoustico-phonétique des locuteurs de différentes langues.

Nous exposons au premier chapitre la méthode de détection des voyelles implantée. Il s'agit d'un module intégrant des algorithmes de segmentation *a priori* du signal, de détection d'activité vocale et de localisation spectrale de voyelles. Plusieurs expérimentations menées sur deux corpus français monolocuteurs et sur le corpus de référence en IAL, le corpus OGI MLTS, sont décrites. Parmi les onze langues qu'il contient, nous en avons retenu cinq présentant des caractéristiques intéressantes au niveau du SV phonologique (cf. chapitre 1, paragraphe 2.2.1). Ces langues sont le coréen, l'espagnol, le français, le japonais et le vietnamien, et elles constituent donc notre corpus de référence en IAL. Au cours du chapitre suivant sont précisés l'algorithme de paramétrisation cepstrale employé ainsi que la méthode statistique MMG (modèle de mélanges de lois gaussiennes) utilisée pour modéliser chaque système vocalique. Cette méthode est couplée à un algorithme intégrant un critère d'information afin de déterminer le nombre de composantes adéquat pour chaque MMG : l'algorithme LBG-Rissanen. Des expériences de modélisation des SV et de classification de voyelles complètent ce second chapitre, de manière à évaluer la modélisation obtenue. Le troisième chapitre est consacré à l'étude du pouvoir discriminant des modèles de SV. Après avoir réalisé un système basé sur un modèle acoustico-phonétique de l'ensemble des sons, nous évaluons l'apport de la modélisation différenciée des SV par rapport à cette méthode de référence afin de valider l'approche phonétique différenciée par des expériences d'IAL pratiquées sur les cinq langues précitées. Nous présentons au cours du quatrième et dernier chapitre de cette partie des expériences complémentaires visant à réaliser un système d'IAL phonétique complet, fondé sur la mise en œuvre d'un modèle consonantique de manière à prendre en compte la totalité des informations phonétiques présentes dans le signal (et non pas uniquement le tiers de la durée du signal correspondant aux voyelles).

Chapitre 1

LA DETECTION AUTOMATIQUE DES SEGMENTS VOCALIQUES

Un préalable nécessaire à la modélisation différenciée des systèmes vocaliques est bien entendu la détection des voyelles dans le signal acoustique. Le terme même de *voyelle* est ambigu dans le cadre de la parole continue et spontanée, et la correspondance entre niveau acoustique et phonétique n'est pas triviale. La stratégie de prononciation spontanée diffère généralement assez nettement de celle de la parole lue, et le comportement des voyelles dépend des langues. Si l'on prend l'exemple de l'anglais, les voyelles en position non accentuée sont particulièrement affaiblies, et la stratégie de production vise à préserver les consonnes, alors qu'en français, la tendance sera plutôt inverse. Cet état de fait nous a amené à employer préférentiellement le terme de *segment vocalique* plutôt que de *voyelle*. Un segment vocalique est un segment acoustique correspondant phonétiquement à une voyelle, mais celle-ci a pu voir ses caractéristiques acoustiques fortement altérées par les phénomènes liés à la parole continue spontanée (coarticulation, réduction, assimilation).

Le premier paragraphe présente l'algorithme de détection des segments vocaliques que nous avons développé au cours de cette thèse, et le second est consacré aux expériences menées sur différents corpus allant de la parole lue à la parole spontanée.

1 LOCALISATION DES SEGMENTS VOCALIQUES

Peu d'expériences visent explicitement à localiser les voyelles dans le signal. Les systèmes de reconnaissance de la parole effectuent implicitement cette tâche en opérant une identification des unités phonétiques prononcées, mais il s'agit comme nous l'avons vu de systèmes basés sur des apprentissages supervisés. Deux études récentes cherchent à effectuer une localisation des voyelles dans une tâche monolingue, indépendante du locuteur. Dans [Sirigos 96], un perceptron multicouche est utilisé pour fournir une classification Voyelle / Non Voyelle à partir de 15 coefficients cepstraux et des quatre premiers formants (fréquence et bande passante). Plusieurs règles de cohérence temporelle (durée des sons, lissage de la sortie du réseau de neurones) sont appliquées pour améliorer la robustesse du système. Les résultats obtenus sur un sous-ensemble du corpus TIMIT (corpus de très bonne qualité) sont excellents : moins de 2,5 % de fausses acceptations et moins de 3,5 % de faux rejets. On peut comparer ces résultats avec ceux publiés dans [Fakotakis 97] et basés sur une discrimination par modèles de Markov cachés (un pour les voyelles, un pour les autres sons). En utilisant 19 coefficients cepstraux ainsi que leurs dérivées et l'énergie, le taux de fausses acceptations sur le corpus TIMIT est inférieur à 3,5 % pour un taux de faux rejets de l'ordre de 30 %. L'objectif des auteurs étant d'utiliser ces voyelles détectées en identification ou en vérification du locuteur (selon une approche apparentée à la nôtre en IAL), il semble qu'ils n'aient pas cherché à optimiser leur module de détection.

Ces deux expériences montrent assez nettement la difficulté à détecter les voyelles alors qu'elles sont réalisées dans des conditions (enregistrement de qualité, cadre monolingue, étiquetage disponible) plus favorables que celles que nous nous sommes fixées. Rappelons en effet que l'objectif poursuivi est de réaliser un algorithme indépendant de la langue, fonctionnant sur des enregistrements bruités et ne nécessitant pas l'utilisation de données étiquetées.

1.1 Un schéma synoptique du système

Nous avons choisi de baser notre module de détection vocalique sur la recherche d'événements caractéristiques des voyelles à partir d'une analyse spectrale. La méthode adoptée a donc consisté à rechercher une fonction spectrale du signal et à établir des critères de localisation des segments vocaliques. Dans ce type d'approche, il est clair que la difficulté consiste à obtenir des règles robustes vis à vis des conditions d'enregistrements, des changements de locuteurs et évidemment des changements de langues. Afin de renforcer cette robustesse (en particulier vis à vis du bruit), deux pré-traitements sont appliqués au signal acoustique avant de procéder à la détection proprement dite (Figure 20). Ces algorithmes visent à fournir une segmentation du signal et à distinguer les zones d'activité vocale des zones de silence.

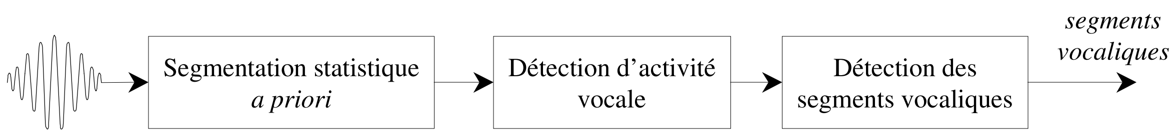

Figure 20 – Schéma bloc du système de détection des segments vocaliques.

Avant de décrire chacun des modules du système de détection, nous allons présenter brièvement le signal qui va illustrer les différents algorithmes (Figure 21) : il est extrait de la base de données OGI MLTS et il s'agit donc d'un signal échantillonné à 8 kHz et enregistré à travers le canal téléphonique. Le locuteur – un homme français – prononce la phrase « cent quatre vingt jours par an, il pleut ». Dans le fichier original, la phrase continue après une pause assez longue. Nous avons choisi cet exemple parce qu'il présente un rapport signal à bruit important et que la lecture en est facilitée.

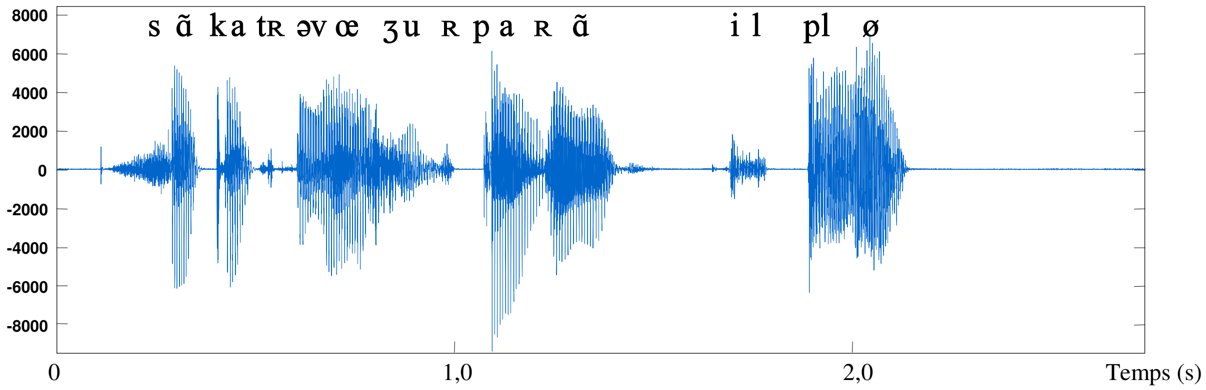

Figure 21 – Exemple de signal de la base de données OGI. Le locuteur prononce la phrase « cent quatre-vingt jours par an, il pleut ».

1.2 La segmentation du signal

L'un des exercices les plus périlleux du TAP consiste à déterminer les frontières des différentes unités phonétiques contenues dans un énoncé. Cette difficulté tient à la nature même de la parole continue : les unités sont fortement coarticulées, et l'on passe souvent de l'une à l'autre de manière continue. Pourtant, on cherche à poser des frontières strictes, puisque la plupart des modèles fonctionnent avec des unités discrètes. Il semble que le problème s'apparente en fait plus à la détermination d'une partition floue, où une trame très courte de parole appartient plus ou moins aux deux unités phonétiques adjacentes. Le problème n'est pas si simple car des phénomènes plus complexes interviennent et modifient la nature même des unités phonétiques produites. Quoi qu'il en soit, on peut distinguer dans le signal de parole des zones stables et des zones transitoires [André-Obrecht 93]. La détermination de ces zones est un pré-traitement qui peut être exploité à plusieurs niveaux, que ce soit pour localiser précisément un événement dans le temps ou pour ajouter une information de durée aux sons.

L'algorithme de segmentation que nous avons utilisé a été développé par Régine André-Obrecht sous le nom d'algorithme de « divergence forward-backward » [André-Obrecht 88] et il a été mis en œuvre sur de nombreux corpus afin d'étudier sa robustesse face à différents environnements (parole propre, bruitée, enregistrée au travers du canal téléphonique).

Le signal de parole est supposé être décrit par une suite de zones quasi-stationnaires ; chacune est caractérisée par un modèle statistique auto-régressif gaussien (AR) paramétré par un vecteur de coefficients de régression et la variance d'un bruit blanc centré. La localisation des frontières entre ces zones se résume à détecter un changement des coordonnées de ce vecteur.

Pratiquement, deux modèles AR M_0 et M_1 sont estimés à chaque instant à partir du signal acoustique. Le premier modèle est calculé sur une fenêtre de longueur croissante, débutant à l'instant de la rupture précédente. Il permet d'établir un modèle adaptatif du segment courant. Le modèle M_1 est quant à lui estimé sur une fenêtre courte glissante ; il permet d'établir un modèle de l'événement courant, c'est-à-dire de la trame de signal étudiée. Lorsque la distance statistique entre ces deux modèles diverge au delà d'un seuil fixé, on considère que l'événement modélisé par M_1 ne correspond plus à la zone homogène modélisée par M_0 : il y a donc rupture. Le critère de rupture est calculé par un test statistique basé sur la divergence de Kullback qui mesure l'entropie mutuelle entre deux lois conditionnelles correspondant à deux modèles AR. La détection d'une rupture est alors liée à un changement de pente de la statistique.

Nous avons utilisé cet algorithme dans sa version *forward -backward* qui présente une meilleure fiabilité en diminuant le nombre de frontières omises en complétant l'analyse directe (*forward*) par une analyse rétrograde (*backward*). Les unités détectées peuvent être regroupées en trois classes :

- ✓ des segments courts (de longueur inférieure à 20ms) appelés segments événementiels. Ils correspondent au chevauchement de gestes articulatoires brefs (amortissement de la structure formantique lors de la fermeture du conduit vocal, explosion d'une plosive...),
- ✓ des segments transitoires entre deux phonèmes,
- ✓ des segments quasi stationnaires qui matérialisent la partie stable des sons, en particulier la partie centrale d'une voyelle.

La Figure 22 donne un aperçu du résultat obtenu par cette technique. Elle met en évidence la localisation correcte des segments attendus dans la majorité des cas, en particulier pour les segments vocaliques. Parmi les erreurs, on peut relever le son /v/ en position intervocalique qui est rattaché au son précédent, et la liquide /l/ située dans le mot *pleut* qui est difficile à localiser. Ces deux omissions font parties des erreurs les plus fréquentes relevées avec l'algorithme de « divergence Forward-Backward ».

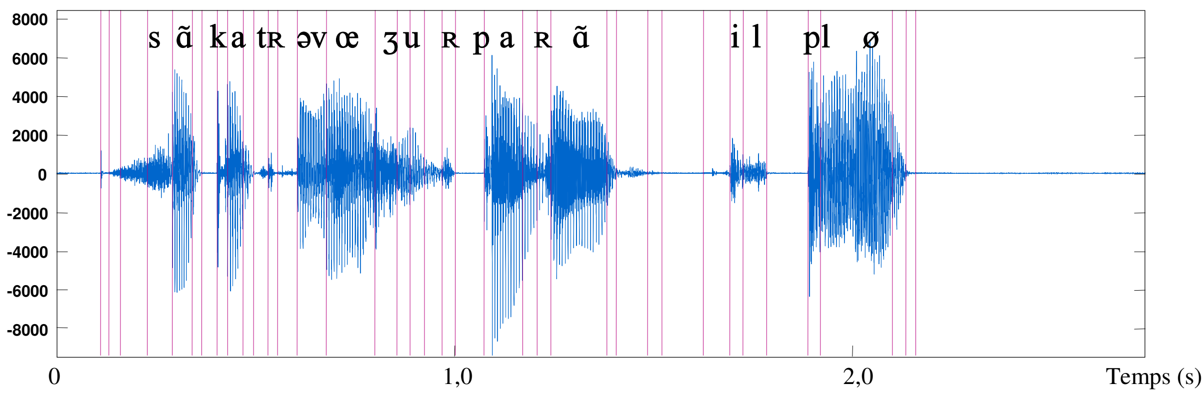

Figure 22 – Résultat de la segmentation automatique (traits verticaux) sur le phrase « cent quatre-vingt jours… ».

Dans le cadre de notre étude, la segmentation en zones quasi stationnaires est principalement exploitée pour donner une dimension temporelle aux voyelles détectées.

Elle est aussi utilisée par le détecteur d'activité vocale, comme nous allons le voir maintenant.

1.3 La détection d'activité vocale

Les travaux présentés dans ce document sont basés sur des expériences menées sur des corpus de parole de nature variée. S'il est vrai que dans certains cas, la qualité du signal est telle que les pré-traitements à apporter sont minimes, ce n'est pas le cas le plus courant. A l'écoute des enregistrements du corpus OGI MLTS, qui a été notre matériau de référence, il est apparu nécessaire de mettre en œuvre un détecteur d'activité vocale (*Speech Activity Detector* ou SAD en anglais) pour éliminer les pauses dans les énoncés. En effet, dans ce corpus, elles peuvent représenter plus d'un tiers de la durée des fichiers, et les conserver aurait entraîné des effets néfastes, tant pour le détecteur de segments vocaliques que pour une éventuelle modélisation des contraintes phonotactiques.

Par contre, et à l'inverse de certaines applications en reconnaissance de la parole, la précision du SAD n'est pas cruciale : une erreur de placement de quelques millisecondes d'une frontière d'activité ne conditionne pas la suite des résultats. Nous avons donc privilégié un algorithme simple aux algorithmes sophistiqués développés en RAP.

La méthode de détection employée est basée sur une analyse statistique du signal dans le domaine temporel. Elle repose sur la segmentation obtenue par l'algorithme de divergence forward-backward : pour chacun des segments obtenus, une décision parole/silence est prise par rapport à un seuil d'activité T_a.

Soit z le signal acoustique et $S = \{s_1, s_2, \ldots, s_N\}$ la suite de N segments. T_a est défini par :

$$T_a = \alpha . \min_{si}(\sigma_{s_i}(z)) \qquad (12.)$$

où $\sigma_{si}(z)$ est l'écart-type du signal z calculé sur le $i^{ème}$ segment. Le coefficient α est expérimentalement fixé à 2,5 dans toutes nos expériences.

De manière à éliminer d'éventuels effets de bord (oscillations d'amortissement…), la décision parole/silence est prise pour chaque segment S_i sur un segment tronqué S'_i ne comportant pas les points de début et de fin du segment[35] (Figure 23). La règle de décision est alors :

- ✓ si $\sigma_{s'_i}(z) > T_a$ alors le segment est étiqueté « Parole »,
- ✓ si $\sigma_{s'_i}(z) \leq T_a$ alors le segment est étiqueté « Silence ».

[35] Le segment S'_i a une longueur de 80 % de la longueur initiale de S_i.

La Figure 24 présente le résultat de l'algorithme de SAD appliqué au signal segmenté précédemment. La Figure 23 propose un agrandissement du silence précédant l'explosion du /p/ dans la séquence « …il pleut ». L'élimination des effets de bord permet de classer le segment comme étant du silence alors que si l'on avait calculé l'écart type sur la totalité du segment, il aurait était reconnu comme étant de la parole.

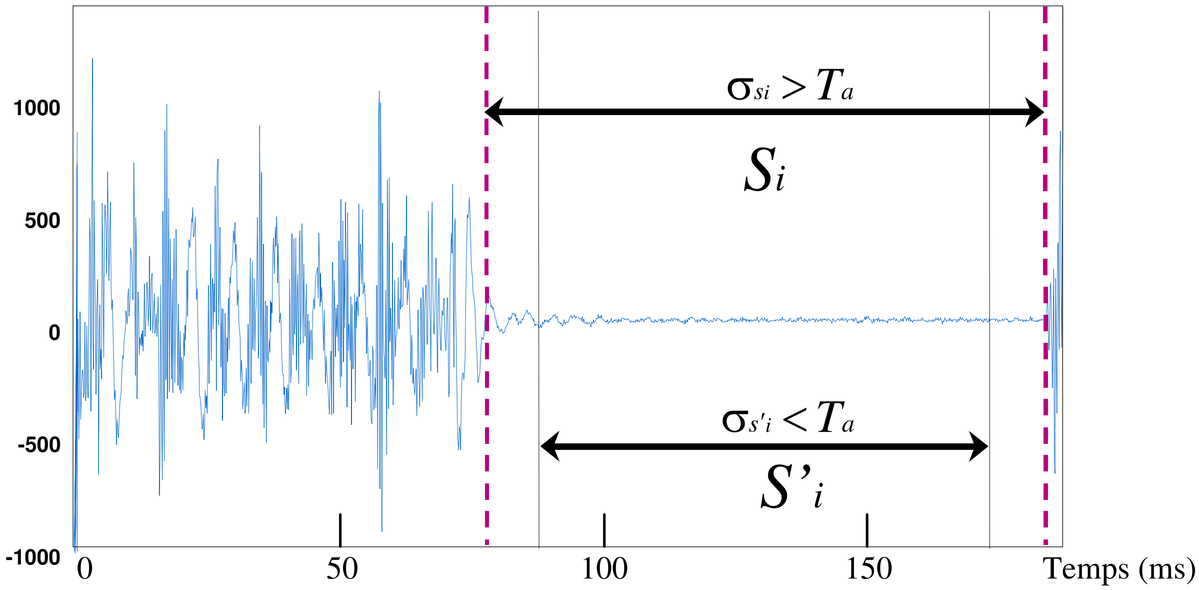

Figure 23 – Prise en compte des effets de bord dans la décision parole/silence. Les traits verticaux pointillés correspondent aux frontières originelles du segment i. Les traits verticaux pleins correspondent au segment tronqué sur lequel la décision est prise.

La durée des segments est également prise en compte puisque les silences d'une durée inférieure à 150 ms sont considérés comme locutoires (silences d'occlusions, pauses courtes…) tandis que les segments plus longs sont assimilés à des zones de non-activité. Notons également que si plusieurs segments consécutifs sont étiquetés comme étant des silences, la durée totale de leur réunion est prise en compte pour effectuer cette distinction.

Le calcul adaptatif du seuil T_a permet de garantir un fonctionnement correct de l'algorithme même en présence d'un bruit de fond important sur l'enregistrement. Un exemple est fourni Figure 25. Il s'agit là encore d'un enregistrement d'un locuteur masculin de la base OGI MLTS mais le bruit de fond est nettement plus important que sur l'exemple précédent. La phrase prononcée est « …euh à proximité d'un petit château… ». Là encore, le détecteur d'activité vocale fonctionne correctement, permettant de localiser la pause initiale. Sur les cinq silences précédant les explosions des occlusives sourdes présentes, seul celui du /k/ n'a pas été détecté alors qu'une analyse plus fine confirme le caractère dévoisé du son produit (le locuteur a réellement prononcé /ks/ et non /gs/) et la présence d'un silence. Cette omission provient de la segmentation qui a produit un segment commun pour le silence et pour l'explosion, ne permettant pas de les dissocier.

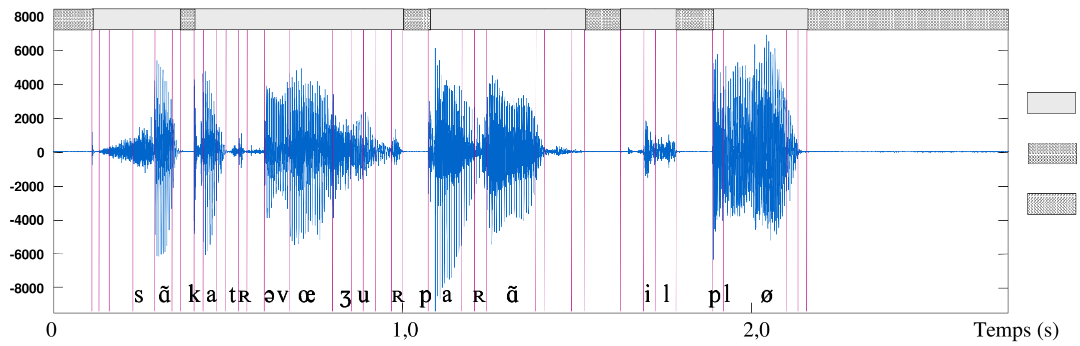

Figure 24 – Résultat de la détection d'activité vocale sur la phrase « cent quatre-vingt jours… ».

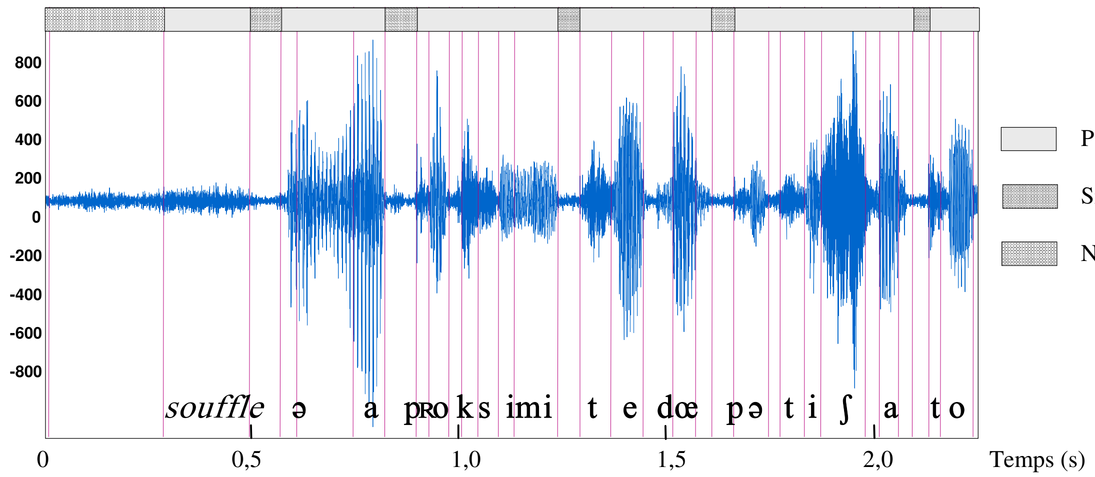

Figure 25 – Exemple de segmentation et de détection parole/silence en milieu bruité.

1.4 La détection des segments vocaliques

Contrairement aux approches citées en début de paragraphe [Sirigos 96, Fakotakis 97], l'algorithme de détection que nous avons implanté ne repose pas sur une classification supervisée mais sur une localisation directe d'événements caractéristiques des voyelles. Ce choix a été opéré de manière à garantir au mieux les deux critères suivants :

- ✓ indépendance vis à vis de la langue,
- ✓ indépendance vis à vis des conditions d'enregistrement.

En effet, il est bien connu que l'efficacité des algorithmes supervisés, qu'ils soient markoviens ou neuromimétiques, est particulièrement liée à l'adéquation des conditions entre l'apprentissage et la classification, en particulier pour ce qui est du canal d'enregistrement. Nous avons donc choisi de rechercher des critères de localisation des voyelles à partir d'une fonction issue de l'analyse fréquentielle du signal de parole.

L'analyse spectrale est pratiquée sur une fenêtre glissante de 32 ms, avec un recouvrement de 16 ms. Nous obtenons par TFR (Transformée de Fourier Rapide) les valeurs de l'énergie ainsi que sa répartition fréquentielle selon l'échelle perceptive de Mel. Il résulte de cette analyse un vecteur d'énergies dans 24 canaux spectraux (centrés sur des fréquences allant de 100 à 5600 Hz) pour chaque trame de signal.

Nous avons vu précédemment (2nde Partie, Chapitre 2) que la plupart des corpus utilisés en IAL sont enregistrés via le canal téléphonique. La bande passante résultante est donc restreinte par rapport à un enregistrement direct, et elle s'étend de 350 à 3500 Hz. C'est bien évidemment dans cette bande fréquentielle que nous avons défini notre fonction spectrale afin de pouvoir utiliser le même algorithme quelles que soient les conditions d'enregistrement (échantillonné à 8 kHz via le téléphone, à 16 kHz en laboratoire…). Cela se traduit dans les équations suivantes par l'adjonction d'un poids α_i à chaque filtre i centré sur une fréquence F_i. Ce poids binaire vérifie les conditions suivantes :

$$\begin{cases} \forall\ i\ /\ 350 \text{ Hz} \leq Fi \leq 3500 \text{ Hz}, & \alpha_i = 1, \\ \forall\ i\ /\ Fi \leq 350 \text{ Hz ou } Fi \geq 3500 \text{ Hz}, & \alpha_i = 0, \end{cases}$$

Nous avons conçu la fonction spectrale à partir de contraintes issues de l'observation du signal et des résultats obtenus au cours de travaux menés précédemment au sein de l'IRIT sur la détection d'activité vocale à partir de la dérivée spectrale et de l'abscisse curviligne temporelle [Puel 93].

L'objectif est de détecter dans le signal les sons possédant une structure formantique vocalique. La Figure 26 (page 121) propose l'analyse spectrale de deux sons issus de l'enregistrement « cent quatre-vingt jours… ». Le son /a/ (Figure 26 – a et b) est extrait du mot *quatre* alors que le son /ʃ/ est issu du mot *chambre* [36] (Figure 26 – c et d). Les valeurs de la fréquence fondamentale et des trois premiers formants relevées pour le /a/ sont respectivement 160, 500, 1600 et 2600 hertz. La représentation de l'énergie dans les différents canaux (Figure 26b) fait apparaître la structure formantique du son. Nous nous sommes donc attachés à établir une fonction révélant une telle structure. Dans un premier temps, nous avons calculé, pour chaque trame, un cumul des énergies par rapport à une moyenne spectrale.

Soient $E_i(t)$ l'énergie calculée dans le $i^{\text{ème}}$ filtre de Mel et $E(t)$ l'énergie totale calculées pour la trame t :

$$E(t) = \sum_{i=1}^{24} \alpha_i . E_i(t) \qquad (13.)$$

Si l'on note $\overline{E}(t)$ la moyenne spectrale de l'énergie contenue dans la bande 350-3500 Hz pour la trame t, on définit le critère *Sbec* (Spectral Band Energy Cumulating) :

[36] Ce mot est prononcé plus tard au cours du même enregistrement. Il ne figure pas sur les figures données en exemple jusqu'à présent.

$$Sbec(t) = \sum_{i=1}^{24} \alpha_i \left| E_i(t) - \overline{E}(t) \right| \qquad (14.)$$

L'étude de cette fonction au cours du temps a montré qu'elle présentait de fortes valeurs pour les voyelles, mais aussi pour certains sons non voisés comme le /ʃ/ (Figure 26d). Nous avons donc été amenés à modifier notre critère spectral en incluant une pondération liée à la distribution spectrale de l'énergie par le biais du taux d'énergie comprise dans la bande de fréquence 350-1000 Hz. De plus, la prise en compte des seuls canaux où l'énergie présente est *supérieure* à la moyenne s'est révélée expérimentalement plus efficace que le cumul des distances à la moyenne de *tous* les canaux.

Soit $E_{BF}(t)$ l'énergie basse fréquence (dix filtres de Mel centrés sur des fréquences inférieures à 1 kHz) :

$$E_{BF}(t) = \sum_{i=1}^{10} \alpha_i . E_i(t) \qquad (15.)$$

La nouvelle fonction spectrale *Rec* (Reduced Energy Cumulating) est alors définie par :

$$Rec(t) = \frac{E_{BF}(t)}{E(t)} \sum_{i=1}^{24} \alpha_i (E_i(t) - \overline{E}(t))^+ \qquad (16.)$$

Cette nouvelle fonction permet de mesurer l'adéquation entre la distribution spectrale d'une trame de signal et la structure vocalique. Ainsi, plus un son se rapproche d'une structure vocalique, plus la valeur de *Rec(t)* sera grande par rapport aux trames voisines. *A fortiori*, les maxima de cette fonction localisent précisément les segments vocaliques présents dans le signal et notre détection est donc basée sur cette localisation.

Afin de renforcer la robustesse de cette détection, un lissage est appliqué à la courbe, de manière à éliminer d'éventuels sommets parasites. Le filtre appliqué est un filtre moyenneur classique s'étendant sur trois trames (soit 48 ms).

La Figure 27 (page 122) représente l'évolution de la fonction *Rec* au cours de la prononciation de la phrase « cent quatre-vingt jours… ». On constate tout d'abord que la fonction Rec détermine des lobes généralement centrés sur les voyelles et dont la hauteur est liée à l'énergie du signal. Malgré le lissage, il persiste certains lobes parasites de faible amplitude (invisibles à cette échelle) pour certains sons voisés et pour certaines voyelles peu énergétiques. En effet, au cours de l'élocution, l'énergie vocalique peut être très variable (phénomène d'accentuation, de réduction vocalique…) et il s'avère délicat de distinguer les lobes significatifs (voyelles faibles) des lobes parasites. Après divers essais de changement de représentation (passage en log(*Rec(t)*), etc.), nous avons placé un seuil sur la courbe *Rec(t)*. Ce seuil *Se* a été expérimentalement fixé au dixième de la valeur moyenne de *Rec(t)* sur l'énoncé de *T* trames :

$$Se = \frac{1}{10T} \sum_{t=1}^{T} Rec(t) \qquad (17.)$$

Seuls les lobes dont la valeur au sommet dépasse ce seuil sont validés comme étant des segments vocaliques. Toujours dans le but d'éliminer des lobes parasites, nous avons uniquement conservé les sommets correspondants à des segments d'une durée supérieure à une durée minimale. En effet, nous avons considéré que les sons d'une durée inférieure à 15 millisecondes ne pouvaient pas être considérés comme étant des voyelles.

Sur l'exemple donné Figure 27, les voyelles /ɑ̃/, /a/, /ə/, /œ̃/ et /u/ sont correctement détectées alors que le /i/ de *il* est ignoré car le lobe d'activité se révèle trop faible. On peut aussi remarquer la présence d'une mauvaise détection dans le mot *pleut*, où un sommet secondaire apparaît après l'explosion du /p/. Il est également intéressant de noter que dans certains cas où la segmentation a omis des frontières (cas de la frontière /əv/), l'algorithme de localisation des segments vocaliques permet de localiser la voyelle dans le segment. Ainsi, l'algorithme de détection de voyelles permet d'indiquer pour chacun des segments obtenu par l'algorithme de divergence forward-backward s'il contient une voyelle ou non. Dans la majorité des cas, l'algorithme de segmentation n'a pas omis de frontière, et il y a correspondance parfaite entre la partie stable de la voyelle détectée et le segment.

La Figure 28 (page 122) représente quant à elle le résultat de la détection sur la phrase « euh à proximité d'un petit château » prononcée en milieu bruité. On constate que les voyelles ont toutes été détectées correctement et qu'aucune fausse détection ne s'est produite.

La mise en œuvre des différents traitements décrits dans ce paragraphe (Figure 29, page 123) aboutit à un étiquetage segmental du signal sous forme de segments homogènes de type :

- ✓ pause – lorsque le locuteur s'interrompt plus de 150 ms,
- ✓ silence – lorsqu'il s'agit d'un silence plus court,
- ✓ consonne – lorsque aucune voyelle n'est présente dans le segment et
- ✓ segment vocalique – lorsqu'une voyelle est présente.

De plus, la position précise (à 16 ms près) du sommet de la fonction spectrale *Rec* dans chaque segment vocalique permet généralement de localiser les voyelles lorsque la segmentation a omis une frontière.

Les différents algorithmes décrits dans ce paragraphe ont été développés en langage C ANSI sous un environnement Unix (système d'exploitation Solaris 2.5). Les traitements sont organisés en modules distincts de manière à permettre une grande souplesse d'utilisation et d'évolution des programmes.

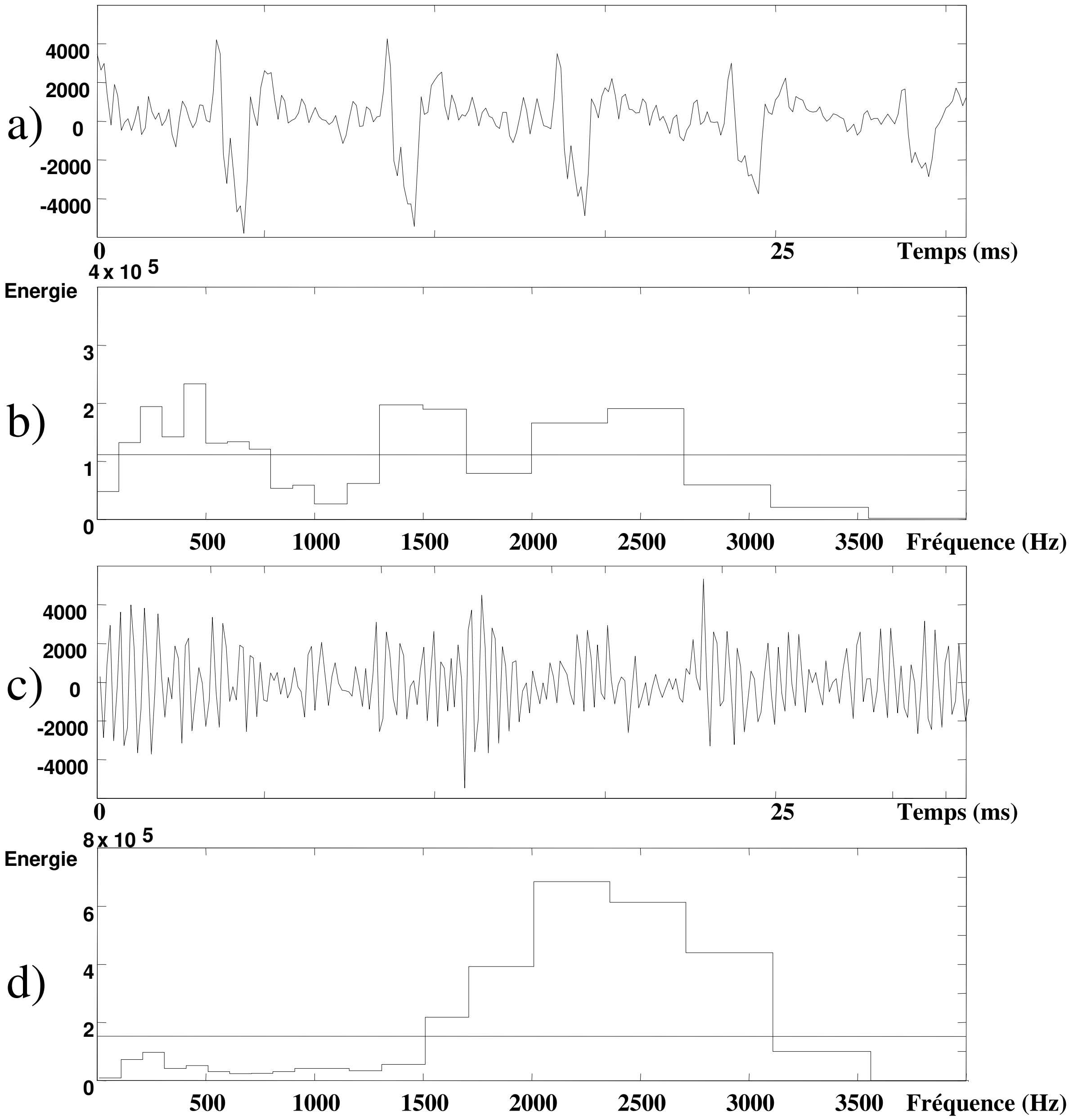

Figure 26 – Analyse spectrale en canaux de Mel pour le /a/ de *quatre* et le /ʃ/ de *chambre*.

a) Fenêtre d'analyse du /a/.
b) Décomposition de l'énergie du /a/ en canaux de Mel.
c) Fenêtre d'analyse du /ʃ/.
d) Décomposition de l'énergie du /ʃ/ en canaux de Mel.

