UNIVERSITE
PAUL
SABATIER
TOULOUSE III

Université Toulouse III **École Doctorale Informatique et Télécommunications**

Formation Doctorale Signal Image & Acoustique

Caractérisation et identification automatique des langues

THÈSE

présentée et soutenue publiquement le 11 Mars 2005

pour l'obtention du

Doctorat de l'Université Toulouse III – Paul Sabatier

(spécialité informatique)

par

Jean-Luc Rouas

Composition du jury

Président :	Daniel Dours
Rapporteurs :	Kamel Smaïli Louis-Jean Boë
Examinateurs :	Martine Adda-Decker Véronique Aubergé
Directeur de thèse :	Régine André-Obrecht
Co encadrant :	François Pellegrino

Institut de Recherche en Informatique de Toulouse — UMR 5505

Remerciements

Je tiens tout d'abord à remercier Régine André-Obrecht, ma directrice de thèse, qui m'a appris à aimer la recherche, et m'a soutenu pendant ces quelques années. Je souhaite également remercier François Pellegrino, mon co-directeur de thèse, pour ses nombreux conseils.

Tous mes remerciements aux membres du jury, Daniel Dours, président, Kamel Smaïli, rapporteur, Véronique Aubergé et Martine Adda-Decker, examinatrices. Je vous remercie d'avoir pris le temps d'évaluer mon travail et d'y avoir porté attention.

Merci à toutes les personnes qui ont accepté de relire ce document, et un merci tout particulier à Sara pour ses nombreuses corrections et ses encouragements encore plus nombreux.

Merci aussi à tous les membres de l'équipe SAMOVA, pour leurs conseils et leur soutien. Merci à Jérôme, qui m'a initié au problème de l'identification automatiques des langues. Merci à Julien, mon camarade de thèse, pour sa bonne humeur et ses blagues. Un merci particulier à ceux qui ont accepté de partager leur bureau avec moi et qui m'ont supporté : Isabelle, Rosa, Zouhir, Yahya, Jorge, Jérômo, José.

Un grand merci aussi à tous mes amis qui ont été à mes cotés pour me remonter le moral, Bertrand, Laura, Greg, Fred, Karine, Loic et LoicSan, Sara, Sylvie, Véro.

Enfin, merci à ma famille, mon frère Joël, et mes parents.

Résumé

Le but de l'identification automatique des langues est de reconnaître la langue parlée par un locuteur inconnu, parmi un ensemble fini de langues, pour des énoncés d'une durée limitée (entre 3 et 50 secondes habituellement). La prosodie, c'est-à-dire l'expression musicale de la parole, constitue une source d'information importante pour l'identification des langues, car elle est présente dans les stratégies humaines de perception et de compréhension de la parole. Cependant sa modélisation et sa mise en œuvre relèvent toujours du défi. La prosodie est un des enjeux majeurs du traitement automatique de la parole. La prosodie est très peu utilisée en identification automatique des langues. L'objectif est d'étudier s'il est également possible, de même que les humains et en accord avec les théories linguistiques, d'utiliser les paramètres prosodiques dans le cadre de l'identification automatique des langues.

Abstract

The aim of Automatic Language Identification is to recognise the language spoken by an unknown speaker, within a finite set of language, for utterances of limited duration (usually between 3 and 50 seconds). Prosody, i.e. the musical expression of speech, is a important information source for language identification, because it is present in human speech perception and understanding strategies. Nevertheless, its modelling and application are not straightforward. Prosody is one of the major stakes of automatic speech processing. Prosody is seldom used in automatic language identification. The objective is to study if it is possible, as humans do and in accordance with linguistic theories, to use prosodic features in the framework of automatic language identification.

Sommaire

Table des figures **xvii**

Liste des tableaux **xix**

Introduction **1**

1 Les enjeux de l'identification automatique des langues 4

1.1 Enjeux applicatifs 4

1.2 Enjeux stratégiques 5

1.3 Enjeux scientifiques 5

2 Les techniques actuelles d'identification automatique des langues 5

3 Le projet de recherche 6

4 Démarche scientifique & organisation du mémoire 7

1 Définitions et motivations sur l'emploi de la prosodie **9**

1.1 La prosodie 12

1.1.1 Définition de la prosodie 12

1.1.2 Interprétation de la prosodie 12

1.2 Extraction de paramètres prosodiques 13

1.2.1 Énergie 13

1.2.2 Durée 14

1.2.3 Fréquence fondamentale 16

1.3 Intérêt de la prosodie pour l'identification des langues 17

1.3.1 Le rythme et la théorie de l'isochronie 18

1.3.2 L'intonation 20

1.3.3 Théories cognitives 21

1.3.4 Expériences en perception 21

1.4 Conclusion 24

2 L'identification automatique des langues : méthodes & approches classiques **25**

2.1 Architecture classique 28
2.2 Approche spectrale 29
2.3 Approche phonétique-phonotactique 30
2.4 Approches syllabiques 31
2.4.1 Méthode syllabotactique 31
2.4.2 Reconnaissance des syllabes 32
2.5 Campagne d'évaluation NIST 2003 [85] 34
2.5.1 Description de la campagne 35
2.5.2 Massachussets Institute of Technology - Lincoln Laboratory [120] . 37
2.5.3 Oregon Graduate Institute 40
2.5.4 Queensland University of Technology 41
2.5.5 R523 (*Department of Defense*) 43
2.5.6 *University of Washington* 43
2.5.7 IRIT/DDL 45
2.5.8 Résultats 46
2.6 Le système du LIMSI [49] 48
2.6.1 Cadre théorique 48
2.6.2 Expériences 49
2.7 Conclusion 50

3 L'identification automatique des langues : méthodes & approches prosodiques **51**

3.1 Systèmes comparatifs 54
3.1.1 Les travaux de Ramus 54
3.1.2 Les travaux de Grabe 55
3.1.3 Les travaux de Galves 55
3.2 Systèmes descriptifs (intonation) 57
3.2.1 Le système ToBI [119] 58
3.2.2 Le système IViE [51] 60
3.2.3 Modèle Intsint [60] 62
3.2.4 Modèle de Fujisaki [42] 63
3.2.5 Modèle de Gårding [48] 67
3.2.6 Modèle de Mertens [89] 68

3.3 Systèmes applicatifs . . . 71
3.3.1 Modèle de Leavers [76] . . . 71
3.3.2 Modèle d'Itahashi [66] . . . 72
3.3.3 Le système de Cummins [24] . . . 74
3.3.4 Le système de Li [78] . . . 76
3.3.5 Modèle d'Adami [2] . . . 77
3.4 Conclusion . . . 79

4 Extraction automatique et caractérisation d'unités prosodiques 81
4.1 Introduction . . . 83
4.2 Extraction automatique d'informations prosodiques . . . 85
4.2.1 Segmentation de la parole . . . 85
4.2.2 Détection de l'activité vocale . . . 86
4.2.3 Localisation des voyelles . . . 87
4.2.4 Conclusion . . . 88
4.3 Cadre expérimental . . . 89
4.3.1 Corpus . . . 89
4.3.2 Protocole expérimental . . . 91
4.3.3 Modélisation : cadre statistique . . . 92
4.4 Adaptation de quelques approches présentées au chapitre précedent . . . 93
4.4.1 Évaluation automatique des paramètres proposés par Ramus (§3.1.1) 93
4.4.2 Évaluation automatique des paramètres proposés par Grabe (§3.1.2) 96
4.4.3 Discussion . . . 98
4.5 Caractérisation d'une unité rythmique : la pseudo-syllabe . . . 99
4.5.1 Localisation de la pseudo-syllabe . . . 99
4.5.2 Modélisation des pseudo-syllabes . . . 100
4.5.3 Modélisation des caractéristiques temporelles des pseudo-syllabes . 100
4.5.4 Modélisation des caratéristiques intonatives des pseudo-syllabes . . 104
4.5.5 Fusion des modèles de durée et d'intonation pseudo-syllabiques . . . 108
4.6 Conclusion . . . 109

5 Un système d'identification automatique des langues par la prosodie 111
5.1 Cadre expérimental . . . 114
5.1.1 Création de classes de segments . . . 114
5.1.2 Modélisation par multigrammes . . . 114

5.1.3 Règle de décision 115
5.2 Modélisation du rythme : prise en compte temporelle 115
5.2.1 Regroupement des pseudo-syllabes en classes 115
5.2.2 Expériences en identification des langues 116
5.3 Modélisation de l'intonation à long terme 118
5.3.1 Traitement de la courbe de fréquence fondamentale 118
5.3.2 Traitement de la courbe d'énergie 125
5.3.3 Identification des langues 126
5.4 Modélisation de la prosodie (rythme et intonation) 127
5.4.1 Ajout d'étiquettes au modèle rythmique 128
5.4.2 Ajout d'étiquettes au modèle intonatif 129
5.4.3 Conclusion 131
5.5 Expériences sur le système d'Adami 131
5.6 Expériences en modélisation de la prosodie à court terme 134
5.7 Fusion : modèle dynamique et modèle statique 136
5.8 Fusion : modèle accentuel et modèle intonatif 137
5.9 Comparaison avec d'autres systèmes d'identification 138
5.9.1 Méthode acoustique 138
5.9.2 Méthode phonotactique (PPRLM) 140
5.9.3 Conclusion 141
5.10 Conclusion 141

6 Quelques pistes pour la parole spontanée 143
6.1 Expériences de discrimination des langues sur OGI 146
6.1.1 Corpus 146
6.1.2 Expériences 146
6.1.3 Conclusion 148
6.2 Mesure du débit 150
6.2.1 Définitions et méthodes d'estimation 150
6.2.2 Analyse des données 151
6.2.3 Evaluation des algorithmes comme estimateurs du débit 154
6.2.4 Conclusion et Perspectives 155
6.3 Conclusion 156

Conclusion et perspectives 159

Annexes **167**

A Exponentielle de Hurst **167**

A.1 Méthode de l'échelle réduite (*Rescaled Range*) . . . 169

A.2 Estimation de l'exponentielle de Hurst . . . 170

B Jeux de données du corpus Multext **173**

B.1 Description du corpus . . . 175

B.2 Enregistrements audio . . . 175

B.3 Répartition des passages par locuteurs . . . 176

B.3.1 Anglais . . . 176

B.3.2 Allemand . . . 176

B.3.3 Espagnol . . . 177

B.3.4 Français . . . 178

B.3.5 Italien . . . 178

B.4 Jeu de données n°1 . . . 180

B.4.1 Données d'apprentissage . . . 180

B.4.2 Données de test . . . 180

B.5 Jeu de données n°2 . . . 181

B.5.1 Données d'apprentissage . . . 181

B.5.2 Données de test (2 locuteurs par langue) . . . 181

B.6 Jeu de données n°3 . . . 182

B.6.1 Données d'apprentissage . . . 182

B.6.2 Données de test (2 locuteurs par langue) . . . 182

C Expériences complémentaires sur les différents jeux **183**

C.1 Expériences complémentaires avec le modèle de rythme . . . 185

C.1.1 Expériences sur les données du jeu n°2 . . . 185

C.1.2 Expériences sur les données du jeu n°3 . . . 186

C.2 Expériences complémentaires avec le modèle de l'intonation (statique) . . . 187

C.2.1 Expériences sur les données du jeu n°2 . . . 187

C.2.2 Expériences sur les données du jeu n°3 . . . 188

D Expériences avec les modèles acoustiques **189**

D.1 Modèles centiseconde . . . 191

D.1.1 Modélisation acoustique globale . . . 191

D.1.2 Modélisation des systèmes vocaliques 192
D.1.3 Modélisation des consonnes 193
D.1.4 Modélisation différenciée consonnes/voyelles 194
D.1.5 Modélisation des consonnes voisées 195
D.1.6 Modélisation des consonnes non voisées 196
D.1.7 Modélisation différenciée consonnes voisées/non voisées/voyelles . . 197
D.2 Modèles segmentaux 198
D.2.1 Modélisation acoustique globale 198
D.2.2 Modèle consonantique 199
D.2.3 Modélisation des systèmes vocaliques 200
D.2.4 Modélisation différenciée Consonnes/Voyelles 201
D.2.5 Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles202

E Algorithme VQ (Quantification Vectorielle) 205
E.1 Objectif 207
E.2 Algorithme des K-means 207
E.3 Algorithme LBG (Linde, Buzo, Gray) 208

F Algorithme EM (Expectation Maximisation) 211
F.1 Rappels 213
F.2 Algorithme de base 213

G Corpus Ogi-mlts 215
G.1 Protocole expérimental 216
G.2 Étiquetage phonétique 217

Bibliographie 221

Table des figures

1.1 Schéma type d'une syllabe 15
1.2 Représentation du signal (en haut), du spectrogramme (spectre large bande à droite et bande étroite à gauche) et des valeurs de F_0 (en bas). 16

2.1 Structure générale d'un système d'identification automatique des langues. . 28
2.2 Exemple de méthode d'apprentissage de MMG par adaptation d'un modèle du monde, et reconnaissance. 30
2.3 Structure d'un système d'identification basé sur une approche phonétique avec plusieurs décodeurs acoustico-phonétiques (Parallel Phone Recognition followed by Language Modelling - PPRLM) [128]. 31
2.4 Système de segmentation automatique en syllabes (image extraite de [94]). 33
2.5 Résultat de la segmentation automatique en syllabes (image extraite de [94]). 34
2.6 Vue d'ensemble du système du MIT 37
2.7 Vue d'ensemble du système du QUT 41
2.8 Vue d'ensemble du système de l'Université de Washington 43
2.9 Vue d'ensemble du système présenté par l'IRIT et le DDL 46
2.10 Résultats pour l'évaluation NIST 2003 47

3.1 Exemple d'étiquetage de la prosodie avec ToBI sur la phrase *Bananas aren't poisonous*. 59
3.2 Exemple d'étiquetage par le système de transcription IViE 62
3.3 Exemple d'étiquetage des contours de la fréquence fondamentale (stylisation MOMEL et codage INTSINT) 63
3.4 Exemple de génération d'un contour de fréquence fondamentale par le modèle de Fujisaki (image extraite de [42]). 65
3.5 Exemple de stylisation de la fréquence fondamentale (seuil $G = 0.32/T^2$) . 70
3.6 Bloc LTSM contenant une seule cellule [24] 75
3.7 Schéma descriptif du système de K.P. Li [78] 77
3.8 Illustration de la segmentation et de l'étiquetage (image extraite de [2]). . . 78

4.1 Fenêtres d'estimation des modèles autorégréssifs de l'algorithme de segmentation automatique. 85
4.2 Résultat de la segmentation sur la phrase « Confirmez le rendez-vous par écrit ». 86

4.3 Résultat de la segmentation en segments consonantiques, vocalique et de silence . . . 89
4.4 Paramètres de Ramus . . . 94
4.5 Paramètres de Grabe . . . 97
4.6 Exemple d'extraction de pseudo-syllabes . . . 100
4.7 Extraction de paramètres de durée . . . 101
4.8 Paramètres de durée extraits des pseudo-syllabes . . . 102
4.9 Paramètres intonatifs extraits des pseudo-syllabes (1) . . . 106

5.1 Histogrammes de répartition des paramètres de pseudo-syllabes . . . 116
5.2 Exemple de segmentation en phrase . . . 119
5.3 Exemple de quantification en demi tons . . . 120
5.4 Exemple de calcul de ligne de base . . . 121
5.5 Exemple d'obtention du résidu sur une phrase . . . 122
5.6 Approximation de F0 sur les pseudo-syllabes . . . 123
5.7 Exemple d'approximation de la fréquence fondamentale sur une phrase . . . 124
5.8 Exemple d'étiquetage de la fréquence fondamentale . . . 125
5.9 Exemple d'approximation et d'étiquetage de la courbe d'énergie . . . 126
5.10 Exemple d'étiquetage des mouvements prosodiques . . . 132
5.11 Exemple d'étiquetage de la fréquence fondamentale . . . 134
5.12 Système d'identification des langues par modélisation différenciée des consonnes et des voyelles . . . 139

6.1 Paramètres de durée extraits des pseudo-syllabes . . . 147
6.2 DP syllabique estimé par le nombre de voyelles par seconde . . . 157
6.3 DP phonémique estimé par le nombre de segments par seconde . . . 158

A.1 Exemple de calcul de R pour des variations de débit hydraulique sur plusieurs années . . . 170
A.2 Estimation de l'exponentielle de Hurst : calcul des différents R/S . . . 171
A.3 Estimation de l'exponentielle de Hurst : régression linéaire . . . 171

Liste des tableaux

1.1 Classification des langues selon différents auteurs dans le cadre de la théorie de l'isochronie 19

2.1 Classes de segments utilisées pour décrire les courbes d'énergie et de fréquence fondamentale pour le système OGI-ASP 41
2.2 Résultats du système du LIMSI 50

3.1 Codage des tons Intsint 63
3.2 Résultats du système prosodique d'Adami et comparaison avec un système phonotactique (% EER (Equal Error Rate)) 79

4.1 Comparaison de différents algorithmes de détection automatique de voyelles. 88
4.2 Description de l'ensemble d'apprentissage du jeu1 (MULTEXT). 91
4.3 Description de l'ensemble de test du jeu1 (MULTEXT). 91
4.4 Paramètres de Ramus : Expériences sur l'ensemble d'apprentissage 95
4.5 Paramètres de Ramus : Expériences sur l'ensemble de test 95
4.6 Paramètres de Ramus : Regroupement en classes rythmiques 96
4.7 Paramètres de Grabe : Expériences sur l'ensemble d'apprentissage 97
4.8 Paramètres de Grabe : Expériences sur l'ensemble de test 98
4.9 Paramètres de Grabe : Regroupement en classes rythmiques 98
4.10 Modèle de durées pseudo-syllabiques : Expériences sur l'ensemble d'apprentissage 103
4.11 Modèle de durées pseudo-syllabiques : Expériences sur l'ensemble de test . 103
4.12 Modèle de durées pseudo-syllabiques : Regroupement selon les classes rythmiques 104
4.13 Modèle intonatif pseudo-syllabique : Expériences sur l'ensemble d'apprentissage 107
4.14 Modèle intonatif pseudo-syllabique : Expériences sur l'ensemble de test . . 107
4.15 Modèle intonatif pseudo-syllabique : Regroupement en classes 107
4.16 Fusion des modélisations rythmiques et intonatives : Expériences sur l'ensemble d'apprentissage 108
4.17 Fusion des modélisations rythmiques et intonatives : Expériences sur l'ensemble de test 109
4.18 Fusion des modélisations rythmiques et intonatives : Regroupement en classes 109
4.19 Récapitulatif des expériences du chapitre 110

5.1 Modèle rythmique : Expériences sur l'ensemble de test 117
5.2 Modèle rythmique : Regroupement en classes rythmiques 117
5.4 Expériences d'identification des langues avec le modèle intonatif 126
5.5 Modèle intonatif : Expériences sur l'ensemble de test 127
5.6 Modèle intonatif : Regroupement en classes rythmiques 127
5.7 Modèle rythmique/intonatif : Expériences sur l'ensemble de test 128
5.8 Modèle rythmique/intonatif · Regroupement en classes rythmiques 128
5.9 Expériences avec les multigrammes intonation/durée 130
5.10 [Modèle intonation/durée : Expériences sur l'ensemble de test 130
5.11 Modèle intonation/durée : Regroupement en classes rythmiques 130
5.12 Expériences avec le modèle d'Adami 132
5.13 Modèle d'Adami : Expériences sur l'ensemble de test 133
5.14 Modèle d'Adami : Regroupement en classes rythmiques 133
5.15 Expériences avec le système segmental durée/intonation 135
5.16 Modèle prosodique (court terme) : Expériences sur l'ensemble de test . . . 135
5.17 Modèle prosodique (court terme) : Regroupement en classes rythmiques . . 136
5.18 Fusion des approches prosodiques statique et dynamique : Expériences sur l'ensemble de test 136
5.19 Fusion des approches prosodiques statique et dynamique : Regroupement en classes rythmiques 137
5.20 Fusion des approches prosodiques à court terme et à long terme : Expériences sur l'ensemble de test 137
5.21 Fusion des approches prosodiques à court terme et à long terme : Regroupement en classes rythmiques 138
5.22 Modèle PPRLM : Expériences sur l'ensemble de test 141
5.23 Modèle PPRLM : Regroupement en classes rythmiques 141
5.24 Récapitulatif des expériences du chapitre 142

6.1 Taux d'identifications correctes dans la tâche de discrimination entre paires de langues sur 10 langues du corpus OGI-MLTS 149
6.2 Description du corpus OGI MLTS 151
6.3 Moyenne et écart-type du DP syllabique (étiquetage manuel) 153
6.4 Moyenne et écart-type du DP phonémique (étiquetage manuel) 153
6.5 Corrélation entre les deux types de débits proposés (avec pauses) 154
6.6 Corrélation entre débits syllabiques réels et estimés par la détection des voyelles (avec pauses) 154
6.7 Corrélation entre débits phonétiques réels et estimés par la segmentation (avec pauses) 155
6.8 Corrélation entre les estimateurs de débits syllabiques et phonémiques (avec pauses) 155

B.1 Répartition des passages par locuteur en anglais sur MULTEXT. 176
B.2 Répartition des passages par locuteurs en allemand sur MULTEXT. 177
B.3 Répartition des passages par locuteur en espagnol sur MULTEXT. 177
B.4 Répartition des passages par locuteurs en français sur MULTEXT. 178

B.5 Répartition des passages par locuteur en italien sur MULTEXT. 179
B.6 Description de l'ensemble d'apprentissage du jeu1 (MULTEXT). 180
B.7 Description de l'ensemble de test du jeu1 (MULTEXT). 180
B.8 Description de l'ensemble d'apprentissage du jeu2 (MULTEXT). 181
B.9 Description de l'ensemble de test du jeu2 (MULTEXT). 181
B.10 Description de l'ensemble d'apprentissage du jeu3 (MULTEXT). 182
B.11 Description de l'ensemble de test du jeu3 (MULTEXT). 182

C.1 Modèle pseudo-syllabes, Expériences sur l'ensemble d'apprentissage (jeu 2) 185
C.2 Modèle pseudo-syllabes, Expériences sur l'ensemble de test (jeu 2) 186
C.3 Modèle pseudo-syllabes, Expériences sur l'ensemble d'apprentissage (jeu 3) 186
C.4 Modèle pseudo-syllabes, Expériences sur l'ensemble de test (jeu 3) 187
C.5 Modèle intonation, Expériences sur l'ensemble d'apprentissage (jeu 2) . . . 187
C.6 Modèle intonation, Expériences sur l'ensemble de test (jeu 2) 188
C.7 Modèle intonation, Expériences sur l'ensemble d'apprentissage (jeu 3) . . . 188
C.8 Modèle intonation, Expériences sur l'ensemble de test (jeu 3) 188

D.1 Modèle acoustique global centiseconde : Expériences sur l'ensemble d'apprentissage 192
D.2 Modèle acoustique global centiseconde ; Expériences sur l'ensemble de test 192
D.3 Modèle vocalique centiseconde : Expériences sur l'ensemble d'apprentissage 192
D.4 Modèle vocalique centiseconde : Expériences sur l'ensemble de test 193
D.5 Modèle consonantique centiseconde : Expériences sur l'ensemble d'apprentissage 193
D.6 Modèle consonantique centiseconde : Expériences sur l'ensemble de test . . 194
D.7 Modélisation différenciée consonnes/voyelles centiseconde : Expériences sur l'ensemble d'apprentissage 194
D.8 Modélisation différenciée consonnes/voyelles centiseconde : Expériences sur l'ensemble de test 195
D.9 Modélisation des consonnes voisées centiseconde : Expériences sur l'ensemble d'apprentissage 195
D.10 Modélisation des consonnes voisées centiseconde : Expériences sur l'ensemble de test 196
D.11 Modélisation des consonnes non voisées centiseconde : Expériences sur l'ensemble d'apprentissage 196
D.12 Modélisation des consonnes non voisées centiseconde : Expériences sur l'ensemble de test 197
D.13 Modélisation différenciée consonnes voisées/non voisées/voyelles centiseconde : Expériences sur l'ensemble d'apprentissage 197
D.14 Modélisation différenciée consonnes voisées/non voisées/voyelles centiseconde : Expériences sur l'ensemble de test 198
D.15 Modèle acoustique global segmental : Expériences sur l'ensemble d'apprentissage 198
D.16 Modèle acoustique global segmental : Expériences sur l'ensemble de test . . 199
D.17 Modèle consonantique segmental : Expériences sur l'ensemble d'apprentissage 199

D.18 Modèle consonantique segmental : Expériences sur l'ensemble de test . . . 200
D.19 Modèle vocalique segmental : Expériences sur l'ensemble d'apprentissage . 200
D.20 Modèle vocalique segmental : Expériences sur l'ensemble de test 201
D.21 Modélisation différenciée Consonnes/Voyelles segmentale : Expériences sur l'ensemble d'apprentissage 201
D.22 Modélisation différenciée Consonnes/Voyelles segmentale : Expériences sur l'ensemble de test 202
D.23 Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles segmentale : Expériences sur l'ensemble d'apprentissage 202
D.24 Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles segmentale : Expériences sur l'ensemble de test 203

G.1 Codage des fichiers dans OGI-MLTS. 216
G.2 Liste des fichiers disposant d'une annotation phonétique sur OGI-MLTS. . . 218
G.3 Durée totale des fichiers disposant d'une transciption phonétique sur OGI-MLTS. 219

Introduction

Le but de l'identification automatique des langues est de reconnaître la langue parlée par un locuteur inconnu, parmi un ensemble fini de langues, pour des énoncés d'une durée limitée (entre 3 et 50 secondes habituellement).

Avec l'ouverture mondiale des télécommunications, les services tels que les serveurs vocaux interactifs (services de transport, spectacle, hôtellerie, météorologie...) doivent évoluer vers la multilingualité. La consultation de bases de données documentaires multilingues (notamment sur internet), l'enseignement ou encore la traduction automatique sont autant de domaines où l'identification automatique des langues doit jouer un rôle.

Une langue est définie comme un regroupement de dialectes partageant un vocabulaire similaire et ayant des systèmes phonologiques et grammaticaux similaires. Par exemple, la langue française est constituée de dialectes comme le français méridional ou provençal (accent de Marseille), le français parisien, le français du sud ouest (accent toulousain), etc. Il existerait actuellement plus de 6000 langues parlées à travers le monde, (6809 sont référencées dans [55]) et plus de 10000 dialectes.

La prosodie, c'est-à-dire l'expression musicale de la parole, constitue une source d'information importante pour l'identification des langues, car elle est particulièrement présente dans les stratégies humaines de perception et de compréhension de la parole. Cependant sa modélisation et sa mise en œuvre relèvent toujours du défi. La prosodie est actuellement un des enjeux majeurs du traitement automatique de la parole.

La prosodie est très peu utilisée en identification automatique des langues. Pourtant, des théories linguistiques, confirmées par des expériences en perception, montrent qu'il existe des différences significatives entre les langues, ou du moins entre les familles de langues. L'objectif est d´étudier s'il est également possible, de même que les humains et en accord avec les théories linguistiques, d'utiliser les paramètres prosodiques dans le cadre de l'identification automatique des langues.

1 Les enjeux de l'identification automatique des langues

1.1 Enjeux applicatifs

L'identification automatique des langues est un thème de recherche apparu aux États-Unis au début des années 1970. La recherche dans ce domaine s'est intensifiée depuis le début des années 1990.

Les raisons de cette émergence sont liées à de nombreux développements de nature essentiellement applicative :
- une demande croissante pour des interfaces homme-machine,
- l'expansion des communications parlées dans un cadre multilingue,
- l'augmentation des performances des systèmes de reconnaissance automatique de la parole,
- l'enregistrement et la mise à la disposition de la communauté scientifique de corpus multilingues conséquents [93].

Notre époque est une ère de communication multilingue, qu'il s'agisse de communications entre humains ou entre humains et machines. Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres.

En ce sens, un des problèmes est la discrimination des langues en amont de systèmes de dialogue multilingue : le système a besoin de savoir quelle langue est parlée pour pouvoir comprendre et répondre à l'utilisateur. On peut distinguer deux approches possibles pour déterminer la langue et dialoguer dans celle-ci : exécuter en parallèle autant de systèmes de reconnaissance de la parole que de langues traitées par le système de dialogue, ou bien utiliser un système dédié à l'identification de la langue, qui permet de lister rapidement les langues les plus probables qui sont ensuite départagées par les systèmes de reconnaissance de la parole adaptés aux langues pré-sélectionnées. La dernière approche, en considérant les contraintes de temps-réel d'un tel système, permet la prise en compte d'un nombre bien plus important de langues.

Les applications envisageables pour l'identification automatique des langues sont multiples :
- Les interfaces homme-machine, pour la dictée vocale ou les serveurs vocaux,
- L'indexation par le contenu, avec la possibilité d'indexer des documents multimédia multilingues,
- L'aide au dialogue assisté par ordinateur, pour les standards téléphoniques, les services d'urgence ou les missions humanitaires.

1.2 Enjeux stratégiques

Les enjeux stratégiques se divisent en deux parties :
- Les communications internationales : pour les instances internationales (ONU, parlement européen) et les missions humanitaires,
- Le renseignement militaire : identification ou vérification de la langue ou du dialecte.

1.3 Enjeux scientifiques

L'identification automatique des langues peut permettre de confirmer ou d'infirmer les théories linguistiques portant sur les différences entre les langues. Ces théories reposent souvent sur un nombre limité d'expériences, certaines sont parfois plus basées sur des intuitions que sur des faits réels. Il est important de procéder à des expérimentations sur de plus larges bases de données, tâche rendue possible par le traitement automatique de la parole. Les retombées sur la connaissance des langues sont nombreuses. À titre d'exemple, signalons deux conséquences immédiates :
- Les typologies linguistiques pourront être comparées aux typologies automatiques.
- De manière corrélée, une distance linguistique entre différentes langues peut émerger et conduire à préciser la distance entre les dialectes d'une même langue.