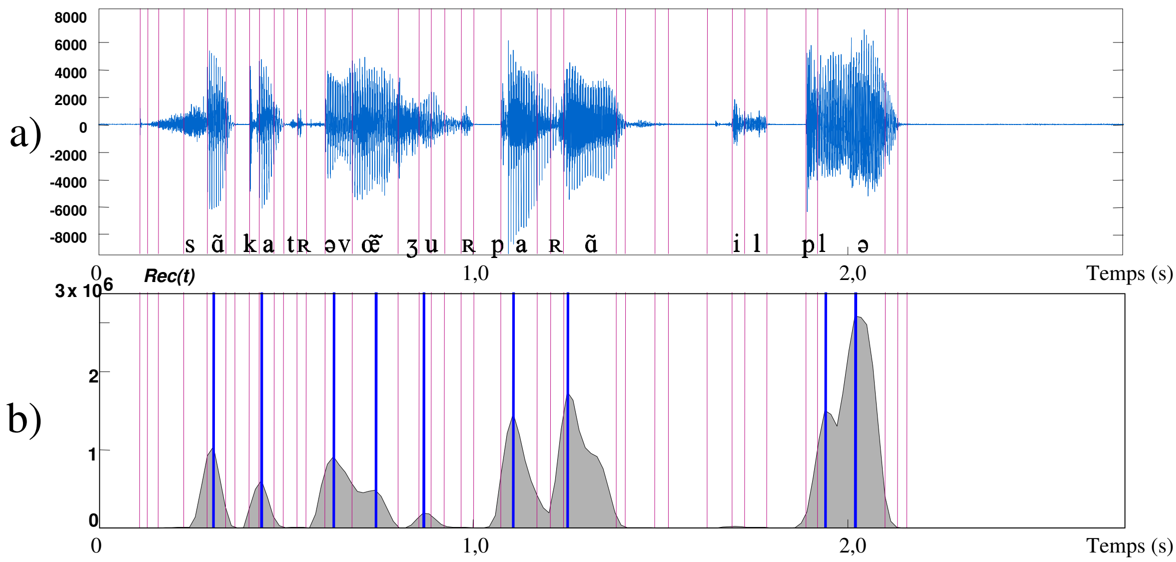

Figure 27 – Détection des voyelles pour la phrase « cent quatre-vingt jours… ».
a) Signal acoustique segmenté automatiquement (traits verticaux fins).
b) Fonction *Rec* et localisation des sons identifiés comme étant des voyelles (traits verticaux épais).

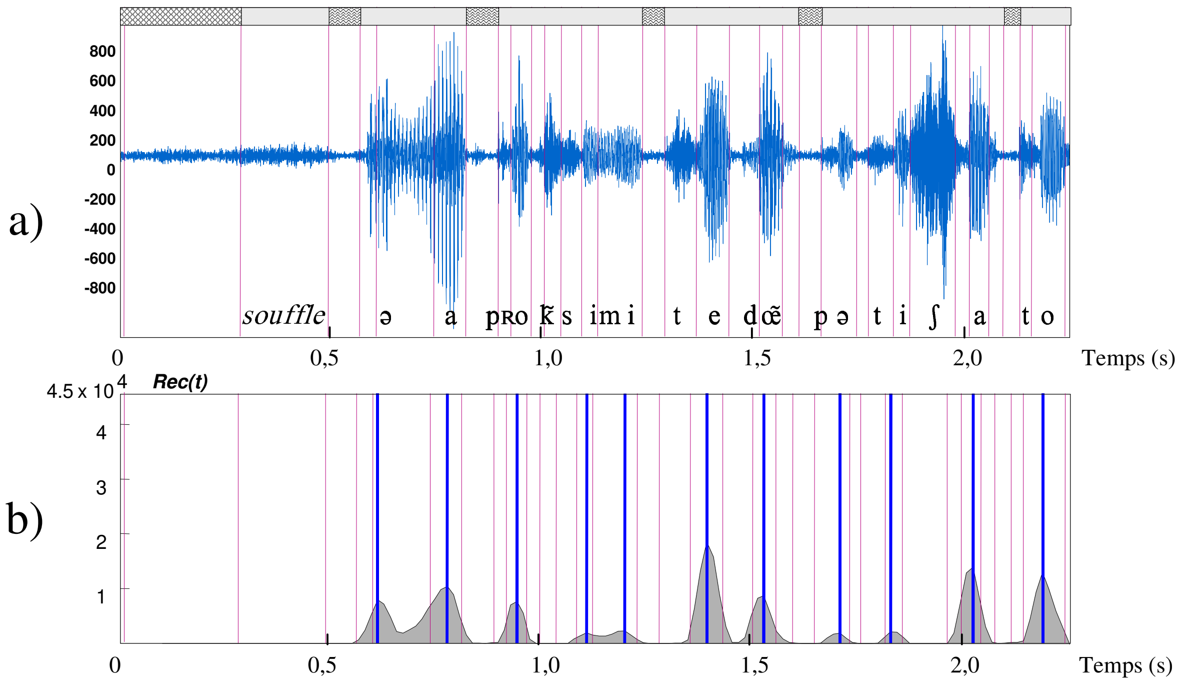

Figure 28 – Détection des voyelles pour la phrase « euh à proximité d'un … ».
a) Signal acoustique segmenté automatiquement.
b) Fonction *Rec* et localisation des sons identifiés comme étant des voyelles (traits verticaux épais).

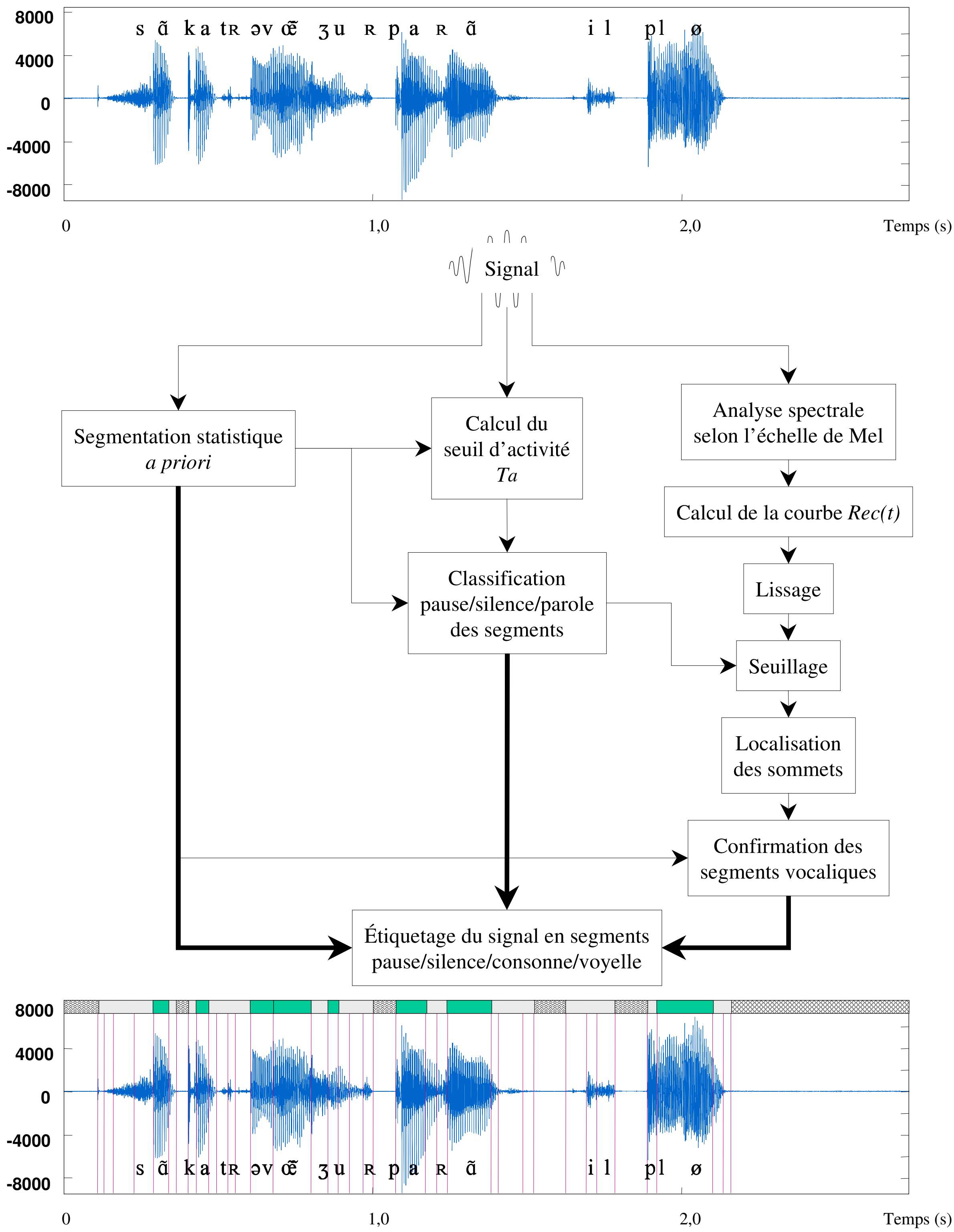

Figure 29 – Récapitulatif des algorithmes mis en œuvre au cours de la détection des segments vocaliques.

1.5 Discussion

La détection des voyelles à partir du signal de manière indépendante de la langue, du locuteur et des conditions d'enregistrement relève de la gageure. Il est certain que, dans le cadre d'une application précise (par exemple la détection des voyelles dans des mots-clefs), les systèmes basés sur un apprentissage markovien ou neuromimétique atteindront les meilleures performances. Par contre, le caractère supervisé de ces techniques ne correspond pas à notre attente et nous a amené à concevoir une approche radicalement différente, même s'il est prévisible que les résultats soient sous-optimaux. Le système que nous avons implanté relève donc le défi de la détection vocalique sans nécessiter d'adaptation au traitement d'une langue inconnue. Nous avons vu sur deux exemples (Figure 27) que l'algorithme développé montre des résultats intéressants. Nous allons maintenant étudier plus précisément son comportement sur des données provenant de divers corpus, de manière à analyser statistiquement les résultats obtenus.

2 EXPERIENCES EN DETECTION DES VOYELLES

Au cours de cette thèse, les algorithmes de détection de voyelles ont été appliqués à différents corpus de nature très diverse (Tableau 7), que l'on peut regrouper en enregistrements de haute qualité (chambre anéchoïque, fréquence d'échantillonnage supérieure à 10 kHz) et en enregistrements téléphoniques (milieu bruité, transmission par réseau téléphonique commuté et fréquence d'échantillonnage de 8 kHz). Historiquement, nos travaux ont été initiés sur un corpus multi-locuteur du CNET, avant d'être portés sur le corpus OGI MLTS lorsqu'il a été disponible à l'IRIT, puis sur d'autres corpus de manière à mener des études complémentaires. Le coefficient α intervenant dans le calcul du seuil de détection d'activité a été fixé à partir des enregistrements OGI MLTS (corpus le plus bruité) à 2,5 tandis que le facteur 0,1 appliqué dans la détermination adaptative du seuil de validation des lobes vocaliques (formule 17) a été déterminé sur le corpus CNET. Ces seuils ont donc été fixés initialement, et leur valeurs sont restées inchangées quel que soit le corpus étudié. Nous allons maintenant étudier plus précisément les résultats obtenus avec deux d'entre eux, à savoir EUROM 1 et OGI MLTS. En effet, le sous-corpus issu de EUROM 1 que nous avons employé, permet d'étudier précisément le comportement des algorithmes de détection dans le cadre de la parole lue en français (grâce à l'étiquetage phonémique), tandis que le corpus OGI MLTS fournit le cadre multilingue et multilocuteur indispensable à nos recherches.

Corpus	Origine	Qualité	Langue	Nombre de locuteurs	Contenu	Etiquetage disponible
Mots-Clefs	CNET	téléphonique	français	100	mots-clefs	non
OGI MLTS	OGI	téléphonique	multilingue	100 / langue	principalement spontané	partiel
Logatomes	ICP	laboratoire	français	1*	logatomes	oui
EUROM 1	ICP	laboratoire	français	1*	textes lus	oui
Parlers arabo-berbères	DDL	laboratoire	arabe et berbère	32	« La bise et le soleil… »	en cours

Tableau 7 – Description des différents corpus utilisés.
* : les corpus comprennent plusieurs locuteurs mais nous n'en avons utilisé qu'un.

2.1 Expériences sur un sous corpus issu de EUROM 1

Le corpus EUROM 1 a été décrit au cours de la seconde partie (page 76), et les enregistrements que nous avons utilisés ici en sont dérivés. La fréquence d'échantillonnage est de 16 kHz et l'enregistrement a été réalisé en environnement anéchoïque. Parmi les données disponibles, nous avons conservé pour un seul locuteur (le locuteur FB) un ensemble de 10 textes d'une vingtaine de secondes chacun, pour un total de 3 minutes 30 de parole.

Chaque texte est étiqueté sous forme de segments phonémiques alignés sur le signal. Cela permet de calculer efficacement des taux de détection des différentes voyelles et d'évaluer la nature des fausses insertions. Le Tableau 8 donne le nombre d'apparitions de chaque étiquette vocalique ou semi-vocalique posée par l'expert. On peut remarquer en particulier que le /œ/ n'a jamais été employé. Il est donc probable que le label /ø/ recouvre une assez grande variation acoustique comme nous le verrons au chapitre suivant.

Phonème	Nombre	Phonème	Nombre	Phonème	Nombre
a	167	ə	-	ɛ̃	-
e	121	ɑ	-	ɔ̃	37
i	105	œ	-	œ̃	32
u	55	ɛ	147	j	46
y	34	ɔ	51	w	29
o	27	ɑ̃	64	ɥ	9
ø	129				

Tableau 8 – Répartition des phonèmes vocaliques et semi-vocaliques dans le corpus.

Les algorithmes de détection décrits au paragraphe précédent sont appliqués à chaque enregistrement et les résultats de la détection des segments vocaliques sont compilés au sein des graphiques suivants.

La Figure 30 indique pour chaque phonème le nombre de fois où il est détecté comme étant un segment vocalique. On constate que le nombre de /a/ et de /e/ détectés est nettement plus important que celui de /i/. Un certain nombre de semi-voyelles sont aussi détectées tandis que les segments vocaliques localisés dans des phonèmes consonantiques sont heureusement rares : il s'agit alors de fricatives voisées (/z/ ou /ʒ/) ou plus généralement de liquides (/l/ et /ʀ/) ou de nasales (/m/ et /n/).

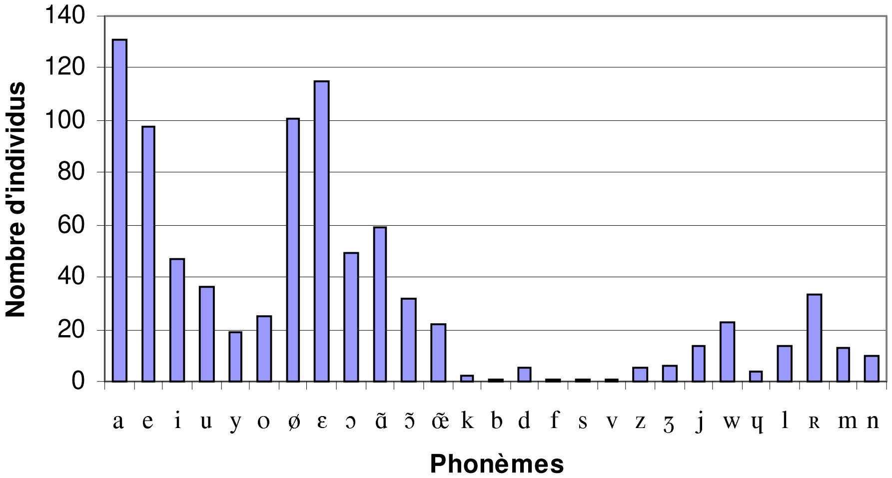

Figure 30 – Ensemble des segments détectés par l'algorithme de détection de voyelles.

Nous pouvons voir (Figure 31) que le taux de détection des voyelles varie entre 45 % (cas du /i/) et plus de 95 % (cas du /ɔ/), avec des valeurs assez fortes pour la plupart d'entre elles (/e, o, ɔ, ã, õ/ sont détectées à plus de 80 %). Les voyelles fermées sont nettement moins bien détectées (/i, y, u/ sont omises à près de 50 %). Cela s'explique par leur premier formant de fréquence faible qui est peu pris en compte dans le calcul du critère *Rec*.

On peut aussi remarquer que le trait de nasalité ne perturbe pas l'algorithme puisque les voyelles nasales sont détectées à plus de 90 % (en nombre). Le comportement de l'algorithme vis à vis des semi-voyelles est variable ; si les segments étiquetés Wé sont détectés dans une assez forte proportion (près de 80 %), ce n'est pas le cas de Yod et Ué (moins de 40 %). Selon que l'on considère ces sons comme étant des fausses détections ou pas, on trouve un taux de pureté de l'ensemble des segments détectés de 84,7 % ou de 89,4 % (Figure 32). Dans les 10,6 % restants, on trouve une forte proportion de /ʀ/ (38 % des erreurs), de /l/ et de /m/ (15 % des erreurs chacun) et) et de /n/ (11 % des erreurs). Ce n'est pas un hasard, et la grande variabilité de ces consonnes en fonction des contextes vocaliques gauche et droit modifie grandement leur caractérisation spectrale. Si l'on compare ces nombres de consonnes détectées au nombre de consonnes présentes dans le corpus, on constate que moins de 15 % des /l, m, n/ et 17,5 % des /ʀ/ ont été détectés.

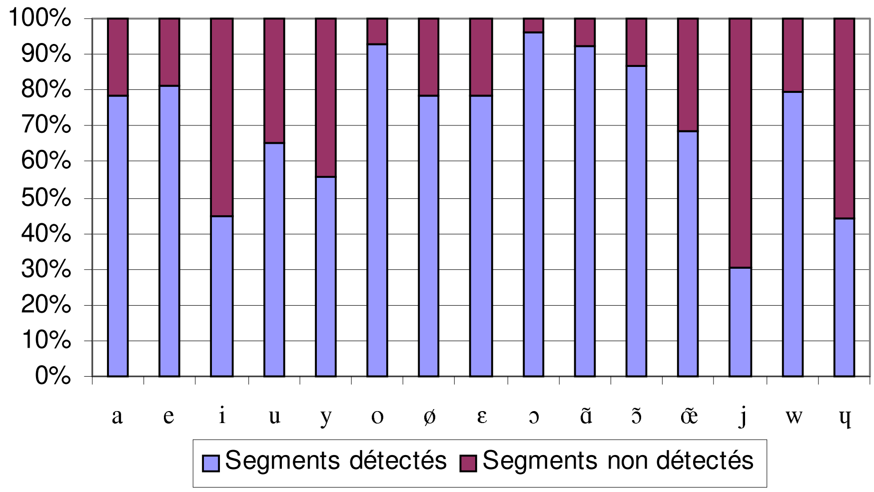

Figure 31 – Taux de détection des voyelles et des semi-voyelles.

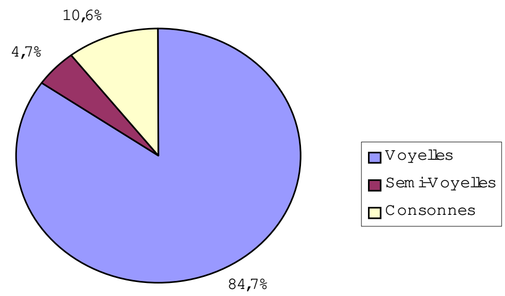

Figure 32 – Répartition des segments par catégorie.

L'analyse des résultats de la détection vocalique sur ces enregistrements de français lu par un locuteur masculin montre que l'algorithme développé enregistre des performances intéressantes : on atteint en effet un **taux de détection de 73,3 %** pour un **taux de pureté de 89,4 %**. Etant donné qu'aucune adaptation n'a été réalisée en fonction des corpus, on peut considérer que ces performances, sans atteindre celles qu'obtiendraient des algorithmes supervisés après adaptation au nouveau corpus, sont bonnes. Les principales faiblesses relevées sont d'une part le faible taux de détection des voyelles fermées, et d'autre part le relativement fort taux d'insertion de la consonne /ʀ/.

Il est maintenant intéressant d'étudier le comportement du détecteur vocalique dans un environnement multilingue et multilocuteur. Il est probable que les taux de détection des voyelles soient dépendants des stratégies d'élocution des différents locuteurs.

2.2 Expériences sur le corpus OGI MLTS

2.2.1 Description des données employées

Le corpus OGI MLTS a déjà été partiellement décrit aux cours des chapitres précédents. Nous rappelons cependant que nous avons travaillé avec 5 des 11 langues qu'il comprend dans sa version définitive. Ces langues sont le français, le japonais, le coréen, l'espagnol et le vietnamien, respectivement notés FR, JA, KO, SP et VI dans les différents schémas et tableaux suivants. Le choix de ces langues a été gouverné par des considérations sur leur systèmes vocaliques (Figure 33). Ils sont en effet de taille et de structure variables (de 5 voyelles pour le japonais et l'espagnol à 18 voyelles pour le coréen).

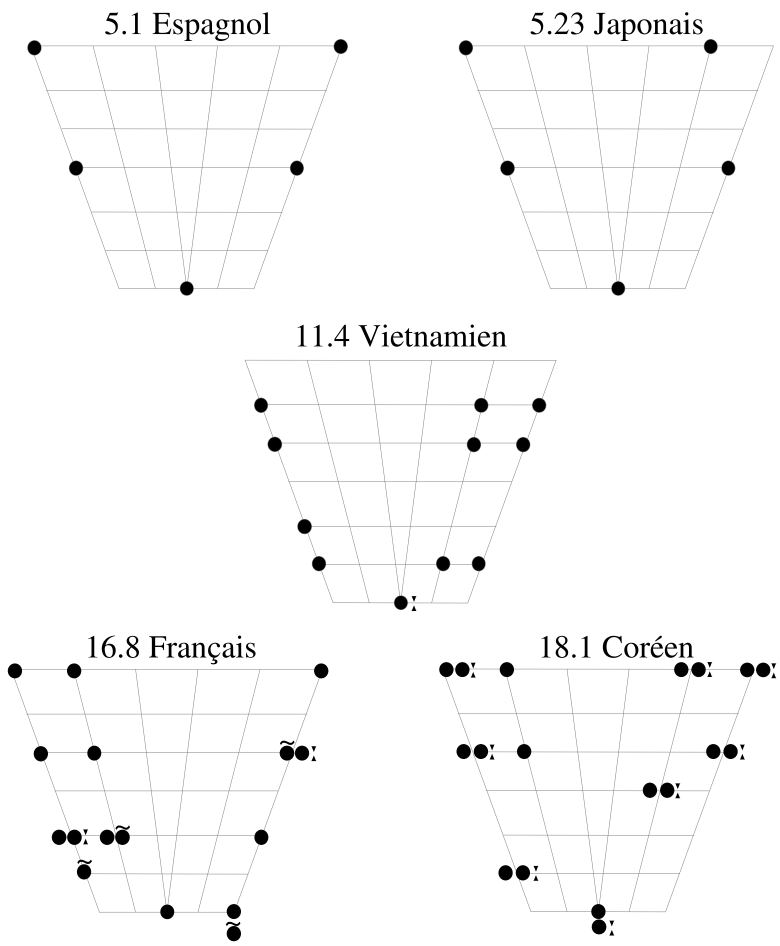

Figure 33 – Description des systèmes vocaliques des cinq langues traitées. La schématisation employée est celle de la typologie établie à partir d'UPSID. (d'après [Vallée 94]).

Le vietnamien présente une complexité intermédiaire puisqu'il se caractérise par un système à 11 voyelles. Le choix du français s'est bien évidemment imposé car il est bien plus facile de contrôler le comportement des algorithmes dans sa propre langue que dans une langue étrangère. Nous traitons ainsi le cas où des voyelles nasales se superposent au système vocalique primaire. L'espagnol et le japonais sont présents tous deux car ils présentent des systèmes vocaliques phonologiquement proches, ce qui va permettre d'évaluer le pouvoir discriminant des modèles dans des conditions difficiles.

Pour chacune des langues, une partie des fichiers est étiquetée en classes majeures de manière semi-automatique : un algorithme supervisé à base de réseaux de neurones classifie chaque trame de signal en l'une des 7 catégories du Tableau 9, puis un expert corrige les frontières et les étiquettes posées.

Etiquette	Sons correspondants
CLOS	silence, pause
STOP	bruit d'explosion
FRIC	consonne fricative, souffle
VOC	voyelle
PRVS	consonne sonante (ou semi-voyelle) en position pré-vocalique
INVS	consonne sonante (ou semi-voyelle) en position intervocalique
POVS	consonne sonante (ou semi-voyelle) en position post-vocalique

Tableau 9 – Description des classes majeures OGI.

On s'aperçoit à la lecture de ce tableau qu'il s'agit plutôt de 5 catégories puisque les labels PRVS, INVS et POVS font référence au même type de sons. On peut aussi remarquer que les fricatives voisées et non voisées sont étiquetées sous le même label FRIC, tout comme les plosives voisées et non voisées (label STOP). Un autre point important que l'on peut soulever est lié à la nature même de l'étiquetage : lorsque deux ou plus phonèmes de même type sont adjacents, ils sont étiquetés dans une même *zone*. La Figure 34 va nous permettre de mieux expliciter cet aspect ; le locuteur, un homme québécois, prononce ici la phrase « j'suis né à Montréal ». On retrouve sur la figure l'étiquetage en classes majeures et les frontières correspondantes (traits verticaux) fournis par OGI. Les labels phonémiques sont posés manuellement par nos soins après écoute et étude spectrographique du signal.

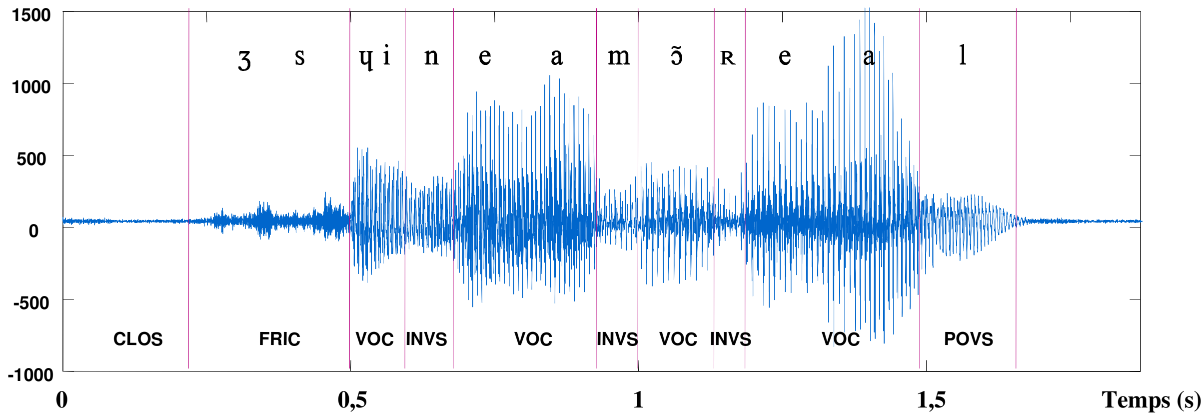

Figure 34 – Exemple d'étiquetage en zones d'un signal OGI. Les labels phonémiques sont posés manuellement et ne sont pas fournis avec OGI.

On constate qu'à une même *zone* peut correspondre plusieurs phonèmes. C'est le cas pour la zone de consonnes fricatives /ʒs/ ainsi que pour trois des quatre zones vocaliques [37] : /ɥi/, /ea/ et encore /ea/. L'un des effets gênants de cet aspect réside dans le calcul des statistiques de détection des segments vocaliques. En effet, ne disposant que de l'étiquetage en classes majeures, il n'est pas possible de connaître le nombre attendu de voyelles et donc le pourcentage d'omissions. On peut uniquement calculer le nombre de *zones* vocaliques détectées, ainsi que le nombre de *zones* vocaliques présentes dans le signal, ce qui conduit à un taux d'omission des zones vocaliques. Pour le calcul des taux d'insertion, le problème est similaire : il est facile de comptabiliser les détections ne se produisant pas dans les zones vocaliques, mais on ignore le nombre de voyelles effectivement détectées. En effet, si deux détections se produisent dans la même zone vocalique, il peut s'agir soit d'une double détection d'une unique voyelle (ce phénomène se produit parfois pour les voyelles nasales), ou de deux détections de deux voyelles distinctes (cas de /ea/ dans l'exemple Figure 34). Plutôt que de faire référence à un taux d'insertion difficile à estimer, nous étudierons dans la suite le taux de pureté de l'ensemble de segments détectés, calculé à partir du nombre total de détections et non du nombre de zones vocaliques détectées.

2.2.2 Résultats de la détection des segments vocaliques

La quantité de données étiquetées est d'environ 4 à 5 minutes par langue, provenant des enregistrements d'une trentaine de locuteurs hommes et femmes dans des proportions diverses. L'algorithme de détection employé met en œuvre les trois modules décrits au paragraphe 1 (segmentation, SAD et détection vocalique).

---

[37] On peut aussi noter que le Ué a été étiqueté comme VOC et non comme PRVS comme les conventions OGI l'indiquent.

Nombre de \ Langue	FR	JA	KO	SP	VI
Zones vocaliques détectées	834	678	727	846	490
Zones vocaliques non détectées	47	47	66	69	28
Détections dans une zone VOC	1037	784	887	980	657
Détections dans les autres zones	125	71	160	107	133
Durée des enregistrements	5'	4'	4'45	4'50	4'10

Tableau 10 – Résultats de la détection de segments vocaliques.

Le Tableau 10 synthétise les résultats obtenus sur les cinq langues. La première constatation est que le comportement de l'algorithme – développé sur un corpus de langue française – semble stable lorsque l'on change de langue. On observe de plus que le français est la langue où l'on détecte le plus de voyelles (1037 détections dans 834 zones vocaliques) tandis que le vietnamien présente moins de zones vocaliques (518 zones présentes), même en tenant compte de la durée des enregistrements. De manière générale, il y a peu d'omissions en terme de zones vocaliques, et les ensembles de segments détectés sont purs à près de 90 %. Nous allons maintenant analyser ces résultats de manière plus précise en fonction des langues.

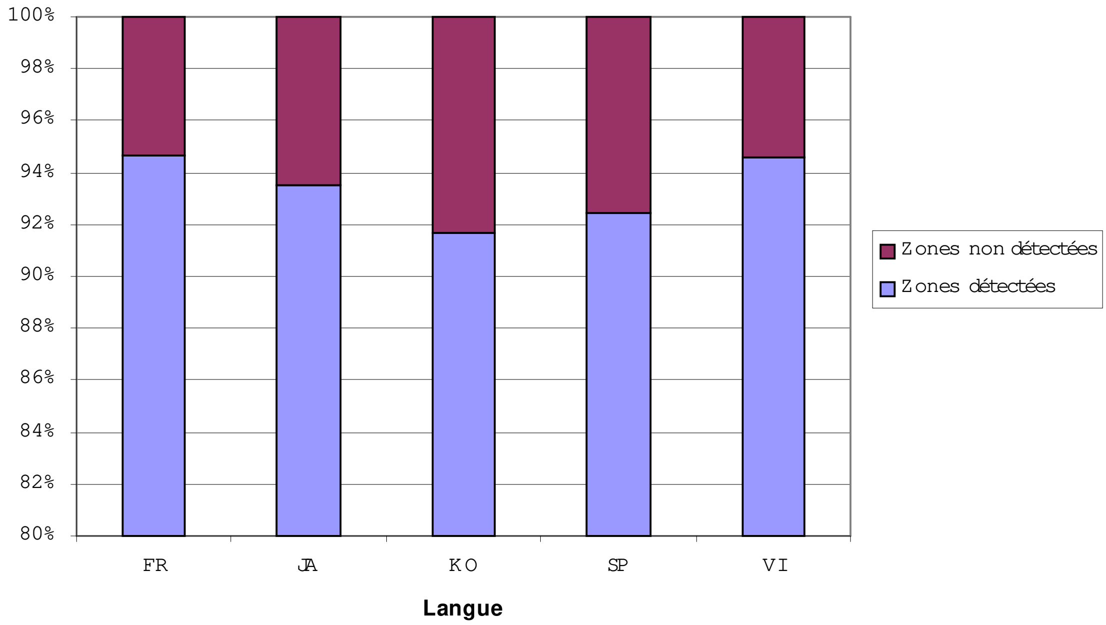

Figure 35 – Taux de détection des zones vocaliques.

La Figure 35 confirme que les taux de zones vocaliques non détectées sont faibles [38], puisque le coréen donne les plus mauvais résultats avec un taux d'omission inférieur à 9 % alors que pour le français et le vietnamien, ce taux dépasse à peine 5 %. Même si on ne peut pas comparer ces taux avec ceux obtenus sur le corpus EUROM (taux calculés en terme de phonèmes), on peut cependant remarquer qu'ils garantissent une assez bonne détection de l'ensemble des segments vocaliques des langues étudiées, sous réserve qu'il ne s'agisse pas d'omissions systématiques d'un phonème vocalique particulier.

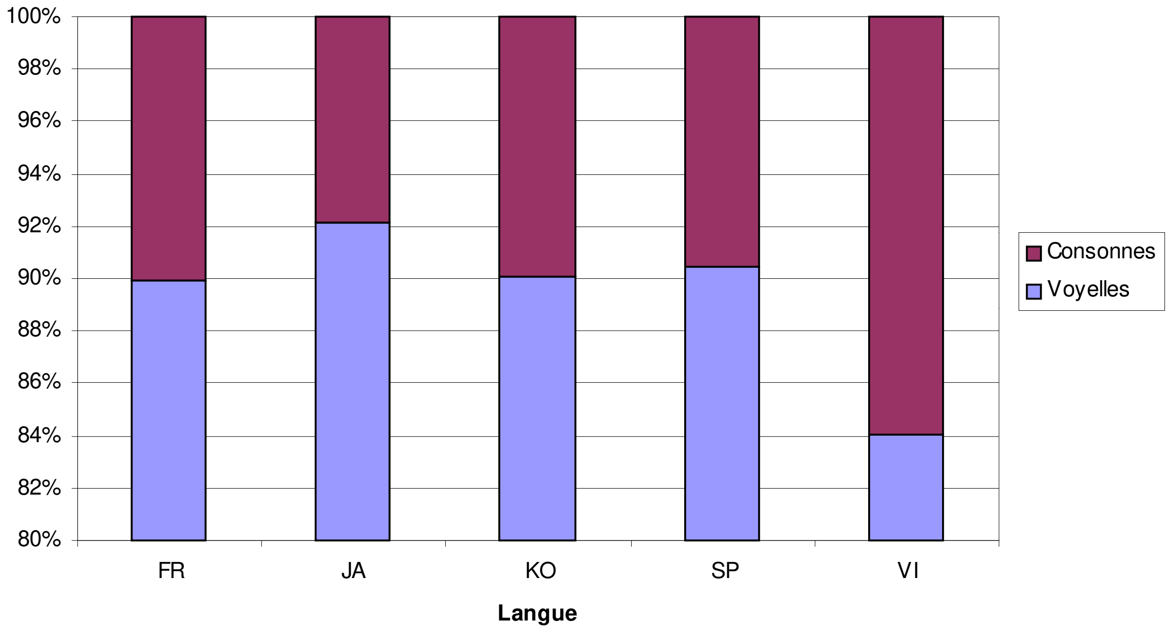

Figure 36 – Composition des ensembles de segments détectés (taux de pureté).

La Figure 36 indique pour chaque langue le taux de pureté de l'ensemble des segments détectés. Il atteint une valeur de plus de 90 % pour quatre des langues traitées, et ce taux descend à 84 % pour le vietnamien. Rappelons que le taux de pureté obtenu pour les enregistrements mono-locuteurs en français lu était de 84,7 %. Cela signifie que le passage à des enregistrements téléphoniques de parole spontanée n'a pas occasionné de dégradation notable et que, dans certains cas, l'algorithme discrimine même mieux les voyelles. Si l'analyse des taux de pureté semble mettre en évidence un comportement similaire du détecteur pour quatre des cinq langues, nous allons voir qu'une analyse plus précise des zones détectées laisse apparaître des différences assez nettes.

La Figure 37 représente sous forme graphique la répartition des erreurs de détection entre les différentes classes majeures OGI. Une première constatation est que

[38] Rappelons que ce taux calculé en zones vocaliques minore le taux effectif d'omissions de phonèmes vocaliques.

le taux de fausses détections portant sur des explosions (plosives voisées et non voisées) est stable pour quatre langues (il est alors proche de 10 %). Ce taux est légèrement supérieur à celui enregistré sur le corpus EUROM en parole lue (environ 8 %). Le japonais présente à l'inverse un taux de 16 % nettement plus élevé que pour les autres langues. Cette langue se caractérise aussi par un taux de fausses détections intervenant dans des zones fricatives nettement plus élevé (35 %) que pour les autres langues mis à part le français (34 %). Ce taux est plus de deux fois plus élevé que celui relevé sur les données EUROM (15 %), où la grande majorité des erreurs provenaient des consonnes sonantes. Cette catégorie d'erreurs reste cependant majoritaire dans les cinq langues : de 49 % pour le japonais à 73 % pour l'espagnol. Parmi ces consonnes, celles se produisant entre deux voyelles (INVS) sont souvent les plus nombreuses et elles représentent près de la moitié des erreurs en espagnol.