L'étude de la prosodie contribue à la compréhension des processus cognitifs impliqués dans la production et la perception du langage.

2 Les techniques actuelles d'identification automatique des langues

L'identification automatique des langues est un domaine de recherche qui a connu de grandes avancées lors des campagnes d'évaluation Nist de 1993 à 1996 [128]. Les systèmes d'identification automatique des langues conçus à cette époque permettent de discriminer une dizaine de langues sur de courts échantillons sonores (environ 45 secondes) avec un taux d'erreur de l'ordre de 10 %. Récemment, la dernière campagne d'évaluation Nist 2003 a montré que les systèmes les plus performants obtiennent un taux d'erreur de l'ordre de 3 % pour des échantillons de 30 secondes.

Parmi les informations disponibles pour identifier une langue, les informations phonétiques et phonotactiques (caractéristiques des sons d'une langue et règles d'enchaînement de ces sons) sont les plus fréquemment utilisées tandis que les informations prosodiques (intonation, rythme, accentuation) sont souvent négligées [92]. Une des principales raisons est la maîtrise technique des modèles phonétiques et phonotactiques issus de la reconnaissance automatique de la parole, domaine bénéficiant de plusieurs décennies d'investissement intellectuel. Pourtant, des expériences en perception (cf. [13]) montrent que l'oreille humaine permet d'identifier les langues à partir de leur seule prosodie mettant ainsi en

avant le pouvoir discriminant de ces traits et l'intérêt manifeste de leur exploitation dans des systèmes d'identification automatique des langues.

3 Le projet de recherche

Le thème de l'identification automatique des langues a été abordé à l'Institut de Recherche en Informatique de Toulouse depuis 1996 dans le cadre du projet « Discrimination automatique multilingue » financé par la DGA. Ce projet réunissait trois instituts de phonétique (ICP Grenoble, ILPGA Paris, DDL Lyon) et l'IRIT, responsable du projet. L'étude était orientée suivant deux axes :

- la recherche d'une typologie des langues selon des indices discriminants et détectables automatiquement,
- une meilleure connaissance des systèmes vocaliques (nombre de voyelles, représentation fréquentielle, fréquence d'apparition) et la recherche d'un modèle probabiliste de l'espace acoustique vocalique pour chaque langue.

Dans le cadre de cette étude, François Pellegrino a montré dans sa thèse, qu'avec une simple détection automatique des noyaux vocaliques présents dans le signal et une modélisation stochastique de type mélange de lois gaussiennes, des performances intéressantes en identification automatique des langues pouvaient être atteintes sur le corpus de la parole téléphonique OGI-MLTS (environ 80% d'identifications correctes sur 6 langues) [98].

Les directions de recherche de l'IRIT sont de deux natures :

- la poursuite de la recherche d'indices discriminants et détectables automatiquement,
- la définition d'un système complet d'identification automatique de la langue, regroupant des aspects acoustiques, phonotactiques et prosodiques, après avoir cerné ces paramètres discriminants. Ce futur système s'organisera en trois modules :
  - un module de décodage acoustico-phonétique généralisé. Il s'agit d'élargir l'approche vocalique initiée par François Pellegrino à l'ensemble des sons de chaque langue en proposant un modèle probabiliste global ou plusieurs modèles différenciés [101],
  - un module phonotactique pour gérer les séquences de sons. Il utilise des modèles probabilistes multigrammes [29] qui sont une extension des modèles bigrammes ou trigrammes,
  - un module prosodique. Ce module exploite les critères discriminants du point de vue prosodique, principalement sur le rythme et l'intonation.

Mes travaux de recherches s'insèrent dans la première direction. Les recherches que j'effectue concernent en priorité la prosodie. Il s'agit d'un travail exploratoire, l'objectif étant d'évaluer la pertinence de l'emploi de paramètres prosodiques dans le cadre de l'identification automatique des langues. Néanmoins, je contribue à la construction du système complet d'identification automatique des langues.

4 Démarche scientifique & organisation du mémoire

Dans le premier chapitre, nous visitons quelques définitions de la prosodie, afin de comprendre le rôle qu'elle joue dans la communication orale. Nous expliquons les difficultés rencontrées lors de l'étude de la prosodie, particulièrement au niveau de l'extraction de paramètres. Afin de déterminer si des critères prosodiques peuvent être caractéristiques des langues, nous étudions ensuite quelques théories linguistiques portant sur les propriétés rythmiques et intonatives des langues. Nous relevons des expériences en perception évoquant les stratégies employées par les humains.

Le deuxième chapitre décrit les approches employées couramment pour l'identification automatique des langues. Nous résumons les différents critères pouvant être utilisés pour l'identification des langues. Quelques méthodes récentes sont ensuite présentées, notamment au travers des résultats de la campagne d'évaluation Nist 2003. La plupart de ces approches, dont le but est avant tout la performance, ne considèrent pas les informations prosodiques.

Le troisième chapitre est consacré aux méthodes employées ou susceptibles de l'être pour identifier les langues par leur prosodie. Nous voyons que le problème majeur rencontré lors de la mise en œuvre de ces méthodes est celui de l'automatisation du traitement, qui permettrait de traiter des bases de données représentatives.

Nous proposons dans le quatrième chapitre des prétraitements automatiques permettant de tester quelques approches vues au chapitre précédent. L'automatisation de ces approches nous permet d'obtenir des résultats préliminaires encourageants. Nous définissons une pseudo-unité prosodique de type syllabique, la pseudo-syllabe. Cette unité est caractérisée par des paramètres relevant de sa nature prosodique. Des modèles statistiques, les modèles de mélanges de lois gaussiennes, sont employés pour modéliser les caractéristiques extraites des pseudo-syllabes pour chaque langue. Les résultats sont comparés à ceux obtenus avec d'autres approches prosodiques.

Toutefois, les modèles statistiques (modèles de mélange de lois gaussiennes) que nous employons pour modéliser les caractéristiques des pseudo-syllabes sont intrinsèquement des modèles statiques. En effet, chaque pseudo-syllabe est caractérisée par un unique vecteur de paramètres, ce qui ne correspond pas à la réalité perceptive de la prosodie, qui est par nature continue. Nous devons donc employer des modèles de nature dynamique afin de prendre en compte cet aspect temporel.

Dans le chapitre cinq, en prenant l'exemple du travail décrit d'Adami [2], nous nous servons de paramètres calculés sur chaque pseudo-syllabe pour leur attribuer des étiquettes indiquant le sens des mouvements de fréquence fondamentale et d'énergie. Les enchaînements de ces étiquettes sont modélisés par des modèles multigrammes, qui permettent d'identifier les séquences les plus fréquentes pour chaque langue et de leur associer une fréquence d'occurrence. La fin du chapitre est consacré aux autres méthodes d'identification des langues développées à l'IRIT. En effet, les recherches menées à l'IRIT sur l'identifica-

tion des langues ne concernent pas que la prosodie. Nous continuons le travail de François Pellegrino et de Jérôme Farinas concernant les modèles acoustiques et phonotactiques. Diverses expériences sont réalisées, et les résultats sont comparés.

Enfin, nous abordons le problème de la parole spontanée dans le dernier chapitre. L'application directe de nos algorithmes - développés sur un corpus de parole lue - à la parole spontanée ne donne pas de résultats très satisfaisants. Ces résultats sont probablement dûs à des problèmes de variabilités liés à la spontanéité du discours. Afin de prendre en compte cet aspect, nous proposons une mesure automatique du débit de parole.

Chapitre 1

Définitions et motivations sur l'emploi de la prosodie

Sommaire

1.1 La prosodie 12
1.1.1 Définition de la prosodie 12
1.1.2 Interprétation de la prosodie 12
1.2 Extraction de paramètres prosodiques 13
1.2.1 Énergie 13
1.2.2 Durée 14
1.2.3 Fréquence fondamentale 16
1.3 Intérêt de la prosodie pour l'identification des langues 17
1.3.1 Le rythme et la théorie de l'isochronie 18
1.3.2 L'intonation 20
1.3.3 Théories cognitives 21
1.3.4 Expériences en perception 21
1.4 Conclusion 24

Le terme « prosodie » fait partie du vocabulaire de nombreuses communautés scientifiques, en sciences humaines comme en sciences pour l'ingénieur. Il est important et nécessaire de le définir, tout d'abord de façon informelle et de manière plus précise, de sorte à comprendre le rôle de la prosodie dans la communication orale. Nous expliquons les difficultés rencontrées dans l'étude de la prosodie du point de vue de l'ingénieur, notamment pour l'extraction de paramètres. Nous étudions finalement l'interêt d'exploiter la prosodie pour l'identification des langues, au travers d'éléments théoriques et d'expériences en perception.

1.1 La prosodie

1.1.1 Définition de la prosodie

La prosodie est liée à l'impression musicale que fournit un locuteur lorsqu'il parle. Différentes informations, linguistiques ou non, sont exprimées simultanément par la prosodie :

- le sens de certains mots (désambiguïsation lexicale),
- le mode de la phrase (interrogation, affirmation, etc.),
- l'état d'esprit du locuteur (émotions, stress, etc.).
- la manière de parler du locuteur (caractéristique du locuteur),
- la prononciation d'une langue.

Les phénomènes observés sont principalement :

- l'intonation,
- l'accentuation,
- le rythme,
- le débit,
- les pauses.

Chacun de ces phénomènes prosodiques se manifeste par des variations au niveau de la fréquence fondamentale du signal de parole émis, de la hauteur, de l'intensité et/ou de la durée des sons.

D'après DiCristo [32], « La prosodie est une branche de la linguistique consacrée à la description et à la représentation formelle des éléments de l'expression orale tels que les accents, les tons, et l'intonation, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F_0), de la durée et de l'intensité (paramètres prosodiques physiques).

Ces variations sont perçues par l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs).

Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation dans le flux du discours. »

1.1.2 Interprétation de la prosodie

L'étude de la prosodie a bénéficié d'un intérêt grandissant au cours des trente dernières années. Les recherches en prosodie concernent d'une part son intégration dans le champ de la linguistique formelle (théories phonologique et "autosegmentale"), et d'autre part son impact dans la "mouvance cognitique" notamment dans les secteurs de la psycholinguistique et des neurosciences (étude de l'acquisition du langage par l'enfant [112] et des

systèmes de traitement qui participent à l'encodage et à la compréhension de la parole chez l'adulte [10]).

La fonction principale de la prosodie est avant tout une fonction d'assistance à l'encodage et au décodage de la parole. La continuité prosodique apporte également une contribution significative à l'intelligibilité et à la perception de la parole.

Outre ces fonctions générales d'assistance à la production et à la perception de la parole, la prosodie assume un ensemble de fonctions linguistiques, paralinguistiques et extralinguistiques qui consistent à structurer la langue et le discours, à contextualiser les énoncés et leurs auteurs, à réguler les interactions verbales, à exprimer les émotions et à caractériser le sujet parlant ainsi que le style discursif qu'il adopte.

L'interprétation de la prosodie est donc rendue difficile puisque les signaux prosodiques sont hautement polysémiques. Cela représente un obstacle majeur à l'explication des relations entre formes et fonctions qui doit être l'aboutissement d'une véritable « cognitique » de la prosodie.

1.2 Extraction de paramètres prosodiques

Les trois paramètres prosodiques classiquement extraits du signal acoustique sont l'énergie, la durée et la fréquence fondamentale. La première étape de l'extraction de paramètres prosodiques est généralement une analyse à court terme du signal, en faisant l'hypothèse que sur une fenêtre de faible longueur le signal est quasi-stationnaire, ce qui signifie que les caractéristiques statistiques du signal évoluent peu.

En général une fenêtre d'analyse de 30 ms, appelée trame, est utilisée. L'analyse est répétée à intervalles réguliers, typiquement toutes les 10 ms. Compte tenu du caractère suprasegmental[1] de la prosodie, une analyse sur 30 ms ne suffit pas.

C'est pourquoi il est nécessaire de calculer d'autres paramètres à partir d'une analyse sur plusieurs trames successives afin de traduire le rythme, l'intonation et l'accentuation.

1.2.1 Énergie

L'énergie est un paramètre couramment utilisé en traitement du signal.

L'énergie à court terme d'un signal échantillonné sur une fenêtre de longueur T $(s_t)_{t=1,T}$ est définie par :

$$E = \frac{1}{T} \sum_{t=1}^{T} s_t^2 \tag{1.1}$$

[1] employé ici dans le sens d'une propriété portée par une unité plus grande que le segment, ou liée aux relations entre segments distants.

Étant donnée sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db} = 10 \times \log_{10}(\frac{1}{T}\sum_{t=1}^{T} s_t^2) \tag{1.2}$$

Pour un signal échantillonné de longueur infinie, on calcule l'énergie à court terme sur des fenêtres glissantes. Ces fenêtres sont étroites, de l'ordre de 5 à 10 ms.

Pour éliminer la variabilité de ce paramètre, due en partie à des conditions d'enregistrement différentes d'un enregistrement à l'autre (une simple variation de la distance entre le locuteur et le microphone suffit pour être source de perturbation de l'énergie), l'énergie peut être normalisée par rapport au maximum observé sur le signal global (plusieurs phrases).

1.2.2 Durée

Les indices de durée classiques supposent généralement la donnée d'une segmentation, c'est-à-dire la connaissance des frontières d'unités dont on désire mesurer la durée.

La durée et la nature de ces unités a fait l'objet de nombreuses études, principalement motivées par la nécessité de la modéliser dans des systèmes de synthèse de la parole. Parmi les unités qui ont servi de base à ces modélisations, on en trouve principalement quatre :

Le phonème : le modèle de Klatt [68] est à l'origine de nombreux développements en prédiction de la durée basée sur le phonème. Ce modèle part du principe que chaque phonème possède une durée intrinsèque qui est modifiée par un coefficient de rétrécissement :

$$Durée_{phonème} = Durée_{min} + \frac{(Durée_{intrinsèque} - Durée_{min}) * Rétrécissement}{100} \tag{1.3}$$

où

- $Durée_{min}$ est calculée en fonction de $Durée_{intrinsèque}$ selon l'accentuation (pour des phonèmes non accentués $Durée_{min} = 0,45 * Durée_{intrinsèque}$ et pour un phonème accentué $Durée_{min} = 2 * Durée_{intrinsèque}$)
- *Rétrécissement* est le pourcentage de rétrécissement, représentant les facteurs qui influent sur la durée segmentale tels que le contexte phonétique et l'environnement syntaxique.

La syllabe : la syllabe se compose d'une attaque et d'une rime [53]. La rime est constituée d'un noyau et d'une coda (figure 1.1). Attaque et coda sont facultatives et peuvent être formées d'une ou plusieurs consonnes, alors que le noyau est typiquement composé d'une voyelle. La durée syllabique peut être calculée à partir de la

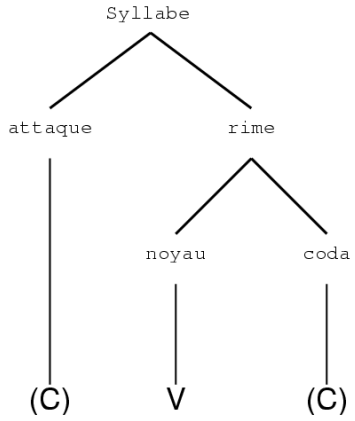


Fig. 1.1 : Schéma type d'une syllabe

durée de la voyelle si l'on se réfère à l'étude psycholinguistique sur la « structure de performance » de Monnin et Grosjean [90]. Elle peut également être calculée en effectuant la somme des durées des phonèmes qui composent la syllabe. Campbell [19] propose de calculer cette durée en introduisant un facteur d'élasticité sur les phonèmes d'une syllabe :

$$Durée_{syllabe} = \sum_{i=1}^{n} \exp(\mu_i + z * \sigma_i) \tag{1.4}$$

où n est le nombre de phonèmes dans la syllabe, la paire (μ_i, σ_i) est liée à la moyenne et à l'écart-type des durées associées au phonème i et z au facteur d'allongement associé à la syllabe (appelé $z - score$).

Le pied : il s'agit du plus petit groupement rythmique formé d'une suite de syllabes inaccentuées suivie d'une syllabe accentuée [72]. Witten [127] a proposé un système de prédiction des durées pour l'anglais et Kohler [69] en a proposé un pour l'allemand, où la durée phonémique est une fonction linéaire de la durée de la syllabe, elle même fonction linéaire de la durée du pied.

Le GIPC : cette unité introduite par Barbosa [12] est basée sur la notion de « Perceptual-Center » [83]. Des expériences perceptives de synchronisation portant sur des syllabes, ont mis en évidence l'existence d'un point qui correspondrait au moment psychologique de perception de chaque syllabe. Ce point est appelé P-Centre. Il serait situé au début de la réalisation vocalique d'une syllabe accentuée [4]. Le Groupe Inter P-Centre est l'ensemble des réalisations phonémiques comprises dans l'intervalle de temps défini entre deux P-Centres consécutifs. Bien qu'il soit difficile à repérer en parole continue, le GIPC constitue une alternative intéressante à la syllabe pour représenter le rythme au niveau de la phrase [91].

Un état de l'art plus exhaustif sur les différentes méthodes d'extraction relatives à ces unités se trouve dans la thèse de Barbosa [12]. La plupart de ces méthodes mettent en œuvre des systèmes d'extraction des unités fortement dépendants de la langue et basés sur une segmentation manuelle des données.

1.2.3 Fréquence fondamentale

La fréquence fondamentale (ou F_0) correspond à la fréquence de vibration des cordes vocales. Les algorithmes d'extraction de F_0 utilisent une représentation temporelle ou spectrale du signal.

Les méthodes temporelles exploitent la similarité du signal d'une période à l'autre pour identifier la période fondamentale. Il est parfois possible de repérer manuellement la période fondamentale T_0 ($F_0 = 1/T_0$) directement sur le signal (par exemple pour les signaux des figures 1.2).

Dans le domaine fréquentiel, les harmoniques de la fréquence fondamentale sont localisées. Cette propriété du signal peut être visualisée sur un spectrogramme avec un spectre à bande étroite (cf. figure de gauche de 1.2) où l'analyse spectrale est réalisée sur une fenêtre de largeur bien supérieure à T_0, ce qui induit une bonne représentation fréquentielle des harmoniques.

Sur un spectrogramme avec un spectre à large bande (cf. figure de droite de 1.2), la périodicité de la fréquence fondamentale se repère selon l'axe temporel. L'analyse spectrale est réalisée sur une fenêtre d'analyse de largeur bien inférieure à T_0, ce qui met en évidence les cycles d'ouverture-fermeture de la glotte.

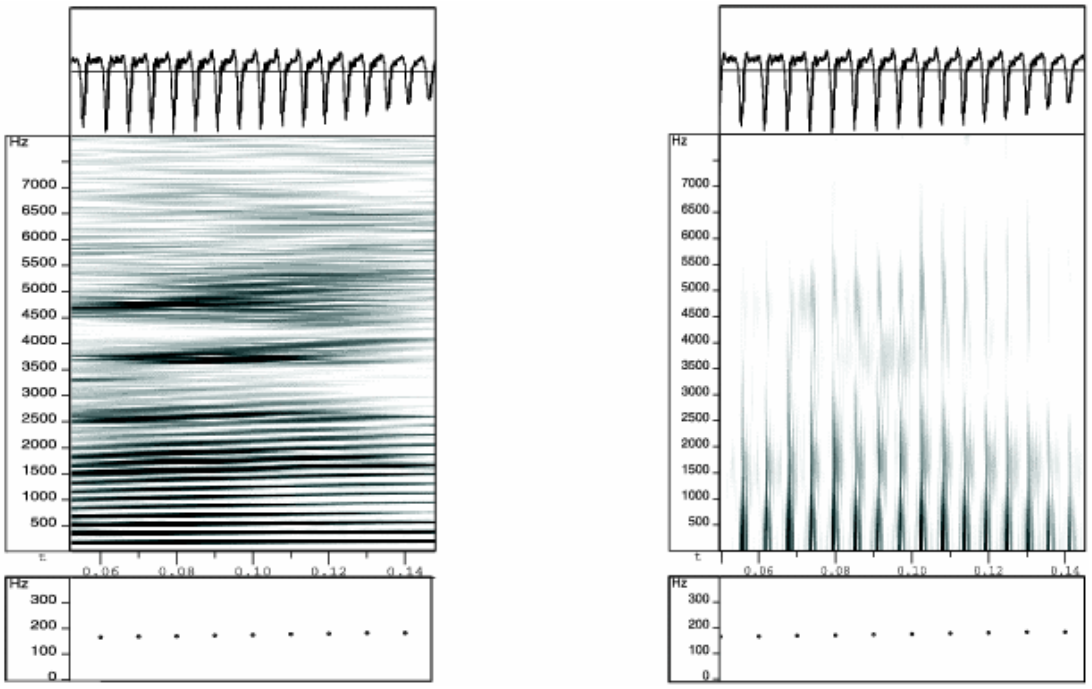


Fig. 1.2 : Représentation du signal (en haut), du spectrogramme (spectre large bande à droite et bande étroite à gauche) et des valeurs de F_0 (en bas).

De nombreuses recherches ont été menées dans le domaine de l'extraction de la fréquence fondamentale des signaux de parole. On peut citer l'ouvrage de référence de Hess [58], où un grand nombre d'algorithmes est détaillé.

Un algorithme d'extraction de la fréquence fondamentale se décompose en trois phases successives :

1. un prétraitement et un changement de représentation,
2. l'extraction du fondamental,
3. un post-traitement visant à corriger les erreurs.

Le prétraitement vise à optimiser les caractéristiques du signal en vue de l'extraction en utilisant un filtrage passe-bas, une pré-accentuation ou un filtrage non linéaire. Ensuite des transformations sont appliquées pour adapter la représentation du signal au domaine du traitement fondamental (temporel, temporel à court terme, fréquentiel...).

La deuxième phase consiste à extraire la fréquence fondamentale et dépend donc du domaine utilisé. Généralement cela revient à optimiser une fonction de la fréquence fondamentale (fonction de coût, résultat d'une transformation, corrélation, densité de probabilité).

La phase de post-traitement a pour but de diminuer les erreurs qui sont de plusieurs types :

- les erreurs de voisement[2] : lorsqu'une valeur de F_0 a été trouvée sur une zone non-voisée, ou lorsque aucune n'a été trouvée sur une zone voisée.
- les erreurs grossières (« gross-errors » en anglais) : la fréquence fondamentale correspond à une harmonique ou une sous-harmonique. Ce type d'erreur peut facilement être corrigé en tenant compte du voisinage ou en effectuant un lissage.
- les erreurs fines : la valeur trouvée est située à plus ou moins 10 % de la valeur réelle.

1.3 Intérêt de la prosodie pour l'identification des langues

Deux aspects des la prosodie motivent l'utilisation de la prosodie dans le cadre de l'identification des langues :
- l'aspect rythmique,
- l'intonation.

Quelques résultats en sciences cognitives et de expériences de perception confortent cet a priori.

[2]une zone de parole est dite voisée quand il y a vibration des cordes vocales

1.3.1 Le rythme et la théorie de l'isochronie

Le rythme des langues est défini comme un effet impliquant la récurrence isochrone, c'est-à-dire à intervalles réguliers, d'un certain type d'unité de discours (*"rhythm has been defined as an effect involving the isochronous recurrence of some type of speech unit"* [50]).

La théorie de l'isochronie

L'isochronie est l'organisation de la parole en portions de durées égales ou équivalentes. Suivant le type d'unité considérée, la théorie de l'isochronie permet de classer les langues en trois grandes classes :

- les langues *stress-timed* ou accentuelles,
- les langues *syllable-timed* ou syllabiques,
- les langues *mora-timed* ou moraïques[3].

Les langues syllabiques partagent la caractéristique de posséder des intervalles réguliers entre les syllabes, tandis que les langues accentuelles ont des intervalles réguliers entre les syllabes accentuées et, pour les langues moraïques, les mora successives sont quasiment égales en termes de durée.

Ce point de vue a été rendu populaire par Pike [74] et plus tard par Abercrombie [1]. De leur point de vue, la distinction entre langues accentuelles et langues syllabiques est strictement catégorique, les langues ne pouvant pas être plus ou moins accentuelles ou syllabiques.

Mise en doute de la théorie

Cette théorie est mise en doute par les expériences (notamment par Roach [113] et Dauer [27]). Malgré sa popularité auprès des linguistes, l'hypothèse des classes rythmiques est contrariée par de nombreuses expériences empiriques. Cet échec oblige quelques chercheurs à passer de l'isochronie « objective » à l'isochronie « subjective ». Ces chercheurs (Beckman [14] par exemple) décrivent la régularité physique de l'isochronie comme une tendance. La véritable isochronie est décrite comme une contrainte, et la réalisation d'unités isochrones est perturbée par les caractéristiques phonétiques, phonologiques et grammaticales des langues.

D'autres chercheurs concluent que l'isochronie est principalement un phénomène perceptuel (par exemple Lehiste [77]). Ils prétendent que les différences de durées mesurées entre les intervalles interaccentuels ou les durées de syllabes sont bien en dessous du seuil de perception. Dans ce cadre, l'isochronie peut être acceptée comme un concept relatif à la perception de la parole.

[3]Une more (ou mora) est une sous-unité de la syllabe constituée par une voyelle courte et les consonnes la précédant.

La solution ?

Le manque de preuves empiriques concernant la théorie de l'isochronie conduit Dauer [27] à proposer un nouveau système de classification rythmique. De son point de vue, les locuteurs n'essaient pas d'égaliser les intervalles interaccentuels ou intersyllabiques, mais les langues sont plus ou moins accentuelles ou syllabiques. Dauer suggère que les syllabes proéminentes interviennent à intervalles réguliers en anglais mais aussi en espagnol, les syllabes accentuées en anglais étant plus saillantes qu'en espagnol. En conséquence, la diversité rythmique résulte de la combinaison de faits phonologiques, phonétiques, lexicaux et syntaxiques. La structure syllabique (complexité), la présence ou l'absence de réduction vocalique, et l'accentuation des mots sont appropriées pour définir les différences rythmiques. Dans les langues accentuelles, les structures syllabiques sont plus variées, tandis que dans les langues syllabiques, la réduction vocalique apparaît plus rarement.

Nespor [95] introduit la notion de langues rythmiquement intermédiaires qui associent des propriétés associées au rythme accentuel et d'autres associées au rythme syllabique. Comme exemple, elle cite le polonais (classé comme accentuel alors qu'il ne possède pas de réduction vocalique) et le catalan (syllabique mais possédant la réduction vocalique).

Récapitulatif de la classification rythmique des langues et dialectes selon la théorie de l'isochronie

Les sources sont données en référence et les articles cités dans les sources entre parenthèses.

Tab. 1.1 : Classification des langues selon différents auteurs dans le cadre de la théorie de l'isochronie

Langue	**Classification**	**Référence**
Anglais américain	Accentuelle	[65]
Anglais britannique	Accentuelle	(Classe 1939, Pike 1946, Abercrombie 1967) [50]
Arabe	Accentuelle	(Abercrombie 1967) [107]
Allemand	Accentuelle	(Kohler 1982) [50]
Hollandais	Accentuelle	(Ladefoged 1975, Smith 1976) [50] & [107]
Polonais	Accentuelle	(Rubach & Booik (1985)) [107]
Russe	Accentuelle Plutôt accentuelle	(Abercrombie 1967)[107] [65]
Thailandais	Accentuelle	(Luangthongkum 1977) [50]
Tamil	Syllabique	(Corder 1973, Asher 1985) [50]
Espagnol	Syllabique	(Pike 1946, Hockett 1958) [50]
Français	Syllabique	(Abercrombie 1967, Catford 1977) [50]
...	...	...

Langue	**Classification**	**Référence**
Italien	Syllabique	(Bertinetto 1981)[107]
Anglais de Singapour	Syllabique	(Tongue 1979, Platt and Weber 1980) [50]
Telegu	Syllabique	(Abercrombie 1967)[107]
Yoruba	Syllabique	(Abercrombie 1967) in [107]
Hindi	Syllabique	(O'Connor 1973 dans Dauer 1983) [107]
Chinois (Shangaï)	Syllabique	[65]
Finnois	Plutôt syllabique	[65]
Farsi (Persan)	Syllabique	[65]
Vietnamien	Plutôt syllabique	[65]
Japonais	Moraïque	(Bloch 1942, Han 1962) [50]
Tamoul	Moraïque	(Steever 1987) [107]
Alyaarra, Aranda and Pauite	Moraïque	(Nazzi 1997) [107]
Polonais	Mélangée	(Dauer 1987, Nespor 1990) [50]
Catalan	Mélangée Syllabique	(Dauer 1987, Nespor 1990) [50] (Mehler et al. 1993) [107]
Estonien	Non classée	[50]
Grec	Non classée Syllabique Plutôt syllabique	[50] (Arvaniti 1994) [107] [65]
Luxembourgois	Non classée	[50]
Malais	Non classée	[50]
Chinois (Mandarin)	Non classée	[50]
Roumain	Non classée	[50]
Gallois	Non classée	[50]

1.3.2 L'intonation

Deux grandes classes de langues sont caractérisées selon leur utilisation de l'intonation :
- les langues à tons, utilisant la fréquence fondamentale pour distinguer plusieurs mots employant la (ou les) même(s) syllabe(s) (en Mandarin par exemple),
- les autres langues, qui ne se servent de l'intonation que pour décrire la modalité des phrases.

D'après Cummins [23], les essais de distinction entre les langues employant uniquement la fréquence fondamentale ont eu un succès modéré. Cela peut s'expliquer de deux façons :
- On peut imaginer une discrimination basée sur l'utilisation de tons lexicaux (Mandarin) ou non (Anglais), mais des cas intermédiaires existent (dialectes Coréens) ; ils sont usuellement considérés comme représentant des états transitoires entre les langues d'une classe et celles d'une autre.
- Les phénomènes liés aux accents et intonations de phrases sont moins exploitables. Il

existe de multiples théories sur l'intonation de phrase qui ne sont pas en accord. La situation est rendue plus complexe encore par les études sur les rôles non-linguistiques de l'intonation, comme par exemple pour rendre compte d'émotions. Plusieurs études s'accordent ici sur une classification par degrés plutôt qu'une séparation en classes distinctes [56].