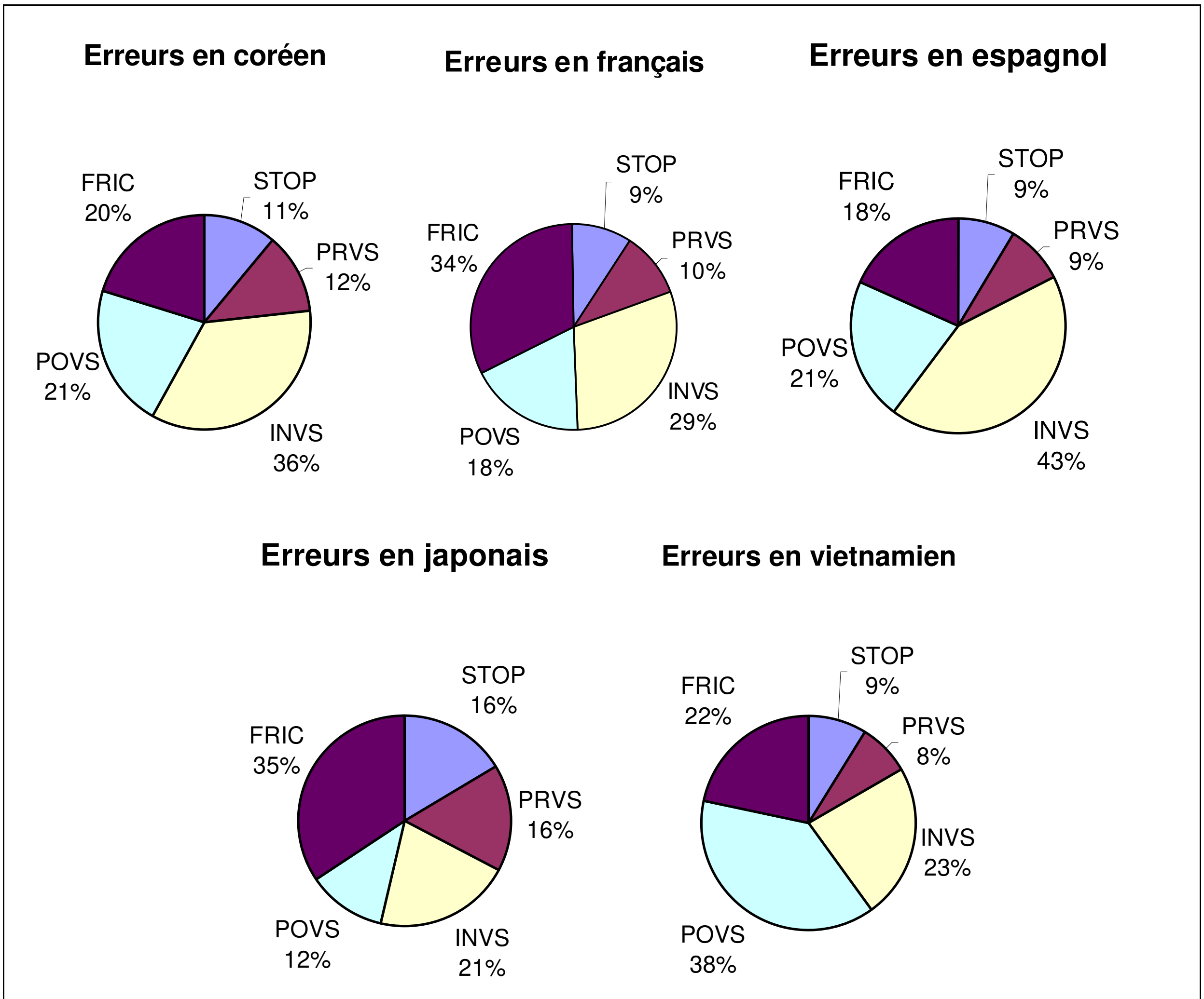

Figure 37 – Répartition des erreurs de détection sur le corpus OGI.

Il ressort de l'analyse des erreurs induites par l'algorithme de détection une relative variabilité de comportement de l'algorithme de détection des segments vocaliques en fonction des langues traitées. Il est cependant difficile d'en tirer des conclusions sans étudier de manière approfondie et systématique les erreurs produites. Ce type d'étude nécessite, pour chacune des langues traitées, une connaissance phonologique poussée que nous n'avons pas, en particulier pour les langues asiatiques. Il sera cependant intéressant de voir les enseignements que des experts pourront en tirer.

2.3 Discussion

Ce chapitre a présenté un ensemble d'algorithmes développés dans le but de localiser les segments vocaliques dans le signal de parole tout en respectant les critères suivants :

- ✓ aucune nécessité de disposer de données étiquetées,
- ✓ indépendance vis à vis des conditions d'enregistrement,
- ✓ indépendance vis à vis de la langue.

Le premier point est garanti par la conception même du détecteur qui fonctionne sans apprentissage. Les points suivants résultent du caractère adaptatif des modules et de la définition de la fonction spectrale *Rec* puisque les différents paramètres intervenant dans les algorithmes sont inchangés lorsque l'on change de corpus. Les expériences montrent que les algorithmes ont un comportement similaire sur des données très différentes (parole mono-locuteur lue, parole téléphonique spontanée) et qu'ils atteignent des performances comparables sur plusieurs langues (taux de pureté de 90 % et taux de détection des zones vocaliques de 93,5 % en moyenne). L'analyse des erreurs montre par ailleurs que la répartition des fausses détections varie d'une langue à l'autre.

Au vu des résultats, on peut dire que les segments collectés pour chacune des langues du corpus OGI sont dans une large mesure représentatives des systèmes vocaliques employés dans ces langues en **parole spontanée**. Il reste évident qu'il ne s'agit pas là d'un inventaire phonologique des voyelles présentes, mais d'une collecte de segments vocaliques, c'est-à-dire de segments présentant les propriétés classiques des voyelles et pouvant résulter de processus tels que coarticulation, réduction, ou assimilation. Bien évidemment, même si nous avons choisi des langues assez différentes d'un point de vue phonologique, il est impossible de généraliser ces résultats à toutes les langues du monde, et bien d'autres expériences doivent être réalisées.

Chapitre 2

LA MODELISATION ACOUSTIQUE DES SYSTEMES VOCALIQUES

La détection des segments vocaliques est une étape préalable à la modélisation acoustique des systèmes vocaliques. Les algorithmes présentés au chapitre précédent fournissent, pour chaque enregistrement, un ensemble de localisations probables pour les voyelles prononcées ainsi que, pour chacune d'entre elles, une indication de durée de sa partie stable. A partir de ces données, il est nécessaire de déterminer un espace de représentation commun puis de choisir une modélisation appropriée. Ce chapitre est consacré à cette problématique. Dans un premier paragraphe, nous proposons plusieurs représentations paramétriques des voyelles et nous expliquons pourquoi notre choix s'est porté sur une paramétrisation cepstrale. Le paragraphe suivant présente les différents modèles statistiques utilisés, ainsi que les algorithmes liés à leur exploitation. Le troisième paragraphe nous permet d'étudier le comportement de ces algorithmes sur des données extraites, comme au chapitre précédent, de plusieurs corpus de nature variée.

1 PARAMETRISATION DES VOYELLES

1.1 Choix de l'espace de représentation

La détermination de l'espace de représentation le plus adapté à la modélisation acoustique des systèmes vocaliques (MSV) n'est pas triviale. En effet, parmi les espaces de représentation classiquement étudiés, il s'en dégage au moins trois ayant des caractéristiques intéressantes : la représentation formantique, la paramétrisation cepstrale et la représentation issue de la mesure par appariement de pic spectraux (APS).

La représentation formantique est – de loin – la plus étudiée et la plus utilisée par les phonéticiens. La description des voyelles en terme de formants est en effet particulièrement lisible puisqu'il s'agit d'un espace de représentation de faible dimension (les trois premiers formants fournissent un espace assez discriminant). De plus, nous avons déjà évoqué au cours de la première partie la dualité de cette représentation avec l'espace articulatoire et son intérêt pour la prédiction et la modélisation des systèmes vocaliques [Schwartz 89, Vallée 94]. La détermination automatique des formants reste cependant une tâche difficile, particulièrement lorsqu'il s'agit de parole spontanée prononcée en milieu bruité.

La représentation cepstrale de la parole est la technique la plus employée dans les applications de RAP. Qu'elle soit obtenue par prédiction linéaire ou par Transformée de Fourier Rapide, elle permet de procéder à une déconvolution particulièrement efficace entre la source du signal et le conduit vocal. Le choix d'une échelle fréquentielle non linéaire (généralement l'échelle de Mel) permet d'obtenir une modélisation rigoureuse dans un espace d'une dizaine de dimensions (classiquement 8 ou 12). La pondération spectrale ainsi effectuée aboutit à une échelle linéaire des coefficients cepstraux, et donc à l'emploi de la distance euclidienne comme mesure de proximité entre stimuli. L'inconvénient majeur de la représentation cepstrale réside dans son manque de lisibilité ; il ne s'agit pas d'une représentation directement liée aux informations qu'un expert peut extraire de la lecture d'un sonagramme, ce qui complexifie l'interprétation des paramètres.

La mesure par Appariement de Pics Spectraux (APS) a été introduite au cours des années 80 [Caraty 87]. Elle est spécialement conçue pour l'étude des voyelles et elle dérive d'une analyse par prédiction linéaire. Les expériences [Yé 89] montrent que cette mesure, de nature perceptive, représente efficacement l'espace vocalique et qu'elle permet une bonne discrimination entre les voyelles (stimuli synthétiques ou logatomes). Par contre, son utilisation sur des enregistrements de parole continue a nécessité d'établir des tables de pondérations des différents paramètres (fréquence, bande passante et amplitude des pics) de manière à en améliorer la robustesse [Caraty 92].

Parmi ces trois représentations, notre choix s'est porté sur la représentation cepstrale, principalement en raison du caractère bruité des enregistrements de la base OGI MLTS. Nous sommes conscients que cela nous prive dans une large mesure de l'expertise de nos collègues phonéticiens, mais il semble que cet espace paramétrique soit le plus robuste dans le cadre de cette étude. Il reste cependant intéressant de décrire ponctuellement certains stimuli en terme de formants, et, une fois que l'on dispose de la description cepstrale de l'espace acoustique d'un locuteur, la recherche d'une méthode d'inversion vers un espace formantique ou articulatoire peut se révéler une tâche intéressante que nous n'avons cependant pas abordée dans le cadre de cette thèse.

1.2 Algorithmes de paramétrisation cepstrale

La paramétrisation est obtenue par une analyse cepstrale classique en RAP (Figure 38) aboutissant à un vecteur de coefficients MFCC (Mel Frequency Cepstral Coefficient). Nous ne reviendrons pas en détail sur ce processus bien connu que nous appliquons de manière standard. Remarquons simplement que l'utilisation d'une fenêtre temporelle de Hamming permet d'éviter la formation d'artefacts liés aux effets de bord durant la transformation du domaine temporel au domaine fréquentiel. La pré-accentuation vise à renforcer l'importance des hautes fréquences par analogie avec le traitement effectué par l'homme au niveau de la cochlée. et l'utilisation de l'échelle non linéaire de Mel prend en compte des connaissances acoustiques sur la perception humaine.

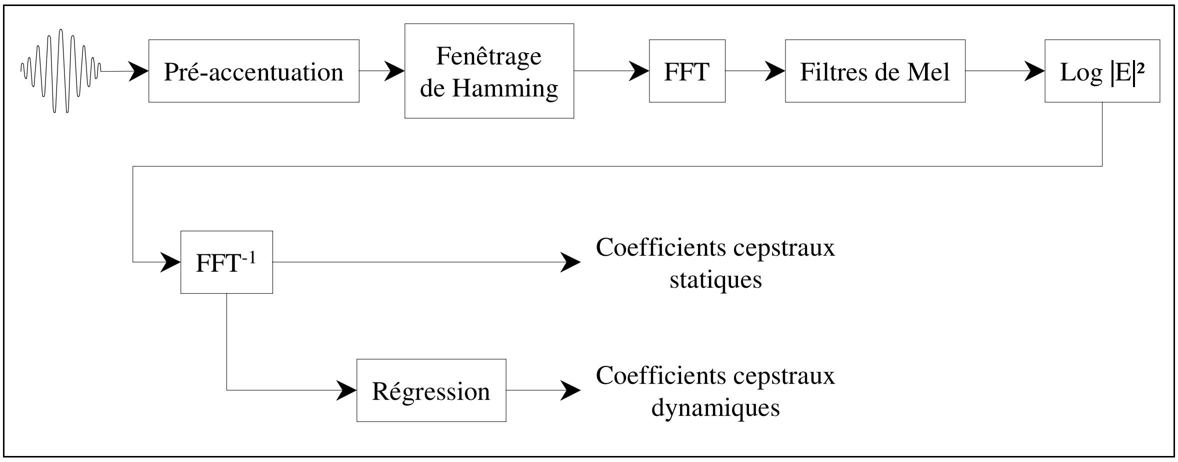

Figure 38 – Schéma synoptique de la chaîne de paramétrisation cepstrale.

Vingt canaux sont répartis sur cette échelle de Mel de manière à réduire le nombre de coefficients utilisés pour décrire le signal tout en conservant une définition suffisante pour les basses fréquences (fréquences où le pouvoir discriminant de l'oreille humaine est important). Ces canaux sont obtenus par l'application de filtres triangulaires se chevauchant et centrés sur les fréquences suivantes : 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1150, 1300, 1450, 1650, 1850, 2050, 2350, 2700, 3100 et 3550 Hertz.

Contrairement au processus mis en œuvre en RAP où la transformée de Fourier du signal est calculée sur des fenêtres glissantes, de manière à traiter tout le signal, nous opérons sur un intervalle de 32 ms centré sur l'instant de détection de la voyelle traitée. Nous conservons pour chaque voyelle un vecteur de 18 coefficients comprenant 9 paramètres statiques (l'énergie E et les 8 premiers coefficients MFCC) ainsi que leurs dérivées temporelles obtenues par calcul d'une régression linéaire sur cinq trames centrées sur la fenêtre modélisée.

Là encore, les différents algorithmes sont implantés en langage C ANSI et employés sous le système d'exploitation UNIX Solaris. Ils sont utilisés au sein de l'équipe « Interface Homme-Machine – Parole & Texte » de l'IRIT depuis plusieurs années.

En appliquant ces différents traitements aux trames centrales des segments vocaliques détectés pour chaque locuteur ou pour chaque langue, on dispose ainsi d'un ensemble de données décrites par 18 coefficients auxquels on adjoint la durée du segment détecté. En effet, dans le cas des voyelles orales, cette information correspond généralement à la durée de la partie stable (lorsque aucune frontière n'est omise) et elle est donc fortement corrélée à la durée du phonème. Dans le cas des voyelles nasales, on ne peut pas être aussi catégorique : la détection se produit généralement sur la partie orale de la voyelle et non sur l'éventuelle queue nasale. Si la segmentation fournit un unique segment pour ces deux parties de la voyelle, sa longueur rend compte de la durée totale du phonème ; dans le cas contraire (segmentation de la voyelle nasale en deux segments) la durée de la queue nasale n'est pas prise en compte.

La durée des segments vocaliques peut donc se révéler pertinente pour la modélisation des systèmes vocaliques (cas des langues à opposition voyelle brève / voyelle longue comme le coréen par exemple). Bien évidemment, cette durée peut aussi se révéler intrinsèquement distinctive entre plusieurs langues en distinguant des voyelles proches sur le plan spectral mais présentant une distribution de durée différente d'une langue à l'autre.

2 MODELISATION DU SYSTEME VOCALIQUE

2.1 Choix de la modélisation

Les méthodes de modélisation de données continues sont nombreuses, qu'elles relèvent de la formulation neuromimétique ou statistique, voire des deux. Dans le cadre de notre étude, le choix d'une modélisation statistique s'est imposé pour plusieurs raisons. La principale d'entre elles tient à l'analogie entre notre approche et les méthodes employées en identification automatique du locuteur, classiquement basées sur une modélisation des observations de chaque locuteur par un mélange de lois gaussiennes (MMG) [39] [Reynolds 95]. Le lecteur trouvera dans [Kambhatla 96] un large éventail des domaines où une telle modélisation est employée.

Dans ce cadre, la fonction de densité d'un vecteur aléatoire X de dimension p est donné par :

$$p(X = x|\lambda) = \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right] \qquad (18.)$$

où le modèle λ est défini par le nombre de lois gaussiennes Q, les poids affectés à chaque terme α_k, ainsi que pour chaque loi de répartition gaussienne k, la moyenne μ_k et la matrice de covariance Σ_k associées.

Il est bien connu que le comportement des modèles statistiques utilisés comme classificateurs dépend fortement de la topologie employée et de la phase d'initialisation des modèles. Dans le cas des MMG, la topologie est définie entièrement par le nombre Q de lois gaussiennes du mélange et l'initialisation est faite au travers de la donnée du vecteur (α_k, μ_k, Σ_k) pour chacune des k composantes du mélange.

Dans le cadre de la modélisation des systèmes vocaliques, deux approches sont envisageables :

[39] Nous emploierons préférentiellement l'acronyme MMG (Modèle de Mélange de lois Gaussiennes) plutôt que le classique acronyme anglais GMM (Gaussian Mixture Model).

- ✓ l'approche supervisée à partir d'un ensemble de données étiquetées manuellement,
- ✓ l'approche non supervisée basée sur un ensemble d'apprentissage non étiqueté.

Le premier cas mène naturellement au choix d'un modèle possédant autant de lois gaussiennes que de voyelles et à l'initialisation de chacune d'entre elles à partir des données étiquetées correspondantes. A l'inverse, dans le cas d'une approche non supervisée, aucune correspondance *a priori* n'est possible entre le nombre de voyelles et le nombre Q de composantes gaussiennes qui est alors déterminé par des techniques heuristiques.

Cette dernière situation correspond à l'application présente puisque nous souhaitons obtenir les modèles sans utiliser aucun étiquetage manuel des données. Cela permet de plus d'établir le nombre de classes à partir des observations, sans être influencé par les connaissances phonologiques et de donner plus de liberté au modèle.

La première solution consiste à choisir Q *a priori* et à initialiser conjointement les paramètres des lois gaussiennes par le biais d'une phase préliminaire de Quantification Vectorielle (QV). Cette étape vise à trouver une partition optimale de l'ensemble d'apprentissage en Q cellules et à initialiser chaque composante gaussienne avec le centroïde et la distorsion de chacune d'entre elles. Notons que la QV aboutit alors à un classificateur associant chaque stimulus à une classe par un calcul de distance minimale (paragraphe 2.2).

Le modèle initial une fois fixé, on applique un algorithme itératif d'optimisation de loi multigaussienne. Il s'agit généralement d'un algorithme de type Expectation-Maximisation (EM). Une reformulation dans un cadre statistique de la modélisation obtenue précédemment par QV permet d'utiliser le modèle multigaussien comme classificateur, associant à chaque stimulus une composante du MMG (paragraphe 2.3).

Insatisfait par le choix heuristique de Q, il nous a paru intéressant de rechercher différemment le nombre de classes optimal. L'analogie entre QV et MMG nous a amené à définir un critère d'information de type Rissanen reliant la vraisemblance des données d'apprentissage au nombre de paramètres du modèle. Sa minimisation conduit à une nouvelle estimation de Q adaptée aux données modélisées (paragraphe 2.4).

2.2 Initialisation du modèle : la quantification vectorielle (QV)

L'objectif de cette phase est de déterminer à partir d'un ensemble de données d'apprentissage un *dictionnaire* représentatif. Le terme de dictionnaire désigne un ensemble fini de Q vecteurs de référence ainsi que la partition de l'espace qui lui est associée [40] au moyen d'une mesure de distance ou de distorsion.

[40] La loi de partition la plus utilisée est celle du plus proche voisin au sens de la distance d(.,.) utilisée.

Si l'on désigne par p la dimension de l'espace paramétrique, par $d(.,.)$ la mesure de distance employée et par $C = (\mu_1, \dots \mu_Q)$ le dictionnaire, le quantificateur vectoriel ainsi défini effectue la projection g associant à tout vecteur x de l'espace de départ R^p un vecteur μ_i de l'ensemble fini C tel que :

$$g(x) = \mu_{\delta(x)} \qquad (19.)$$

avec

$$\delta(x) = \arg \min_{k=1}^{Q} d(x, \mu_k) \qquad (20.)$$

Le quantificateur vectoriel réalise donc une compression de l'ensemble d'apprentissage en un ensemble fini C. Cette compression est réalisée avec pertes puisqu'elle induit une distorsion D généralement non nulle entre les données de l'ensemble d'apprentissage et leurs projections. Si la distance utilisée est la distance euclidienne, on a :

$$D = E[d(x, g(x))] = E\left[\min_{k=1}^{Q} d(x, \mu_k)\right] = E\left[\min_{k=1}^{Q} \|x - \mu_k\|^2\right] \qquad (21.)$$

L'objectif des algorithmes d'apprentissage est de déterminer le dictionnaire C minimisant cette distorsion D. Si l'on prend l'estimateur par maximum de vraisemblance (ou par maximum a posteriori) de l'espérance de l'expression de la distorsion, on obtient comme dictionnaire optimal C^* :

$$C^* = \arg \min_{C} \frac{1}{N} \sum_{i=1}^{N} \left[\min_{k=1}^{Q} \|x - \mu_k\|^2\right] \qquad (22.)$$

Les études [Gersho 92] montrent que le dictionnaire ainsi obtenu est au moins *localement* optimal et qu'il n'existe pas de formulation évidente de contraintes suffisantes pour obtenir un dictionnaire globalement optimal.

La détermination d'un dictionnaire localement optimal à Q constant s'appuie sur l'algorithme connu en statistiques sous le nom des nuées dynamiques, dont la forme la plus utilisée en traitement de parole est l'algorithme de Lloyd. Il est couplé à une phase de détermination du nombre de classes Q par *splitting* au sein de l'algorithme LBG [Linde 80] .

Son principe général est de minimiser la distance – ou distorsion – de l'ensemble d'apprentissage au modèle par partitionnement successif de l'espace paramétrique, en respectant à chaque étape les critères nécessaires d'optimalité et en diminuant la distorsion résultante. Chaque itération consiste donc en deux phases :

- ✓ phase 1 : augmentation du nombre de classes (phase de *splitting* en anglais),

- ✓ phase 2 : détermination des centroïdes optimaux pour le nouveau nombre de classes et en fonction des positions des centroïdes à l'itération précédente (optimisation de Lloyd).

Dans le cas où le nombre de cellules désiré n'est pas connu *a priori*, des variantes de l'algorithme permettent de le déterminer automatiquement en imposant un seuil minimal sur la distorsion relative des données d'apprentissage par rapport au dictionnaire. L'augmentation du nombre de classes se poursuit tant que ce seuil n'est pas franchi. L'algorithme LBG est décrit plus précisément en Annexe.

Le modèle obtenu peut être utilisé comme classificateur par décision prise sur la base « du plus proche voisin ». Un stimulus y est associé à la classe $c(y)$ vérifiant :

$$c(y) = \arg\min_{k=1}^{Q} \left\| y - \mu_k \right\|^2 \qquad (23.)$$

dans le cas de la distance euclidienne.

2.3 Modèles par mélange de lois gaussiennes (MMG)

2.3.1 Méthode d'apprentissage : l'algorithme EM

Les MMG sont bien connus pour avoir un comportement très dépendant des conditions initiales. L'application d'un algorithme de QV permet d'éviter l'initialisation aléatoire qui peut amener les algorithmes d'apprentissage à être piégés vers des optima locaux de piètre qualité.

L'apprentissage des différents paramètres d'un MMG $\lambda = (\alpha_k, \mu_k, \Sigma_k, \forall\ k\ /\ 1 \le k \le Q)$ est classiquement réalisé par un algorithme de type EM (Expectation-Maximisation en anglais). Cet algorithme, décrit en Annexe, est basé sur une approche par maximum de vraisemblance. Sa maximisation se fait en introduisant une fonction intermédiaire sur laquelle porte l'optimisation. L'algorithme converge en un temps fini vers un extremum local, et là encore, il n'existe pas de critère absolu permettant la convergence vers un maximum global. Contrairement aux algorithmes de type LBG qui réalisent une augmentation du nombre Q de composantes du modèle, l'algorithme EM n'effectue qu'une optimisation des composantes gaussiennes à Q constant. L'algorithme, de nature itérative, suppose que les données d'apprentissage sont indépendantes et identiquement distribuées, et que chaque observation est générée uniquement par une composante gaussienne du mélange et une seule. Il se décompose alors à chaque itération en deux phases :

- ✓ Phase d'estimation : la probabilité *a posteriori* que la donnée i soit générée par la loi gaussienne k est calculée, pour toutes les observations, $1 \le i \le N$, et pour toutes les lois k, $1 \le k \le Q$.
- ✓ Phase de maximisation : une réévaluation des paramètres du modèle est effectuée à partir des probabilités calculées durant la phase d'estimation.

2.3.2 Relation entre QV et MMG

Les modèles obtenus par QV et MMG sont *a priori* différents dans le cas général : la QV effectue une classification basée sur un critère de distance alors que les MMG mettent en œuvre un critère de vraisemblance. En fait, si l'on reformule le critère de classification de la QV en terme de vraisemblance calculée sur un modèle gaussien à matrice de covariance sphérique constante, il devient clair que les deux approches sont extrêmement liées. Une telle correspondance est déjà mentionnée dans [Duda 73], et les travaux de S. Nowlan [Nowlan 91] permettent de préciser leur relation.

Rappelons l'expression de la vraisemblance d'une observation de l'ensemble d'apprentissage dans un modèle MMG λ :

$$p(x) = p(X = x|\lambda) = \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right] \qquad (24.)$$

Supposons que pour chaque x, cette somme peut-être approchée par son terme le plus grand (hypothèse W *inner-Take-All* ou WTA en anglais) :

$$p(x) = \max_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right] \quad (25.)$$

Cette hypothèse nous ramène à un modèle où l'espace des observations est partitionné en classes gaussiennes. Si, de plus, on suppose que les lois gaussiennes respectent une symétrie sphérique (i.e. $\Sigma_k = \sigma^2 I_p,\ \forall\ k\ /\ 1 \le k \le Q$) et que les poids des lois gaussiennes sont égaux (i.e. $\alpha_k = 1/Q,\ \forall\ k\ /\ 1 \le k \le Q$), l'équation précédente se ramène à :

$$p(x) = \max_{k=1}^{Q} \frac{1}{Q(2\pi)^{p/2}\sqrt{(\sigma^2)^p}} \exp\left[-\frac{\|x-\mu_k\|^2}{2\sigma^2}\right] = \frac{1}{Q(2\pi)^{p/2}\sqrt{(\sigma^2)^p}} \exp\left[-\frac{\min_{k=1}^{Q}\|x-\mu_k\|^2}{2\sigma^2}\right] \qquad (26.)$$

Nous allons maintenant montrer que l'estimateur du modèle λ^* obtenu en maximisant la vraisemblance de l'ensemble d'apprentissage X se ramène à l'estimation d'un dictionnaire de QV par équation des moindres carrés (équation 22).

Si l'on suppose les observations indépendantes et identiquement distribuées, l'estimateur du maximum de vraisemblance donne :

$$\hat{\lambda} = \arg\max p(X) = \arg\max \prod_{i=1}^{N} p(x_i) = \arg\max \sum_{i=1}^{N} \log(p(x_i)) \qquad (27.)$$

Les expressions 26 et 27 nous donnent donc pour l'estimateur du modèle :

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{N} \left[-\log\left(Q(2\pi)^{p/2}\sqrt{(\sigma^2)^p}\right) - \frac{\min_{k=1}^{Q}\|x_i-\mu_k\|^2}{2\sigma^2}\right] \qquad (28.)$$

L'estimateur du maximum de vraisemblance de la variance de X est classiquement donné par :

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \left[\frac{\min_{k=1}^{Q} \|x_i - \mu_k\|^2}{Np} \right]$$ et il est donc indépendant du modèle λ. L'expression 28 se ramène alors à :

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{i=1}^{N} \left[\min_{k=1}^{Q} \|x_i - \mu_k\|^2 \right] \qquad (29.)$$

Les paramètres obtenus pour le modèle λ^* par maximum de vraisemblance sont donc identiques à ceux calculés par équation des moindres carrés dans le cadre de la QV (équation 22).

Cette reformulation de la QV sous forme statistique relie la partition obtenue par un des algorithmes LBG ou LBG-Rissanen vu précédemment à un modèle de type MMG. Cela permet d'une part d'utiliser les centroïdes calculés par QV comme initialisation des modèles MMG, et d'autre part de réintroduire la notion de classes à l'issue d'une classification par MMG (hypothèse WTA). En effet, en supposant que l'observation est produite par la loi gaussienne ayant la plus forte participation à la somme p(x), il est possible d'obtenir pour chaque segment vocalique la « classe » à laquelle elle appartient.

2.4 L'algorithme LBG-Rissanen

Il est naturel de relier le nombre de classes Q aux données de l'ensemble d'apprentissage. Fixer un seuil d'arrêt sur la distorsion relative de l'ensemble d'apprentissage est une méthode simple pour y parvenir, mais le réglage du seuil pose problème comme nous allons brièvement le rappeler.

De par la formulation de l'algorithme LBG, la distorsion des données décroît de manière monotone lorsque l'on augmente le nombre de classes Q. La Figure 39 montre la variation de la distorsion des données [41] calculée en fonction du nombre de classes Q du modèle obtenu par l'algorithme LBG. L'étude de la figure révèle que la distorsion décroît rapidement tant que Q est inférieur à 10 puis tend vers une valeur asymptotique. Comme dans tout problème de ce type, fixer le seuil au delà duquel la variation de distorsion n'est plus significative est difficile : une faible variation du seuil dans la zone asymptotique induit une variation de Q significative. Il est donc naturel de vouloir quantifier, à chaque augmentation du nombre de classes, le gain d'information apporté par l'augmentation de la taille du dictionnaire et de le comparer à la diminution de distorsion induite.

[41] Ces données ont été obligeamment fournies par l'ICP de Grenoble, et sont décrites plus précisément au paragraphe 3.1 de ce chapitre sous le nom de corpus JLVoc.

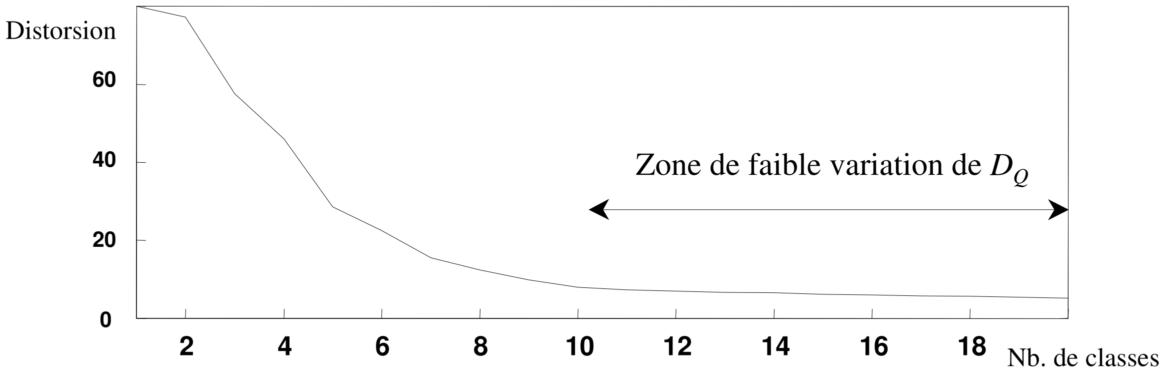

Figure 39 – Variation de la distorsion d'un ensemble de données en fonction du nombre de classes du modèle.

Cette approche peut être reliée à la notion de quantité d'information (via la distorsion) et de coût de quantification (via la taille du dictionnaire). L'algorithme de LBG-Rissanen [André-Obrecht 93] a pour but de mesurer ces deux paramètres et d'en faire la synthèse afin d'obtenir sur la courbe représentant la variation du critère un point d'inversion. De cette manière, on s'affranchit de la contrainte de sélection du seuil.

Dans l'algorithme de LBG-Rissanen [André-Obrecht 93], avant chaque phase d'augmentation du nombre de classes, (phase de *splitting*), un critère de type Rissanen *I(Q)* est calculé [Rissanen 83]. Il prend en compte la distorsion des données d'apprentissage par rapport au modèle, et la complexité du modèle, par l'intermédiaire des paramètres p et Q, correspondant respectivement à la dimension de l'espace des données et au nombre de classes du modèle ; le produit de ces deux entités représente le nombre de paramètres indépendants décrivant le modèle, il est pondéré par le logarithme du carré du cardinal de l'ensemble des données :

$$I(Q) = D_Q + 2p.Q.\log N \qquad (30.)$$

où : - D_Q est la distorsion des données d'apprentissage calculée par la formule 21,

- p est la dimension de l'espace des paramètres,

- Q est la taille actuelle du dictionnaire,

- N est le cardinal de l'ensemble de données.

Le terme de distorsion diminue lorsque Q augmente alors que la complexité du modèle augmente proportionnellement à sa taille. Le nombre de classes optimal Q_{opt} au sens du critère de Rissanen correspond au minimum de la fonction *I(Q)* :

$$Q_{opt} = \arg\min_Q (I(Q)) \qquad (31.)$$

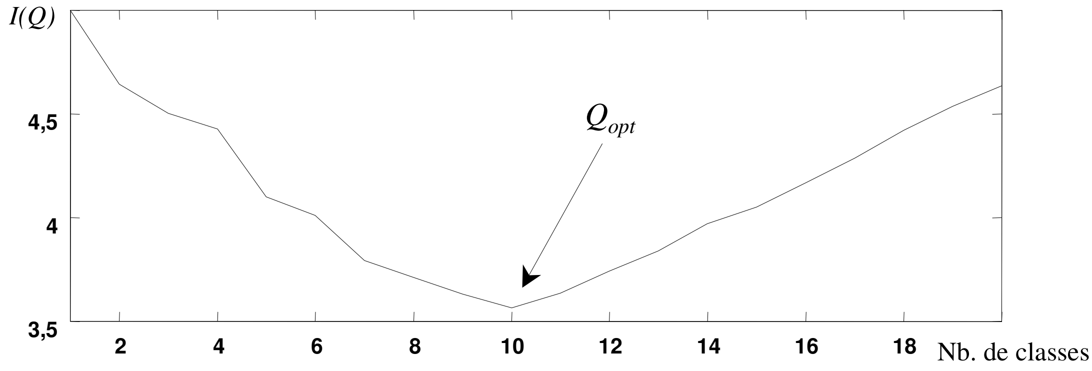

Figure 40 – Variation du critère *I(Q)* en fonction du nombre de classes du modèle.

La Figure 40 représente la variation du critère *I(Q)* calculé sur les mêmes données que précédemment pour différentes tailles de dictionnaire. Lorsque la courbe décroît, l'augmentation de la taille du dictionnaire s'accompagne d'un gain au sens du critère de Rissanen, alors que lorsqu'elle croît, le gain en terme de diminution de la distorsion ne compense plus l'augmentation de la complexité du modèle.

2.5 Discussion

Nous avons vu au cours de ce second paragraphe les différentes modélisations statistiques que nous nous proposons d'employer dans un but de classification des systèmes vocaliques acoustiques. Il s'agit de méthodes utilisées dans de nombreux domaines, en particulier en reconnaissance ou en vérification du locuteur. Aux classiques algorithmes de QV et de MMG, nous avons ajouté l'algorithme LBG-Rissanen qui permet de fixer le nombre de composantes gaussiennes en fonction des données d'apprentissage. L'implantation des différents algorithmes cités a été réalisée sous Matlab [42], offrant ainsi une grande souplesse d'utilisation.

3 EXPERIENCES DE MODELISATION DES SYSTEMES VOCALIQUES

Chacune des méthodes proposées étant non supervisée, on peut s'attendre à obtenir des modèles présentant des différences notables par rapport à ce qu'un expert aurait choisi. Pour étudier de manière précise le comportement des algorithmes employés ainsi que les modélisations qui en résultent, nous avons mené plusieurs séries d'expériences sur des corpus variés, tout comme cela a été présenté pour les algorithmes de détection des segments vocaliques. Aux corpus présentés dans le Tableau 7 s'ajoute une autre série d'enregistrements de très bonne qualité. Ce corpus, que nous désignons

[42] Matlab est un logiciel de programmation scientifique développé par The MathWorks Company (cf. http://www.mathworks.com).

sous le nom de JLVoc, nous a là encore été fourni par l'ICP dans le cadre du projet DGA n° 95/118.

3.1 Expériences sur le corpus JLVoc

3.1.1 Description des données employées

Le corpus JLVoc se compose d'un ensemble de 10 voyelles du français prononcées par un seul locuteur. Il s'agit d'un corpus particulièrement intéressant car les voyelles sont prononcées de manière isolée. Ainsi chaque fichier du corpus correspond à une répétition d'une voyelle hors contexte, tenue sur une durée supérieure à 64 ms. Les seuls facteurs de variation sont donc intra-locuteur.

L'ICP a mis à notre disposition 100 répétitions de chacune des 10 voyelles orales [i, y, u, e, ɛ, ø, œ, o, ɔ, a] enregistrées en chambre acoustique anéchoïque.

Le signal est échantillonné à 16 kHz et paramétré selon l'approche présentée au paragraphe 1.2 de ce chapitre, en centrant la fenêtre d'analyse cepstrale sur le milieu de l'enregistrement. Au cours des expériences qui sont présentées ici, seuls les 8 coefficients cepstraux statiques ont été employés. En effet, étant donnée la nature du corpus, nous avons jugé que les paramètres d'énergie et les coefficients cepstraux dynamiques étaient peu représentatifs.

Les 1000 stimuli sont partagés en un ensemble d'apprentissage (noté APP) de 700 occurrences et un ensemble de test (noté TST) composé des 300 autres. Les deux corpus sont équilibrés, et les fréquences d'apparition des voyelles sont donc égales dans chaque corpus.

La Figure 41 représente les 700 stimuli du corpus APP dans le premier plan factoriel obtenu par analyse en composantes principales (ACP) à partir des 8 coefficients cepstraux. Cette représentation à deux dimensions permet de visualiser de manière claire les données multidimensionnelles mais elle ne peut cependant refléter qu'une partie de l'information présente dans les 8 coefficients initiaux[43]. Chaque ensemble de points est schématisé par une ellipse de manière à préserver la clarté de la projection. On remarque que la plupart des ellipses sont quasiment disjointes, ce qui signifie que la représentation cepstrale est bien discriminante. Bien évidemment, cela met surtout en avant le fait que des voyelles prononcées de manière isolée sont bien discriminables même si l'existence de zones de recouvrement indique que la variabilité intra-locuteur peut introduire des ambiguïtés malgré la qualité des stimuli. Il faut cependant rester prudent puisque la représentation adoptée se base sur deux dimensions et qu'elle représente donc uniquement 74 % de l'information présente dans les 8 MFCC. Une représentation dans un espace de dimension supérieure permet peut-être de dissocier

[43] L'ACP opère une compression avec pertes lorsque l'on ne conserve pas le nombre de dimensions initial.

totalement les nuages de points. On peut constater aussi que certaines voyelles présentent une variabilité plus grande que d'autres. En particulier les nuages des voyelles / y/ et / œ/ sont peu étendus alors que ceux correspondant aux voyelles / e/, /o/ et / ɔ/ le sont nettement plus. Si l'on cherche à établir une relation entre cet espace et l'espace articulatoire classique, la variabilité de / e/ semble se situer surtout au niveau de l'ouverture (le nuage est étiré entre / i/ et / ɛ/) alors que pour les voyelles postérieures, cet axe ne semble pas privilégié. Il est bien évidemment difficile de tirer des conclusions d'une telle représentation car elle n'indique pas clairement la répartition des stimuli à l'intérieur de chaque ellipse. Les représentations statistiques permettent heureusement de faire apparaître ce type d'information comme nous allons le voir à présent.

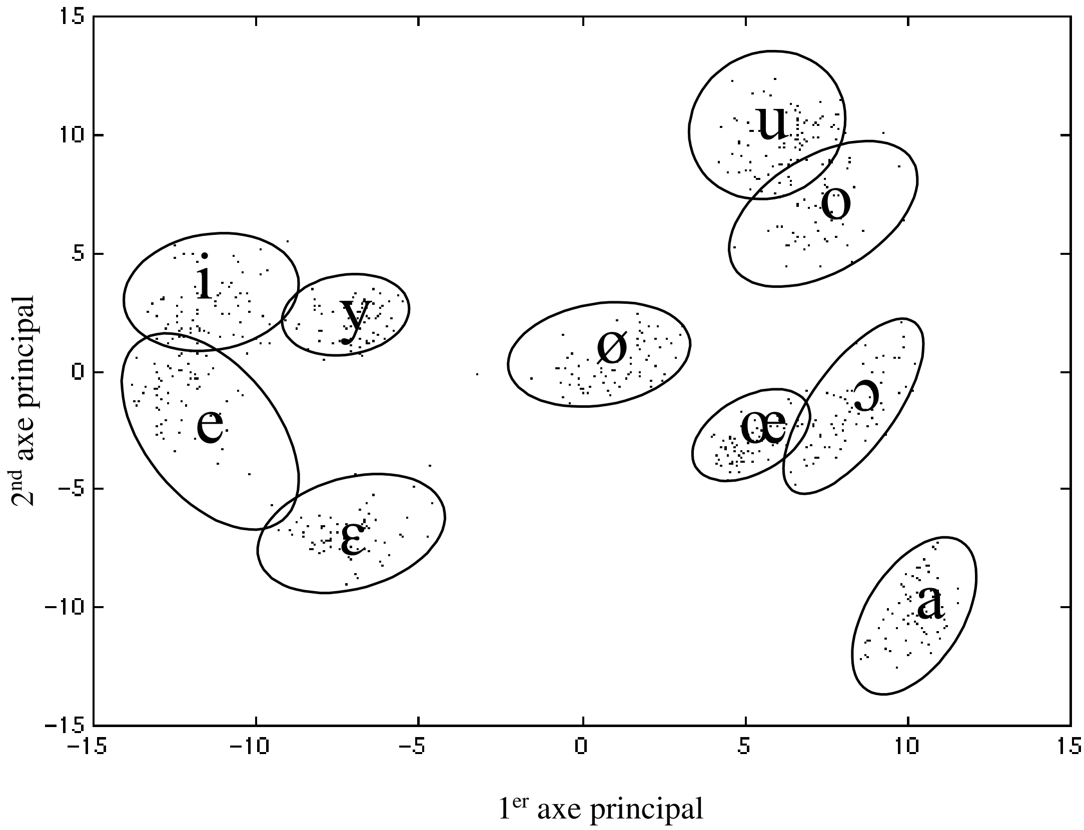

Figure 41 – Répartition des stimuli du corpus JLVoc dans le plan principal dérivé de la modélisation cepstrale.

Si l'on cherche à étudier de manière plus précise la variabilité des différentes voyelles au niveau cepstral, il peut être judicieux d'adopter une autre présentation, par exemple sous forme de « boîtes à moustaches », généralement désignées par le terme anglais de *boxplot*. La Figure 42 explicite la lecture de ce type de diagramme avec l'exemple du /a/. L'axe horizontal correspond aux 8 coefficients MFCC et l'axe vertical correspond à leurs valeurs pour les données APP. Chaque boîte permet de visualiser la répartition des données autour de la médiane par le biais des valeurs extrêmes atteintes et surtout des premier et troisième quartiles, entre lesquels se concentrent la moitié de données. Cette représentation se retrouve pour les 9 autres voyelles avec le même axe vertical. La Figure 43 représente pour chacune des voyelles la « boîte à moustaches » – ou *boxplot* – correspondant aux données APP.

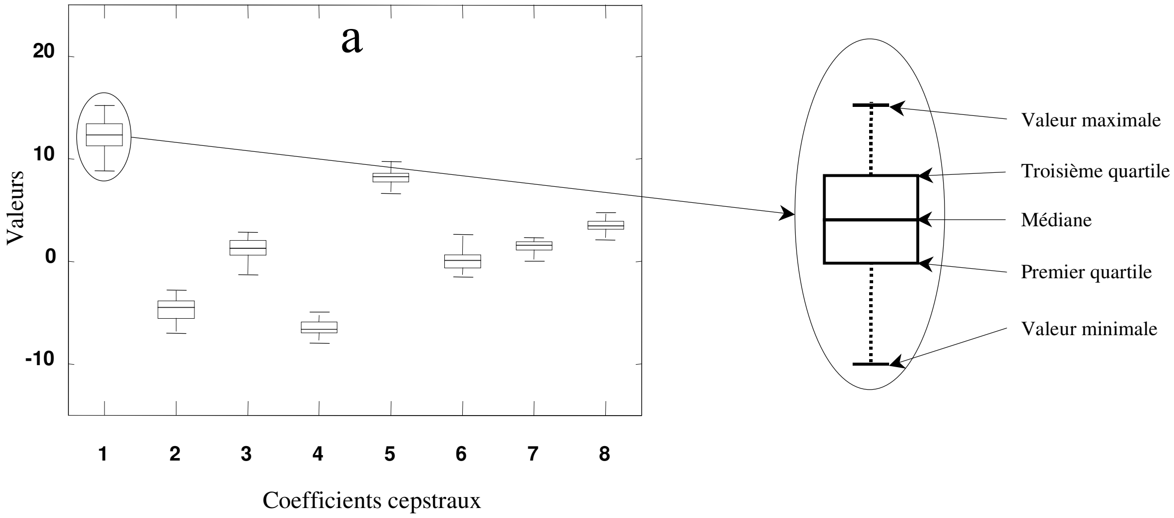

Figure 42 – Exemple de représentation sous forme de *boxplot ;* la voyelle /a/.

L'analyse de ces figures montre que pour un même locuteur, la variabilité des voyelles peut être assez importante, même si ses stimuli sont regroupés dans le premier plan factoriel. L'exemple du /y/ est intéressant car son ellipse est peu étendue sur la Figure 41 alors que le diagramme *boxplot* correspondant indique une variabilité assez importante, répartie sur tous les coefficients excepté le 5ème MFCC. Cela semble indiquer que l'axe de variabilité principal du /y/ n'est pas l'un des deux axes principaux du triangle /a, i, u/. Il s'agit vraisemblablement dans ce cas précis de variations de protrusion et donc d'arrondissement. A l'inverse, certaines voyelles présentent des variations assez importantes sur certains axes et faibles sur d'autres.

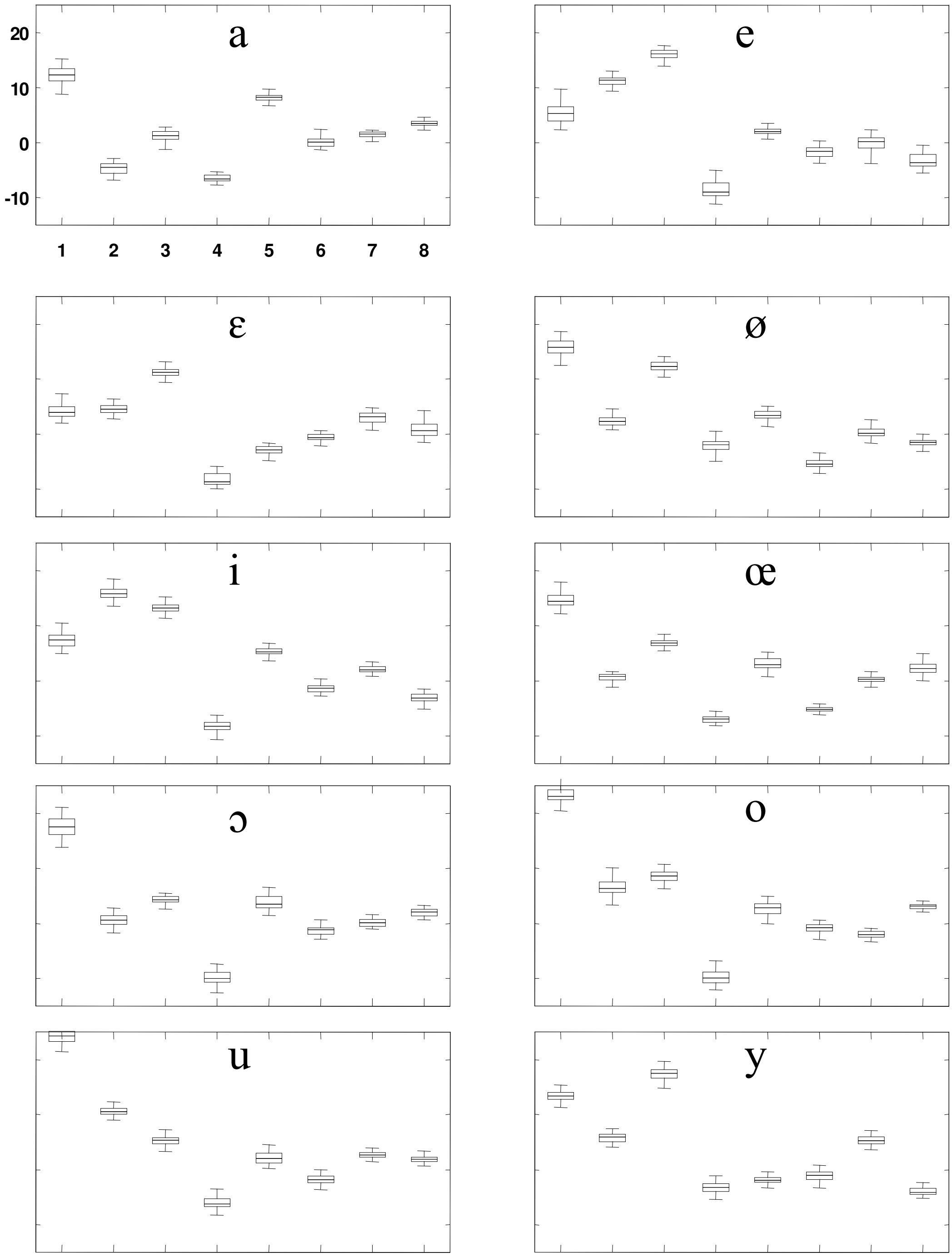

Figure 43 – Répartition des données du corpus APP selon les 8 MFCC. Les axes sont identiques sur chaque *boxplot*.

3.1.2 Modélisation du système vocalique et classification des voyelles

Le corpus JLVoc permet d'étudier la modélisation obtenue à partir des algorithmes de QV et de MMG présentés précédemment. En particulier, nous étions désireux de voir si l'algorithme LBG-Rissanen déterminait correctement le nombre de classes présentes à partir des données APP et s'il était possible d'utiliser les modèles obtenus par cet algorithme et l'algorithme EM dans une tâche de classification des voyelles.

✓ Expériences de modélisation - Algorithme LBG-Rissanen

L'algorithme LBG-Rissanen est utilisé dans une tâche de modélisation du corpus APP, sans qu'aucune contrainte ne soit imposée sur le nombre de classes souhaité. Le résultat obtenu est excellent puisque cet algorithme détermine un nombre de classes Q_{opt} égal à 10, ce qui correspond au nombre de qualités vocaliques présentes (cf. la Figure 40 utilisée comme exemple au paragraphe précédent). A partir de la répartition des différentes voyelles du corpus APP dans ces classes, nous étudions la manière dont les stimuli sont rassemblés. Il s'avère que chaque classe regroupe la totalité des stimuli correspondants à une qualité vocalique, donnant en quelque sorte une matrice de « confusion » classe/voyelle diagonale.

D'autres expériences sont menées dans le but de mieux étudier l'algorithme LBG-Rissanen. En particulier, des sous-ensembles du corpus APP sont extraits et modélisés par cette méthode. A chaque fois, le nombre de classes obtenu correspond au nombre de qualités vocaliques présentes, comme le montre le Tableau 11. Les sous-ensemble constitués des voyelles /o, u/ et /e, ɛ, ø, œ, o, ɔ/ sont employés pour déterminer si la discrimination demeure lorsque les voyelles extrêmes /i, a, u/ sont absentes.

Qualités vocaliques (nb)	Nombre de classes	Qualités vocaliques (nb)	Nombre de classes
/o, u/ (2)	2	/e, ɛ, ø, œ, o, ɔ/ (6)	6
/i, a, u/ (3)	3	/a, e, i, o, u, ø, y/ (7)	7
/a, e, i, o, u/ (5)	5	/a, e, i, o, u, ø, y, ɔ, œ/ (9)	9

Tableau 11 – Nombre de classes déterminé par LBG-Rissanen en fonction du nombre de qualités vocaliques du corpus.

✓ Expériences de classification des voyelles

A partir du moment où l'algorithme LBG-Rissanen permet d'obtenir une modélisation de chaque qualité vocalique du corpus, il est possible de procéder à la classification des stimuli du corpus TST (Figure 44). Il s'agit alors de vérifier que la classe la plus proche (cas de la QV) ou la plus vraisemblable (cas des MMG) d'un stimulus de test correspond bien à sa valeur phonétique.

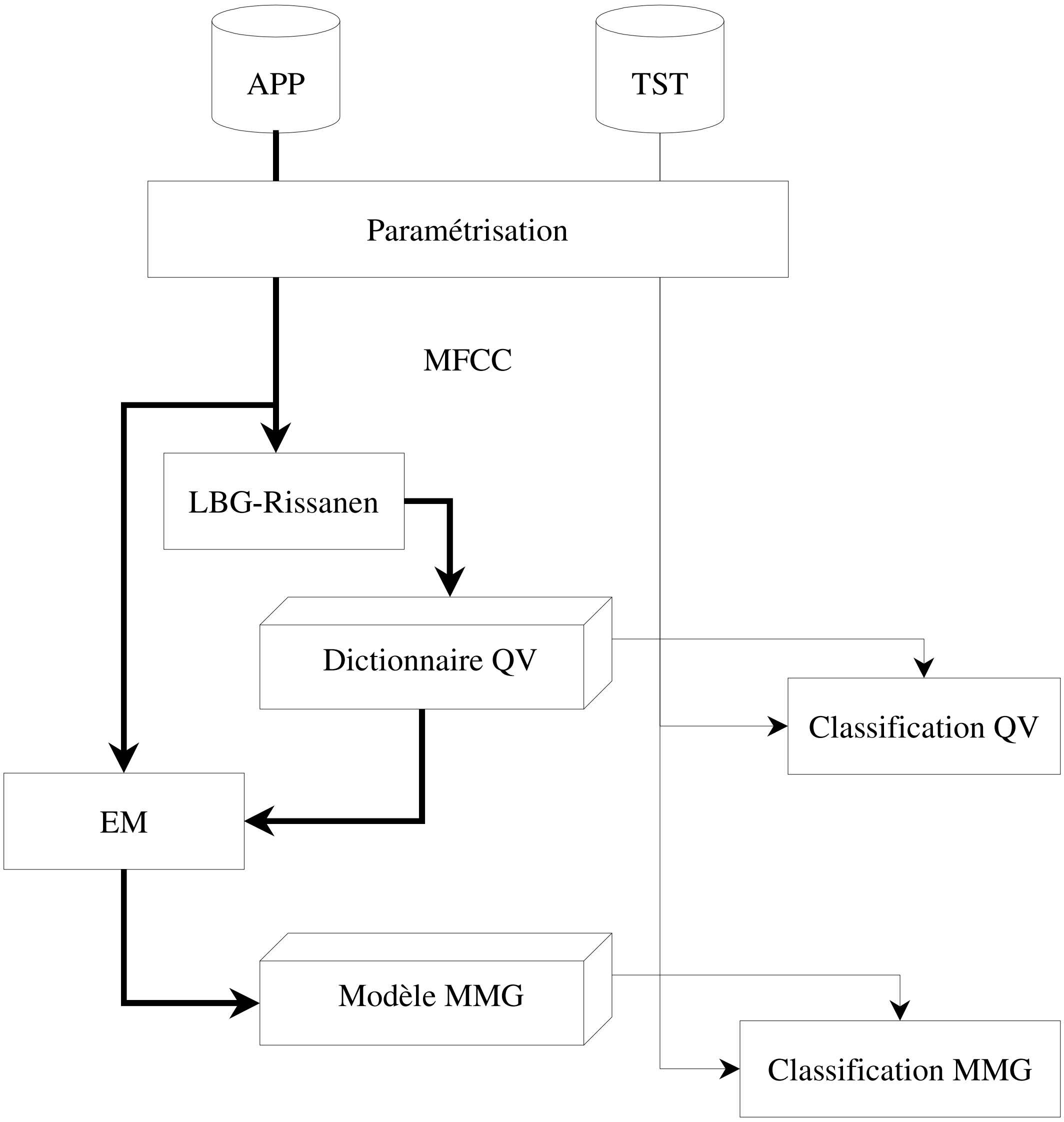

Figure 44 – Schéma synoptique de la modélisation (traits épais) et de la classification des voyelles (trais fins).

La classification basée sur la QV se fait par sélection du centroïde le plus proche au sens de la distance euclidienne. Il s'agit là d'une classification stricte puisque à un stimulus est associé une classe unique. La classification MMG se fait sur la base du calcul de vraisemblance du stimulus par rapport aux différentes lois gaussiennes constituant le mélange. Il s'agit là d'une classification stricte si l'on se place dans l'hypothèse WTA (*Winner-Take-All*).

Modèle	Nombre de stimuli	Nombre d'erreurs	Taux de classification
Dictionnaire QV – 10 classes	300	4	98,7 %
Modèle MMG – 10 gaussiennes	300	-	100 %

Tableau 12 – Résultat de la classification des voyelles.

Les résultats obtenus sont excellents (Tableau 12), puisque les seules erreurs enregistrées en utilisant une classification par distance euclidienne portent sur quatre stimuli de /o/ classés parmi les /u/. En utilisant un critère de vraisemblance et un modèle MMG, ces erreurs disparaissent.

3.2 Expériences sur le sous corpus issu de EUROM

Les données employées pour ces expériences proviennent de la détection effectuée sur les données extraites du corpus Eurom décrites au chapitre précédent. Tous les segments ayant été détectés comme vocaliques sont utilisés. Il s'agit donc d'une situation réelle en ce sens que des fausses détections (consonnes ou semi-voyelles) sont aussi présentes dans les données modélisées.

La Figure 45 représente dans le premier plan factoriel l'ensemble des segments détectés et paramétrés par 8 coefficients cepstraux. Chaque label est placé au centre de gravité du nuage correspondant. On constate que la présence des voyelles nasales et la nature de l'enregistrement (parole continue) modifie assez nettement les positions des voyelles par rapport au système issu du corpus précédent ; on note en particulier l'apparition de la voyelle / ã/. De manière générale, il est évident que le processus de coarticulation augmente fortement la distance entre les individus d'une même qualité vocalique (distance inter-classe). Remarquons également que l'utilisation par l'expert du seul symbole phonétique / ø/ pour tous les sons [ø, œ, ə] aboutit a une position centrale de cette classe.

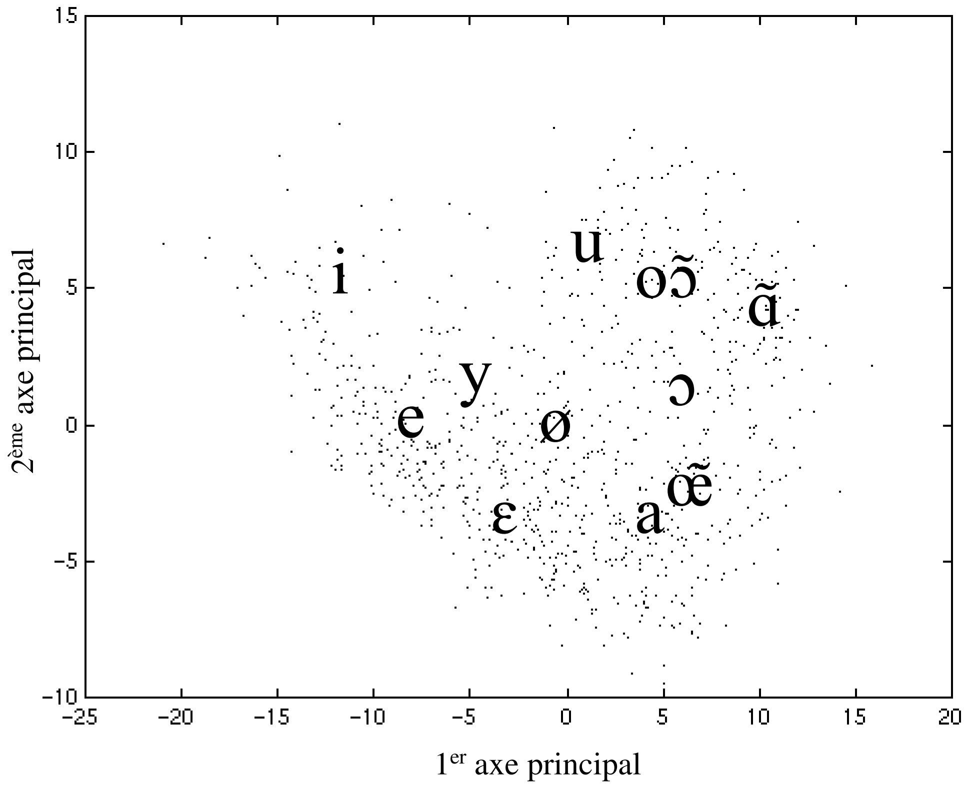

Figure 45 – Répartition des segments vocaliques détectés dans le premier plan factoriel.

✓ Expériences de modélisation - Algorithme LBG-Rissanen

L'algorithme LBG-Rissanen aboutit à un modèle à 5 classes. Cela signifie donc que dans cette tâche complexe, le critère d'information employé augmente plus vite que la distorsion ne baisse lorsque le nombre de classes croît. Cela est vraisemblablement dû au recouvrement des différentes qualités vocaliques orales et nasales dans l'espace cepstral. En observant la répartition des segments vocaliques détectés (Tableau 13 et Tableau 14) dans les différentes classes, on constate que certaines qualités se retrouvent très majoritairement dans une classe (c'est le cas de /i/ et des voyelles nasales) alors que d'autres se trouvent réparties dans deux, voire trois classes (le / ɛ/ et le / ɔ/ par exemple).

Cet examen révèle les attributions grossières suivantes :

- ✓ la classe 1 correspond principalement à /i/,
- ✓ la classe 2 correspond principalement à /y/, /e/ et / ø/,
- ✓ la classe 3 correspond principalement à /œ̃/ et / a/,
- ✓ la classe 4 correspond principalement à /ɑ̃/,
- ✓ la classe 5 correspond principalement à /u/, /o/ et / ɔ̃/

Le fait que l'une des classes corresponde principalement à une voyelle nasale (classe 4) indique que ce trait est bien pris en compte par le modèle. Les classes 1, 3 et 5 correspondent plutôt aux voyelles périphériques (respectivement avant-fermé, avant-ouvert et arrière-fermé) tandis que la classe 2 comporte des voyelles intérieures et périphériques antérieures : le /e/ se répartit entre les classes 1 et 2 et le / ɛ/ occupe de manière équilibrée les espaces définis par les classes 2 et 3. La voyelle / ɔ/ se répartit quant à elle sur tout le front arrière (classes 3, 4 et 5).

La première conclusion de cette étude est qu'il est impossible d'employer cette approche non supervisée en classification des voyelles, en particulier à cause du faible nombre de classes déterminé par le critère de Rissanen. La seconde conclusion est plus satisfaisante ; en effet, si l'on se reporte à la Figure 30 concernant la composition de l'ensemble des segments détectés, on constate que le nombre de segments /i/ du corpus est quatre fois plus faible que le nombre de représentants du /a/. Cela vient en grande partie du faible taux de détection des sons fermés, comme nous l'avions expliqué au chapitre précédent. Malgré ce handicap assez net hérité de la détection, l'algorithme LBG-Rissanen regroupe ces segments /i/ au sein de la classe 1. On retrouve donc une classe avant-fermé (de même, on retrouve une classe arrière-fermé) alors que les données étaient relativement faiblement représentées dans l'ensemble modélisé.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
i	43	6	1	-	1
y	3	11	-	-	5
e	32	72	2	-	-
ø	-	54	27	1	22
ɛ	12	57	50	-	-
a	-	8	94	31	1
œ̃	-	-	18	6	-
ã	-	-	1	58	2
ɔ	-	6	13	22	10
o	-	-	3	3	23
õ	-	-	-	5	28
u	1	5	-	1	30

Tableau 13 – Répartition des voyelles détectées dans les 5 classes obtenues par LBG-Rissanen.

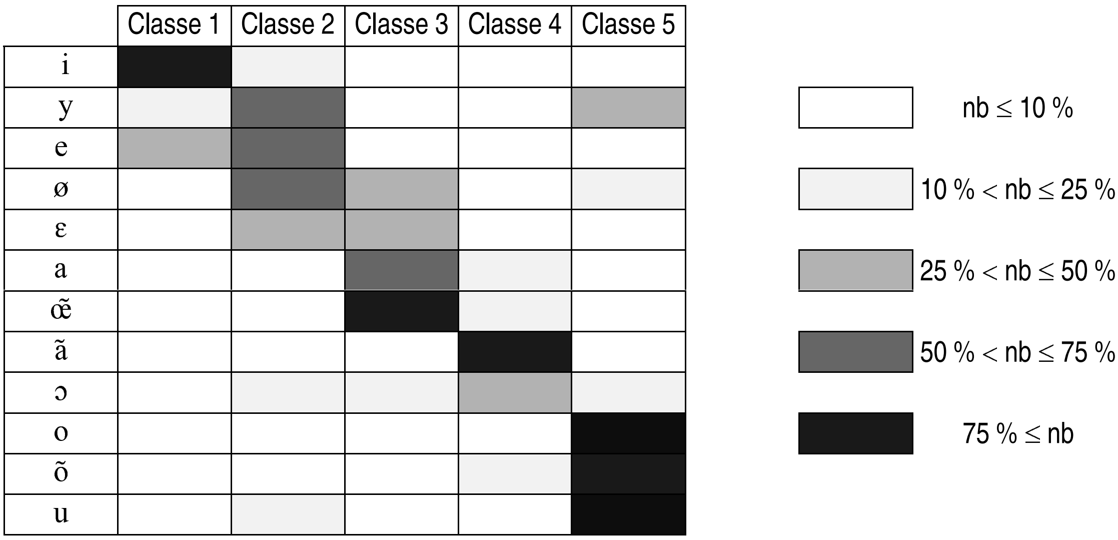

Tableau 14 – Répartition en pourcentage des voyelles détectées dans les cinq classes.

✓ <u>Expériences de modélisation - Algorithme LBG</u>

A titre de comparaison, nous étudions la répartition des stimuli lorsque le nombre de classes correspond au nombre théorique de qualités vocaliques présentes dans le corpus. Pour cela, un modèle est calculé par algorithme LBG standard en contraignant le nombre de classes à une valeur de 12. Le Tableau 15 indique comme précédemment la répartition des segments vocaliques dans chaque classe. On peut noter que là encore, la voyelle nasale /ã/ constitue une classe unique (la classe 9) alors que les voyelles orales /u/ et /o/ ne sont guère discriminées. Une meilleure discrimination entre les voyelles intérieures et avant est observée : en particulier, la voyelle /ø/ correspond à la classe 8 alors que la voyelle /e/ en est bien dissociée (classes 11 et 12). La classe 12 regroupe des stimuli /i, e/, même si la majorité des /i/ constituent là encore une classe (classe 1). On retrouve par contre une séparation des voyelles /a/ en deux ensembles, « plutôt avant » et « plutôt arrière » (classes 7 et 2) tout comme pour la voyelle nasale /œ̃/. Cela peut correspondre à des distinctions phonologiques entre /a/ et /ɑ/ et entre /œ̃/ et /ɛ̃/ faites par le locuteur mais non relevées par l'expert ou à l'inverse à une grande dispersion acoustico-phonétique des voyelles /a/ et /œ̃/. On peut constater enfin qu'il existe une classe « fourre-tout » pour les voyelles fermées (classe 4).

	1	2	3	4	5	6	7	8	9	10	11	12
i	25	-	-	8	-	-	-	1	-	3	1	13
y	-	-	-	2	3	-	-	6	-	-	6	2
e	1	-	-	2	-	-	-	1	-	14	56	32
ø	-	4	-	6	14	-	5	56	-	4	15	-
ɛ	3	-	-	-	-	-	27	4	-	54	22	9
a	-	43	-	1	1	-	67	7	1	12	2	-
œ̃	-	13	-	-	-	-	8	1	-	2	-	-
ã	-	2	1	-	1	4	-	1	52	-	-	-
ɔ	-	12	-	1	3	11	2	12	7	-	3	-
o	-	1	3	-	14	7	2	2	-	-	-	-
õ	-	-	24	-	6	1	-	-	2	-	-	-
u	-	-	8	6	16	5	-	-	1	-	-	1

Tableau 15 – Répartition des voyelles détectées dans les 12 classes obtenues par LBG.

L'étude de la modélisation du système vocalique acoustique du locuteur FB de la base de données Eurom confirme d'une part que la variabilité des voyelles prononcées en parole continue est sans commune mesure avec les stimuli du corpus JLVoc, et d'autre part que, si certains contrastes demeurent suffisants pour effectuer une discrimination entre qualités vocaliques (cas du /ø/ et du /ɑ̃/ par exemple), d'autres se révèlent trop faibles pour que des modèles établis de manière non supervisée ne les détectent (cas du /u/ et du /o/ par exemple). Même si l'objectif de notre approche n'est pas de modéliser individuellement chaque qualité vocalique, ce résultat permet d'évaluer les limites d'une représentation acoustique des systèmes vocaliques. Par contre, le fait que l'algorithme LBG permette de modéliser correctement un ensemble de données même si les qualités vocaliques présentes ne sont pas équilibrées nous paraît être une conclusion importante et encourageante.

3.3 Expériences sur le corpus OGI MLTS

Les expériences effectuées sur les corpus JLVoc et EUROM sont réalisées dans un environnement monolocuteur. La modélisation du système vocalique acoustique d'une langue nécessite bien évidemment la prise en compte de la variabilité inter-locuteurs. Les expériences présentées ici ont donc été réalisées à partir des segments vocaliques détectés sur les enregistrements des locuteurs correspondants aux cinq langues extraites du corpus OGI MLTS (notées toujours FR, JA, KO, SP et VI). Avant d'aller plus loin dans cette description, il est nécessaire de préciser la manière dont sont organisés les enregistrements du corpus OGI. Le corpus est organisé, selon les recommandations du NIST, en trois sous-corpus dédiés : à l'apprentissage des modèles (APP), à leur amélioration et à leur test (DEV) ainsi qu'à leur évaluation finale (TST).

Bien évidemment, les ensembles sont disjoints en terme de locuteurs, ce qui garantit que le système développé ne réalise pas une identification du locuteur ou de la ligne téléphonique mais bien de la langue.

Le Tableau 16 indique pour chacune des langues traitées le nombre de locuteurs masculins et féminins composant les trois corpus. Comme on peut le constater, les corpus ne sont pas équilibrés en terme d'hommes et de femmes. Etant données les différences nettes existant au niveau du système vocalique entre les locuteurs masculins et féminins, il ne nous semble pas raisonnable de procéder à une modélisation commune de ces espaces. En conséquence, les expériences réalisées l'ont généralement été avec les locuteurs masculins du corpus. En terme d'identification des langues, cela suppose soit d'effectuer une détection du sexe du locuteur avant de procéder à l'identification de la langue, soit de mettre en parallèle deux systèmes de reconnaissance, l'un entraîné avec les locuteurs masculins et l'autre avec les locuteurs féminins selon une approche semblable à celle employée en RAP.

		FR	JA	KO	SP	VI
APP	Hommes	40	30	32	34	31
	Femmes	10	20	17	16	19
DEV	Hommes	15	15	18	16	16
	Femmes	5	5	2	4	4
TST	Hommes	12	11	15	11	13
	Femmes	8	8	5	8	7

Tableau 16 – Répartition des locuteurs du corpus OGI MLTS.

Les expériences de modélisation menées sur ces données constitue le cœur du système d'IAL développé au cours de cette thèse. Elles seront donc présentées de manière approfondie au cours du prochain chapitre, intitulé fort justement « Expériences en IAL ». Nous allons donc plutôt présenter ici quelques observations établies sur les données issues de la détection des segments vocaliques et de leur paramétrisation cepstrale.

Chaque segment vocalique détecté est paramétré par ses 8 coefficients MFCC, ses 8 coefficients dynamiques ΔMFCC auxquels est ajouté sa durée.

3.3.1 Normalisation et représentation cepstrale

Dès lors que l'on s'intéresse à des données provenant de plusieurs locuteurs et de plusieurs canaux d'enregistrement, il peut être judicieux voire indispensable de procéder à une normalisation des données. Dans le cadre du traitement des données OGI MLTS, deux traitements ont été appliqués. Le premier, effectué au niveau de l'ensemble des données correspondant à un appel téléphonique, consiste en une soustraction cepstrale. Ce traitement, classiquement employé en TAP, vise à éliminer l'influence du canal téléphonique, considérée comme additive dans le plan cepstral. Le second réalise une normalisation inter-locuteurs, de manière à représenter les segments vocaliques détectés dans un espace commun à tous les locuteurs. En effet, il est possible que les distances inter-vocaliques d'un locuteur donné soient plus importantes que pour un autre, ou encore que le niveau d'enregistrement soit si différent entre deux appels qu'un facteur d'échelle non négligeable sépare les deux espaces vocaliques cepstraux résultants.

La soustraction cepstrale permet de ramener les nuages de points correspondants à chaque locuteur autour de l'origine dans l'espace des paramètres. Le second traitement consiste alors, pour chaque locuteur, à normaliser chaque paramètre en le divisant par son écart-type calculé sur tous les segments vocaliques du locuteur. Les données détectées pour chaque locuteur sont donc représentées dans un espace paramétrique ayant des bornes communes. Ces traitements supposent cependant qu'il n'existe pas de phénomène de rotation entre les espaces vocaliques des locuteurs, puisqu'ils fonctionnent uniquement sur les principes de translation et d'homothétie. De nombreuses recherches sont menées sur la normalisation des locuteurs ; la plupart d'entre elles sont basées sur une transformation de l'échelle spectrale (*warping* en anglais) à partir de caractéristiques extraites des formants. On pourrait également envisager une normalisation des sous-espaces cepstraux en déterminant les axes principaux du système vocalique de chaque locuteur, mais nous n'avons mené aucune expérience dans ce sens.