1.3.3 Théories cognitives

MacNeilage and Davis (2000) [81] ont introduit, il y a quelques années, la théorie « *Frame/Content* » de l'évolution de la production de parole. Cette théorie est d'après ces auteurs un antidote à la tendance relativement admise de considérer le langage comme un « embarrassement pour la théorie évolutionnaire ». Elle contrarie les théories développées par Chomsky qui impliquent une apparition instantanée du langage, qui résulterait d'une mutation brutale dotant le cerveau de la capacité de manier sans apprentissage les composants du langage.

L'hypothèse de base est que la parole diffère des communications vocales des autres mammifères par le fait que seuls les humains surimposent une alternance rythmique continuelle entre une bouche ouverte et fermée (la *frame*) dans le mécanisme de production de parole. Des mouvements de mastication auraient évolué vers des mouvements destinés à la commmunication visio-faciale chez les primates non humains évolués. Par la suite, l'évolution de certaines zones du cerveau aurait permis de moduler ces mouvements cycliques pour aller vers la production de différentes consonnes et voyelles (le *content*).

Les mêmes auteurs trouvent également des échos ontologiques à leur théorie en étudiant les babillage des enfants. Durant la période s'étalant de 7 à 18 mois, ils montrent que la production de babillages redupliqués, de babillages variés et plus tard des premiers mots, implique une base, fournie par l'oscillation mandibulaire, et graduellement améliorée par le développement d'un contrôle plus précis des autres articulateurs.

Pour conclure, la théorie Frame/Content, confirmée par les résultats expérimentaux, outre le fait d'expliquer l'origine du rythme syllabique, souligne son rôle dans le procédé de production de la parole.

1.3.4 Expériences en perception

Depuis les deux dernières décennies, de nombreuses expériences montrent l'efficacité des êtres humains pour la reconnaissance des langues (voir [13]). Trois types majeurs de paramètres aident les humains à identifier les langues :

1. Les paramètres segmentaux (les propriétés acoustiques des phonèmes et leur fréquence d'occurrence),
2. Les paramètres supra-segmentaux (phonotactique, prosodie),

3. Les paramètres de haut niveau (lexique, morpho-syntaxe).

En ce qui concerne les paramètres prosodiques, de nombreuses expériences tentent de mettre en avant les capacités humaines à distinguer les langues en n'en gardant que les propriétés rythmiques ou intonatives. Les sujets sont des adultes naïfs ou entraînés, des enfants, des nouveau-nés, et même des primates [112].

Il s'agit en général de dégrader un enregistrement de parole au moyen de filtrage ou de resynthèse en ne laissant que peu d'indices aux sujets qui doivent identifier la langue. Par exemple, toutes les syllabes sont remplacées par une syllabe unique "/sa/" dans les expériences de Ramus [109], [110] ou [112]. Dans d'autres cas, le passage de la parole à travers un filtre passe-bas (fréquence de coupure 400 Hz) est utilisé pour dégrader volontairement le signal de parole [41].

D'autres auteurs (Komatsu [70]) proposent différentes méthodes de dégradation du signal de parole afin de ne garder que des indications ciblées (intensité ou intonation ou rythme).

Quelques expériences sont décrites plus en détail ci-dessous.

Expériences de Ramus : discrimination des langues par les nouveau-nés

Un ensemble d'expériences, décrites dans [112], prouvent que les nouveau-nés sont capables de faire la différence entre leur langue maternelle et une autre langue dès leurs premiers jours, pour peu qu'il y ait des différences entre ces deux langues au niveau suprasegmental. Que les nouveau-nés prennent les informations uniquement du rythme ou à la fois du rythme et de l'intonation est une question importante. Cependant la réponse n'est pas évidente à donner. Il est probable que les deux informations fournissent des éléments permettant de déterminer la langue et qu'ils sont pondérés en fonction des conditions expérimentales (langue, bruit, débit de parole) et peut être en fonction de stratégies individuelles.

Expériences de Frota [41] : discrimination Portugais Européen / Portugais du Brésil

Dans les expériences décrites dans [41], le corpus est composé de phrases déclaratives courtes enregistrées dans de bonnes conditions. Toutes les phrases sont passées à travers un filtre passe-bas de fréquence de coupure 400 Hz. Deux versions des phrases filtrées ont été créées :

- la première est la sortie directe du filtre,
- la deuxième est la sortie du filtre pour laquelle le contour de la fréquence fondamentale est rendu plat.

Les expériences de discrimination entre le portugais brésilien et le portugais européen ont été menées sur un panel d'étudiants ayant pour langue maternelle le portugais eu-

ropéen. Lors de cette expérience, les deux langues ne sont discriminées avec succès que lorsque le contour de la fréquence fondamentale n'est pas altéré. Avec une intonation plate, les résultats ne sont pas significativement différents de la chance.

Dans une deuxième expérience, l'ensemble des phrases de test est composé d'enregistrements en hollandais , en espagnol, en portugais brésilien et en portugais européen. Les sujets doivent décider si la langue entendue est du hollandais ou de l'espagnol. Les résultats montrent que le portugais du brésil et le portugais européen sont considérés comme de l'espagnol et sont bien discriminés par rapport au hollandais. Un autre résultat de cette expérience montre la non-discrimination entre les deux types de portugais. Les résultats obtenus lors des expériences avec les signaux possédant un contour de la fréquence fondamentale plat montrent que la discrimination entre le portugais européen et le hollandais ne dépend pas de la présence ou de l'absence de ce paramètre prosodique.

Pour résumer, les sujets parviennent à distinguer le portugais du brésil du portugais européen : les expériences montrent que les sujets arrivent à distinguer ces deux langues grâce aux motifs intonatifs. Dans une autre expérience, les portugais européens et brésiliens sont facilement distingués du hollandais, mais les deux variétés de portugais ne sont plus discriminées.

Expériences de Komatsu

Dans les expériences décrites dans [70], les langues considérées sont le chinois (mandarin), l'anglais, le japonais et l'espagnol. Elles sont choisies pour leur représentativité des différents types prosodiques :

- le chinois est une langue à rythme accentuel et à accent lexical tonal,
- l'anglais est une langue à rythme accentuel et à accent lexical non tonal,
- l'espagnol est une langue à rythme syllabique et à accent lexical non tonal,
- le japonais est une langue à rythme moraïque et possédant un accent lexical non tonal.

Les enregistrements utilisés sont soit extraits de la base de données MULTEXT [20] (voir §4.3.1), soit enregistrés suivant le même protocole (pour le mandarin et le japonais).

Six types de stimuli sont crées à partir des enregistrements, avec différents niveaux de dégradation (en essayant de ne conserver que l'intensité ou que l'intonation, ...).

Les sujets sont des étudiants et des chercheurs spécialisés en linguistique ou traitement de la parole. Lors des tests, les sujets écoutent une paire de langues, en connaissant les deux langues, et ils doivent déterminer l'ordre dans lequel les langues sont présentées.

Bien évidemment, les résultats augmentent avec la quantité d'information disponible. Les paires de langues les plus faciles à discriminer sont les paires chinois/japonais et chinois/espagnol. Les paires anglais/chinois et anglais/japonais ne sont pas évidentes à discriminer, mais les plus difficiles sont les paires anglais/espagnol et japonais/espagnol.

Considérant ces résultats, les types rythmiques théoriques semblent être importants pour déterminer la difficulté de la discrimination, ainsi que le contraste entre accent tonal ou non.

Pour conclure, les humains sont en général capables de distinguer plusieurs langues d'après les types d'accent lexicaux et les types rythmiques. Cependant la distinction entre deux langues proches rythmiquement sans prendre en compte d'information intonative n'est pas possible (par exemple entre l'espagnol et le japonais). De même la distinction entre l'anglais et l'espagnol (possédant le même type d'accent lexical mais des rythmes différents) est difficile lorsque l'information d'amplitude n'est plus disponible.

1.4 Conclusion

Les principales difficultés rencontrées suite à la définition que nous avons donné de la prosodie ont été évoquées, notamment la nature polysémique des informations véhiculées.

Après un inventaire des possibilités offertes pour l'extraction automatique de paramètres liés à la prosodie à partir du signal de parole brut, il apparaît que le faible nombre de paramètres disponibles par rapport à la quantité d'information véhiculée n'est pas très encourageant de prime abord. Cependant, nous avons ensuite vu quelles sont les potentialités offertes par la prosodie en vue de l'identification automatique des langues, notamment comment l'étude de la prosodie permettrait de confirmer ou d'infirmer les théories proposées par les linguistes, et s'inscrire dans le cadre de théories cognitives liées à la compréhension et à l'évolution du langage.

L'ensemble des expériences en perception ainsi que les théories linguistiques sous-jacentes montrent que les humains sont capables de distinguer des langues uniquement à partir de leurs propriétés prosodiques. L'objectif est d'étudier s'il est également possible, de même que les humains et en accord avec les théories linguistiques, d'utiliser les paramètres prosodiques dans le cadre de l'identification automatique des langues.

Chapitre 2

L'identification automatique des langues : méthodes & approches classiques

Sommaire

2.1 Architecture classique 28
2.2 Approche spectrale 29
2.3 Approche phonétique-phonotactique 30
2.4 Approches syllabiques 31
2.4.1 Méthode syllabotactique 31
2.4.2 Reconnaissance des syllabes 32
2.5 Campagne d'évaluation NIST 2003 [85] 34
2.5.1 Description de la campagne 35
2.5.2 Massachussets Institute of Technology - Lincoln Laboratory [120] 37
2.5.3 Oregon Graduate Institute 40
2.5.4 Queensland University of Technology 41
2.5.5 R523 (*Department of Defense*) 43
2.5.6 *University of Washington* 43
2.5.7 IRIT/DDL 45
2.5.8 Résultats 46
2.6 Le système du LIMSI [49] 48
2.6.1 Cadre théorique 48
2.6.2 Expériences 49
2.7 Conclusion 50

L'architecture des systèmes d'identification automatique des langues se déduit directement des sources d'information retenues pour la discrimination, accessibles d'un point de vue automatique. Compte tenu des dernières avancées en traitement automatique de la parole et plus particulièrement en reconnaissance automatique de la parole, l'avantage est actuellement donné aux systèmes employant les informations acoustiques (approche spectrale, §2.2) et phonotactiques (reconnaissance de phonèmes et modèle de langage, §2.3). L'approche dite « syllabique » (§2.4) est abordée comme une alternative.

Les systèmes présentés lors de la campagne d'évaluation NIST 2003 (§2.5) illustrent ces approches et donnent une idée des performances actuelles et des défis restants. Nous noterons que la prosodie reste très peu utilisée.

2.1 Architecture classique

Les sources d'information discriminantes pour l'identification des langues sont liées à quatre domaines linguistiques :

- la phonologie : les espaces acoustiques des langues sont différents, les inventaires phonétiques sont distincts suivant les langues. Même s'il existe des recoupements pour certaines langues, les fréquences d'apparition des phonèmes peuvent être caractéristiques. De plus, les règles d'enchaînements (phonotactiques) de ces unités varient d'une langue à l'autre.
- la morphologie : les lexiques sont différents suivant les langues, chaque langue a son propre vocabulaire et sa propre manière de former les mots,
- la syntaxe : les phrases sont structurées différemment selon les langues,
- la prosodie : le rythme, l'intonation et l'accentuation varient suivant les langues.

Dans une perspective de l'augmentation du nombre de langues à reconnaître par les systèmes d'identification, il est important de prendre en compte le maximum de sources possible.

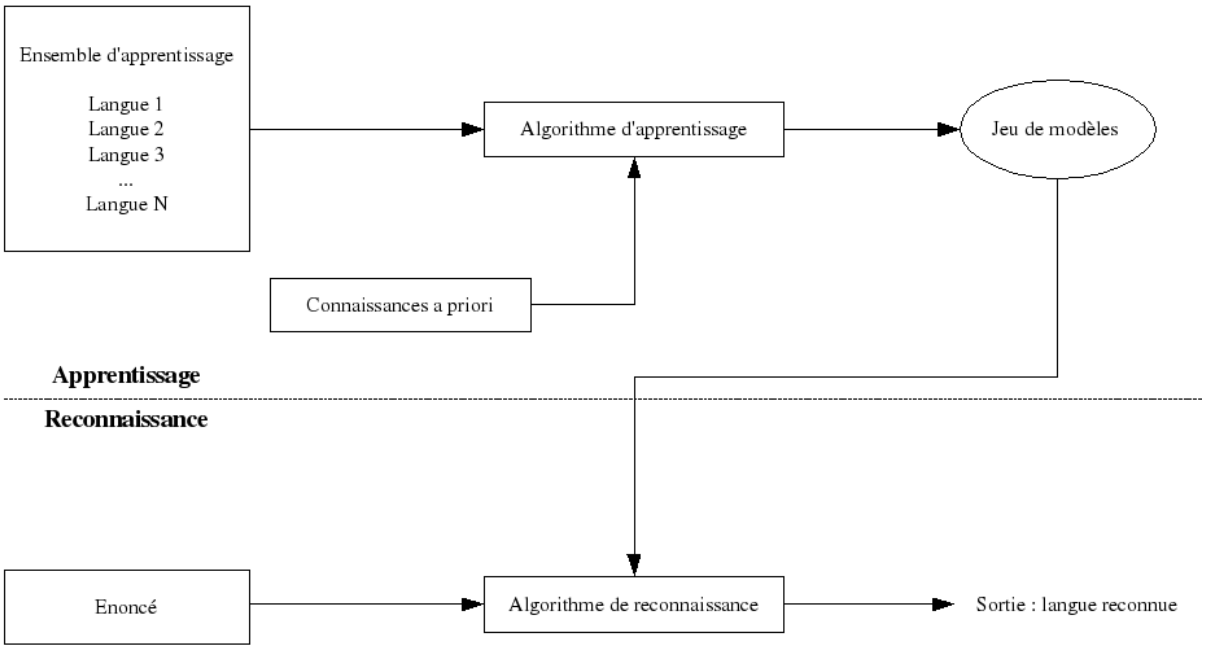


Fig. 2.1 : Structure générale d'un système d'identification automatique des langues.

Le système général s'apparente à un système statistique classique de reconnaissance des formes utilisant un apprentissage supervisé. Son fonctionnement se décompose en deux phases (figure 2.1) :

- Apprentissage : des paramètres sont extraits pour les signaux de parole de chaque langue. Pour chaque source d'information prise en compte, un modèle spécifique à

chaque langue est appris à partir de ces paramètres.
– Reconnaissance : les paramètres sont extraits pour un signal de parole d'une langue inconnue. La langue la plus vraisemblable est déterminée en fonction des modèles issus de la phase d'apprentissage. Se greffe un problème de fusion si plusieurs sources d'information sont modélisées engendrant plusieurs scores.