La Figure 46 présente, pour chacune des cinq langues, l'ensemble des segments vocaliques détectés pour les locuteurs masculins du corpus APP. Le nombre de points des différents nuages est donc fortement lié au nombre de voyelles présentes dans chaque langue pour une durée comparable. Il y a environ 14000 points pour le français, 10000 pour le japonais et le coréen, 13000 pour l'espagnol et seulement 8000 pour le vietnamien. L'espace de représentation est commun à toutes les projections puisqu'il s'agit de l'espace factoriel calculé par analyse en composantes principales sur l'ensemble des données.

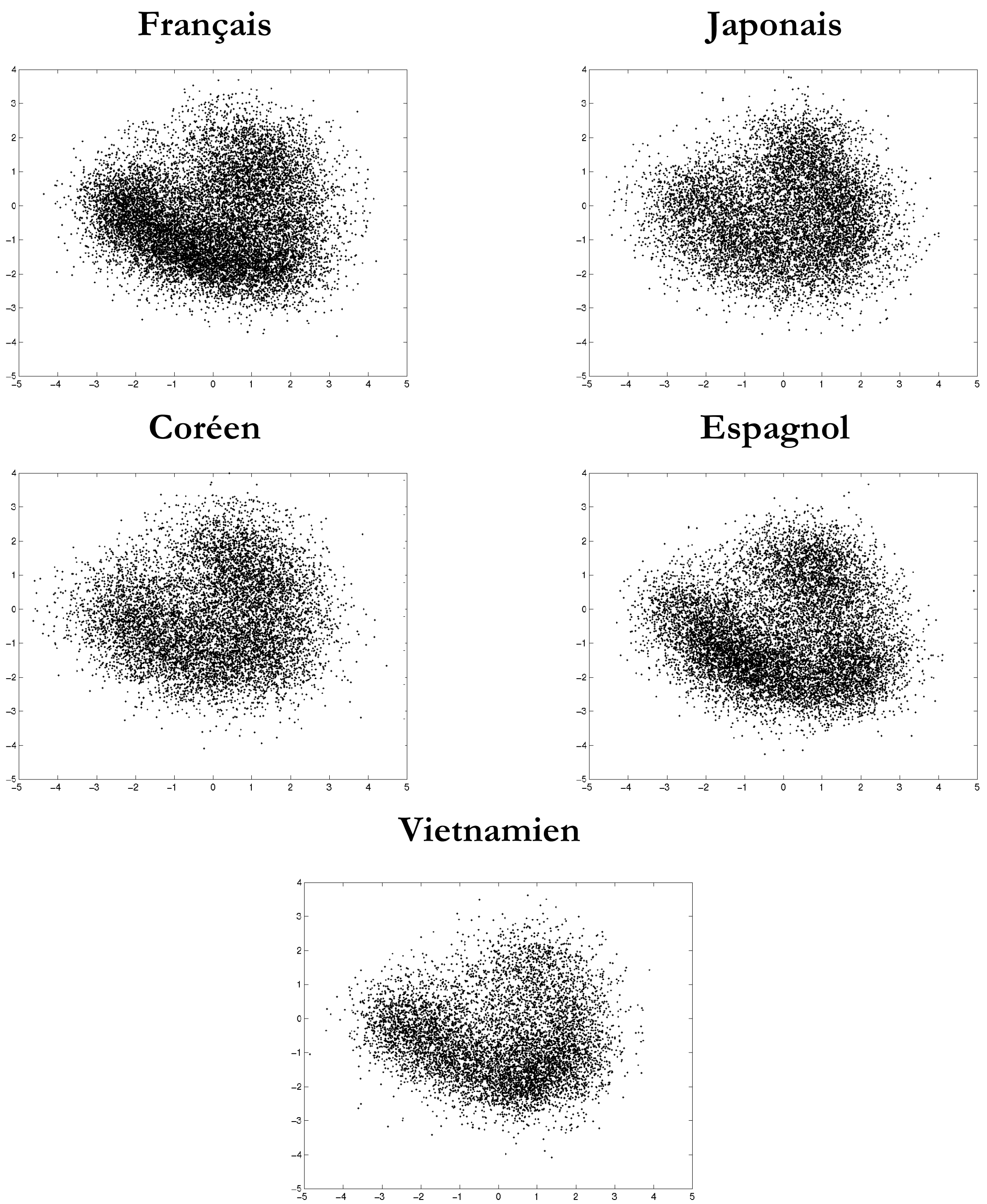

Figure 46 – Répartition dans le premier plan factoriel des segments vocaliques extraits des locuteurs masculins APP pour les cinq langues.

Chaque nuage de points présentent des caractéristiques générales proches quant à sa forme : on retrouve en particulier la structure en « fer à cheval » classique [Boë 98], plus ou moins marquée selon les langues. Elle correspond à la faible représentation des voyelles centrales fermées observée dans la plupart des langues du monde. L'écoute d'extraits du signal permet de vérifier en effet que les axes sont proches de ceux déterminés sur les corpus JLVoc et EUROM. L'observation des différents nuages permet cependant de faire apparaître des différences individuelles sur la répartition des segments vocaliques. La densité de points dans le nuage du français indique donc une assez forte proportion de voyelles antérieures. On peut aussi remarquer que, si les systèmes vocaliques phonologiques du japonais et de l'espagnol sont proches, le nombre d'occurrences des différentes qualités vocaliques doit sensiblement différer car les nuages ne sont pas identiques. Le nuage correspondant au coréen ne présente pas de déséquilibre majeur visible à cette échelle, tandis que celui du vietnamien laisse supposer une assez forte proportion de voyelles antérieures. Ces observations sont bien évidemment préliminaires, et, comme il a été dit précédemment, rien n'assure que les algorithmes de modélisation n'obtiennent des conclusions similaires en prenant en compte la totalité des informations présentes dans les paramètres.

Chapitre 3

LA DISCRIMINATION DES SYSTEMES VOCALIQUES APPLIQUEE A L'IAL

Les études préliminaires menées sur les voyelles détectées dans le corpus OGI MLTS montrent que l'on retrouve des différences statistiques assez nettes entre les systèmes vocaliques acoustiques des cinq langues étudiées. Les expériences menées en modélisation des systèmes vocaliques du français laissent supposer qu'il est possible de modéliser efficacement ces systèmes acoustiques, mais pas de les relier aisément à leur contrepartie phonologique (l'expérience sur le corpus JLVoc reste un cas très particulier, de par la nature des données). Une grande inconnue demeure donc à la fois sur la possibilité de modéliser un système vocalique acoustique issu des enregistrements de plusieurs dizaines de locuteurs et sur le pouvoir discriminant des modèles obtenus. L'objectif de ce chapitre est de lever cette inconnue en comparant les performances d'un système d'IAL basé sur la discrimination automatique des systèmes vocaliques avec celles issues d'un système de référence plus classique.

Nous décrivons tout d'abord les enregistrements issus de la base OGI MLTS qui ont servi de corpus de test dans ce chapitre. Le second paragraphe présente le système de référence, basé sur une modélisation de l'ensemble des sons présents dans le corpus d'apprentissage de chaque langue par un GMM. Cette approche est inspirée de l'une de celles proposées par Zissman [Zissman 96] à laquelle nous intégrons la dimension temporelle des unités via la segmentation a priori du signal.

Le troisième paragraphe est consacré à notre système d'IAL basé sur la discrimination des modèles de systèmes vocaliques. L'influence des différents facteurs intervenant au cours de la modélisation et du processus d'identification est étudiée au cours de nombreuses expériences. Nous examinons particulièrement le comportement de l'algorithme LBG-Rissanen et son influence sur les résultats en IAL.

1 DESCRIPTION DES ENSEMBLES D'APPRENTISSAGE ET DE TEST

Les enregistrements employés pour toutes les expériences en IAL sont issus du corpus OGI, organisé en trois sous-ensembles comme il a été précisé au chapitre précédent (Tableau 16). Rappelons que lorsque le contraire n'est pas explicitement précisé, les expérimentations sont réalisées uniquement avec les locuteurs masculins. A l'apprentissage, la totalité des enregistrements APP est utilisée, qu'il s'agisse de phrases fixes (les jours de la semaine, les chiffres) ou d'énoncés spontanés ou semi-spontanés (description du climat, du repas…). Ainsi, pour chaque langue, on dispose en apprentissage d'une durée totale d'environ 50 minutes.

En phase de discrimination, les expériences sont réalisées avec deux ensembles de fichiers de parole spontanée ou semi-spontanée issus du corpus DEV (Tableau 17). Il s'agit des fichiers *story-bt* qui ont une durée proche de 45 secondes et de l'ensemble des phrases ayant une durée voisine de 10 secondes. Dans la suite de ce chapitre, ces deux sous-corpus sont respectivement appelés DEV_ST et DEV_10. Certaines études complémentaires sont menées avec l'ensemble des données disponibles pour chacun des locuteurs du corpus DEV, pour une durée totale d'environ 2 minutes par locuteur. Ce troisième ensemble de test est désigné par DEV_ALL.

	DEV_ST		DEV_10	
	Nombre de fichiers	Durée moyenne	Nombre de fichiers	Durée moyenne
Français	14[44]	47	54	8
Japonais	13	48	59	7
Coréen	17	41	68	8
Espagnol	15	45	64	8
Vietnamien	15	45	63	7

Tableau 17 – Description des ensembles de test DEV_ST et DEV_10. Les durées sont indiquées en secondes.

Devant leur faible nombre (74), nous conservons délibérément la totalité des fichiers du corpus DEV_ST même si la proportion d'enregistrements de chaque langue n'est pas équilibrée. Pour cette raison, nous donnons dans la plupart des cas le score d'identification global calculé sur l'ensemble des 74 fichiers plutôt qu'un score par langue.

2 LE SYSTEME DE REFERENCE

2.1 Synoptique du modèle MMG segmental global

Le système de référence est basé sur une modélisation de tous les segments de parole présents dans le corpus d'apprentissage par un MMG global segmental par langue (Figure 47). Mise à part la détection des segments vocaliques qui n'est pas employée ici, les traitements appliqués au signal sont identiques à ceux présentés au cours des deux chapitres précédents : segmentation *a priori*, détection d'activité vocale, paramétrisation cepstrale et modélisation MMG. Les vecteurs cepstraux sont calculés pour chaque segment en son milieu (fenêtre de 32 ms centrée). Notre approche segmentale diffère de

[44] A l'origine, il y a quinze locuteurs de langue française, mais il s'avère que l'un d'eux parle en latin dans les enregistrements.

l'approche classique basée sur une analyse cepstrale pratiquée sur des trames de longueur fixe (approche centi-seconde). Plusieurs expériences en RAP ont montré l'intérêt de l'approche segmentale [Suaudeau 94], et il est probable qu'elle se révèle également efficace en IAL.

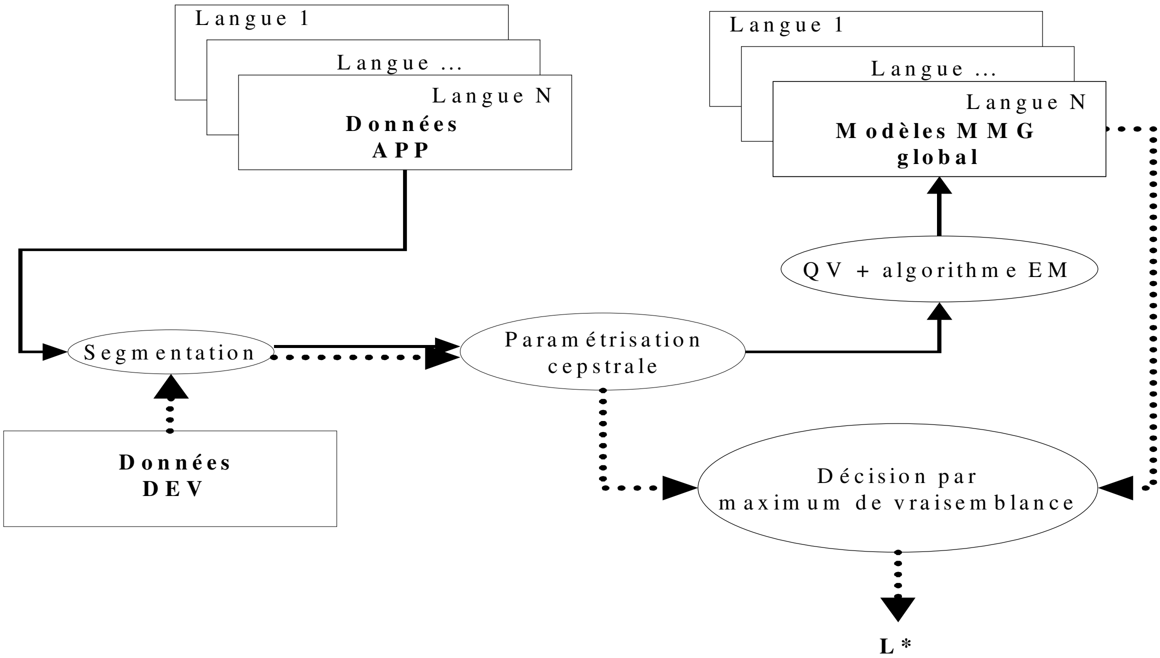

Figure 47 – Description du fonctionnement du système de référence. Le processus d'apprentissage est spécifié en flèches pleines, et le processus d'identification en flèches pointillées.

2.2 Topologie du modèle et résultats expérimentaux

2.2.1 Choix de l'espace de paramètres

Au cours d'une première expérience le système ainsi défini est utilisé en identification des langues sur le corpus DEV_ST avec quatre jeux de paramètres intégrant ou non les paramètres cepstraux dynamiques, l'énergie et sa dérivée ou encore la durée du segment. Pour chaque langue, le nombre de composantes gaussiennes est fixé arbitrairement à 20, puis le modèle est initialisé par algorithme LBG et appris par algorithme EM à partir des enregistrements APP. En phase d'identification, la langue obtenant la vraisemblance cumulée maximale (calculée dans l'hypothèse *Winner-Take - All* ou WTA) est retenue et aucune procédure de rejet n'est intégrée.

Le système obtient un taux d'identification correcte maximal de 73 % avec l'ensemble des 19 paramètres (Tableau 18). Le français est la langue la mieux reconnue (100 %) alors que le japonais est confondu plus d'une fois sur deux avec une autre langue.

Paramètres	Score d'identification (%)					
	français	japonais	coréen	espagnol	vietnamien	Moyenne
8 MFCC	100	31	76	73	60	69
8 MFCC + D	93	54	71	73	67	72
8 MFCC + 8 ΔMFCC +D	100	54	47	87	73	72
8 MFCC + E + 8 ΔMFCC + ΔE + D	100	38	71	80	73	73

Tableau 18 – Taux d'identification correcte obtenus sur le corpus DEV_ST avec le modèle global segmental de 20 lois gaussiennes.

2.2.2 Influence du nombre de composantes du modèle

L'expérience présentée ci-dessus est complétée par une évaluation de l'influence de la taille Q des modèles gaussiens sur le taux d'identification des langues. Les expériences sont toujours réalisées avec le même nombre de composantes pour chacune des langues. On constate (Tableau 19) que pour les espaces d'observations ne prenant pas en compte les coefficients cepstraux dynamiques, les scores d'identification varient peu lorsque Q augmente. A l'inverse, lorsque ces coefficients sont pris en compte, avec ou sans les coefficients d'énergie, le nombre de composantes du modèle est un facteur de variation important et on peut déterminer un nombre optimal de classes.

Paramètres \ Q	20	30	40	50	60	70	80
8 MFCC + D	72	74	72	72	70		73
8 MFCC + 8 ΔMFCC +D	72	80	78	85	80	77	66
8 MFCC + E + 8 ΔMFCC + ΔE + D	72	78	84	76	77	73	70

Tableau 19 – Taux d'identification correcte obtenus sur le corpus DEV_ST avec le modèle global segmental en faisant varier la taille des modèles gaussiens.

Le meilleur taux d'identification correct est de 85 %, et il est obtenu avec des modèles gaussiens comportant cinquante composantes, et une modélisation à 17 paramètres (8 MFCC + 8 ΔMFCC +D). La prise en compte de l'énergie et de sa dérivée ne montre quant à elle pas de changement significatif du taux d'identification correcte.
La matrice de confusion obtenue avec le meilleur système (Tableau 20) montre de bons taux d'identification pour quatre des cinq langues (taux proches de 90 %). Seul le japonais révèle un taux de confusion important. Le taux d'identification correcte de cette langue n'atteint en effet que 54 %, et les locuteurs japonais sont identifiés comme parlant espagnol dans 30 % des cas.

Langue \ Modèle	FR	JA	KO	SP	VI
français	13			1	
japonais		7		4	2
coréen			15	2	
espagnol		1		14	
vietnamien	1				14

Tableau 20 – Matrice de confusion obtenue sur le corpus DEV_ST avec 17 paramètres (8 MFCC + 8 ΔMFCC + D) et des modèles à 50 composantes.

Ces expériences sont complétées par une étude de l'influence du type de décision prise lors de la classification : il s'avère que l'hypothèse WTA, utilisée dans les expériences décrites ci-dessus, aboutit à une classification significativement plus efficace que la prise en compte de l'ensemble des composantes gaussiennes de chaque modèle dans le calcul de la vraisemblance.

2.3 Résultats et discussion

Peu de résultats obtenus avec une approche MMG sont disponibles dans la littérature. On peut citer le système “GMM” décrit par Zissman évalué sur le corpus OGI. Ce système est basé sur la fusion des vraisemblances obtenues par deux MMG de 40 gaussiennes, l'un traitant un flux de 12 paramètres cepstraux statiques et l'autre un flux de 12 paramètres cepstraux dérivés, calculés sur des trames de 20 ms. Un pré-traitement (détection d'activité vocale, filtrage RASTA) est appliqué au signal. Les résultats donnés par Zissman portent sur une tâche d'identification d'une langue parmi trois (anglais, espagnol et japonais). Les résultats obtenus avec le corpus d'enregistrements de 45 secondes sont de **65 %** d'identification correcte.

Nous avons donc comparé les résultats obtenus par Zissman aux taux calculés avec notre système de référence testé également sur trois langues (le français remplaçant l'anglais). En prenant en compte uniquement les locuteurs masculins, nous obtenons un taux d'identification correcte de 86 %. Nous avons également introduit les locuteurs féminins dans l'ensemble de test, de manière à évaluer la dégradation résultant de l'inadaptation entre les modèles (appris sur les locuteurs masculins) et les locuteurs à identifier (féminins et masculins). Le taux d'identification correcte obtenu (phrases de 45 s) est alors de **76 %**. Cela signifie que, même si les modèles utilisés ne sont pas adaptés aux locuteurs féminins, le taux d'identification obtenu par modélisation segmentale est significativement plus élevé que celui obtenu par Zissman dans son approche centi-seconde. Le fait que l'une des langue diffère dans chaque expérience (anglais ou français) reste bien évidemment un élément à prendre en compte. Cependant, il semble que l'utilisation de la segmentation introduit une paramétrisation efficace puisque, en extrayant un seul vecteur d'observation par segment, les scores obtenus sont meilleurs que ceux atteints par Zissman avec une approche centiseconde.

3 LA DISCRIMINATION AUTOMATIQUE DES SYSTEMES VOCALIQUES (SV)

3.1 Description du système

Ce paragraphe présente les expériences menées en discrimination automatique des SV sur les données issues du corpus OGI telles qu'elles sont présentées précédemment. Le protocole opératoire est donc proche de celui appliqué dans le système de référence à la différence notable que l'algorithme de détection des segments vocaliques présenté au Chapitre 1 est appliqué. Dans le but d'étudier la validité de l'approche différenciée et le pouvoir discriminant des SV, seuls les segments vocaliques sont modélisés (Figure 48). Aucun modèle consonantique n'est employé dans ces expériences, contrairement à ce qui sera envisagé au chapitre suivant et il est important de noter qu'il s'agit donc là d'une modélisation partielle des informations présentes dans le signal, et que les segments pris en compte ne représentent au mieux qu'un tiers de la durée totale des énoncés employés.

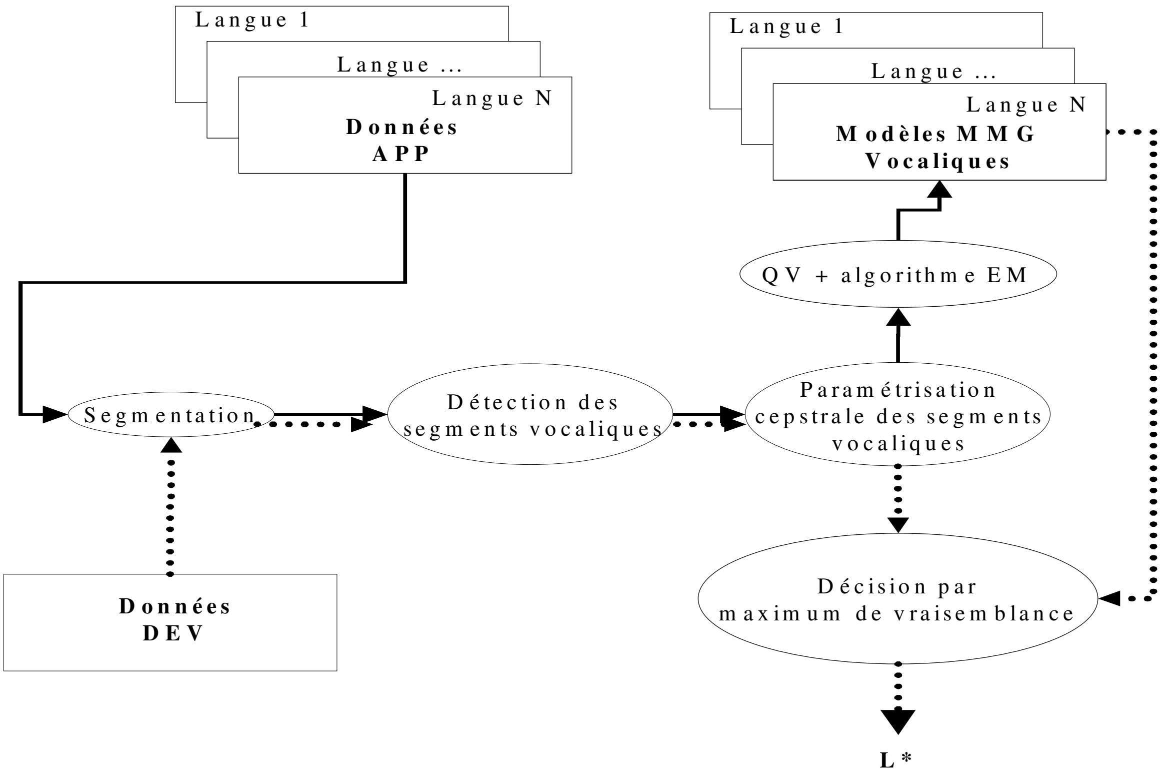

Figure 48 – Description du fonctionnement du système de discrimination des SV. Le processus d'apprentissage est spécifié en flèches pleines, et le processus d'identification en flèches pointillées.

3.2 Expériences en discrimination des SV

De nombreuses expériences ont été menées au cours de ces trois ans sur la modélisation des SV. Bien évidemment, certaines d'entre elles se sont avérées plus intéressantes que d'autres et elles sont donc privilégiées dans ce document. Il a cependant fallu opéré un choix de manière à ne pas ensevelir le lecteur sous une masse de matrices de confusion entre langues. D'une part, les résultats obtenus par approche du plus proche voisin avec les modèles obtenus par QV (c'est-à-dire sans appliquer d'algorithme EM) ne sont pas donnés. Ils sont en effet toujours inférieurs à ceux obtenus par calcul du maximum de vraisemblance avec les modèles MMG dérivés. D'autre part, seul le score d'identification global sur les cinq langues est donné pour chacune des expériences, et les matrices de confusion ne sont fournies que dans certains cas.

3.2.1 Influence du type de décision (WTA ou multigaussienne)

Nous avons vu au cours du chapitre précédent que le critère du maximum de vraisemblance peut être appliqué directement par rapport à l'ensemble du modèle multigaussien ou plus spécifiquement, sous l'hypothèse *Winner-Take-All*, à la composante gaussienne générant la vraisemblance maximale. Cette hypothèse a l'avantage de générer une partition de l'espace vocalique en classes discrètes qui correspond bien à la notion de système vocalique. Par contre, nous avons vu avec les expériences réalisées en modélisation sur les corpus EUROM et OGI MLTS qu'il est difficile de relier la réalisation acoustico-phonétique dans l'espace cepstral à une quelconque qualité vocalique au sens phonologique du terme. Il est donc nécessaire d'évaluer l'influence de l'hypothèse WTA sur le score d'identification des langues. Deux séries d'expériences ont été réalisées dans ce but avec à chaque fois un nombre Q de composantes gaussiennes fixé *a priori*.

Les deux ensembles de paramètres utilisés sont respectivement les 8 coefficients cepstraux augmentés de la durée du segment, et les 19 coefficients obtenus par la paramétrisation (8 coefficients cepstraux et leurs 8 dérivées, l'énergie et sa dérivée, ainsi que la durée du segment).

Décision \ Q	5	10	20	30	40	50	60	80
WTA	54	64	61	60	65	61	60	56
Multigaussienne	52	64	62	62	64	58	61	54

Tableau 21 – Influence du type de décision (hypothèse W *inner-Take-All* WTA ou vraisemblance multigaussienne). Le vecteur de paramètres est 8 MFCC + D.

Décision \ Q	5	10	20	30	40	50	60	80
WTA	58	66	70	65	66	60	60	56
Multigaussienne	58	69	66	65	64	58	57	57

Tableau 22 – Influence du type de décision (hypothèse WTA ou vraisemblance multigaussienne). Le vecteur de paramètres est 8 MFCC E + 8 ΔMFCC + ΔE + D.

Les résultats obtenus sur le corpus DEV_ST (Tableau 21 et Tableau 22) montrent que les différences entre les hypothèses WTA et multigaussienne sont faibles. De manière générale, l'hypothèse WTA tend à donner de meilleurs résultats en identification, confirmant ainsi les résultats déjà observés avec l'approche segmentale globale. Ces expériences justifient notre utilisation exclusive, par la suite, de l'approche WTA. Dans la suite de ce chapitre, *il sera donc fait indifféremment référence* au mot *classe*, *composante* ou *gaussienne* pour désigner la composante du MMG donnant la valeur de vraisemblance monogaussienne la plus grande pour une observation donnée.

Nous pouvons noter également que le score obtenu avec un modèle à 20 gaussiennes et le jeu complet de 19 paramètres est proche de 70 %. La matrice de confusion obtenue au cours de cette expérience est donnée à titre de référence (Tableau 23). Les taux d'identification obtenus vont de 100 % pour le français à 53 % pour l'espagnol. Le japonais présente un fort taux de confusion (46 % d'erreurs). On constate que l'espagnol a tendance à être confondu avec les modèles de toutes les autres langues, ce qui laisse supposer un modèle peu discriminant pour cette langue. De la même façon, un assez grand nombre de stimuli sont classés de manière erronée dans les modèles japonais ou espagnol.

Langue \ Modèle	FR	JA	KO	SP	VI
français	14				
japonais		7		4	2
coréen	1	2	13	1	
espagnol	3	1	2	8	1
vietnamien		2	1	2	10

Tableau 23 – Matrice de confusion – (DEV_ST / 19 coefficients / 20 gaussiennes).

3.2.2 Influence de la topologie du modèle et de l'espace d'observation

L'une des principales difficultés que l'on rencontre dans une tâche de modélisation consiste à trouver le nombre de composantes gaussiennes et l'espace des observations optimaux.

Nous avons donc réalisé une étude des scores d'identification obtenus en faisant varier le nombre Q de composantes des modèles de SV de 5 à 40, et cela pour différents ensembles de paramètres (Tableau 24). Les vecteurs de paramètres #1 et #2 sont qualifiés de jeux de référence car ils correspondent à l'espace cepstral tel que nous l'avons représenté jusqu'à présent (#1), auquel est ajouté la dimension temporelle D (#2). Les modèles #3 et #4 correspondent à ces deux espaces de référence augmentés du coefficient d'énergie E, tandis que les jeux de paramètres #5 à #8 sont construits par analogie dans le domaine cepstral dynamique (le paramètre de durée, intrinsèquement dynamique, est utilisé tel quel dans toutes les représentations, statiques ou dynamiques). L'ensemble de paramètres #9 correspond à l'espace cepstral global, statique et dynamique, et intégrant la durée des segments. Le jeu #10 est le plus complet puisqu'il prend en compte tous les paramètres calculés.

Paramètres \ Q		5	10	20	30	40	50	60	80
#1	**8 MFCC**	54	58	58	**60**	58			
#2	**8 MFCC + D**	54	64	61	60	**65**	61	60	56
#3	**8 MFCC + E**	56	58	**64**	57	53			
#4	**8 MFCC + E + D**	56	64	62	**65**	61			
#5	**8 ΔMFCC**	49	44	46	48	**50**			
#6	**8 ΔMFCC + D**	44	48	30	**56**	50			
#7	**8 ΔMFCC + ΔE**	42	48	45	48	**49**			
#8	**8 ΔMFCC + ΔE + D**	52	**54**	53	45	53			
#9	**8 MFCC + 8 ΔMFCC +D**	61	**68**	**68**	65	65	57	62	62
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	58	66	**70**	65	66	60	60	56

Tableau 24 – Pourcentage d'identification correcte sur le corpus DEV_ST dans l'hypothèse WTA. Le meilleur score obtenu pour chaque vecteur de paramètres est affiché en gras.

Pour le jeu de paramètres de référence #2 ou lorsque la dimension de l'espace des paramètres était importante (#9 et #10), nous avons poursuivi les expériences jusqu'à 80 gaussiennes par modèle. Ces expériences sont réalisées avec l'algorithme LBG-splitting classique, ce qui implique que le nombre Q est fixé *a priori* ; nous l'a vons choisi constant pour les modèles des cinq langues.

- **Influence de l'espace des observations**

Dans un premier temps, nous allons analyser le comportement des modèles en recherchant, pour chaque jeu de paramètres, la taille de modèle Q donnant le meilleur score.

Si l'on s'intéresse aux espaces de paramètres, on constate que les coefficients cepstraux statiques (modèles #1 à #4) donnent des scores d'identification de l'ordre de 60-65 % selon que l'on prenne en compte la durée des segments et leur énergie. Les modèles à base de coefficients dynamiques (modèles #5 à #8) atteignent quant à eux des scores proches de 50 %. Le jeu de paramètres complet permet d'atteindre de scores proches de 70 %.

De manière générale la prise en compte de la durée améliore toujours les résultats tout comme l'ajout du coefficient d'énergie aux paramètres cepstraux statiques (modèle #3). L'apport de la dérivée de l'énergie dans les modèles dynamiques s'avère moins marqué (modèle #7).

Le couplage de l'énergie et de la durée (#4, #8 et #10) se montre plus efficace que la prise en compte de la durée seule, uniquement lorsque les coefficients cepstraux statiques et dynamiques sont modélisés ensemble (modèle #10). A l'inverse, si l'on se place dans des espaces de modélisations homogènes (coefficients cepstraux statiques et dynamiques séparés), la prise en compte de la durée seule (#2 et #6) semble préférable.

- **Influence du nombre de composantes gaussiennes**

Si l'on s'intéresse à l'influence de la taille Q du modèle, il est plus difficile de formuler des conclusions. On peut cependant faire apparaître quelques tendances générales. On constate tout d'abord que le nombre de gaussiennes permettant d'obtenir les meilleurs résultats dépend des ensembles de paramètres mis en œuvre. Ce nombre a tendance à être plus élevé pour les espaces de paramètres homogènes, c'est-à-dire ceux où les coefficients cepstraux statiques et dynamiques ne sont pas fusionnés. Cela peut signifier que les ensembles non homogènes présentent une meilleure séparation des composantes gaussiennes en particulier grâce à la plus grande dimension de l'espace paramétrique, mais cela peut tout aussi bien indiquer que, pour ces modèles là (#9 et #10), on ne dispose pas d'une quantité de données suffisante pour apprendre un plus grand nombre de composantes. En effet, pour chacun des modèles, si l'espace paramétrique est de taille p, il est nécessaire d'apprendre $1 + p \times (p+1)$ coefficients par gaussienne du mélange (ce nombre se décompose en 1 coefficient α, p coefficients de moyennes μ et p^2 coefficients de variance σ puisque nous utilisons des matrices de covariances pleines).

Si N est le nombre de segments vocaliques du corpus APP, chaque paramètre est appris à partir d'un nombre moyen R de valeurs calculé comme suit :

$$R = \frac{N.p}{Q\left(p.(p+1)+1\right)} \approx \frac{N.p}{Q.p.(p+1)} \approx \frac{N}{Q.(p+1)} \qquad (32.)$$

Le cardinal de l'ensemble d'apprentissage étant de l'ordre de 10000, si l'on prend l'exemple du modèle à 40 gaussiennes correspondant aux paramètres #2, on obtient R = 25. Ce rapport descend à 12,5 si l'on conserve le même nombre de gaussiennes mais que l'on calcule un modèle des paramètres #10. Il est vraisemblable que dans ce second cas le modèle soit mal appris.

On retrouve là des résultats classiques en RAP, et il est donc nécessaire d'être prudent sur l'interprétation des différences de résultats obtenus en fonction du nombre de composantes gaussiennes. Par contre, le Tableau 24 permet de mettre en évidence le rôle fondamental des initialisations dans l'algorithme LBG. Les variations observées peuvent aussi provenir d'un mauvais partitionnement entraînant l'algorithme vers un piètre optimum local. Il semble bien que cette situation se produise pour le modèle #6 à 20 classes, qui donne un score d'identification très en deçà de ceux obtenus avec d'autres tailles.

- **Discussion**

Les expériences présentées ici permettent d'établir dans une certaine mesure l'importance des différents paramètres dans une tâche d'IAL. On constate ainsi que la durée des segments est un paramètre important de la modélisation et que l'énergie semble apporter moins. Une autre conclusion est que la détermination du nombre de gaussiennes optimal est ardue voire impossible, en partie à cause du caractère imprévisible quant à l'optimalité de la QV.

Un point essentiel que ne permettent cependant pas de déterminer ces expériences est de savoir s'il est plus efficace de modéliser chaque système vocalique avec le même nombre de gaussiennes ou non.

3.2.3 Modèle à nombre de gaussiennes variable : application de l'algorithme LBG-Rissanen

Au vu des expériences présentées au chapitre précédent, l'algorithme de LBG-Rissanen est prometteur en discrimination des systèmes vocaliques. En effet, si le critère d'information de Rissanen permet d'obtenir le nombre de gaussiennes optimal pour chaque langue, il est très probable que les résultats en identification des langues soient améliorés. Nous avons donc réalisé une modélisation MMG de chaque système vocalique avec un nombre de gaussiennes $Q(L)$ fixé par l'algorithme LBG-Rissanen pour chaque langue L.

De manière à clarifier les tableaux de résultats, nous avons sélectionné parmi les 10 jeux de paramètres présentés précédemment les plus significatifs. En fait, cela a consisté à écarter les modèles #3, #4, #7 et #8 en considérant que l'apport de la durée était supérieur à celui de l'énergie excepté avec le jeu de paramètres global #10.

- **Etude de la taille des modèles en fonction de la langue**

Le Tableau 25 indique pour chacune des langues L le nombre de gaussiennes $Q(L)$ obtenu, en fonction du jeu de paramètres utilisé. La première constatation est bien évidemment que ce nombre varie en fonction des langues. Pour les jeux de paramètres homogènes, le nombre moyen de composantes se situe entre 15 et 20 tandis que pour les modèles #9 et #10, ce nombre chute aux environs de 6. On constate donc une différence importante de comportement de l'algorithme en fonction de l'ensemble des paramètres.

On peut aussi observer que les langues latines – et particulièrement le français – ont un plus grand nombre de classes que les langues orientales composant notre corpus, et cela sans rapport évident avec le nombre de qualités vocaliques attendues au niveau phonologique (cf. le paragraphe 2.2.1 du premier chapitre). De manière générale, la prise en compte de la durée s'accompagne d'une baisse du nombre de classes.

Il est également intéressant de remarquer que dans le cas du jeu de paramètres #6, l'algorithme LBG-Rissanen a obtenu le même nombre de classes pour les cinq langues. Dans ce cas précis, aucune information discriminante supplémentaire n'est apportée par rapport aux modèles obtenus par LBG, mais le nombre de classes n'est pas fixé *a priori* , mais à partir des données.

Paramètres \ Modèle		FR	JA	KO	SP	VI
#1	**8 MFCC**	26	21	19	23	17
#2	**8 MFCC + D**	26	15	15	21	16
#5	**8 ΔMFCC**	23	19	18	22	17
#6	**8 ΔMFCC + D**	14	14	14	14	14
#9	**8 MFCC + 8 ΔMFCC +D**	9	5	6	8	4
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	7	4	5	7	4

Tableau 25 – Nombre de gaussiennes déterminés par algorithme LBG-Rissanen.

- **Etude du pouvoir discriminant des modèles obtenus**

Les modèles établis par LBG-Rissanen sont employés dans une tâche d'identification similaire à celle réalisée avec un nombre de gaussiennes fixe. Le Tableau 26 indique les scores d'identification obtenus en fonction du jeu de paramètres employé.

Paramètres		Taux d'identification	Paramètres		Taux d'identification
#1	**8 MFCC**	64	#6	**8 ΔMFCC + D**	56
#2	**8 MFCC + D**	68	#9	**8 MFCC + 8 ΔMFCC +D**	56
#5	**8 ΔMFCC**	54	#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	57

Tableau 26 – Pourcentage d'identification correcte sur le corpus DEV_ST. Le nombre de composantes gaussiennes est fixé par LBG-Rissanen.

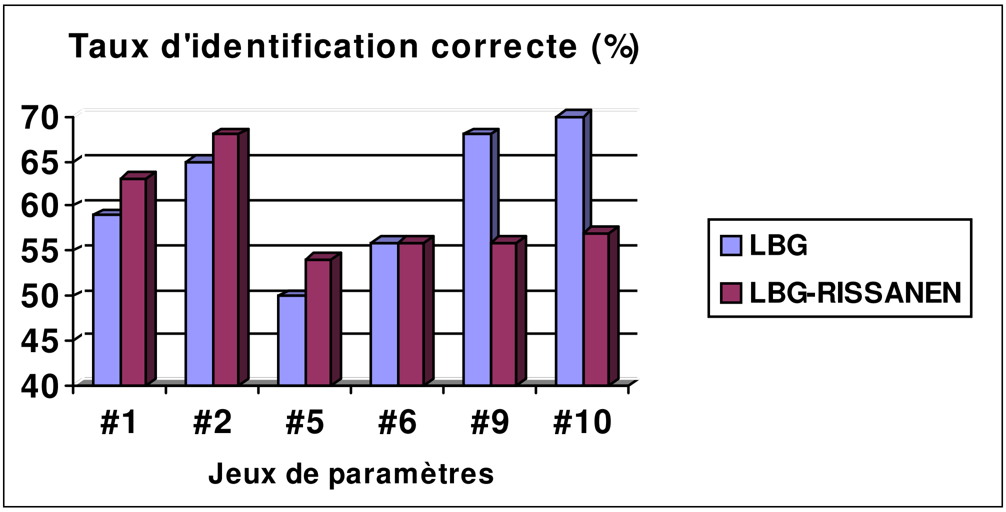

Figure 49 – Comparaison des taux d'identification correcte entre les modèles de taille constante pour les 5 langues (LBG) donnant le meilleur score et les modèles de taille variable (LBG-Rissanen).

La Figure 49 compare les taux atteints avec les modèles issus de l'algorithme LBG-Rissanen aux meilleurs scores obtenus avec les modèles à Q constant sur les ensembles de paramètres correspondants. L'utilisation de l'algorithme LBG-Rissanen s'accompagne d'une amélioration du score d'identification lorsque l'on est dans l'espace des coefficients cepstraux statiques. Les modèles issus des paramètres dynamiques voient aussi leur score progresser (#5) ou au pire rester inchangé (#6). Lorsque l'algorithme LBG-Rissanen détermine un nombre de composantes inférieur à 10 (#9 et #10), on observe une nette dégradation du score obtenu.

Le Tableau 27 présente la matrice de confusion obtenue avec les modèles issus de l'algorithme LBG-Rissanen et l'ensemble de paramètres #2. Les scores d'identification par langue varient de 47 % pour l'espagnol, qui est encore mal reconnu à 82 % pour le coréen.

Modèle / Langue	FR	JA	KO	SP	VI
français	11	1			2
japonais		8	2	2	1
coréen	1	2	14		
espagnol		4	3	7	1
vietnamien	2	2		1	10

Tableau 27 – Matrice de confusion – (DEV_ST / 8 MFCC + D / LBG-Rissanen).

La dégradation des résultats obtenus avec les modèles #9 et #10 résulte vraisemblablement du faible nombre de composantes gaussiennes déterminées par l'algorithme LBG-Rissanen. Il est possible que la dimension de l'espace des observations soit trop importante vis à vis du nombre de données d'apprentissage disponibles. La répartition des données dans un tel espace est alors parcellaire et l'augmentation du nombre de classes ne s'accompagne pas forcément d'une baisse sensible de la distorsion. Dans ces conditions, la complexité croissante du modèle n'est pas compensée par la baisse de distorsion lorsque l'on augmente sa taille. Le nombre de classes fixé par l'algorithme LBG-Rissanen est alors nettement sous-optimal. Puisque ce phénomène apparaît lorsque la dimension de l'espace d'observation est importante, nous avons décidé d'implanter une variante de l'algorithme LBG-Rissanen pour les modèles #9 et #10. Cette variante, nommée LBG-log-Rissanen repose sur la minimisation du critère $J(Q)$, défini par :

$$J(Q) = D_Q + 2p.Q.\log(\log N) \quad (33.)$$

Par rapport au critère de Rissanen $I(Q)$ (équation 30), ce critère pondère le nombre de paramètres indépendants du modèle par le logarithme du carré du logarithme du cardinal de l'ensemble d'apprentissage. L'effet principal de cette pondération est d'augmenter la taille du modèle optimal par rapport à celle donnée par le critère $I(Q)$.

- **Expériences avec l'algorithme LBG-log-Rissanen**

On constate (Tableau 28) que l'algorithme LBG-log-Rissanen détermine effectivement pour chaque langue un nombre de classes plus en accord avec les résultats obtenus par l'algorithme LBG que dans le cas de l'algorithme LBG-Rissanen.

Paramètres \ Modèle		FR	JA	KO	SP	VI
#9	**8 MFCC + 8 ΔMFCC +D**	29	24	23	22	21
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	32	21	21	22	18

Tableau 28 – Nombre de gaussiennes déterminés par algorithme LBG-log-Rissanen.

Paramètres \ Algorithme		LBG (20 classes)	LBG-Rissanen	LBG-log-Rissanen
#9	**8 MFCC + 8 ΔMFCC +D**	68	56	77
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	70	57	68

Tableau 29 – Pourcentage d'identification correcte sur le corpus DEV_ST.

Le pouvoir discriminant des modèles obtenus (Tableau 31) se révèle intéressant puisque, dans le cas de l'espace d'observation #9, on constate un gain de près de 10 % par rapport aux modèles ayant un nombre de composantes constant. Le fait que ce bénéfice ne se retrouve pas avec l'ensemble complet de paramètres (#10) peut être lié au fort déséquilibre entre les modèles générés : plusieurs fichiers sont en effet improprement reconnus comme étant du français, vraisemblablement à cause du grand nombre de gaussiennes du modèle par rapport aux autres langues.

- **Discussion**

Les résultats obtenus avec les algorithmes LBG-Rissanen ou LBG-log-Rissanen prouvent que l'utilisation d'un critère d'information adéquat est généralement supérieure au choix *a priori* du nombre de composantes gaussiennes des modèles. La prise en compte des coefficients cepstraux statiques et dynamiques, ainsi que de la durée des segments est particulièrement efficace (77 % d'identification correcte). Par contre, on constate également qu'un fort déséquilibre entre les tailles des modèles peut entraîner un biais durant la phase de classification.

3.3 Expériences complémentaires

Les expériences décrites au paragraphe précédant montrent que les informations prises en compte dans les modèles de systèmes vocaliques sont pertinentes en IAL puisque elles permettent d'atteindre un taux d'identification de 77 %. Ce taux est certes plus faible que celui atteint par le modèle global (85 %), mais il faut garder en mémoire que près des deux-tiers du signal (les sons non vocaliques) ne sont pas pris en compte. Ces expériences montrent également que l'algorithme LBG-Rissanen permet d'obtenir une information d'ordre topologique (un nombre de gaussiennes variable) adaptée à chaque langue lorsqu'on l'applique à un espace d'observation homogène. Cette information topologique obtenue par une approche de codage (critère de type *Minimum Description Length*) se révèle de pluspertinente en classification. Nous allons maintenant mettre en œuvre plusieurs algorithmes complémentaires dans le but d'améliorer les résultats déjà obtenus.

3.3.1 Procédures de normalisation

De nombreuses méthodes de normalisation des scores ou des locuteurs ont été développées, tant en identification des langues que du locuteur [Payan 93, Wegmann 96, Zhan 97]. Nous en avons implanté deux parmi les plus usitées. La première vise à éliminer un éventuel biais dû aux modèles, tandis que la seconde vise à sélectionner les segments vocaliques les plus discriminants parmi ceux présents dans le signal.

- **Prise en compte du biais d'apprentissage**

Dans [Zissman 96] l'auteur préconise d'opérer, dans le cadre d'une modélisation acoustico-phonétique et en phase d'identification, une normalisation des scores de vraisemblance par soustraction de la log-vraisemblance moyenne obtenue par chacun des modèles à l'apprentissage. Cette procédure vise à supprimer un éventuel biais qui altérerait les vraisemblances calculées en phase d'identification. Nous testons cette approche en utilisant les mêmes jeux de paramètres que précédemment, et, pour chacun d'eux, en se basant sur le modèle ayant donné les meilleurs résultats jusqu'à présent. Il s'agit dans le cas des paramètres #1, #2, #5 et #6 des modèles obtenus par LBG-

Rissanen, et pour les paramètres #9 et #10 des modèles obtenus par LBG avec 20 composantes ou par LBG-log-Rissanen.

Les expériences sont réalisées sur le corpus DEV_ST sous l'hypothèse WTA. Les résultats (Tableau 30) montrent que de manière générale, la prise en compte de la vraisemblance moyenne obtenue avec les données d'apprentissage ne permet pas d'améliorer la discrimination. Seul le modèle #9 à 20 classes présente une amélioration significative. Le score obtenu (74 %) reste cependant en deçà du taux d'identification correcte atteint avec les modèles calculés par l'algorithme LBG-log-Rissanen (77 %).

Il apparaît donc que, lorsqu'un biais existe entre des modèles de taille variable, il n'est pas efficace de le supprimer puisqu'il contient une information discriminante sur la topologie des SV modélisés. Par contre, lorsque le nombre de classes est commun à tous les modèles, la suppression du biais peut se révéler plus ou moins appropriée.

Paramètres		Modèle de référence	Taux d'identification (%)	
			Standard	Normalisé
#1	**8 MFCC**	LBG-Rissanen	64	60
#2	**8 MFCC + D**	LBG-Rissanen	68	64
#5	**8 ΔMFCC**	LBG-Rissanen	54	54
#6	**8 ΔMFCC + D**	LBG-Rissanen	56	60
#9	**8 MFCC + 8 ΔMFCC +D**	LBG (20 classes)	68	74
#9	**8 MFCC + 8 ΔMFCC +D**	LBG-log-Rissanen	77	77
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG (20 classes)	70	70
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG-log-Rissanen	68	69

Tableau 30 – Pourcentages d'identification correcte sur le corpus DEV_ST avec et sans normalisation par soustraction du biais calculé à l'apprentissage.

- **Normalisation des vraisemblances par calcul d'une fonction discriminante**

Les segments vocaliques extraits d'un enregistrement ne sont pas tous porteurs de la même information caractéristique sur la langue parlée. En effet, il est vraisemblable que dans les cas des voyelles extrêmes comme /i/, présentes dans plus de 90 % des langues du monde, la variabilité inter-locuteur soit comparable à la variabilité inter-langue. Il paraît donc raisonnable de vouloir prendre en compte uniquement les segments vocaliques les plus discriminants de l'énoncé.

Le procédé mis en œuvre est basé sur une normalisation du score obtenu pour chaque segment vocalique dans chacun des modèles en appliquant la fonction discriminante définie dans [Fukunaga 90] et utilisée en identification du locuteur dans

[Besacier 98]. Soit o_k le vecteur d'observation correspondant à la $k^{\text{ième}}$ voyelle de la phrase à identifier. Si $\Pr_V(o_k| L_i)$ est sa vraisemblance calculée par rapport au modèle de SV de la langue L_i, le score associé à ce segment vocalique est donné par :

$$h_{L_i}(o_k) = \log \Pr_V\left(o_k \middle| L_i\right) - \max_{Lj \neq Li}\left[\log \Pr_V\left(o_k \middle| L_j\right)\right] \qquad (34.)$$

et le score vocalique total de l'énoncé dans le modèle de la langue L_i est alors :

$$H_{L_i}(O) = \sum_{k=1}^{N} h_{L_i}(o_k) \qquad (35.)$$

Les résultats (Tableau 31) montrent que, là encore, on observe plutôt une dégradation des performances lorsque le nombre de composantes diffère d'une langue à l'autre et une légère amélioration lorsque qu'il s'agit de modèles possédant le même nombre de gaussiennes.

Paramètres		Modèle de référence	Taux d'identification (%)	
			Standard	Discriminant
#1	**8 MFCC**	LBG-Rissanen	64	61
#2	**8 MFCC + D**	LBG-Rissanen	68	66
#5	**8 ΔMFCC**	LBG-Rissanen	54	50
#6	**8 ΔMFCC + D**	LBG-Rissanen	56	60
#9	**8 MFCC + 8 ΔMFCC +D**	LBG (20 classes)	68	73
#9	**8 MFCC + 8 ΔMFCC +D**	LBG-log-Rissanen	77	77
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG (20 classes)	70	70
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG-log-Rissanen	68	65

Tableau 31 – Pourcentages d'identification correcte sur le corpus DEV_ST avec et sans normalisation des scores par fonction discriminante.

Ces tendances, déjà observées avec la normalisation par rapport aux données d'apprentissage confirment que les modèles obtenus par LBG-Rissanen ou LBG-log-Rissanen prennent implicitement en compte des informations topologiques que les procédures de normalisation font disparaître (on retrouve alors des scores proches de ceux obtenus avec un nombre fixe de gaussiennes).

- **Sélection des segments vocaliques les plus discriminants par élagage**

Une autre méthode d'analyse visant à augmenter la robustesse de l'identification en retenant uniquement les segments vocaliques les plus représentatifs peut être dérivée de la technique d'élagage temporel proposée dans [Besacier 98]. Cette technique consiste à conserver, pour chacun des modèles, les trames obtenant les meilleurs scores et donc à ne pas prendre en compte les autres dans le calcul de la vraisemblance. L'objectif est en fait d'éliminer les observations ponctuelles anormales qui dégradent fortement la vraisemblance de toute la phrase. Nous avons appliqué cette méthode à la discrimination des systèmes vocaliques en testant différents niveaux d'élagage (Tableau 32) tout en conservant la normalisation par la fonction discriminante décrite dans l'équation 34.

Paramètres \ % d'élagage		Modèle de référence	0 %	5 %	10 %	20 %	30 %
#1	**8 MFCC**	LBG-Rissanen	61	61	64	60	58
#2	**8 MFCC + D**	LBG-Rissanen	66	69	69	69	65
#5	**8 ΔMFCC**	LBG-Rissanen	50	44	42	44	44
#6	**8 ΔMFCC + D**	LBG-Rissanen	60	54	52	50	50
#9	**8 MFCC + 8 ΔMFCC +D**	LBG (20 classes)	73	72	72	72	73
#9	**8 MFCC + 8 ΔMFCC +D**	LBG-log-Rissanen	77	77	78	76	76
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG (20 classes)	70	70	70	72	70
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	LBG-log-Rissanen	65	72	69	68	68

Tableau 32 – Pourcentages d'identification correcte sur le corpus DEV_ST avec normalisation des scores par fonction discriminante et élagage des segments les moins discriminants.

Les résultats confirment que dans la plupart des cas, éliminer environ 10 % des segments vocaliques les moins vraisemblables amène une légère amélioration de la robustesse des modèles.

Les expériences ont été refaites en **supprimant la normalisation discriminante** (de manière à garder le maximum d'information pour les modèles de taille variable) et en conservant uniquement la procédure d'élagage. Si les résultats obtenus sont similaires à ceux rapportés pour la quasi-totalité des modèles, pour le modèle #9 obtenu par l'algorithme LBG, ils sont meilleurs (Tableau 33) puisqu'ils atteignent un taux d'identification correcte de 78 %. Les deux modèles #9 possédant un nombre de composantes fixe (algorithme LBG) ou variable (LBG-log-Rissanen) atteignent donc des performances similaires, mais, dans le cas de l'algorithme LBG, il est nécessaire de procéder à un élagage important des segments vocaliques pris en compte, ce qui peut se révéler critique lorsque la longueur des enregistrements de test diminue.

Elagage (%)	0	5	10	20	25	30	35	40	45	50
Score d'identification (%)	68	72	73	77	**78**	**78**	77	74	72	70

Tableau 33 – Influence de la proportion d'élagage sur le score obtenu en identification avec le modèle #9 à 20 classes **sans** normalisation des scores.

- **Prise en compte des segments vocaliques discriminants et du biais d'apprentissage**

Les conclusions des expériences précédentes montrent qu'il est difficile d'améliorer les scores obtenus dans le cas de modèles de taille variable, mais que ce processus peut se révéler efficace lorsque les différents modèles possèdent un nombre de composantes identique.

La plupart des expériences montrent également un apport inexistant des paramètres d'énergie par rapport au modèle prenant en compte uniquement les coefficients cepstraux, leurs dérivées et la durée du segment.

Nous avons donc implanté pour le modèle #9 à 20 gaussiennes un algorithme de double normalisation, à la fois par élagage des segments les moins vraisemblables et par soustraction du biais calculé sur les données d'apprentissage puisque les scores obtenus en identification par cette modélisation sont améliorés par les deux techniques utilisées séparément (Tableau 34).

Description du modèle	Standard	Elagage de :				
		0 %	10 %	20 %	30 %	40 %
Score d'identification (%)	68	74	77	**78**	76	72

Tableau 34 – Influence de la proportion d'élagage utilisé en conjonction avec la suppression de biais sur le modèle #9 à 20 classes **sans** normalisation des scores.

Comme on peut le constater, les améliorations portées par la prise en compte du biais et l'utilisation d'un système d'élagage ne se conjuguent pas puisque le score obtenu reste de 78 %. On peut tout au plus noter qu'il est atteint pour un élagage moins important que sans la prise en compte du biais (20 % au lieu de 30 %) ce qui peut être un avantage lorsque la durée des stimuli de test est plus faible. Cet élagage reste cependant supérieur à celui requis dans le cas du modèle #9 obtenu par LBG-log-Rissanen (10 %).

- **Discussion**

Les expériences menées jusqu'à présent en discrimination des SV nous permettent de formuler plusieurs remarques. Tout d'abord, les différents résultats obtenus tendent à montrer que les paramètres d'énergie apportent peu ou pas d'information supplémentaire par rapport aux coefficients cepstraux. La durée des segments, par contre, s'avère un paramètre très important des différents modèles. Il est intéressant de noter par ailleurs que les algorithmes LBG-Rissanen et LBG-log-Rissanen permettent d'intégrer aux modèles une information très discriminante par le biais de leur nombre de composantes. De plus, il s'avère que certaines techniques de normalisation ou d'élagage sont partiellement efficaces sur les modèles possédant un nombre de classes fixe. Il apparaît intéressant de fusionner les décisions prises avec d'une part des modèles de taille variable (information topologique) et des modèles de taille fixe, plus aptes à tirer profit des algorithmes discriminants (normalisation et élagage) étudiés.

3.3.2 Fusion de décisions

Plusieurs indices laissent supposer que le nombre de composantes fixé par l'algorithme LBG-Rissanen est un paramètre discriminant important. Nous avons donc étudié la possibilité de fusionner les décisions prises par les modèles appris par cette méthode sur des espace homogènes (statique ou dynamique) avec les décisions prises par le meilleur classificateur dont on dispose, à savoir celui basé sur des modèles MMG à 20 gaussiennes, appris avec un vecteur de 17 paramètres (8 MFCC + 8 ΔMFCC +D) et optimisé avec les méthodes de normalisation présentées ci-dessus (ce modèle est appelé dans la suite le modèle #9 amélioré). En réalisant une fusion statistique, le score de vraisemblance pris en compte pour une phrase à identifier est sa vraisemblance conjointe par rapport aux deux modèles fusionnés. En faisant l'hypothèse forte (et fausse) que les deux modèles sont indépendants, cette vraisemblance s'exprime comme le produit des deux vraisemblances obtenues par rapport à chacun des modèles.

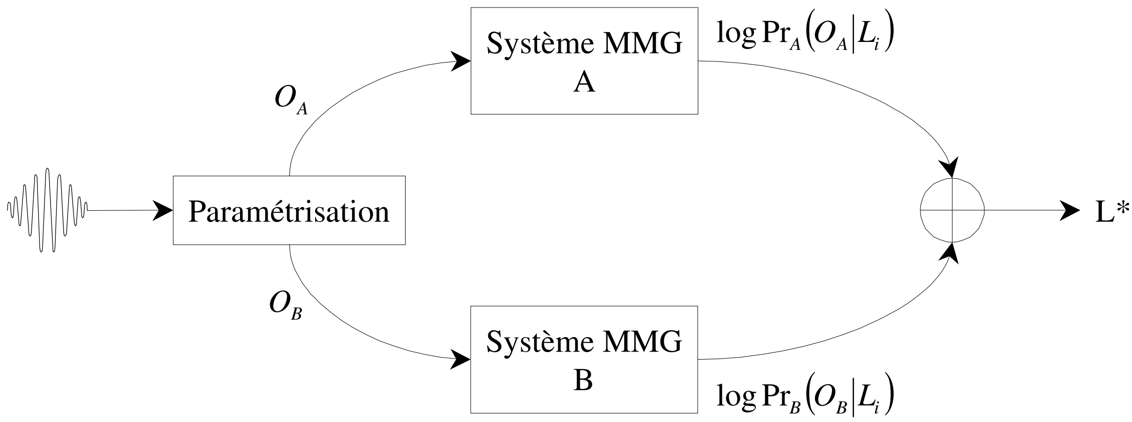

Figure 50 – Schéma de la fusion statistique de décisions issues de deux modèles A et B.

Au cours d'une première expérience, nous cherchons à savoir si la fusion d'un modèle statique (#1 ou #2) et d'un modèle dynamique (#5 ou #6) appris tous deux par LBG-Rissanen amène un score comparable à celui obtenu avec les mêmes coefficients dans un modèle global (#9). Les résultats obtenus (Tableau 35) indiquent que ce n'est pas le cas. En effet, les scores obtenus par fusion se révèlent dégradés par rapport à ceux atteints par les modèles statiques seuls (respectivement 64 % et 68 % d'identification correcte). Ils restent donc nettement inférieurs à ceux obtenus dans l'espace global #9 (77 %) par algorithme LBG-log-Rissanen.

Modèle dynamique \ Modèle statique	8 MFCC	8 MFCC + D
8 ΔMFCC	62	65
8 ΔMFCC + D	64	66

Tableau 35 – Scores d'identification obtenus en fusionnant les modèles LBG-Rissanen statiques et dynamiques.

Une autre expérience est réalisée en fusionnant les scores obtenus avec le meilleur modèle appris par LBG-Rissanen (modèle #2) et le modèle #9 amélioré. De manière à contrôler l'apport du paramètre « nombre de composantes » déterminé par le LBG-Rissanen, nous testons également la fusion du modèle #9 amélioré avec chacun des modèles #2 appris avec un nombre fixe de gaussiennes.

Modèle #2	Sans	LBG Rissanen	LBG 20 classes	LBG 30 classes	LBG 40 classes	LBG 50 classes	LBG 60 classes
#9 amélioré	78	**80**	70	72	77	74	74

Tableau 36 – Scores d'identification obtenus en fusionnant le modèle #9 amélioré avec différents modèles #2.

L'expérience (Tableau 36) montre que seule la fusion des décisions avec le modèle #2 obtenu par algorithme LBG-Rissanen est intéressante. Dans tous les autres cas (nombre de composantes fixé *a priori*) les résultats sont dégradés par la fusion. De manière à confirmer l'apport du paramètre topologique sur la taille des modèles, nous renouvelons l'expérience précédente en n'utilisant plus le modèle #9 amélioré possédant un nombre de gaussiennes fixe, mais le modèle #9 obtenu par LBG-log-Rissanen. Cela revient à fusionner un modèle #9 où l'information topologique liée à la taille du modèle est déjà prise en compte avec les modèles issus de l'espace d'observation #2. Les résultats (Tableau 37) confirment que dans ce cas là, la fusion des décisions n'apporte pas d'information discriminante supplémentaire.

Modèle #2	Sans	LBG Rissanen	LBG 20 classes	LBG 30 classes	LBG 40 classes	LBG 50 classes	LBG 60 classes
#9 LBG-log-Rissanen	77	76	75	77	77	75	73

Tableau 37 – Scores d'identification obtenus en fusionnant le modèle #9 obtenu par LBG-log-Rissanen avec différents modèles #2.

- **Matrice de confusion**

Dans le cas de la fusion du modèle #9 amélioré avec le modèle #2 LBG-Rissanen, les scores d'identification par langue calculés à partir de la matrice de confusion (Tableau 38) s'étendent de 67 % d'identification correcte pour l'espagnol à 94 % pour le coréen et à 100 % pour le français. Le japonais est confondu avec l'espagnol dans plus de 20 % des cas (rappelons que ces deux langues ont des inventaires vocaliques phonologiques proches). Le vietnamien est quant à lui confondu une fois avec chacune des autres langues.

Langue \ Modèle	FR	JA	KO	SP	VI
français	14				
japonais		9		3	1
coréen	1		16		
espagnol		2	2	10	1
vietnamien	1	1	1	1	11

Tableau 38 – Matrice de confusion obtenue sur le corpus DEV_ST par fusion des modèles #9 amélioré et #2 LBG-Rissanen.

4 DISCUSSION

Nous avons présenté au cours de ce chapitre les résultats les plus intéressants obtenus en IAL par discrimination des SV. D'autres expériences ont été menées, en particulier en procédant à une analyse en composantes principales des données avant de procéder à la modélisation mais aucune amélioration n'a été obtenue. Nous avons également testé l'influence d'une réduction du nombre de coefficients des modèles en évaluant les modèles avec une matrice de covariance diagonale et non pleine pour chacune des lois gaussiennes les constituant ; nous avons alors constaté une dégradation des résultats.

Si l'on dresse un bilan des expériences menées (Tableau 39), on constate que le meilleur taux d'identification correct est obtenu sur le corpus DEV_ST en prenant en compte les paramètres cepstraux statiques et dynamiques, la durée des segments vocaliques ainsi que les paramètres topologiques issus de la modélisation de l'espace cepstral par LBG-Rissanen ou LBG-log-Rissanen.

Le meilleur score (80 % d'identification correcte) est obtenu en fusionnant les résultats calculés par deux modèles :

- ✓ Le modèle basé sur les 8 paramètres MFCC et la durée des segments, initialisé par algorithme LBG-Rissanen,
- ✓ le modèle basé sur les 8 coefficients MFCC, leurs 8 dérivées et la durée des segments, initialisé par algorithme LBG avec 20 classes. Lors de la phase d'identification, une procédure d'élagage de 20 % des segments testés et une soustraction du biais calculé à l'apprentissage sont appliquées.

Modèle	#2 40 gaussiennes	#2 LBG-Rissanen	#9 20 classes	#9 LBG-log-Rissanen	#9 amélioré 20 classes	Fusion #9 amélioré - #2 LBG-Rissanen
DEV_10	52	51	60	61	59	62
DEV_ST	65	68	68	77	78	80
DEV_ALL	71	72	76	83	85	82

Tableau 39 – Bilan des expérimentations en discrimination des SV sur les corpus masculins.

Lorsque l'on se limite à des énoncés d'une durée moyenne inférieure à 10 secondes (DEV_10), le taux d'identification est d'environ 60 % alors que la durée des voyelles prises en compte ne dépasse pas 3 secondes en moyenne. Lorsqu'au contraire, on dispose d'une durée de signal plus longue (DEV_ALL), le taux d'identification atteint avec les mêmes modèles des valeurs supérieures à 80 %. Il ressort de ces expérimentations une différence de performances significative entre les modèles « simples » (#2 LBG et LBG-Rissanen et #9 LBG) et les modèles améliorés par normalisation, élagage ou fusion d'information lorsque la durée des stimuli est suffisante (corpus DEV_ST et DEV_ALL). Les différences entre ces différentes techniques d'amélioration sont par contre peu ou pas significatives.

Ce score est proche, quoique inférieur, au score obtenu par le modèle segmental global de référence (85 %). La modélisation des systèmes vocaliques permet donc d'extraire une information particulièrement discriminante du signal. Il apparaît donc indispensable de compléter le modèle de SV par un modèle consonantique, de manière à comparer de manière rigoureuse les résultats obtenus par les approches globale et différenciée.

Chapitre 4

VERS UN SYSTEME D'IAL PHONETIQUE COMPLET

Les résultats obtenus en discrimination des systèmes vocaliques appliquée à l'IAL sont probants : on obtient un taux d'identification correct de l'ordre de 80 % avec les locuteurs masculins de cinq langues. Cette étude lève donc, au moins partiellement, l'inconnue présentée en introduction du chapitre précédent sur la possibilité de modéliser de manière efficace les systèmes vocaliques des langues traitées à partir d'enregistrements acoustiques. Bien évidemment, les réponses apportées s'accompagnent de nouvelles questions portant principalement sur la possibilité de tirer profit de la modélisation des systèmes vocaliques au sein d'un système d'IAL plus complet. Ce terme de « système complet » fait référence à l'intégration d'autres sources d'information classiques, en particulier à la modélisation acoustico-phonétique des consonnes. Ce chapitre est consacré à des expériences préliminaires menées selon cet axe en conservant le cadre de l'approche non supervisée. Nous appliquons donc la modélisation phonétique par MMG à l'ensemble des consonnes de chaque langue de manière à étudier l'apport d'une modélisation voyelles / consonnes différenciée par rapport à la modélisation de référence présentée au début du chapitre précédent. Un paragraphe récapitulatif détaille ensuite les résultats obtenus avec le système obtenu dans des tâches de discrimination du français contre chacune des autres langues et dans des tâches d'identification d'une langue parmi quatre.

1 DISCRIMINATION DIFFERENCIEE DES SYSTEMES VOCALIQUES ET CONSONANTIQUES

1.1 Description du système d'identification phonétique

Les algorithmes de détection de segments vocaliques présentés au premier chapitre fournissent un étiquetage du signal en segments vocaliques, consonantiques et en silences. Ce paragraphe est consacré à la modélisation par MMG des segments consonantiques ainsi qu'à la fusion des vraisemblances consonantiques et vocaliques en IAL. Le système mis en place est celui qui a été proposé en fin de seconde partie (Figure 19). L'étiquetage voyelle / consonne n'agit plus alors comme un filtre (i. e. on modélise les segments vocaliques et on ne prend pas en compte les segments consonantiques) mais comme un aiguillage (un modèle consonantique et un modèle vocalique sont appris pour chaque langue). Rappelons simplement que, dans ce cadre, la langue la plus vraisemblable L^* correspondant à une suite d'observations $O = \{o_1, o_2, \ldots, o_T\}$ est donnée par la relation :

$$L^* = \arg\max_{1 \le i \le N_L} \left(\prod_{\phi k = V} \Pr_V(o_k | L_i) . \prod_{\phi k = C} \Pr_C(o_k | L_i) \right) \quad (36.)$$

où l'étiquette Φ_k est fournie par le module de détection vocalique.

- **Paramétrisation des consonnes**

La nature de la détection des voyelles (localisation d'une trame de signal à l'intérieur d'un segment) permet d'appliquer l'analyse cepstrale de manière précise à un instant donné pour les segments vocaliques. Ne disposant pas de la même information pour les segments consonantiques, on est amené à la réaliser au milieu du segment, tout comme dans le cas du système segmental global. Mis à part cette différence, les traitements appliqués sont identiques à ceux mis au point pour la modélisation des SV. En particulier, une normalisation cepstrale est appliquée par soustraction, à tous les segments d'un appel, du vecteur cepstral moyen issu de cet appel.

- **Introduction aux expérimentations**

Les expériences sont présentées ci-après en deux parties. Dans un premier temps, nous évaluons, comme nous l'avons fait pour les voyelles, l'influence des divers paramètres sur la discrimination obtenue par modélisation des systèmes consonantiques, puis nous étudions les résultats obtenus par fusion des modélisations consonantiques et vocaliques.

Les conclusions formulées sur la modélisation des SV nous ont amené à ne conserver que trois jeux de paramètres sur les dix initialement testés. Ces trois espaces d'observation sont, en conservant la terminologie définie au chapitre précédent, les ensembles #2 (8 MFCC + D), #9 (8 MFCC + 8 ΔMFCC +D) et #10 (8 MFCC + E + 8 ΔMFCC + ΔE + D).

1.2 Expériences en discrimination des systèmes consonantiques

1.2.1 Influence de la topologie du modèle et de l'espace d'observation

Tout comme dans le cas des modèles vocaliques, nous avons étudié les taux d'identification correcte obtenus avec chacun des espaces d'observation en faisant varier le nombre de composantes gaussiennes du mélange (Tableau 24).

	Paramètres \ Q	20	30	40	50	60	80
#2	**8 MFCC + D**	69	**73**	69	62	68	65
#9	**8 MFCC + 8 ΔMFCC +D**	69	**76**	66	61	65	57
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	**76**	66	72	**76**	64	54

Tableau 40 – Pourcentage d'identification correcte sur le corpus DEV_ST dans l'hypothèse WTA.

On constate tout d'abord que les meilleurs scores obtenus avec chacun des jeux de paramètres sont supérieurs à ceux que l'on obtenait avec les modèles de SV. On notera également que les coefficients d'énergie ne semblent pas apporter d'information supplémentaire par rapport aux coefficients cepstraux.

Sans préjuger du pouvoir discriminant intrinsèque des systèmes vocaliques et consonantiques, on peut constater que, de manière générale, la durée totale des segments consonantiques est très nettement supérieure à la durée des segments vocaliques pris en compte (Tableau 41). En tenant compte de la durée de signal utilisée pour effectuer l'identification, il est plus juste de comparer les taux obtenus en identification consonantique sur 45 secondes avec ceux issus de la discrimination vocalique sur les corpus de deux minutes. Dans le Tableau 41 figurent les durées consonantiques et vocaliques moyennes calculées sur le corpus APP_ST. La durée consonantique (resp. vocalique) est estimée en calculant la somme des durées de chaque segment consonantique (resp. vocalique) d'un enregistrement. Les taux sont alors proches (de l'ordre de 70 à 75 % pour les consonnes sur 45 secondes, et de l'ordre de 80 à 85 % pour les voyelles sur 2 minutes).

Langue	français	japonais	coréen	espagnol	vietnamien
Durée vocalique	11	9	9	10	8
Durée consonantique	27	25	21	28	23

Tableau 41 – Valeur moyenne (en secondes) des durées consonantiques et vocaliques sur les fichiers d'apprentissage de 45 secondes (APP_ST).

1.2.2 Modèles à nombre de gaussiennes variable

L'application de l'algorithme LBG-Rissanen (cas du modèle #2) ou LBG-log-Rissanen (cas des modèles #9 et #10) permet, tout comme dans le cadre de la modélisation des SV, de déterminer un nombre de composantes gaussiennes adapté à chaque langue. Cependant, un nouveau phénomène biaise de manière importante le nombre de classes obtenu par ces algorithmes. En effet, le nombre de segments consonantiques – et donc d'observations – obtenus varie dans une proportion importante en fonction de la langue. Bien que le critère ait pour but de compenser ces différences, on obtient un nombre de classes pouvant aller du simple au double (Tableau 42). Ce biais se retrouve durant la phase de classification et si un modèle présente un nombre de composantes bien plus important que les autres, il aura tendance à être préférentiellement reconnu.

Langue	FR	JA	KO	SP	VI
Nombre de segments consonantiques	26 869	18 108	17 242	24 366	14 947
Nombre de classes fixé par LBG-log-Rissanen (modèle #10)	31	26	24	46	20

Tableau 42 – Nombre de segments consonantiques du corpus d'apprentissage et nombre de classes fixé par LBG-log-Rissanen dans l'espace d'observation #10.

Cela nous a amené, en phase de détermination du nombre de composantes des modèles, à pratiquer une uniformisation assez brutale de la taille de l'ensemble d'apprentissage de chaque langue puisque nous avons retenu 14000 segments consonantiques par tirage au sort. Il est évident que cette technique peut être améliorée [45], mais elle permet d'uniformiser la taille des dictionnaires (Tableau 25). Il est important de noter que lors de la ré-estimation des modèles MMG, la totalité des données disponibles est de nouveau employée (et non plus les 14000 segments consonantiques tirés au sort).

Paramètres	**Modèle**	**FR**	**JA**	**KO**	**SP**	**VI**	***Méthode***
#2	**8 MFCC + D**	18	16	22	18	15	LBG-Rissanen
#9	**8 MFCC + 8 ΔMFCC +D**	22	23	24	25	27	LBG-log-Rissanen
#10	**8 MFCC + E + 8 ΔMFCC + ΔE + D**	28	28	23	18	19	LBG-log-Rissanen

Tableau 43 – Nombre de gaussiennes déterminé par algorithme LBG-Rissanen et LBG-log-Rissanen avec des corpus équilibrés.

Les résultats obtenus en IAL (Tableau 44) sont comparables à ceux précédemment obtenus avec des modèles de taille fixe (Figure 49) même si l'on observe une légère dégradation. L'absence d'amélioration constatée peut provenir de plusieurs causes. Tout d'abord, l'uniformisation des tailles des dictionnaires (dans le cas du français, on ne prend pas en compte près de 50 % des segments consonantiques disponibles) altère la taille optimale déterminée par les algorithmes de type LBG-Rissanen de manière importante. De plus, il est possible que la modélisation globale de consonnes très différentes au plan spectral et cepstral (fricatives non voisées, consonnes liquides…) se révèle inadaptée à la détermination d'un dictionnaire de taille optimale.