Un comparatif des différentes approches est disponible dans [128].

2.2 Approche spectrale

Les premiers systèmes d'identification automatique des langues étaient fondés sur des différences de contenu spectral entre les langues, en exploitant le fait que les langues possèdent différents ensembles de phonèmes.

Pour effectuer l'apprentissage, des paramètres représentant le spectre à court terme sont extraits des signaux de parole. Ils sont modélisés pour chaque langue, souvent par des modèles statistiques. Durant la phase de reconnaissance, les mêmes paramètres sont extraits sur les phrases de test et comparés aux prototypes issus de l'apprentissage.

Les paramètres employés peuvent varier quelque peu : les coefficients spectraux peuvent être utilisés directement comme paramètres, ou peuvent servir à calculer des paramètres tels que les formants ou les coefficients cepstraux.

Des différences entre les approches s'observent au niveau des mesures de similarité (distances Euclidienne, Mahalanobis, ...). Les approches récentes (notamment celles employées pour la campagne NIST 2003 [85], par exemple [120]) calculent une distance cumulée entre chaque vecteur de test et les exemples d'apprentissage. La distance prise en compte est alors une distance « globale » entre les exemples et la phrase de test. Une généralisation très employée de cette approche est l'utilisation de Modèles de Mélange de lois Gaussiennes (MMG). Dans ce cas, chaque vecteur suit hypothétiquement une loi de distribution dont la densité de probabilité est une somme pondérée de lois gaussiennes multidimensionnelles. Plusieurs variantes existent au niveau de l'apprentissage des modèles : le cas le plus classique consiste à apprendre un modèle par langue en utilisant les données d'apprentissage spécifiques à chaque langue, et un des cas rencontrés fréquemment consiste à apprendre un modèle dit "du monde" ou *Universal Background Model* avec toutes les données de toutes les langues, puis à adapter les paramètres de ce modèle afin de créer des modèles spécifiques à chaque langue. Une telle méthode d'apprentissage – représentée sur la figure 2.2 – permet d'obtenir ce que l'on appelle couramment des modèles MMG-UBM.

Ce type d'approche est toujours largement employé, les différents espaces sonores pouvant être l'objet de modélisations séparées, comme dans [100] où les espaces vocaliques et consonantiques sont séparés pour la modélisation.

D'autres auteurs (Dutat [35] par exemple) ont appliqué des variantes pour essayer de

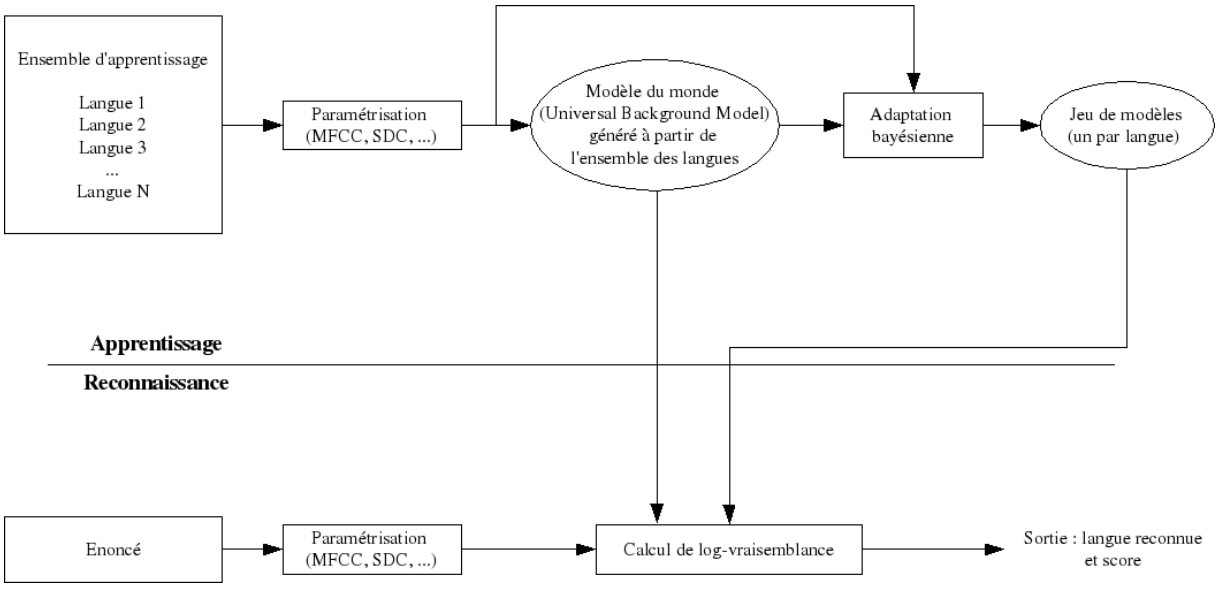


Fig. 2.2 : Exemple de méthode d'apprentissage de MMG par adaptation d'un modèle du monde, et reconnaissance.

prendre en compte les évolutions au cours du temps.

Une étape de fusion avec les approches décrites ci-dessous permet d'améliorer les performances globales.

2.3 Approche phonétique-phonotactique

Les systèmes donnant les meilleurs résultats sont pour l'heure ceux qui se basent principalement sur l'aspect phonologique et phonotactique. Ce sont les systèmes les plus présents dans la littérature (on peut citer par exemple [59], [128] ou [84]).

En général, ces systèmes possèdent un ou plusieurs systèmes de reconnaissance de phonèmes (souvent appelés décodeurs acoustico-phonétiques) constitués de Modèles de Markov Cachés. Ces décodeurs acoustico-phonétiques sont le plus souvent spécifiques à l'inventaire phonétique d'une langue, c'est pourquoi il faut en employer plusieurs en parallèle afin d'obtenir la meilleure couverture possible de l'ensemble des sons. Un décodeur acoustico-phonétique unique capable de reconnaître des phonèmes de plusieurs langues peut également être employé [15, 21].

Les séquences de phonèmes sont ensuite modélisées à l'aide de modèles probabilistes n-grammes. Ainsi, les enchaînements (de 2 ou 3 phonèmes le plus souvent) les plus caractéristiques des langues sont retrouvés.

Un exemple d'un tel système est proposé sur la figure 2.3.

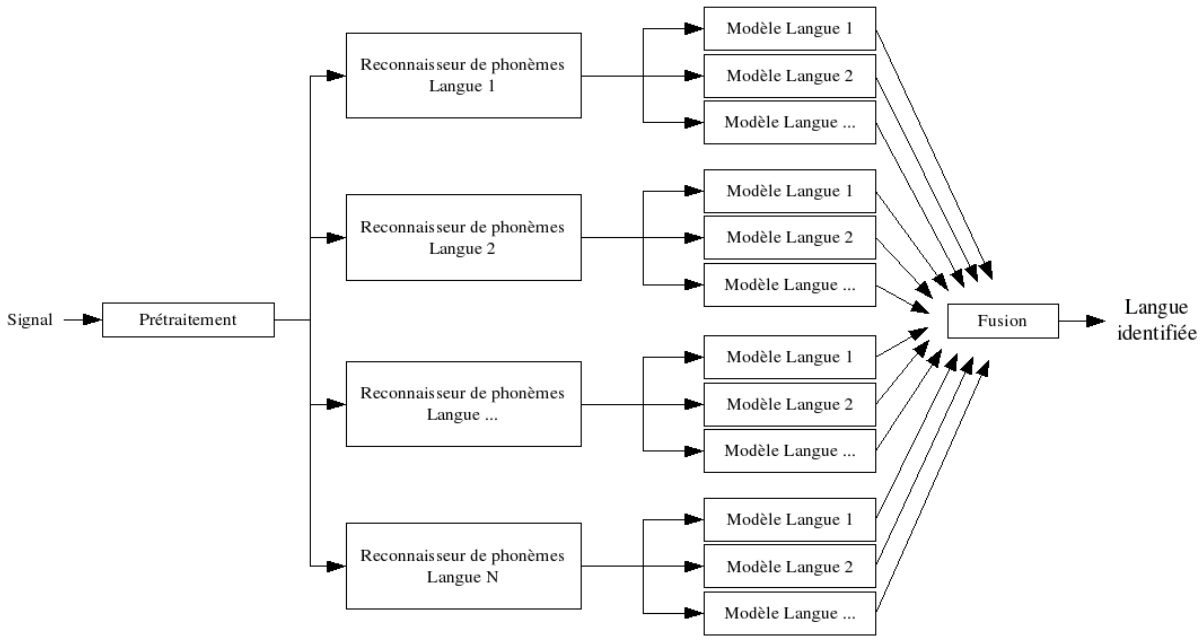


Fig. 2.3 : Structure d'un système d'identification basé sur une approche phonétique avec plusieurs décodeurs acoustico-phonétiques (Parallel Phone Recognition followed by Language Modelling - PPRLM) [128].

2.4 Approches syllabiques

Les approches novatrices les plus récentes emploient des unités qui différent par rapport à celles employées dans les systèmes décrits précédemment. Pour les systèmes acoustiques (§2.2) et phonotactiques (§2.3), les unités utilisées étaient respectivement la trame et le phonème. L'apparition de bases de données plus importantes a permis de prendre en compte des unités de taille plus importante, certainement plus corrélées à la perception humaine, les syllabes.

Les méthodes présentées dans [8] et [94] montrent deux techniques différentes pour prendre en compte des unités de taille syllabique.

2.4.1 Méthode syllabotactique

La méthode décrite dans [8] est basée sur les modèles phonotactiques décrits précédemment. Il s'agit de considérer les enchaînements de phonèmes, mais en tenant compte des inventaires syllabiques des différentes langues. Au départ, les phonèmes sont obtenus à partir d'un corpus de données étiquetées orthographiquement. Ces transcriptions sont alors phonétisées et alignées au moyen d'outils automatiques. Un inventaire phonétique est réalisé pour chaque langue ; la mise en commun de ces inventaires résulte en un jeu de

phonèmes commun de 74 phones pour l'ensemble des langues.

La méthode de syllabation utilisée ensuite est basée sur le *principe de sonorité* (ordonnancement des phonèmes selon une échelle de sonorité correspondant au degré d'intensité perçue, les consonnes d'attaque doivent se succéder selon une intensité décroissante) et le *principe d'attaque maximale* (la frontière syllabique entre deux voyelles séparées par des consonnes est placée de façon à maximiser le nombre de consonnes en attaque de la deuxième syllabe).

Un inventaire des syllabes est réalisé pour chaque langue. Le taux de couverture syllabique est évalué pour chaque langue, et un nombre de syllabes est choisi pour chaque langue afin d'atteindre un taux de couverture d'au moins 95%. Ces inventaires syllabiques réduits sont alors fusionnés pour obtenir un inventaire syllabique multilingue comportant 5380 syllabes.

Une fois cet inventaire obtenu, le décodage acoustico-phonétique est réalisé sur les données d'apprentissage. Les modèles syllabotactiques spécifiques à chaque langue sont appris à partir des séquences de phonèmes décodées.

Lors de la phase de reconnaissance, les syllabes sont identifiées au moyen du décodeur syllabique multilingue. Des mesures de vraisemblance sont associées aux syllabes reconnues par les modèles syllabotactiques. La langue identifiée est la langue pour laquelle la séquence de syllabes est la plus vraisemblable.

Des expériences ont été menées sur un corpus de données radiophoniques constitué de 7 langues (arabe standard, chinois mandarin, anglais américain, allemand, italien, français et portugais européen). Le taux d'identification correcte obtenu avec des modèles trigrammes est de 79 % avec des échantillons de test de 10 secondes, et passe à 86 % pour des échantillons de test de 20 secondes. Bien qu'inférieurs à ceux obtenus avec une approche phonotactique, ces résultats sont prometteurs pour les développements futurs.

2.4.2 Reconnaissance des syllabes

La méthode décrite dans [94] est basée sur une segmentation automatique et indépendante des langues en syllabes, ne requérant aucun étiquetage manuel préalable. Compte tenu de l'objectif de cette thèse, nous détaillons cette méthode. La procédure employée est la suivante :

1. Calcul de l'énergie à court terme
2. Construction de la partie symétrique de la séquence par rapport à l'axe des ordonnées. Cette nouvelle séquence est vue comme un spectre d'amplitude arbitraire, noté $E(k)$.
3. Calcul de l'inverse de la fonction $E(k)$. On notera la fonction résultante $E^i(k)$.
4. Calcul de la transformée de Fourier inverse de $E^i(k)$. La séquence résultante $c(n)$ est le cepstre racine.

5. Calcul du temps de propagation de groupe $\phi(k)$ sur la partie causale fenêtrée.

Une illustration de cette procédure est représentée sur la figure 2.4.

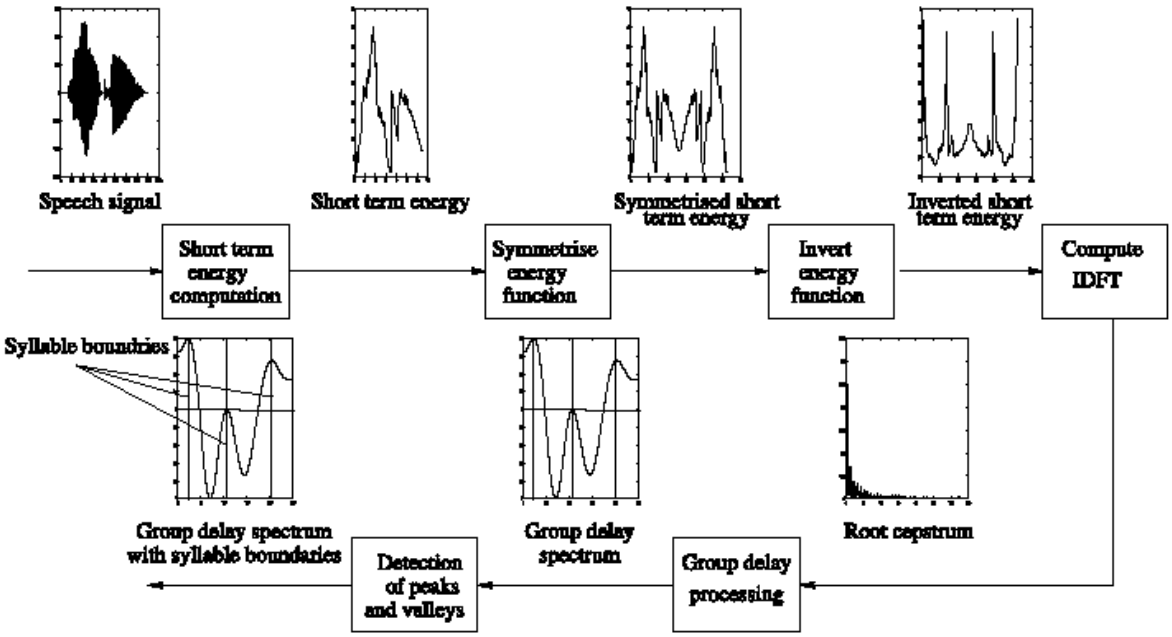


Fig. 2.4 : Système de segmentation automatique en syllabes (image extraite de [94]).