[45] en effectuant par exemple une partition de l'espace avec un nombre de classes constant et en éliminant des observations dans chaque classe en respectant la répartition des données.

Paramètres		Taux d'identification	Méthode
#2	8 MFCC + D	66	LBG-Rissanen
#9	8 MFCC + 8 ΔMFCC +D	72	LBG-log-Rissanen
#10	8 MFCC + E + 8 ΔMFCC + ΔE + D	73	LBG-log-Rissanen

Tableau 44 – Pourcentage d'identification correcte sur le corpus DEV_ST pour les modèles de taille variable.

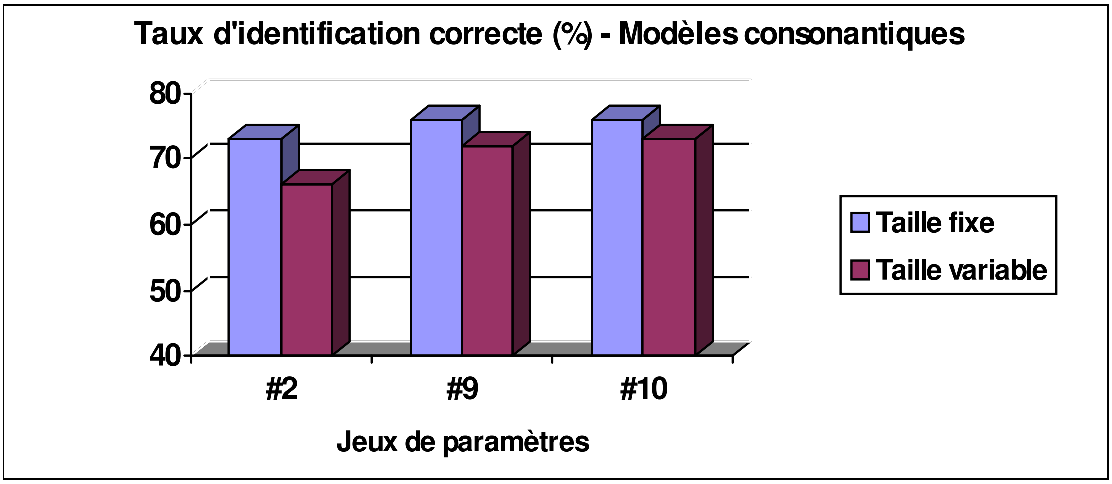

Figure 51 – Comparaison des taux d'identification correcte entre les modèles de taille constante (algorithme LBG) donnant le meilleur score et les modèles de taille variable (algorithmes LBG-Rissanen ou LBG-log-Rissanen).

1.2.3 Procédure d'élagage

Parmi les différentes techniques de normalisation et de discrimination présentées au chapitre précédent, seule la technique d'élagage apportait une réelle amélioration des résultats obtenus avec les modèles de SV. Nous avons donc étudié l'impact de l'élagage sur la discrimination obtenue avec les systèmes consonantiques.

Elagage (%)	0	5	10	15	20	25	30	35	40	45	50
#9 – LBG 30 classes	76	73	72	73	73	76	77	78	78	77	76
#9 – LBG log-Rissanen	72	72	72	72	69	70	70	70	70	70	69
#10 – LBG 20 classes	76	73	72	74	73	72	73	72	69	69	69
#10 – LBG log-Rissanen	73	72	73	74	73	74	72	73	73	74	72

Tableau 45 – Influence de la proportion d'élagage sur le score obtenu en identification avec les modèles #9 à 30 classes et #10 log-Rissanen.

Les résultats (Tableau 33) montrent que la discrimination augmente uniquement dans le cas des modèles ayant 30 composantes dans l'espace de paramètres #9. Dans tous les autres cas, on observe une dégradation des performances.

1.2.4 Discussion

En gardant à l'esprit que cette thèse porte principalement sur la modélisation des systèmes vocaliques, nous n'avons pas mené autant d'expériences sur les modèles consonantiques. Les résultats obtenus en IAL par modélisation des segments consonantiques sont comparables à ceux obtenus par modélisation des segments vocaliques alors que la durée de signal prise en compte est doublée. Cela signifie d'une part que la modélisation obtenue n'est sans doute pas la meilleure, et d'autre part que la prise en compte de toutes les consonnes dans un espace global n'est probablement pas l'approche la plus discriminante. Nous aborderons d'ailleurs le thème de la modélisation différenciée des consonnes dans les perspectives clôturant ce manuscrit. Les expériences rapportées dans ce paragraphe ne constituent en somme qu'une approche du problème de la modélisation consonantique, et elles visent uniquement à évaluer par la suite si la fusion des modèles vocaliques et consonantiques obtenus peut se révéler efficace ou non.

1.3 Discrimination des systèmes vocaliques et consonantiques

A partir des modèles les plus discriminants établis pour les systèmes vocaliques et consonantiques, nous étudions l'impact de la prise en compte de la vraisemblance conjointe des segments vocaliques et consonantiques dans une tâche d'IAL. Les modèles employés ayant abouti aux meilleurs résultats dans les expériences séparées sont rappelés dans le Tableau 46 .

Paramètres	Code du modèle	Normalisation	Elagage
Consonnes - #9	LBG-log-Rissanen	-	-
Consonnes - #9	LBG (30 Classes)	-	-
Consonnes - #9	LBG (30 Classes) élagage	-	40 %
Consonnes - #10	LBG-log-Rissanen	-	-
Consonnes - #10	LBG (20 Classes)	-	-
Voyelles - #9	LBG-log-Rissanen	-	-
Voyelles - #9	LBG (20 Classes) amélioré	Soustraction du biais	20 %
Voyelles - #9	LBG-log-Rissanen élagage	-	10 %

Tableau 46 – Description des modèles employés en discrimination des systèmes consonantiques et vocaliques.

Les résultats sont synthétisés dans le Tableau 47 . Il s'agit d'un tableau à double entrée, où les colonnes correspondent au modèle consonantique dans lequel la vraisemblance des segments consonantiques est calculée, et les lignes correspondent aux modèles vocaliques employés. Pour chacun des modèles vocaliques (resp. consonantiques), le taux d'identification correcte obtenu durant les expériences de discrimination vocalique (resp. consonantique) est rappelé dans la case intitulé *Score seul*. Les scores obtenus en tenant compte des vraisemblances vocaliques et consonantiques sont affichés à l'intersection des lignes et des colonnes correspondant aux modèles décrits.

<table>
<tr><td></td><td colspan="2">Modèle</td><td colspan="3">Consonantique #9</td><td colspan="2">Consonantique #10</td></tr>
<tr><td rowspan="2">Modèle</td><td colspan="2">Algorithme</td><td>LBG-log-Rissanen</td><td>LBG (30 classes)</td><td>LBG (30 classes) élagage</td><td>LBG-log-Rissanen</td><td>LBG (20 classes)</td></tr>
<tr><td></td><td>Score seul</td><td>72</td><td>76</td><td>78</td><td>73</td><td>76</td></tr>
<tr><td rowspan="3">Vocalique #9</td><td>LBG-log-Rissanen</td><td>77</td><td>81</td><td>84</td><td>84</td><td>81</td><td>78</td></tr>
<tr><td>LBG (20 classes) amélioré</td><td>78</td><td>74</td><td>78</td><td>84</td><td>82</td><td>80</td></tr>
<tr><td>LBG-log-Rissanen élagage</td><td>78</td><td>76</td><td>82</td><td>85</td><td>81</td><td>81</td></tr>
</table>

Tableau 47 – Taux d'identification correcte obtenu par fusion des scores consonantiques et vocaliques.

La fusion des modèles résulte pratiquement toujours en une amélioration du score d'identification obtenu. La seule exception se présente lorsque l'on calcule les vraisemblances consonantiques par rapport au modèle obtenu par LBG-log-Rissanen dans l'espace d'observation #9. Le meilleur score est de 85 % d'identification correcte, et il est obtenu avec le système constitué des meilleurs modèles (LBG-log-Rissanen avec élagage pour les voyelles et LBG (30 classes) avec élagage pour les consonnes). Il est cependant intéressant de noter que ce score n'est pas significativement plus élevé que ceux obtenus avec plusieurs autres configurations. En particulier, l'utilisation des modèles vocaliques (resp. consonantiques) obtenus par LBG-log-Rissanen (resp. LBG (30 classes)) aboutit à un score de 84 %, et ce, sans nécessiter de techniques d'élagage ou de normalisation. Cela signifie que ce score d'identification est obtenu sans réaliser d'adaptation aux données du corpus DEV_ST.

2 UN APPORT DE LA MODELISATION DIFFERENCIEE ?

A la lumière des expériences décrites précédemment, il apparaît que la modélisation différenciée consonnes/voyelles et la modélisation globale permettent d'obtenir des performances comparables en terme d'identification des langues (85 % d'identification correcte). Il demeure nécessaire de déterminer si les informations extraites du signal par les deux approches sont redondantes ou complémentaires. Pour cela, il nous faut disposer pour chacune d'entre elles de modèles comparables : les algorithmes mis au point pour la modélisation différenciée (LBG-Rissanen et élagage) sont donc appliqués au modèle global, de manière à procéder à son optimisation. Ensuite, des expériences de fusion de décision sont proposées de manière à mettre en évidence une éventuelle complémentarité des modèles global (MG) et différencié (MD).

2.1 Optimisation du modèle segmental global

Dans le cas de la modélisation segmentale globale, le prétraitement acoustique se limite à une segmentation automatique du signal et une paramétrisation cepstrale pour chaque segment de longueur variable. Le modèle acoustico-phonétique de référence que nous avons étudié jusqu'ici est un MMG initialisé à partir de l'algorithme LBG. Le système d'IAL basé sur ce modèle atteint 85 % d'identification correcte, pour un nombre de gaussiennes égale à 50, nombre constant pour chaque langue.

Le travail d'optimisation fait dans le cadre de l'approche vocalique est maintenant appliqué au modèle segmental global.

Le premier travail consiste à optimiser, si cela est possible, le nombre de lois gaussiennes du MMG, en appliquant l'algorithme LBG-Rissanen lors de l'initialisation. Nous limitons les expériences aux trois familles de paramètres #2 (8MFCC + D), #9 (8MFCC + 8 ΔMFCC + D) et #10 (8MFCC + E + 8 ΔMFCC + ΔE + D), qui ont donné les meilleurs résultats dans le cadre classique. Le Tableau 48 donne le nombre de lois gaussiennes optimal, au sens du critère de Rissanen, pour chaque cas : il est facile de remarquer que plus le nombre de paramètres augmente, plus le nombre de lois diminue et que le Vietnamien est toujours sous représenté, et ceci quel que soit l'espace de paramètres. Il est à noter que, pour les paramétrisations #9 et #10, il faut utiliser l'algorithme LBG-log-Rissanen pour atteindre un nombre de classes raisonnable. Les taux d'identification correcte qui sont obtenus avec ces différentes configurations (Figure 52) sont tout à fait prévisibles : plus le nombre de classes diminue, plus les taux chutent. Dans tous les cas, la configuration avec un nombre équivalent de lois gaussiennes pour chaque langue donne le meilleur résultat. Il semble que l'algorithme de Rissanen ne parvient pas à trouver la topologie optimale dans l'espace global de tous les sons.

Langue / Jeu de paramètres	FR	JA	KO	SP	VI
#2	45	33	38	37	21
#9	15	12	12	20	10
#10	13	8	8	12	9

Tableau 48 – Nombre de composantes gaussiennes fixé par LBG-Rissanen (paramètres #2) et LBG-log-Rissanen (paramètres #9 et #10) pour la modélisation globale segmentale.

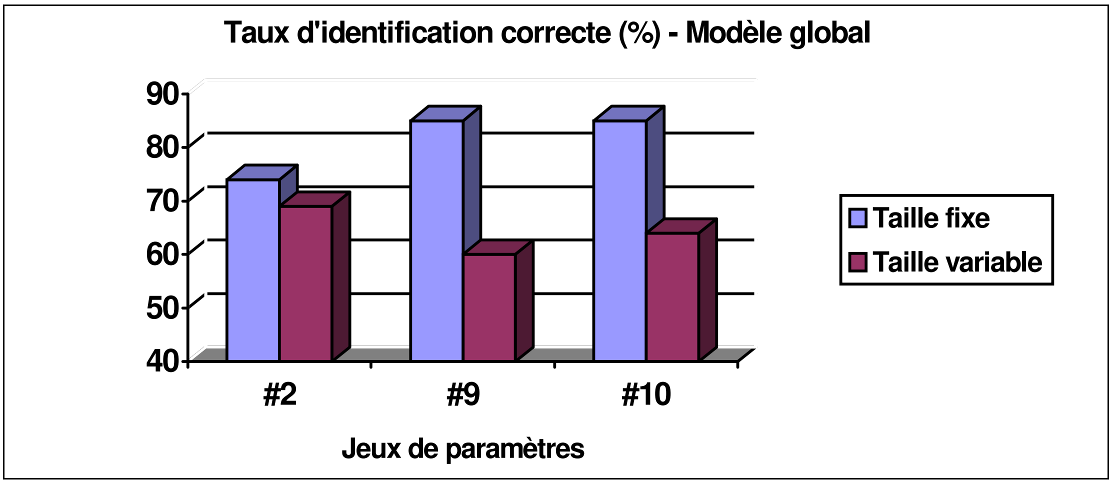

Figure 52 – Comparaison des taux d'identification correcte entre les modèles de taille constante (algorithme LBG) donnant le meilleur score et les modèles de taille variable (algorithmes LBG-Rissanen ou LBG-log-Rissanen).

Une deuxième expérience est réalisée en appliquant la technique d'élagage aux modèles segmentaux globaux les plus performants, à savoir #9 (LBG 50 classes) et #10 (LBG 40 classes). Quel que soit le pourcentage d'élagage utilisé (Tableau 49), les taux d'identification restent très stables et se maintiennent autour de 85 % (tableau 49) : on note 86 % pour le modèle #9 et 20 % d'élagage. Un score similaire peut être atteint avec le modèle #10. Les modifications apportées par cette heuristique demeurent non significatives.

Modèle \ Elagage (%)	0	5	10	15	20	25	30	35	40
#9 – LBG 50 classes	85	85	85	85	86	85	85	85	84
#10 – LBG 40 classes	84	84	86	86	84	84	82	82	82

Tableau 49 – Influence de la proportion d'élagage sur le score obtenu en identification avec la modélisation globale.

2.2 Fusion des modèles global et différencié

Compte tenu des résultats précédents, nous cherchons à fusionner le modèle global #9-LBG-50 classes avec successivement le modèle vocalique #9-LBG-Rissanen, le modèle consonantique #9-LBG-30 classes et le modèle différencié (modèle vocalique #9-LBG-Rissanen + modèle consonantique #9-LBG-30 classes). Les résultats sont données dans le Tableau 50 et dans la Figure 53.

	Modèle	Consonantique	Vocalique	Différencié (Consonantique + Vocalique)
Modèle	Score seul	78	78	84
Global	86	84	91	88

Tableau 50 – Taux d'identification obtenus par fusion du modèle global et des modèles issus de la modélisation différenciée.

La fusion du modèle global et du modèle consonantique (MG + MC) se révèle infructueuse. Il semble donc que les informations consonantiques modélisées soient redondantes. A l'inverse, la fusion des modèles global et vocalique (MG + MV) est efficace puisque le taux d'identification passe de 86 % (MG seul) à 91 % (MG + MV). Les intervalles de confiance (α = 95 %) correspondants sont respectivement de 8 % et 5 %. Cela met en évidence l'intérêt d'une prise ne compte du système vocalique dans un modèle homogène. La fusion des modèles global et différencié (MG + MD) donne logiquement un score intermédiaire où l'apport des modèles vocaliques est atténué par l'influence des systèmes consonantiques.

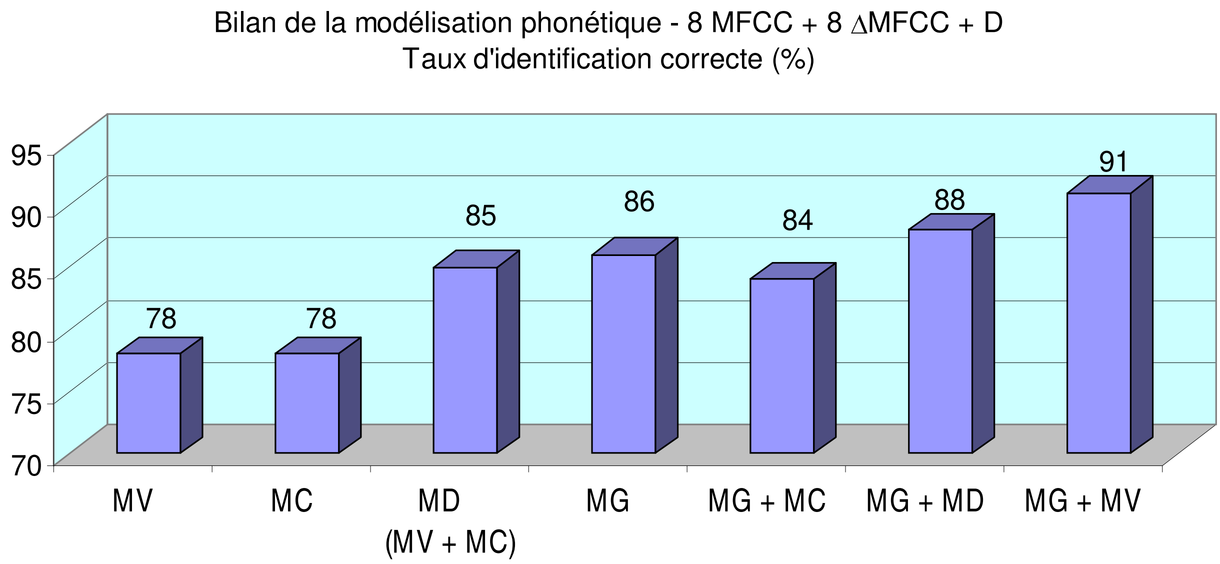

Figure 53 – Bilan des expériences réalisées avec les modèles différencié et global sur le corpus DEV_ST.

La modélisation du système vocalique par un mélange de lois gaussiennes dont le nombre de lois est adapté à chaque langue à l'aide de l'algorithme LBG-Rissanen semble efficace et pertinente. Le même effet n'est pas observé pour le système consonantique : dans ce cas, tout comme dans le cas de la modélisation globale, la complexité du modèle augmente bien plus vite que la distorsion des données ne décroît. L'algorithme LBG-Rissanen (ou LBG-log-Rissanen) n'est alors pas adapté à la structure complexe de l'espace d'observations (données hétérogènes).

3 RESULTATS DETAILLES DU SYSTEME DERIVE DES MODELES PHONETIQUES GLOBAL ET VOCALIQUE

En conclusion de ce travail, nous présentons de manière plus complète que précédemment les résultats obtenus avec la modélisation différenciée des consonnes et des voyelles dans différentes tâches d'IAL dérivées du système d'identification des cinq langues. Le système que nous avons choisi d'approfondir est celui basé sur :

- ✓ les modèles de SV obtenus par l'algorithme LBG-log-Rissanen à partir du jeu de paramètres #9,
- ✓ les modèles issus de la modélisation segmentale globale obtenus par l'algorithme LBG (50 classes) à partir du même jeu de paramètres.

Dans les deux cas les procédures d'élagage sont appliquées. Ce système aboutit au taux d'identification correcte de 91 % sur les enregistrements du corpus DEV_ST.

3.1 Matrices de confusion obtenues pour DEV_10 et DEV_ST

Le Tableau 51 indique la matrice de confusion obtenue avec le système précédemment présenté sur les locuteurs masculins du corpus DEV_ST. Si le français, le coréen et le vietnamien sont parfaitement reconnus (100 % d'identification correcte), les performances obtenues sur le japonais restent nettement en deçà de celles des autres langues. Nous n'avons pas trouvé de raisons justifiant ce manque de discrimination chronique de nos modèles MMG japonais ; cela dit, les auteurs de la présente étude ne comprenant pas le japonais, aucune étude approfondie n'a pu être menée.

Langue \ Modèle	FR	JA	KO	SP	VI	Taux
français	14					100 %
japonais		8		4	1	62 %
coréen			17			100 %
espagnol		1	1	13		87 %
vietnamien					15	100 %

Tableau 51 – Matrice de confusion obtenue sur le corpus DEV_ST par fusion du modèle vocalique #9 initialisé par LBG-log-Rissanen avec le modèle global #9 appris par LBG (50 classes).

Le Tableau 52 indique la matrice de confusion obtenue avec le même système sur les locuteurs masculins du corpus DEV_10. Le taux d'identification correct moyen est alors de 69 %. Le score du français reste nettement supérieur à la moyenne (80 %), et les problèmes rencontrés avec le japonais se trouvent accentués par la brièveté des stimuli employés.

Langue \ Modèle	FR	JA	KO	SP	VI	Taux
français	43	2	3	6		80 %
japonais	9	25	9	13	3	42 %
coréen	4	5	51	7	1	75 %
espagnol	5	6	4	45	4	70 %
vietnamien	2	7	5	2	47	75 %

Tableau 52 – Matrice de confusion obtenue sur le corpus DEV_10 par fusion du modèle vocalique #9 initialisé par LBG-log-Rissanen avec le modèle global #9 appris par LBG (50 classes).

3.2 Prise en compte des hommes et des femmes

Les expériences présentées ici sont réalisées en ne prenant pas en compte les locuteurs féminins et cela rend difficile toute comparaison de notre système avec ceux présentés au cours de la seconde partie. Etant bien conscients de ce fait, nous avons aussi testé le système appris sur les **seuls locuteurs masculins** sur l'ensemble des locuteurs, **masculins et féminins**, des corpus DEV_10 et DEV_ST. Il est évident que le système obtenu est sous-optimal pour de nombreuses raisons, parmi lesquelles nous relevons les suivantes :

- ✓ on rajoute durant la phase d'identification une grande variabilité non prise en compte durant l'apprentissage,
- ✓ il est connu que les systèmes vocaliques féminins et masculins diffèrent singulièrement [Calliope 89, Henton 95].

Cela signifie que les résultats que nous présentons peuvent être améliorés en adaptant la topologie du système et en intégrant un détecteur homme/femme, comme le fait Zissman [Zissman 96]. Ces remarques étant formulées, le Tableau 53 présente les résultats obtenus avec le même système que celui utilisé au paragraphe précédent.

	DEV_10	DEV_ST
Locuteurs masculins	69 %	91 %
Locuteurs masculins + féminins	61 %	80 %

Tableau 53 – Taux d'identification correcte obtenus avec les locuteurs féminins et masculins et les modèles appris sur les locuteurs masculins seuls.

L'inadaptation entre les modèles et les enregistrements de test aboutit à une chute de performances de 5 à 10 % par rapport aux tests réalisés avec les locuteurs masculins seuls.

3.3 Expériences en discrimination FR-L

Nous avons évalué notre système d'identification phonétique dans une tâche de discrimination entre une langue de référence (dans notre cas le français) et une autre langue, notée L. Les résultats obtenus (Tableau 54) indiquent un taux de discrimination moyen supérieur à 97 % sur le corpus DEV_ST. On constate que les meilleurs taux d'identification sont obtenus avec le vietnamien et le coréen, ce qui est confirme le caractère très discriminant des modèles de ces langues. Le passage à des stimuli d'une durée de 10 secondes s'accompagne d'une baisse de performance d'environ 10 %.

Langues	Taux d'identification	
	DEV_10	DEV_ST
FR-JA	86	93
FR-KO	93	100
FR-SP	85	97
FR-VI	94	100
Moyenne FR-L	89,5	97,5

Tableau 54 – Taux d'identification correcte obtenus en discrimination FR-L sur les corpus DEV_10 et DEV_ST.

3.4 Expériences en identification de quatre langues

Une autre expérience complémentaire consiste, à partir des langues disponibles, à évaluer l'influence de l'ajout d'une langue dans la tâche d'identification. Pour cela, nous évaluons les performances du système avec quatre langues parmi les cinq dont nous disposons (Tableau 55). Le passage d'un système de quatre langues à un système plus complexe de 5 langues induit une dégradation des performances sur le corpus DEV_10 (passage de 74 % en moyenne à 69 %). Cette dégradation est surtout sensible lor sque l'on ajoute le japonais au système puisqu'il s'agit de la langue la moins discriminée (on chute alors de 80 % d'identification correcte à 69 %).

Langues	Taux d'identification	
	DEV_10	DEV_ST
FR JA KO SP	69	90
FR JA KO VI	77	95
FR JA SP VI	72	88
FR KO SP VI	80	98
JA KO SP VI	70	88
Moyenne	74	92

Tableau 55 – Taux d'identification correcte obtenus en identification de quatre langues en ensemble fermé sur les corpus DEV_10 et DEV_ST.

Sur le corpus DEV_ST, l'augmentation de la complexité de la tâche n'induit pas de dégradation significative des performances (passage de 92 % en moyenne à 91 %). Les taux d'identifications obtenus avec les cinq systèmes de quatre langues s'échelonnent de 88 % (lorsqu'on ne traite pas le coréen) à 98 % (lorsque le japonais n'est pas considéré). Les meilleurs taux sont obtenus en ne prenant pas en compte l'une des deux langues les plus confondues, à savoir l'espagnol et le japonais (respectivement 95 % et 98 %).

CONCLUSION

Le travail présenté au cours de ce chapitre a pour but de répondre à la principale question posée : une modélisation acoustico-phonétique des SV est-elle pertinente en IAL ? Le problème abordé se décompose en plusieurs phases ayant pour point commun notre volonté de développer un système ne nécessitant pas l'emploi de données étiquetées manuellement.

La première phase consiste à extraire du signal les voyelles qui nous permettront par la suite d'identifier la langue parlée. Les expériences menées avec les algorithmes présentés au premier chapitre montrent que l'approche choisie, basée sur une localisation spectrale d'événements vocaliques, est adaptée à l'objectif poursuivi. En effet, si elle n'atteint peut-être pas les scores des systèmes à base de modèles de Markov ou de réseaux de neurones, elle assure une relative indépendance vis à vis des conditions d'enregistrements et de la langue parlée tout en obtenant de bons taux de détections en environnement multilingue (environ 90 % de détection correcte).

La seconde phase du problème vise à obtenir une modélisation discriminante des systèmes vocaliques de chaque langue et le second chapitre présente les algorithmes mis en œuvre en modélisation au niveau acoustico-phonétique. A partir du cadre classique de la modélisation par MMG, nous avons introduit l'algorithme LBG-Rissanen, qui, en intégrant un critère de type MDL (*Minimum Description Length*), permet d'adapter la topologie des modèles aux données d'apprentissage. Les expériences proposées au cours de ce chapitre sur des corpus francophones confirment le caractère prometteur de cette approche.

La phase suivante de notre travail a bien évidemment consisté à étudier la robustesse des modèles de SV dans une tâche d'IAL. Les nombreuses expériences décrites au cours du troisième chapitre montrent que les segments vocaliques sont porteurs d'une information importante sur l'identité de la langue puisque nous obtenons par discrimination des SV un taux d'identification correcte de 80 % avec les locuteurs masculins de cinq langues (contre un taux de 86 % avec l'ensemble des segments) alors que seulement 10 secondes de parole environ sont exploitées sur les 45 secondes des enregistrements. Si les expériences menées au cours de ce chapitre montre l'intérêt très net de la modélisation des SV en IAL, il reste indispensable d'étudier son impact dans un système d'IAL prenant en compte plus d'information que les quelques secondes de segments vocaliques extraites des fichiers. Le quatrième chapitre présente des expériences complémentaires visant à répondre à cette dernière question. La prise en compte des segments vocaliques et consonantiques au sein de deux modèles s'avère plus efficace que l'approche vocalique seule, même si l'apport d'une modélisation commune de toutes les consonnes reste faible (on obtient cependant un taux d'identification correcte de 85 % sur les mêmes enregistrements).

L'algorithme LBG-Rissanen permet d'obtenir des modèles de SV particulièrement discriminants mais il se révèle inefficace dans le cadre de la modélisation des systèmes consonantiques. Cette tendance se confirme lorsqu'un modèle de l'ensemble des sons de chaque langue est recherché par cette méthode.

Le fait que la modélisation des SV apporte une information supplémentaire par rapport à la modélisation globale est indéniable puisque la prise en compte conjointe des deux modèles aboutit à un taux de 91 % d'identification correcte. Afin de comparer ces résultats avec ceux disponibles dans la littérature qui portent généralement sur les locuteurs masculins et féminins, nous avons effectué une évaluation des modèles appris sur les locuteurs masculins avec l'ensemble des locuteurs : les résultats se trouvent logiquement dégradés par l'inadaptation entre les ensembles d'apprentissage et de test ; le taux d'identification correct s'élève alors à 80 %. Il reste supérieur à celui rapporté par Berkling (74 % sur trois langues du corpus OGI MLTS [Berkling 94]) ; il est également comparable à celui obtenu par Andersen sur quatre langues du même corpus (84 % [Andersen 97]) au moyen de modèles phonotactiques. Une autre comparaison intéressante peut être faite avec les résultats fournis par Kadambe : elle obtient 91 % d'identification correcte sur trois langues du corpus OGI MLTS en procédant à une modélisation phonétique markovienne suivie d'une modélisation phonotactique. Lorsqu'elle ne prend pas en compte ce dernier module, le score d'identification descend à 72 % [Kadambe 94]. Ces résultats montrent d'une part que la modélisation segmentale, différenciée ou non, permet souvent d'obtenir une meilleure discrimination phonétique que l'approche centi-seconde, et d'autre part que la modélisation acoustico-phonétique des SV est efficace. Ces expériences tendent à démontrer que l'information acoustico-phonétique peut être utilisée dans un système d'IAL, sans qu'il faille pour autant écarter les modèles phonotactiques qui sont au cœur des systèmes les plus performants.

(1628) - © Heritage Map Museum

Conclusion générale et perspectives

Le travail de thèse présenté dans ce manuscrit a pour objectif principal la recherche d'une discrimination efficace entre plusieurs langues par modélisation de leur système vocalique au niveau acoustico-phonétique. Cette problématique est née d'une double réflexion sur les techniques d'ingénierie classiquement mises en œuvre au sein des systèmes d'Identification Automatique des Langues (IAL) et sur les études typologiques menées en linguistique.

La première partie du document offre, par le biais d'un premier chapitre d'introduction à la linguistique, un éclairage sur la diversité des langues et sur leurs traits distinctifs. Cette introduction amène à s'intéresser à la caractérisation des langages parlés, et plus précisément à en étudier une composante phonologique : les systèmes vocaliques. Les études citées, basées sur des travaux typologiques à partir de la base de données UPSID, montrent qu'il est possible d'établir une typologie des langues à partir de la substance de leur Système Vocalique (SV). Dès lors, en poursuivant dans la direction du contenu acoustico-phonétique, la possibilité de modéliser automatiquement les SV à partir du signal acoustique, sans faire appel au « filtrage linguistique » des experts, s'avère prometteuse, tant en linguistique qu'en IAL.

Ce dernier domaine ayant été peu abordé en France, la seconde partie de notre exposé en présente les motivations et inventorie la plupart des approches développées à travers le monde. L'étude des méthodes employées montre qu'elles reposent généralement sur la modélisation des contraintes phonotactiques spécifiques aux langues traitées, en négligeant généralement leurs caractéristiques phonétiques. Ce constat nous amène à envisager une exploitation de l'information phonétique au sein de modèles différenciés. A notre sens, cette approche, tout en permettant éventuellement de réaliser un apprentissage de type non supervisé, peut faire émerger des caractéristiques discriminantes et exploitables dans un système d'IAL.

Les réflexions menées sur l'aspect linguistique et sur l'aspect ingénierie conduisent à réaliser une application d'IAL basée sur la modélisation acoustico-phonétique des SV : cette étude fait l'objet de la troisième partie. Nous y présentons tout d'abord les algorithmes développés pour localiser les voyelles dans le signal acoustique. La nécessité de disposer d'un système multilingue et indépendant du locuteur nous a orienté vers la mise au point d'un algorithme d'analyse spectrale plutôt que sur une modélisation markovienne ou neuromimétique. Les résultats obtenus sur des données enregistrées au travers du canal téléphonique sont satisfaisants : les segments vocaliques détectés comportent environ 10 % d'insertions, tandis que moins de 10 % des zones vocaliques sont omises. Le second chapitre présente ensuite les méthodes employées pour faire émerger des modèles de SV caractéristiques de chaque langue à partir des segments vocaliques détectés pour chacun des locuteurs de la base d'apprentissage. L'approche choisie est basée sur une modélisation des SV par Modèles de Mélanges de lois Gaussiennes (MMG) et elle intègre un critère de type MDL (*Minimum Description Length*) par le biais de l'algorithme LBG-Rissanen. Des expériences réalisées sur des corpus variés permettent d'évaluer la robustesse des algorithmes implantés, mais c'est surtout au cours des expériences en IAL du troisième chapitre qu'est mis en évidence le pouvoir discriminant de ces modèles, obtenus de manière automatique et entièrement non supervisée.

Dans une tâche d'identification d'une langue parmi cinq, à partir d'enregistrements de 45 secondes de parole spontanée prononcée par des locuteurs masculins, le taux d'identification correcte obtenu par discrimination des SV est de 80 % alors qu'un système de référence basé sur une modélisation segmentale globale de tous les sons du corpus d'apprentissage aboutit à un score de 86 %. Ces résultats montrent qu'une exploitation judicieuse de certains segments de parole permet de prendre en compte des informations très discriminantes, même si une partie du signal n'est pas modélisée (en l'occurrence, les segments vocaliques ne représentent en moyenne que 10 secondes de durée pour chaque enregistrement de 45 secondes). Ces expériences valident l'approche par modélisation des SV et confirment qu'une approche linguistique basée sur la substance des systèmes phonologiques peut être intégrée à un système automatique. Des expériences complémentaires visant à coupler cette modélisation vocalique à une modélisation consonantique sont présentées au dernier chapitre de cette thèse ; elles aboutissent à un taux d'identification correcte de 85 % dans la même tâche que précédemment. La relativement faible amélioration obtenue avec le modèle consonantique global (toutes les segments non vocaliques sont modélisés ensemble) semble confirmer qu'il est plus efficace de modéliser chaque catégorie de sons d'une langue dans un espace différencié homogène. Une modélisation commune de segments aussi différents que des fricatives ou des consonnes liquides est en effet sous-optimale. Si l'apport de la modélisation phonétique différenciée n'est pas démontrée par ces expériences (les taux atteints par les approches différenciée et globale sont similaires), elles permettent cependant de confirmer le caractère discriminant des modèles de SV, puisque l'utilisation conjointe d'un modèle global et d'un modèle de SV permet d'atteindre un taux d'identification correcte de 91 %. Une comparaison de nos résultats avec ceux disponibles dans la littérature montre que notre système est l'un des meilleurs parmi ceux basés sur une modélisation phonétique et non phonotactique. Ce constat confirme l'apport de la modélisation différenciée des SV et nous amène à envisager plusieurs évolutions de notre système, tant au niveau de la modélisation phonétique différenciée que de l'intégration de plusieurs sources d'information. Ces perspectives concernent le perfectionnement de la modélisation différenciée, son extension à de nouvelles classes phonétiques et l'exploitation du décodage ainsi réalisé au sein d'un modèle phonotactique adapté.

Nous avons relevé, au cours des expériences réalisées en IAL, que les algorithmes employés pour estimer les modèles acoustico-phonétiques sont particulièrement sensibles aux initialisations. Les techniques visant à pallier ce défaut bien connu sont nombreuses, tant au niveau de la quantification vectorielle que des mélanges MMG. Citons parmi elles la méthode proposée par Bernd Fritzke, inspirée des algorithmes d'apprentissage des réseaux neuromimétiques auto-organisés [Fritzke 97], que nous implanterons prochainement. Une autre modification plus fondamentale de notre système porte sur l'étiquetage du signal en segments vocaliques ou consonantiques. Il est vraisemblable que le caractère déterministe du détecteur actuellement employé représente un point faible du système et que l'on gagne à probabiliser cet étiquetage, par exemple au moyen d'une mesure dérivée du critère *Rec* utilisé au cours de la détection. Cela permettra d'augmenter la robustesse du modèle tout en intégrant un critère de « qualité » des voyelles, de manière à éventuellement privilégier les segments vocaliques les moins co-articulés. La mise en place d'un espace probabilisé pour la détection des voyelles nous amène à envisager une modélisation différenciée plus complexe, à base de Modèles de Markov Cachés (MMC) à deux niveaux. Un premier modèle MMC (Modèle Maître) prend en compte les résultats des différents détecteurs, et commande des modèles MMC différenciés (Modèles esclaves) pour chacune des classes phonétiques détectées. Cette approche, déjà exploitée avec succès dans le cadre de la reconnaissance de parole audio-visuelle [André-Obrecht 97], constitue une extension de notre système actuel en probabilisant les résultats du détecteur vocalique et en modélisant plus finement les systèmes vocaliques au niveau cepstral.

Les améliorations décrites ci-dessus seront couplées à la mise en place d'une modélisation différenciée pour chaque classe consonantique. En effet, l'observation des résultats obtenus avec les SV nous conforte dans l'opinion que la modélisation différenciée peut aussi être appliquée avec succès aux différentes classes de consonnes (fricatives, plosives, consonnes sonantes…). Un tel système permettra de modéliser les consonnes présentant des caractéristiques spectrales proches dans des espaces adaptés afin d'en augmenter le pouvoir discriminant et donc d'améliorer la discrimination phonétique des systèmes consonantiques.

La mise au point des modèles phonétiques des différentes classes consonantiques est également intimement liée à une perspective des plus prometteuses : l'intégration d'un module phonotactique à notre système. En effet, la séquence de segments produite par décodage dans les différents modèles phonétiques différenciés (sous l'hypothèse *Winner-Take-All*) est indéniablement porteuse d'une part des contraintes phonotactiques des langues. Il sera intéressant de modéliser ces contraintes au sein d'un modèle adapté et d'évaluer si la conception d'un système entièrement non supervisé intégrant ces deux modules, phonétique et phonotactique permet de rivaliser avec les meilleurs systèmes supervisés.

A ces perspectives s'ajoute une considération plus générale sur la réalité de la parole. Comme cela avait été dit en introduction de la première partie, le signal acoustique de parole est porteur d'informations sur le *trium vira* constitué du contenu sémantique, du locuteur et de la langue. On peut considérer que le message acoustique est obtenu par filtrage d'un énoncé cognitif par le locuteur dans l'espace de la langue, mais il est difficile d'extraire les caractéristiques propres à chaque membre du trio. La modélisation des SV que nous appliquons à l'IAL effectue une normalisation visant à minimiser les spécificités de chaque locuteur, de manière à se ramener à un espace commun et caractéristique de la langue. On peut envisager, à l'inverse, d'exploiter ces caractéristiques spécifiques au locuteur, de manière à obtenir pour chacun d'entre eux un espace correspondant au SV de la langue filtré par le locuteur. Cette approche, duale de la nôtre, peut alors être exploitée dans un but de caractérisation du locuteur et exploitée dans une application de reconnaissance du locuteur (identification ou vérification).

Annexes

ANNEXE 1

LISTE DES LANGUES DES CORPUS CITES

1 CALLFRIEND (12 LANGUES & 3 DIALECTES)

- Allemand
- Anglais américain – dialecte du sud
- Anglais américain – autres dialectes
- Arabe égyptien
- Coréen
- Espagnol – dialecte des Caraïbes
- Espagnol – autres dialectes
- Farsi
- Français canadien
- Hindi
- Japonais
- Mandarin – dialecte chinois du continent
- Mandarin – dialecte chinois de Taiwan
- Tamoul
- Vietnamien

2 CALLHOME (6 LANGUES)

- Allemand
- Anglais américain
- Arabe égyptien
- Espagnol
- Japonais
- Mandarin

3 EUROM_1 (11 LANGUES)

- Allemand
- Anglais
- Danois
- Espagnol
- Français
- Grec
- Hollandais
- Italien
- Norvégien
- Portugais
- Suédois

4 GLOBAL PHONE (9 LANGUES)

- Arabe
- Chinois
- Coréen
- Croate
- Espagnol
- Japonais
- Portugais
- Russe
- Turc

5 IDEAL (4 LANGUES)

- Allemand
- Anglais britannique
- Espagnol
- Français

6 OGI 22 LANGUAGES (22 LANGUES)

- Allemand
- Anglais
- Arabe (oriental)
- Cantonais
- Coréen
- Espagnol
- Farsi
- Français
- Hindi
- Hongrois
- Italien
- Japonais
- Malais
- Mandarin
- Polonais
- Portugais
- Russe
- Suédois
- Swahili
- Tamoul
- Tchèque
- Vietnamien

7 OGI MLTS (11 LANGUES)

- Allemand
- Anglais
- Coréen
- Espagnol
- Farsi
- Français
- Hindi
- Japonais
- Mandarin
- Tamoul
- Vietnamien

ANNEXE 2

LES MMG ET L'ALGORITHME EM

Dans le cadre de la modélisation de la fonction de densité de probabilité d'un vecteur aléatoire X par un mélange de lois gaussiennes, on a :

$$p(X = x|\theta) = \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right]$$

où $\theta = \{\alpha_k, \mu_k, \Sigma_k\}_{1\le k\le Q}$ est l'ensemble de paramètres définissant le MMG : les coefficients α_k sont les poids des différentes composantes, μ_k est la moyenne de la $k^{\text{ème}}$ gaussienne et Σ_k est la matrice de covariance associée.

On dispose d'un ensemble de données d'apprentissage $\mathcal{X} = \{x_1, \ldots, x_N\}$ observées pour la variable X et on suppose que ces observations sont indépendantes et identiquement distribuées. L'objectif de l'algorithme EM (Expectation-Maximization) est d'obtenir l'ensemble de paramètres θ^* qui maximise la vraisemblance de $\mathcal{X}$:

$$\theta^* = \arg\max_{\theta} p(\mathcal{X}|\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i|\theta)$$

L'algorithme EM est basé sur la supposition que chaque vecteur observé a été généré par une et une seule composante gaussienne du modèle. On introduit alors la variable aléatoire Z qui correspond au numéro de la gaussienne émettrice de X. Cette variable aléatoire est cachée car non directement observée.

L'algorithme, dans sa version de base, se déroule alors comme suit :

- **<u>Etape 1 – Initialisation</u>**
  - Initialisation des moyennes μ_k par Q points extraits aléatoirement de $\mathcal{X}$.
  - Initialisation de toutes les matrices de covariances Σ_k à la matrice unité I_p.
  - Initialisation équiprobable des poids des composantes : $\alpha_k = 1/Q$.
- **<u>Etape 2 – Itération</u>**
  - ***Phase d'Estimation***

    Calcul de la probabilité *a posteriori* $h_{ij}(t)$ que le vecteur x_i soit généré par la gaussienne j, pour tout i et pour tout j :

$$h_{ij}(t) \equiv E\left[z_{ij} \middle| x_i, \theta\right] = \frac{\dfrac{\alpha_j(t)}{(2\pi)^{p/2}\sqrt{\left|\Sigma_j(t)\right|}} \exp\left[-\dfrac{1}{2}\left((x_i - \mu_j(t))^T \Sigma_j(t)^{-1}(x_i - \mu_j(t))\right)\right]}{\sum_{k=1}^{Q} \dfrac{\alpha_k(t)}{(2\pi)^{p/2}\sqrt{\left|\Sigma_k(t)\right|}} \exp\left[-\dfrac{1}{2}\left((x_i - \mu_k(t))^T \Sigma_k(t)^{-1}(x_i - \mu_k(t))\right)\right]}$$

où $\theta(t) = \left\{\alpha_k(t), \mu_k(t), \Sigma_k(t)\right\}_{1 \le k \le Q}$ correspond aux paramètres du modèles après la $t^{\text{ième}}$ itération.

- ***Phase de Maximisation***

Ré-estimation des paramètres à partir de $\mathcal{X}$ et des probabilités $h_{ij}(t)$:

$$\alpha_j(t+1) = \frac{1}{N}\sum_{i=1}^{N} h_{ij}(t)$$

$$\mu_j(t+1) = \frac{\sum_{i=1}^{N} h_{ij}(t).x_i}{\sum_{i=1}^{N} h_{ij}(t)}$$

$$\Sigma_j(t+1) = \frac{\sum_{i=1}^{N} \left\{h_{ij}(t)\left(x_i - \mu_j(t+1)\right)\left(x_i - \mu_j(t+1)\right)^T\right\}}{\sum_{i=1}^{N} h_{ij}(t)}$$

- Incrémentation de t à $t+1$ et retour à la phase d'estimation.

- <u>**Etape 3 – Arrêt de l'algorithme**</u>

L'étape 2 est itérée jusqu'à ce que la variation des paramètres de moyennes des composantes gaussiennes descendent en dessous d'un seuil fixé.

L'étape 1 est souvent remplacée par une initialisation des gaussiennes à partir des résultats obtenus avec un algorithme de quantification vectorielle de type LBG. D'autre part, lorsque l'on ré-estime les matrices de covariances, il est courant de leur ajouter une matrice diagonale de petite valeur de manière à éviter qu'elles ne deviennent singulières (ce cas peut se produire lorsqu'on dispose de peu de données par exemple) [Kambhatla 96].

Références bibliographiques

[Andersen 97] O. Andersen & P. Dalsgaard, "Language-Identification Based on Cross-Language Acoustic Models and Optimised Information Combination", *Proc. of Eurospeech '97*, Rhodes, pp. 67-70, (1997)

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Speech Segmentation", *IEEE Trans. on ASSP*, Vol. 36, n° 1, pp 29-40, (1988)

[André-Obrecht 93] R. André-Obrecht, *Segmentation et parole ?*, Habilitation à diriger des recherches, Rennes, (1993)

[André-Obrecht 97] R. André-Obrecht & B. Jacob, "Direct Indentification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition", *Proc. of ICASSP '97*, Munich, pp. 999-992, (1997)

[Barkat 97] M. Barkat, J.M. Hombert & C. Taine-Cheikh, « Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes : Etats d'[a]mes vocaliques », *Actes des Journées d'Etudes Linguistiques – La voyelle dans tous ses états*, Nantes, pp. 141-146, (1997)

[Bell 1867] A. Bell, *Visible Speech*, Edited by Simpkin & Marshall, London, (1867)

[Berkling 95] K. M. Berkling, T. Arai & E. Barnard, " Theoretical Error Prediction for a Language Identification System using Optimal Phoneme Clustering", *Proc. of Eurospeech '95*, Madrid, pp. 351-354, (1995)

[Besacier 98] L. Besacier & J.F. Bonastre, « Système d'élagage temps-fréquence pour l'identification du locuteur », *Actes des XXII èmes Journées d'Etude sur la Parole*, Martigny, pp. 1-4, (1998)

[Bickerton 95] D. Bickerton, *Language and Human Behavior*, Univ. Of Washington Press, Seattle, (1995)

[Bladon 81] A. Bladon & B. Lindblom, "Modeling the Judgments of Vowel Quality Differences", *Journal of the Acoustical Society of America* 69, Vol. 5, pp. 1414-1422, (1981)

[Boë 86] L. J. Boë & C. Abry, « Nomogrammes et systèmes vocaliques », *Actes des 15èmes Journées d'Etude du GALF-GCP*, Aix en Provence, pp. 303-305, (1986)

[Boë 89] L. J. Boë, P. Perrier, B. Guérin & J. L. Schwartz, "Maximum Vowel Space", *Proc. of Eurospeech '89*, Paris, pp. 281-284, (1989)

[Boë 92] L. J. Boë, P. Perrier & A. Morris, « Une prédiction de l'audibilité des gestes de la parole à partir d'une modélisation articulatoire », *Actes des 19èmes Journées d'Etude sur la Parole*, Bruxelles, pp. 151-157, (1992)

[Boë 95] L. J. Boë, B. Gabioud, P. Perrier, J. -L. Schwartz & N. Vallée, « Vers une unification des espaces vocaliques », dans *Levels in Speech Communication – Relations and Interactions*, Edited by C. Sorin et al., Elsevier Science B. V., pp. 63-71, (1995)

[Boë 98] L.J. Boë, S. Maeda, C. Abry & J.L. Heim, « [i a u] ? à portée d'un conduit vocal de Néandertal », *Actes des 22èmes Journées d'Etude sur la Parole*, Martigny, pp. 245-248, (1998) [Botha 97] E.C. Botha & L.C.W. Pols, “Modeling the Acoustic Differences Between L1 and L2 Speech: The Short Vowels of Africaans and South-African English”, *Proc. of Eurospeech '97*, Rhodes, pp. 1035-1038, (1997)

[Bottéro 93] J. Bottéro & M.J. Stève, *Il était une fois la Mésopotamie*, Collection Découvertes, Edition Gallimard, Paris, (1993)

[Calliope 89] Calliope, La parole et son traitement automatique, Editeur principal J. P. Tubach, Masson, Editions CNET-ENST, Paris, (1989)

[Caraty 87] M.J. Caraty, *Contribution au décodage acoustico-phonétique étude de distances inter-spectres et reconnaissance de cycles vocaliques*, Thèse de 3 ème cycle de l'Université Paris 6, (1987)

[Caraty 92] M.J. Caraty & C. Montacié, « Intégration de la décomposition temporelle généralisée dans un système d'apprentissage symbolique. Application à la reconnaissance des voyelles », *Actes des 19èmes Journées d'Etude sur la Parole*, Bruxelles, pp. 387-392, (1992)

[Carré 95] R. Carré & M. Mrayati, “Vowel Transitions, Vowel Systems, and the Distinctive Region Model”, in *Levels in Speech Communication – Relations and Interactions*, Edited by C. Sorin et al., Elsevier Science, pp. 73-89, (1995)

[Chomsky 68] N. Chomsky & M. Halle, *The Sound Pattern of English*, Edited by Harper & Row, New York, (1968)

[Cimarusti 82] D. Cimarusti & R. B. Ives, “Development of an Automatic Identification System of Spoken Languages: Phase 1”, *Proc. of ICASSP '82*, Paris, pp. 1661-1663, (1982)

[Corredor-Ardoy 97] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker & L. Lamel, “Language Identification with Language-Independent Acoustic Models”, *Proc. of Eurospeech '97*, Rhodes, pp. 55-58, (1997)

[Crothers 78] J. Crothers, “Typology and Universals of Vowel Systems”, in *Universals of Human Language*, Edited by J. Greenberg, Stanford University Press, USA, (1978)

[Dalsgaard 94] P. Dalsgaard & O. Andersen, “Application of Inter-Language Phoneme Similarities for Language Identification”, *Proc. of ICSLP '94*, Yokohama, pp. 1903-1906, (1994)

[Damasio 97] A. Damasio & H. Damasio, « Le cerveau et le langage », *Dossier Pour La Science : Les langues du monde*, HS 17, pp. 8-15, (1997)

[Deacon 97] T. W. Deacon, *The Symbolic Species: The co-evolution of language and the brain*, Norton & Company Editors, New-York, (1997)

[Delattre 48] P. Delattre, « Un triangle acoustique des voyelles orales du français », *The French Review* 21, Vol. 6, pp. 477-484, (1948)

[Deligne 96] S. Deligne, *Modèles de séquences de longueurs variables : application au traitement du langage naturel et de la parole*, Thèse de 3 ème cycle, ENST, Paris, (1996) [Deligne 98] S. Deligne & Y. Sagisaka, « Modélisation statistique du langage avec un modèle bi-multigramme », *Actes des 22èmes Journées d'Etude sur la Parole*, Martigny, pp. 355-358, (1998)

[Dorr 98] B. Dorr & D. W. Oard, "Evaluating Resources for Query Translation in Cross-Language Information Retrieval", *Proc. of 1st International Conference on Language Resources & Evaluation*, Granada, pp. 759-764, (1998)

[Duda 73] R.O. Duda & P.E. Hart, *Pattern Classification and Scene Analysis*, Edited by Wiley-Interscience, (1973)

[Fakotakis 97] N. Fakotakis, K. Georgila & A. Tsopanoglou, "A continuous HMM text-independent speaker recognition system based on vowel spotting", *Proc. of Eurospeech '97*, Rhodes, pp. 2347-2350, (1997)

[Fant 60] G. Fant, *Acoustic Theory of Speech Production*, Edited by Mouton, The Hague, (1960)

[Farinas 98] J. Farinas, *La prosodie en identification automatique des langues*, mémoire de DEA, Univ. de Toulouse, non publié, (1998)

[Foil 86] J. T. Foil, " Language Identification using Noisy Speech ", *Proc. of ICASSP '86*, Tokyo, pp. 861-864, (1986)

[Fritzke 97] B. Fritzke, *The LBG-U method for vector quantization – an improvement over LBG inspired from neural networks*, Internal Report 97-01, Institut für Neuroinformatik, Ruhr-Universität Bochum, Kluwer Academic Publishers, (1997)

[Fukunaga 90] K. Fukunaga, *Statistical Pattern Recognition*, Second Edition, edited by Academic Press, San Diego,USA, (1990)

[Gauvain 94] J.L. Gauvain et al., *Identification automatique de la langue à travers le réseau téléphonique*, Rapport du contrat CNET n° 94 1B 089, n° 1-6, (1994-97)

[Gersho 92] A. Gersho & R.M. Gray, *Vector Quantization and Signal Compression*, Edited by Kluwer Academic Publishers, Norwell, USA, (1992)

[Glass 93] J. Glass, D. Goodine, M. Phillips, S. Sakai, S. Seneff & V. Zue, "A bilingual voyager system", *Proc. of the 1993 ARPA Human Language Technology Workshop*, Princeton, (1993)

[Goddijn 97] S.M.A.Goddijn & G. de Krom, "Evaluation of Second Language Learners' Pronunciation Using Hidden Markov Models", *Proc. of Eurospeech '97*, Rhodes, pp. 2331-2334, (1997)

[Goodman 89] F. J. Goodman, A. F. Martin & R. E. Wohlford, " Improved Automatic Language Identification in Noisy Speech ", *Proc. of ICASSP '89*, Glasgow, pp. 528-531, (1989)

[Hagège 82] C. Hagège, *Les structures des langues*, Collection Que sais-je, Edition Presses Universitaires de France, Paris, (1982) [Halloran 97] J. A. Halloran, *Sumerian Language Page*, http://www.primenet.com/~seagoat/sumerian, (1997)

[Hazen 97] T. J. Hazen, & V. W. Zue, “Segment-based automatic language identification”, *Journal of the Acoustical Society of America*, Vol. 101, No. 4, pp. 2323-2331, April, (1997)

[Hellwag 1781] C. Hellwag, traduction M. P. Monin, , "De Formatione Loquelae", Inauguralis Physiologico Medica, Univ. De Tübigen, *Bulletin de la Communication Parlée* n°1, Grenoble, (1991)

[Henton 95] C. Henton, “Cross-language Variation in the Vowels of Female and Male Speakers” *Proc. of 13 th International Congress of Phonetic Sciences*, Stockholm, pp. 420-423, (1995)

[Hieronymous 97] J. Hieronymous & S. Kadambe, “Robust Spoken Language Identification using Large Vocabulary Speech Recognition”, *Proc. of ICASSP '97*, Munich, pp. 1111 -1114, (1997)

[Hirst 98] D. J. Hirst, A. Di Cristo & R. Espesser (sous presse) “Levels of representation and levels of analysis for intonation” dans *Prosody : Theory and Experiments*, Edited by M. Horne, Kluwer Academic Publishers, Dordrecht, (1998)

[Hombert 98] J.M. Hombert & I. Maddieson, “A linguistic approach to Automatic Language Recognition”, *Actes du Congrès International des Linguistes*, Paris, (1997)

[House 77] A. S. House & E. P. Neuberg, , “Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations”, *Journal of the Acoustical Society of America* 62, Vol. 3, pp. 708-713, (1977)

[Itahashi 95] S. Itahashi & L. Du, “Language Identification Based on Speech Fundamental Frequency”, *Proc. of Eurospeech '95*, Madrid, pp. 1359- 1362, (1995)

[Ives 86] R. B. Ives, “A Minimal Rule AI Expert system for real-Time Classification of Natural Spoken Languages”, *Proc. of 2 nd Artificial Intelligence Advanced Computer Technology*, Long Beach, pp. 337-340, (1986)

[Jardino 96] M. Jardino, “ Multilingual Stochastic N-Gram Class Language Models ”, *Proc. of ICASSP '96*, Atlanta, pp. 161-164, (1996)

[Jean 87] G. Jean, *L'écriture, mémoire des hommes*, Collection Découvertes, Editions Gallimard, Paris, (1987)

[Jones 1786] Sir W. Jones, « Troisième dis cours anniversaire : On the Hindus », Reproduit dans *The Collected Works of Sir William Jones III*, J. Stockdale, London, pp. 23-46, (1807)

[Jones 18] D. Jones, *An Outline of English Phonetics*, Edited by W. Heffer & Sons LTD, Cambridge, (1918).

[Jusczyk 93] P. Jusczyk, A. Friederici, J Wessels, V. Svenkerud & A. Jusczyk, “Infants' Sensitivity to the Sound Pattern of Native Language Words”, *Journal of Memory and Language* 32, pp. 402-420, (1993)

[Kadambe 94] S. Kadambe & J. L. Hieronymous, “Spontaneous Speech Language Identification with a Knowledge of Linguistics”, *Proc. of ICSLP '94*, Yokohama, pp. 1879-1882, (1994)

[Kambhatla 96] N. Kambhatla, *Local Models and Gaussian Mixture Models for Statistical Data Processing*, PhD. in Computer science and Engineering, OGI, USA, (1996)

[Kingston 91] J. Kingston, “Phonetic Underresolution in UPSID”, *Proc. of 12 th International Conference of Phonetic Sciences*, Aix en Provence, pp. 359-362, (1991)

[Krauss 92] M. Krauss, K. Hale, L. Watahomigie, A. Yamamoto, C. Craig, L. Masayesva Jeanne & N. England, “Endangered languages”, in *Language*, Vol. 68, n° 1, (1992)

[Kumpf 97] K. Kumpf & R.W. King, “Foreign Speaker Accent Classification Using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Perception Benchmarks”, *Proc. of Eurospeech '97*, Rhodes, pp. 2323-2326, (1997)

[Kwan 95] H. Kwan & K. Hirose, “Recognized Phoneme-Based N-Gram Modeling in Automatic Language Identification”, *Proc. of Eurospeech '95*, Madrid, pp. 1367-1370, (1995)

[Kwan 97] H. Kwan & K. Hirose, “ Use of Recurrent for Unknown Language Rejection in Language Identification System”, *Proc. of Eurospeech '97*, Rhodes, pp. 63-66, (1997)

[Ladefoged 82] P. Ladefoged & A. Bladon, “Attempts by Human Speakers to Reproduce Fant's Nomograms”, *Speech Communication* 1, pp. 185-198, (1982)

[Lamel 94] L. F. Lamel & J.L. Gauvain, “ Language Identification using Phone-Based Acoustic Likelihood ”, *Proc. of ICASSP '94*, Adelaide, pp. 293-296, (1994)

[Lamel 98] L. F. Lamel, M. Adda-Decker, C. Corredor, J.J. Gargolf & J.L. Gauvain, , “A Multilingual Corpus for Language Identification”, *Proc. of 1 st International Conference on Language Resources & Evaluation*, Granada, pp. 1118-1122, (1998)

[Langaney 97] A. Langaney, « La génétique des populations à l'appui de la linguistique », *Dossier Pour La Science : Les langues du monde*, HS 17, pp. 50-52, (1997)

[Lass 84] R. Lass, “Vowel System Universals and Typology: Prologue to Theory”, *Phonology YearBook*, Vol. 1, Cambridge, pp. 75-111, (1984)

[Leonard 80] R. G. Leonard, *Language Recognition Test and Evaluation*, Technical Report RADC-TR-80-83, RADC/Texas Instruments Inc., Dallas, 1980

[Li 80] K. P. Li & T. J. Edwards, “Statistical Models for Automatic Language Identification ”, *Proc. of ICASSP '80*, Denver, pp. 884-887, (1 980)

[Li 94] K. P. Li, “Automatic Language Identification using Syllabic Spectral Features”, *Proc. of ICASSP '94*, Adelaide, pp. 297-300, (1994)

[Lieberman 91] P. Lieberman, *Uniquely Human: the Evolution of Speech, Thought and Selfless Behavior*, Harvard University Press, Cambridge, (1991)

[Liljencrants 72] J. Liljencrants & B. Lindblom, “Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast”, *Language* 48, pp. 839-862, (1972)

[Lindblom 75] B. E. F. Lindblom, “Experiments in Sound Structure”, *Proc. of 8th International Conference of Phonetic Sciences*, Leeds, (1975)

[Lindblom 86] B. E. F. Lindblom, “Phonetic Universal in Vowel Systems”, in *Experimental Phonology*, Edited by J. J. Ohala, Academic Press, Orlando, pp. 13-44, (1986)

[Lindblom 92] B. E. F. Lindblom, D. Krull & J. Stark, “Use of Place and Manner dimensions in the SUPERB-UPSID Database: Some Patterns of In(ter)Dependence”, *Fonetik*, pp. 39-42, (1992)

[Linde 80] Y. Linde, A. Buzo & R. M. Gray, “An Algorithm for Vector Quantier Design”, *IEEE Trans. on Com*., Vol. 28, January 80, pp. 84-95, (1980)

[Lund 95] M. A. Lund & H. Gish, “Two Novel Language Model Estimation Techniques for Statistical Language Identification”, *Proc. of Eurospeech '95*, Madrid, pp. 1363-1366, (1995)

[Maddieson 84] I. Maddieson, *Patterns of sounds*, Edited by Cambridge Univ. Press, Cambridge, USA, (1984)

[Maddieson 86] I. Maddieson, *Patterns of sounds*, 2nd Edition, Edited by Cambridge Univ. Press, USA, (1986)

[Maeda 90] S. Maeda, “Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of vocal-tract Shapes using an Articulatory Model”, in *Speech Production and Speech Modelling*, Edited by W. J. Hardcastle & A. Marchal, Kluwer Academic Publishers, Netherlands, pp. 131-149, (1990)

[Magnan 98] G. Magnan, « Le père de l'humanité », *Science et Vie*, n°967, pp. 88-92, (1998)

[Maidment 83] J. A. Maidment, “Language recognition and prosody: further evidence”, dans *Speech, hearing and language: Work in progress*, Vol. 1, Edited by University College, London, pp. 133-141, (1983)

[Mallory 89] J. P. Mallory, *Search of the Indo-Europeans: Language, Archeaology and Myth*, Edited by Thames and Hudson, London, (1989)

[Martinet 68] A. Martinet, *La linguistique synchronique, études et recherches*, Collection Le linguiste, Edition Presses Universitaires de France, Paris, (1968)

[Martinet 70] A. Martinet, *Eléments de linguistique générale*, Editions Armand Colin, Paris, (1970) [Mehler 86] J. Mehler, G. Lambertz, P. W. Jusczyk & C. Amizl-Tison, « Discrim ination de la langue maternelle par le nouveau-né », *Comptes-rendus de l'Académie des Sciences de Paris* 303, Série III(15), pp. 637-640, (1986)

[Mehler 95] J. Mehler & E. Dupoux, Naître Humain, 2nde Ed., Collection Opus, Editions Odile Jacob, Paris, (1995)

[Moon 93] C. Moon, R. Panneton-Cooper & W. P. Fifer, "Two-day-olds prefer their native language", *Infant Behavior and Development* 16, pp. 495-500, (1993)

[Morimoto 93] T. Morimoto, T. Takezawa, F. Yato, S. Sagayama, T. tashiro, M. Nagata & A. Kurematsu, "ATR's Speech Translation System: ASURA", ", *Proc. of Eurospeech '93*, Berlin, pp. 1291-1294, (1993)

[Muthusamy 92] Y. K. Muthusamy, R. A. Cole & B. T. Oshika, "The OGI Multilingual Telephone speech Corpus", *Proc. of ICSLP '92*, Banff, pp. 895-898, (1992)

[Muthusamy 93] Y. K. Muthusamy, *A Segmental Approach to Automatic Language Identification*, Ph. D. Thesis, Oregon Graduate Institut of Science & Technology, (1993)

[Muthusamy 94a] Y. K. Muthusamy, N. Jain & R. A. Cole, "Perceptual Benchmarks for Automatic Language Identification", *Proc. of ICASSP '94*, Adelaide, pp 333-336, (1994)

[Muthusamy 94b] Y. K. Muthusamy, E. Barnard & R. A. Cole, "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazine*, 10/94, pp 33-41, (1994)

[Navrátil 97] J. Navrátil & W. Zuhlke, "Phonetic-Context Mapping in Language Identification", *Proc. of Eurospeech '97*, Rhodes, pp. 71-74, (1997)

[Nazzi 98] T. Nazzi, J. Bertoncini & J. Mehler, "Language Discrimination by Newborns: Towards an Understanding of the Role of Rythm", *Journal of Experimental Psychology: Human Perception and Performance*, (1998)

[Nichols 92] J. Nichols, *Linguistic Diversity through Space and Time*, University of Chicago Press, Chicago, (1992)

[Nowlan 91] S. Nowlan, *Soft Competitive Adaptation: Neural Network Learning Algorithm based on fitting Statistical Mixtures*, PhD in School of Computer Science, Carnegie Mellon Univ., (1991)

[Nyland 97] E. Nyland, *Edo Nyland's Homepage*, http://www.islandnet.com/~edonon, (1997)

[Ohala 79] J. J. Ohala, & J. B. Gilbert, "Listeners' ability to identify languages by their prosody", dans *Problèmes de prosodie*, Vol. II, Edité par P. Léon & M. Rossi, Ottawa, pp. 123-131, (1979)

[Ohala 91] J. J. Ohala, "The Integration of Phonetics and Phonology", Proc. of 12th *International Conference of Phonetic Sciences*, Aix en Provence, pp. 2-16, (1991) [Parris 97] E. S. Parris, H. Lloyd-Thomas, M. J. Carey & J. H. Wright, “Bayesian Methods for Language Verification”, *Proc. of Eurospeech '97*, Rhodes, pp. 59-62, (1997)

[Payan 93] Y. Payan & P. Perrier, “Vowel normalization by articulatory normalization: Preliminary results”, *Proc. of the ASA 125th Meeting*, Ottawa, (1993)

[Pinker 94] S. Pinker, *The language Instinct: How the mind creates language*, Edited by W. Morrow, New-York, (1994)

[Pellegrini 92] B. Pellegrini, *Origine des Hommes Modernes : Revue Critique des Faits, des Modèles et des Hypothèses de la Paléontologie, de l'Archéologie et de la Génétique des Populations*, Thèse de 3ème cycle, Univ. de Genève, (1992)

[Phillipson 94] D. W. Phillipson, *African Archaeology*, 2nd Edition, Cambridge University Press, Cambridge, (1994)

[Puel 93] J.B. Puel & R. André-Obrecht, « Détection des début et fin de parole en environnement difficile », Actes du GRETSI, (1993)

[Ramesh 94] P. Ramesh & D. B. Roe, “Language Identification with Embedded Word Models”, *Proc. of ICSLP '94*, Yokohama, pp. 1887-1890, (1994)

[Ramus 98] F. Ramus & J. Mehler (in press), “Language identification with suprasegmental cues: A study based on speech resynthesis”, *Journal of the Acoustical Society of America*, (1998)

[Rayner 93] M. Rayner, , “A Speech to speech Translation system built from Standard Components”, *Proc. of the 1993 ARPA Human Language Technology Workshop*, Princeton, (1993)

[Renfrew 90] C. Renfrew, L'énigme indo-européenne. Archéologie et langage, Collection Histoires, Flammarion, (1990)

[Reynolds 95] D.A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models”, *Speech Communication*, Vol. 17, Nos 1-2, 08/95, pp. 91-108, (1995)

[Ruhlen 97] M. Ruhlen, traduction P. Bancel, *L'origine des langues – sur les traces de la langue mère*, Edition Belin, (1997)

[Ruhlen 98] M. Ruhlen, traduction P. Bancel, « Toutes parentes, toutes différentes », *La Recherche*, n° 306, pp. 68-75, (1998)

[Rissanen 83] J. Rissanen, “A universal Prior for Integers and Estimation by Minimum Description Length”, *The Annals of Statistics*, Vol. 11, No 2, pp. 416-431, (1983)

[Ross 97] Ph. Ross, « L'histoire du langage », *Dossier Pour La Science : Les langues du monde*, HS 17, pp. 20-27, (1997)

[Samouelian 96] A. Samouelian, “Automatic language identification using inductive inference”, *Proc. of Sixth Australian Int. Conf. on Speech, Science and Technology*, Adelaide, pp 251-256, (1996).

[Sanchez-Mazas 92] A. Sanchez-Mazas, L. Graven & B. Pellegrini, « Génétique, linguistique et préhistoire du peuplement subsaharien », *Bulletin du Centre Genevois d'Antropologie* 3, pp. 3-21, (1991-92)

[Savic 91] M. Savic, E. Acosta & S. K. Gupta, "An Automatic Language Identification System", *Proc. of ICASSP '91*, Toronto, pp. 817-820, (1991)

[Schultz 97] T. Schultz & A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets", *Proc. of Eurospeech '97*, Rhodes, pp. 371-374, (1997)

[Schwartz 89] J. L. Schwartz, L. J. Boë, P. Perrier, B. Guérin & P. Escudier, "Perceptual Contrast and Stability in Vowel systems: A 3-D Simulation Study", *Proc. of Eurospeech '89*, Paris, pp. 63-66, (1989)

[Schwartz 89b] J. L. Schwartz & P. Escudier, "A Strong Evidence for the Existence of a Large-scale Integrated Spectral Representation in Vowel Perception", *Speech Communication* 8, pp. 235-259, (1989)

[Schwartz 97] J. L. Schwartz, L. J. Boë, N. Vallée & C. Abry, "Major Trends in Vowel System Inventories", *Journal of Phonetics* 25, pp. 233-253, (1997)

[Siguan 96] M. Siguan, *L'Europe des langues*, Editeur P. Mardage, Sprimont, 1996

[Sirigos 96] J. Sirigos, V. Darsinos, N. Fakotakis & G. Kokkinakis, "Vowel-Non Vowel Classification of Speech using an MLP and Rules", *Proc. of EUSIPCO'96*, Trieste, pp. 1059 –1062, (1996).

[Stevens 55] K. N. Stevens & A. S. Halle, "Development of a Quantitative Description of Vowel Articulation", *Journal of the Acoustical Society of America* 27, Vol. 5, pp. 484-493, (1955)

[Stevens 72] K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data", in *Human Communication: a Unified View*, Edited by P. B. Denes & J. R. Davis, McGraw-Hill, New-York, pp. 51-66, (1972)

[Suaudeau 94] N. Suaudeau & R. André-Obrecht, "An Efficient Combination of Acoustic and Supra-segmental Informations in a Speech Recognition System", *Proc. of ICASSP '94*, adelaide, pp. 65-68, (1994)

[Taylor 97] R. Taylor, Entretien, Revue en ligne *DiversCité*, http://www.teluq.uquebec.ca/diverscite/SecEntre/09ct_fen.htm, (1997)

[ten Bosch 95] L. F. M. ten Bosch, "On the Lexical Aspects of Vowel Dispersion Theory: Dutch Case", *Proc. of 13 th International Congress of Phonetic Sciences*, Stockholm, pp. 416-419, (1995)

[Tucker 94] R. C. F. Tucker, M. J. Carey & E. S. Parris , "Automatic Language Identification using Sub-Words Models", *Proc. of ICASSP '94*, Adelaide, pp. 301-304, (1994)

[Vaissière 98] J. Vaissière et O. Piot, « Discrimination multilingue automatique à partir des caractéristiques prosodiques», Rapport final de l'ILPGA, convention DGA 95/118, (1998) [Vallée 94] N. Vallée, S *ystèmes vocaliques : de la typologie aux prédictions*, Thèse de 3ème cycle, Univ. Stendhal, Grenoble, (1994)

[Vallée 98] N. Vallée, L.J. Boë & M. Stefanuto, « Les systèmes consonantiques – des tendances universelles à l'ontogenèse », *Actes des 22èmes Journées d'Etude sur la Parole*, Martigny, pp. 241-244, (1998)

[Victorri 97] B. Victorri, « Débat sur la langue mère », », *Dossier Pour La Science : Les langues du monde*, HS 17, pp. 28-32, (1997)

[Waibel 96] A.Waibel, M. Finke, D. Gates, M. Gavaldà, T. Kemp, A. Lavie, L. Levin, M. Maier, L. Mayfield, A. McNair, I. Rogina, K. Shima, T. Sloboda, M. Woszczyna, T. Zeppenfeld & P. Zhan, "JANUS II: Translation of Spontaneous Conversational Speech ", *Proc. of ICASSP '96*, Atlanta, pp. 409-412, (1996)

[Wegman 96] S. Wegman, D. McAllaster, J. Orloff & B. Peskin, "Speaker Normalization on Conversational Telephone Speech", ", *Proc. of ICASSP '96*, Atlanta, pp. 339-342, (1996)

[Willems 82] N. Willems, *English intonation from a Dutch point of view*, Kluwer Academic Publishers, Dordrecht, (1982)

[Wood 79] S. A. J. Woods, "A Radiographic Analysis of Constriction Locations for Vowels", *Journal of Phonetics* 7, pp. 23-43, (1979)

[Yan 96] Y. Yan, E. Barnard & R. A. Cole, "Development of An Approach to Automatic Language Identification based on Phone Recognition", *Computer Speech and Language*, Vol. 10, n° 1, pp 37-54, (1996)

[Yé 89] H. Yé, M.J. Caraty, L.J. Boë & D. Tuffelli, "Structural (Phonetic) Evaluation of Dissimilarities Functions used in Speech Recognition", *Proc. of Eurospeech '89*, Paris, pp. 404-407, (1989)

[Zhan 97] P. Zhan, M. Westphal, M. Finke & A. Waibel, "Speaker Normalization and Speaker Adaptation – A Combination for Conversational Speech Recognition", *Proc. of Eurospeech '97*, Rhodes, pp. 2087-2090, (1997)

[Zissman 96] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *IEEE Trans. on Speech and Audio Processing*, Vol. 4, n° 1, pp 31-44, (1996)

[Zue 90] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni & S. Seneff, "Recent Progress on the SUMMIT System", *3rd DARPA Speech and Natural Language Workshop*, pp. 380-384, (1990)