Les pics dans la fonction de temps de propagation de groupe correspondent approximativement à des frontières syllabiques. Le résultat de la segmentation en syllabes est montré sur la figure 2.5.

Les données de chaque langue sont alors segmentées selon cette méthode. Les syllabes sont caractérisées par des coefficients cepstraux. Un ensemble de Modèles de Markov Cachés est entraîné sur ces données. Le nombre de modèles à apprendre est déterminé automatiquement par une procédure de regroupement des syllabes en classes. Chaque modèle syllabique est un modèle de Markov à 5 états avec 3 lois gaussiennes par état. N_l modèles de classes de syllabes spécifiques à chaque langue sont obtenus de cette manière ($M_l = \{m_1, ..., m_{N_l}\}$).

Lors de la phase de test, l'énoncé X est segmenté automatiquement et paramétré ($X = \{x_1, ..., x_k\}$). Les syllabes ($S(X) = \{s_1, ..., s_k\}$) sont ensuite identifiées grâce aux modèles de syllabe :

$$p(s_i|L_l) = \max_{n=1...N_l} p(s_i|m_n)$$

La somme des probabilités relatives à chaque langue est calculée sur la phrase :

$$p_l(S) = \sum_{i=1}^{k} log(p(s_i|m_n))$$

La langue est alors déterminée par le maximum de p_l selon les langues.

Les expériences ont été menées sur les onze langues du corpus Ogi Mlts (anglais, farsi, français, allemand, hindi, japonais, coréen, espagnol, tamil et vietnamien). Le taux d'identification correcte , en moyenne sur les onze langues, est de 75,9 %.

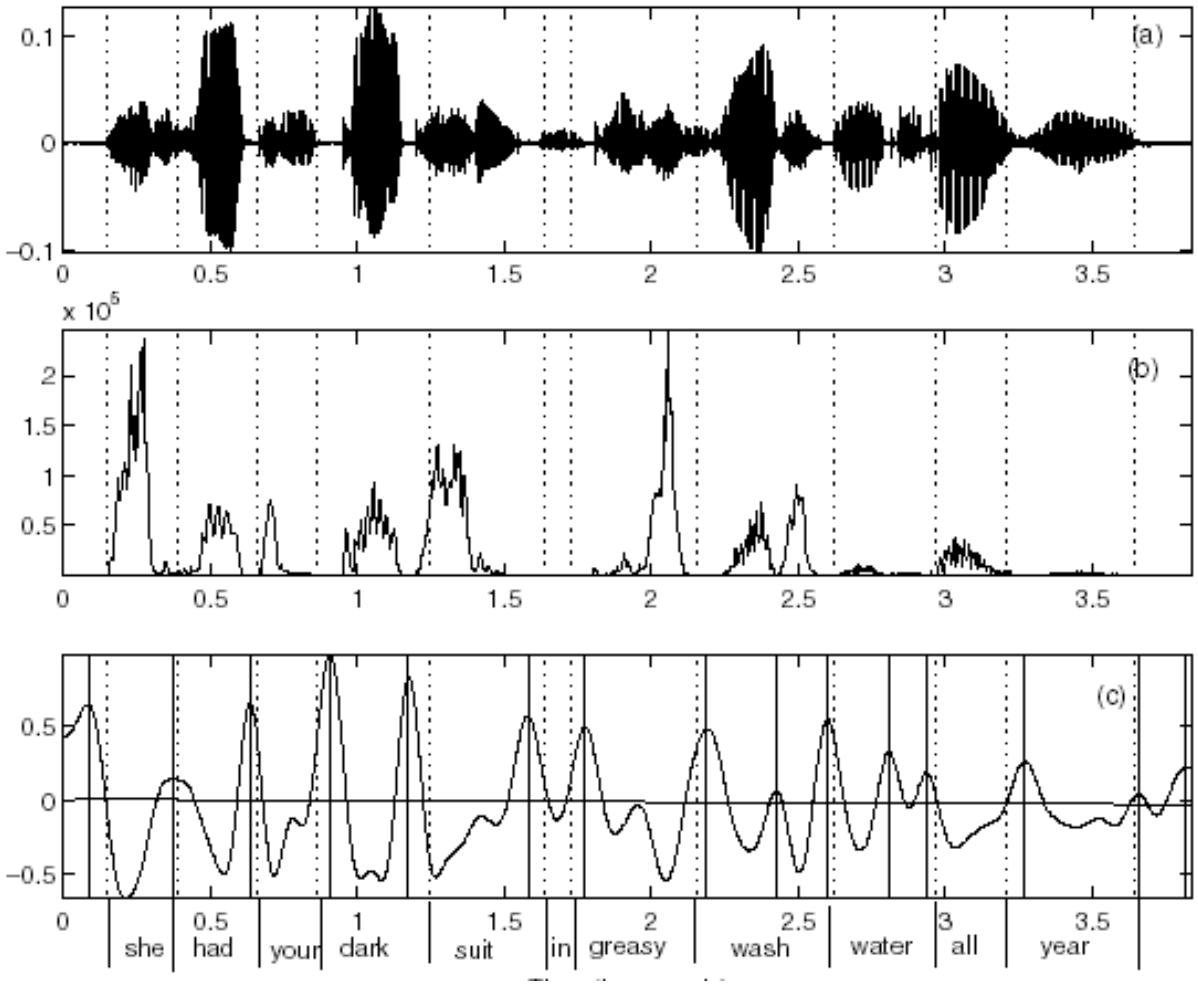


Fig. 2.5 : Résultat de la segmentation automatique en syllabes (image extraite de [94]).

2.5 Campagne d'évaluation NIST 2003 [85]

La campagne d'évaluation NIST 2003 a eu pour but d'établir les performances actuelles des systèmes dédiés à la reconnaissance de la langue sur des données conversationnelles de qualité téléphonique. Six sites, en Amérique du Nord, Europe et Australie, ont participé à cette campagne :

– le *Lincoln Laboratory*, *Massachussets Institute of Technology*, USA,

- le *Center for Spoken Language Understanding*, *Oregon Graduate Institute*, USA,
- le *Speech Research Lab*, *Queensland University of Technology*, Autralie,
- le *Department of Electrical Engineering*, *University of Washington*, USA,
- le R523, *Department of Defense (DoD)*, USA,
- l'Institut de Recherche en Informatique de Toulouse (IRIT), associé au laboratoire Dynamique Du Langage (DDL), Lyon.

2.5.1 Description de la campagne

La tâche d'évaluation demandée consiste à vérifier la présence d'une langue hypothèse, étant donné un segment de parole conversationnelle enregistrée au téléphone. Les 12 langues à reconnaître sont les suivantes : arabe (égyptien), anglais (américain), farsi, français (canadien), allemand, hindi, japonais, coréen, mandarin, espagnol (amérique latine), tamoul, vietnamien. Les participants ont également été prévenus qu'une langue intruse (le russe), ne faisant pas partie de l'ensemble d'apprentissage, ferait partie de l'ensemble de test.

Un segment de parole est un enregistrement d'une partie d'une conversation et est encodé en 8-bit 8 kHz mu-law digital (format téléphone). Chaque segment est préparé en utilisant une détection automatique de l'activité vocale pour identifier les parties de parole de la conversation qui sont ensuite concaténées pour former le segment.

Les segments ont une durée nominale de 3 secondes (de 2 à 4 secondes), 10 secondes (de 7 à 13 secondes) et 30 secondes (de 25 à 35 secondes). Ils sont choisis par jeux de trois : trois segments de 3 secondes sont contenus dans un segment de 10 secondes et trois segments de 10 secondes sont contenus dans un segment de 30 secondes.

Un segment est présenté au système de détection qui doit confirmer ou infirmer la langue hypothèse.

Évaluation

La métrique d'évaluation est basée sur la performance du système de détection, qui est caractérisée par ses probabilités de fausse alarme et de fausse détection. Le coût attendu de prendre une décision de détection, noté C_{Det}, est défini comme :

$$C_{Det} = (C_{Miss}.P_{Miss|Target}.P_{Target}) + (C_{FalseAlarm}.P_{FalseAlarm|Non-Target}.P_{Non-Target})$$

avec C_{Miss} et $C_{FalseAlarm}$ représentant les coûts relatifs d'une fausse détection et d'une fausse alarme. Pour cette évaluation, ces deux coefficients sont choisis égaux à 1, et P_{Target}, la probabilité *a priori* de la langue cible est toujours égale à 0,5.

L'évaluation est déduite d'un grand nombre de segments de test. Pour chacun de ces segments, il y a 12 essais, correspondant à chacune des langues cibles.

Pour chaque essai, le système donne deux informations. La première est la décision (Vrai ou Faux) de savoir si la langue parlée dans le segment de test est ou n'est pas la langue cible. La seconde sortie est un score de vraisemblance, indiquant la probabilité que la langue du segment de test corresponde à la langue cible.

Corpus

La source de données primaire pour cette évaluation est le corpus multilingue CallFriend de données conversationnelles téléphoniques, collectées il y a plusieurs années par le Linguistic Data Consortium (LDC, `http://www.ldc.upenn.edu/`). Ce corpus se compose d'appels téléphoniques passés en Amérique du Nord par des locuteurs de langue native. Les langues collectées incluent les 12 langues spécifiées pour cette campagne.

Le corpus se décompose en trois sous-ensembles :

- Ensemble d'apprentissage : Les données d'apprentissage peuvent venir de n'importe quelle source. Cependant, les données de 20 conversations d'une demi-heure pour chacune des langues du corpus CallFriend, qui étaient disponibles pour la précédente évaluation en 1996, sont distribuées aux participants.
- Ensemble de développement : Les ensembles de développement et de test distribués lors de la campagne 1996 sont disponibles comme ensemble de développement pour cette évaluation. Tous ces ensembles contiennent 2 segments de 3, 10 et 30 secondes, pour chaque locuteur dans un total de 20 conversations pour chacune des 12 langues cibles.
- Ensemble d'évaluation : Les segments de test sont composés de 80 fichiers pour chacune des durées envisagées et pour les 12 langues cibles similaires à ceux de l'ensemble de développement. Ces données proviennent de conversations collectées pour le corpus CallFriend, qui n'étaient pas inclues dans la version distribuée de ce corpus. En addition, il y a 4 jeux additionnels de 80 segments pour chaque durée sélectionnés dans d'autres sources de parole conversationnelles (conversations en Russe, en Japonais, en Anglais). Au total, il y a 3840 segments de test.

Règles

Les participants pouvaient se limiter à des tests n'impliquant qu'une partie des 12 langues. Cependant, tous les participants ont choisi de faire les tests sur l'ensemble des 12 langues. Les règles et restrictions suivantes s'appliquent à tous les participants :

- Chaque segment de test doit être traité séparément, indépendamment, et sans connaissance des autres segments. La normalisation sur plusieurs segments de test n'est pas autorisée.
- L'usage de connaissances sur l'ensemble des langues cibles est autorisé. La normalisation sur plusieurs langues cibles, telle que la limitation du nombre de langues pour laquelle est prise la décision « Vrai » est autorisée. Toutefois, il est possible que certains segments de test proviennent de langues cibles inconnues. L'utilisation

de connaissances sur ces langues n'est pas autorisée.
- La connaissance du sexe ou d'autres caractéristiques des locuteurs de test (sauf celles obtenues par des moyens automatiques) n'est pas autorisée.
- L'écoute des données d'évaluation, ou tout autre tentative d'expérimentation avec ces données, n'est pas permise avant que les résultats soient soumis au NIST.

2.5.2 Massachussets Institute of Technology - Lincoln Laboratory [120]

Le laboratoire Lincoln du *Massachussets Institute of Technology* a présenté un système pour la campagne d'évaluation NIST 2003, obtenu après fusion de trois systèmes élémentaires (figure 2.6). Les deux premiers relèvent de deux approches classiques :
- l'approche phonotactique emploie des modèles de langage qui capturent les enchaînements de séquences phonétiques issues d'un jeu de décodeurs acoustico-phonétiques.
- l'approche acoustique utilise les caractéristiques acoustiques des langues, qui sont modélisées par des modèles statistiques.

La troisième approche employée est une approche acoustique discriminative, contrairement aux deux précédentes qui s'appuient sur des techniques génératives. Elle est basée sur l'utilisation de SVM (Machine à vecteurs supports) [18].

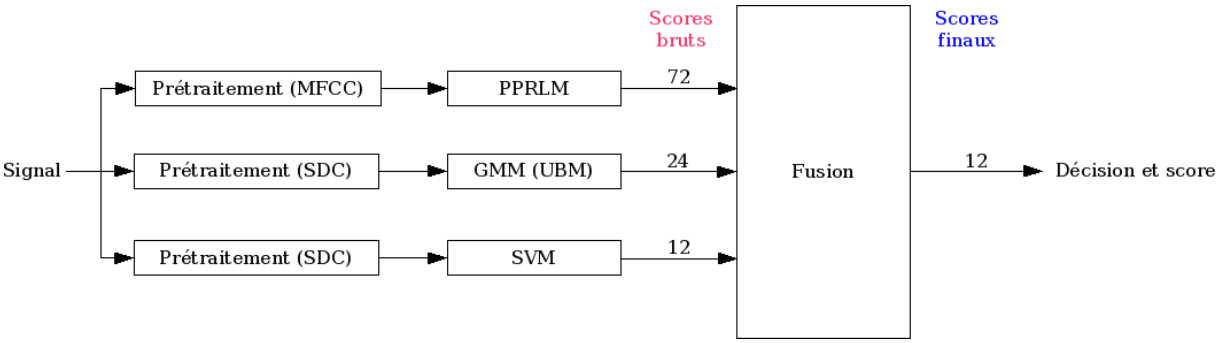


Fig. 2.6 : Vue d'ensemble du système du MIT

Approche phonotactique

Le système phonotactique proposé par le MIT est similaire au PPRLM (Parallel Phone Recongnition and Language Modelling) (figure 2.3). Cette technique utilise un banc de décodeurs acoustico-phonétiques, qui permettent de phonétiser un signal de parole selon les inventaires phonétiques de différentes langues ; six jeux de phonèmes sont employés et proviennent de six langues : Anglais, Allemand, Hindi, Japonais, Chinois Mandarin et Espagnol.

Les séquences phonétiques décodées sont ensuite utilisées pour calculer des modèles de langage indépendants du genre pour chacune des langues à reconnaître. Il en résulte que 6 * 12 modèles de langage sont appris.

Au cours des tests, un vecteur de 72 scores est produit pour chaque fichier de test.

Approche acoustique générative (MMG)

Le système acoustique est basé sur des paramètres innovants. Les *Shifted Delta Cepstra* ou *SDC* sont créés à partir des dérivées des coefficients cepstraux calculées sur plusieurs trames [122].

Si le coefficient cepstral d'ordre j pour la trame t est noté $c_j(t)$, l'estimation de la dérivée de ce coefficient (δ) est :

$$\delta_j(t) = c_j(t+d) - c_j(t-d) \tag{2.1}$$

avec $d = 1$ pour la plupart des applications. Le calcul des *Shifted Delta Cepstra* dépend de quatre paramètres : N, d, P et k, avec N le nombre de coefficients cepstraux calculés pour chaque trame ; d représente le délai pris en compte pour le calcul des dérivées ; k est le nombre de trames utiles pour définir l'observation à l'instant t, et P est le décalage ($shift$) entre les trames. Pour chaque décalage envisagé $(i-1)P$,

$$\delta_j(t+(i-1)P) = c_j(t+(i-1)P+d) - c_j(t+(i-1)P-d) \tag{2.2}$$

avec $1 \leq j \leq N$, et $1 \leq i \leq k$.

De cette manière, on obtient pour chaque indice j et chaque observation t un vecteur de dimension k, appelé le « *Shifted Delta Cepstrum* » :

$$SDC_j(t) = \begin{bmatrix} \delta_j(t) \\ \delta_j(t+P) \\ \vdots \\ \delta_j(t+(k-1)P) \end{bmatrix} \tag{2.3}$$

La dimension du vecteur d'observation $SDC(t)$ est égale à $N * k$. Pour la campagne d'évaluation, les SDC sont calculés avec les paramètres : $N = 7$, $d = 1$, $P = 3$ et $k = 7$. Un vecteur de 49 paramètres est ainsi obtenu pour chaque trame.

Des modèles de mélange de lois gaussiennes (MMG) d'ordre 2048 sont définis pour chacune des 12 langues sur cet espace, en utilisant un modèle du monde (*Universal Background Model* ou UBM) indépendant des langues, appris sur la totalité de l'ensemble d'apprentissage (toutes les langues réunies). Les modèles dépendants des langues sont adaptés à partir du modèle du monde, pour chacun des deux sexes. Il en résulte 24 modèles de mélanges de lois gaussiennes dépendants des langues et du sexe du locuteur.

Approche acoustique discriminative (SVM)

Cette approche emploie des Machines à Vecteurs Supports (ou SVM), avec un noyau dit « *Generalized Linear Discriminant Sequence* » [18]. L'objectif des SVM est de déterminer une frontière non linéaire entre deux classes. Les données issues des deux classes sont en réalité projetées, implicitement ou explicitement, dans un espace de dimension supérieure dans lequel sera déterminée la frontière.

La méthode décrite ici est légèrement différente de l'approche « classique » des SVM car elle permet de traiter directement des séquences d'observations. La projection est explicite, puisque le système traite des séquences de longueur variable. Cela augmente le temps nécessaire pour l'apprentissage (par rapport à une projection implicite), mais permet de réduire la taille du modèle à l'expression d'un simple vecteur (et non l'ensemble des vecteurs supports). Il en résulte un gain en temps de calcul lors de la phase de test.

Les paramètres employés sont les mêmes que pour l'approche acoustique, il s'agit de *SDC* de dimension 49 (paramètres (7,1,3,7)) avec soustraction cepstrale et normalisation en variance. La première étape consiste à projeter chaque vecteur de paramètres en procédant à une expansion polynomiale de degré 3.

À titre d'exemple, si on note $X = [x_1, x_2]$ (en dimension 2) un vecteur correspondant à une observation, l'expansion polynomiale de degré 3 (notée $b(X)$) sera :

$$b(X) = [x_1, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3]$$

Dans cette approche la dimension de X est 49, la dimension de l'expansion polynomiale $b(X)$ sera 22100. La moyenne de cette expansion $\bar{b}(X)$ est calculée sur chaque séquence.

Lors de la phase d'apprentissage, pour chaque langue cible, les données de la langue sont projetées et moyennées sur les séquences. Les données de l'ensemble des autres langues sont également projetées dans le même espace. Une frontière linéaire est alors déterminée dans l'espace de projection. Comme elle est supposée linéaire (hyperplan), elle est représentée par un vecteur dans cet espace. Il est ainsi obtenu un vecteur par langue, soit 12 au total.

Lors de la phase de test, une séquence d'observation est projetée par expansion polynomiale. Pour chaque langue cible, le score est obtenu en mesurant la distance entre la moyenne de la projection de la séquence d'observation et la frontière par le calcul d'un produit scalaire. En sortie, un vecteur de scores de dimension 12 est obtenu.

Fusion

Les scores en sortie de chacun des modèles sont fusionnés en utilisant un classifieur Gaussien. Le vecteur d'entrée est de dimension 108 (24 scores MMG, 72 scores PPRLM, et 12 scores SVM). Les 12 scores en sortie sont convertis en rapports de log-vraisemblance.

2.5.3 Oregon Graduate Institute

Le laboratoire *Center for Spoken Language Understanding* de l'*Oregon Graduate Institute* a proposé trois systèmes, avec deux architectures différentes : deux systèmes basés sur une reconnaissance de phonèmes (figure 2.3) et un système utilisant des paramètres prosodiques.

Systèmes phonotactiques

Deux systèmes phonotactiques sont proposés par OGI :
- le score pour chaque langue est obtenu à partir de modèles de mélange de lois gaussiennes appliqués sur les scores en sortie de chaque modèle de langage et chaque décodeur phonétique (similaire au système du MIT).
- le score pour chaque langue est obtenu à partir de réseaux de neurones appliqués aux mêmes données.

Système prosodique

Le but de ce système est de convertir le signal de parole en une suite d'unités de longueur inférieure au mot qui caractérisent la langue.

L'approche employée code la variation temporelle de la fréquence fondamentale et de l'énergie afin d'obtenir la séquence d'unités. Les suites d'unités discrètes sont ensuite modélisées grâce à des modèles de langage N-grammes.

Le procédé employé pour segmenter le signal de parole, initié dans [2], est détaillé dans [3], ainsi que ses applications à la reconnaissance du locuteur et de la langue. En voici les grandes lignes :

La segmentation de la parole est décomposée en 5 étapes :

1. Calcul de la trajectoire temporelle de la fréquence fondamentale et de l'énergie.
2. Calcul de la courbe dérivée pour chaque trajectoire.
3. Détection des points d'inflexion pour chaque trajectoire.
4. Segmentation du signal de parole selon les points d'inflexion et les débuts/fins de voisement.
5. Conversion des segments en séquences de symboles à partir des dérivées de chacune des trajectoires.

Les symboles ou classes résultants de la segmentation et de l'étiquetage sont résumés dans le tableau 2.1.

Classe	Description
1	F_0 montant et énergie montante
2	F_0 montant et énergie descendante
3	F_0 descendant et énergie montante
4	F_0 descendant et énergie descendante
5	Non voisé

Tab. 2.1 : Classes de segments utilisées pour décrire les courbes d'énergie et de fréquence fondamentale pour le système OGI-ASP

.

Une information de durée est ajoutée à la description : les segments courts (d'une durée inférieure à 80 ms pour les segments voisés, 140 ms pour les segments non voisés) sont étiquetés S, les autres segments sont considérés longs (L). On obtient ainsi 10 classes.

Les séquences d'étiquettes obtenues pour chaque phrase de chaque langue sont ensuite modélisées au moyen de modèles tri-grammes.

Malheureusement, ce système, moins performant que les approches acoustiques et phonotactiques, n'a pas été retenu lors de la phase de fusion et ne fait donc pas partie du système global proposé par OGI. Cette approche est intéressante pour la modélisation de la prosodie, nous la reprendrons plus en détail dans le chapitre consacré aux approches prosodiques (§3.3.5).

2.5.4 Queensland University of Technology

Le *Speech Research Lab* de *Queensland University of Technology* (Australie) a présenté trois systèmes de reconnaissance de la langue dérivés de trois approches classiques (figure 2.7).

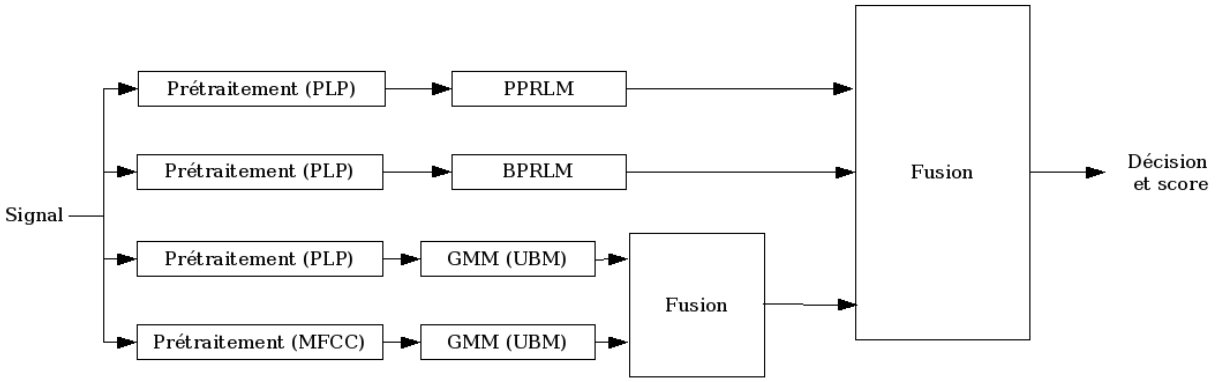


Fig. 2.7 : Vue d'ensemble du système du QUT

Système phonotactique de type PPRLM

Ce système a la même architecture que les deux approches phonotactiques citées précédemment (figure 2.3), avec 6 décodeurs acoustico-phonétiques (anglais, allemand, hindi, japonais, mandarin et espagnol). Les paramètres employés sont 12 coefficients PLP (Perceptual Linear Predictive coefficients [57]) avec leurs dérivées et leurs accélérations. Les modèles de langage employés sont des bi-grammes. Le score final est obtenu en faisant la somme des log-vraisemblances obtenues avec chaque modèle de langage. La langue reconnue est celle qui a le score le plus élevé.

Système phonotactique de type BPRLM (Broad Phonetic Recognition followed by Language Modelling)

Il est identique au système précédent à la différence que ce ne sont plus des phonèmes qui sont directement reconnus, mais de grandes classes phonétiques (voyelles, diphtongues, semi-voyelles, occlusives, fricatives, nasales). Cela permet d'avoir un taux de reconnaissance de grandes classes phonétiques supérieur au taux de reconnaissance des phonèmes. Les paramètres employés sont les mêmes que dans le cas précédent. Les modèles de langage utilisés ici sont d'un ordre plus important, il s'agit de tri-grammes. La décision est obtenue de la même manière que pour le système PPRLM ci-dessus.

Système acoustique (MMG)

Les paramètres employés sont des PLP (d'ordre 5) avec leurs dérivées et leurs accélérations ainsi que la dérivée de l'énergie, et des « *Linear Prediction Cepstral Coefficients* » ou LPCC (12^{e} ordre) avec leurs dérivées, accélérations et la dérivée de l'énergie. Une normalisation en fonction de la longueur du conduit vocal est opérée : chaque personne possède un conduit vocal différent, et sa longueur influe de manière inverse sur les fréquences formantiques.

Un modèle de mélange de lois gaussiennes, le modèle du monde (ou UBM) est appris sur les données d'apprentissage de toutes les langues réunies (figure 2.2). Pour chaque langue, une adaptation bayésienne est faite en vue d'obtenir des modèles spécifiques.

Lors de la phase de test, seules les 5 composantes les plus représentatives du mélange de lois gaussiennes (celles qui ont les poids les plus importants) sont prises en compte pour le modèle de monde, ainsi que les composantes correspondantes du modèle de la langue hypothèse.

Les scores obtenus avec les deux types de paramètres (PLP et LPCC) sont fusionnés.

Fusion

Au final, les scores obtenus avec les systèmes PPRLM et BPRLM sont fusionnés avec le score MMG (déjà fusionné) puis normalisés pour obtenir le score final.

2.5.5 R523 (*Department of Defense*)

Le modèle du R523 du *Department of Defense* utilise des Shifted Delta Cepstra comme paramètres. Les modèles sont des modèles de mélange de lois gaussiennes (1024 composantes). Ce modèle est quasiment identique au modèle du MIT décrit dans la section 2.5.2. Les différences se situent au niveau du calcul des *SDC* (voir §2.5.2) : les paramètres de fonctionnement diffèrent ($(N, d, P, k) = (7, 1, 3, 7)$ pour le MIT et $(6, 1, 3, 3)$ pour le R523). pour le MIT, il s'agit de (7,1,3,7) alors qu'ici il s'agit de (6,1,3,3). De plus les modèles de mélanges de lois gaussiennes du MIT sont dépendants du genre (hommes et femmes), alors que ce n'est pas le cas ici. Le nombre de composantes dans les modèles est inférieur (1024 ici contre 2048 pour le MIT).

2.5.6 *University of Washington*

Le système de reconnaissance de la langue développé à l'Université de Washington est décomposé en 6 étapes (voir figure 2.8) :

1. un prétraitement acoustique,
2. un système basé sur un décodage acoustico-phonétique suivi de modèles de langage,
3. un système original, basé sur des paramètres articulatoires,
4. un module de combinaison des scores,
5. un module de décision,
6. un module de rejet de décision associé au module de décision.

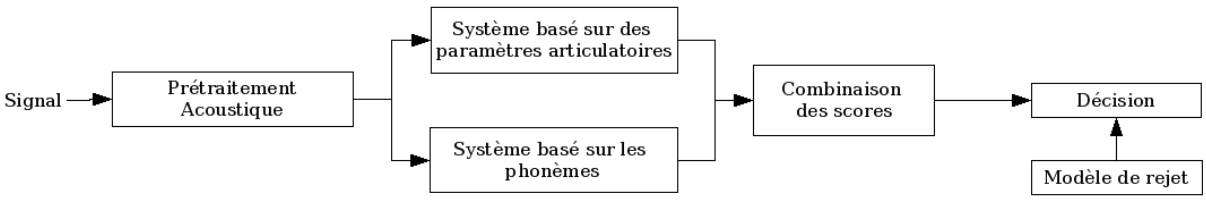


Fig. 2.8 : Vue d'ensemble du système de l'Université de Washington

Prétraitement acoustique

Le prétraitement acoustique est composé des étapes suivantes :
- extraction de paramètres : 12 MFCC, énergie et dérivées,
- détection d'activité vocale (avec un réseau de neurones appris sur le corpus OGI-TS, 10 langues, parole téléphonique, voir [93]),
- segmentation du signal et concaténation en morceaux d'environ 10 secondes,
- normalisation en moyenne et en variance sur chaque fichier.

Système phonotactique

La constitution des décodeurs acoustico-phonétiques diffère quelque peu de la méthode classique. 133 Modèles de Markov Cachés de phonèmes indépendants des langues sont appris sur le corpus OGI-TS. L'ensemble d'apprentissage de CALLFRIEND est décodé à l'aide de ces modèles. Pour chaque langue, seuls sont conservés les 50 phonèmes ayant les plus fortes occurences. Les modèles correspondants sont alors ré-estimés pour chaque langue sur le même corpus. Au final, il y a 12 décodeurs de phonèmes dépendants des langues recherchées. Les modèles de phonèmes sont des Modèles de Markov Cachés à 3 états, avec 2 lois gaussiennes par état.

Les modèles de langage sont des tri-grammes pour les phrases de 3 et 10 secondes, et des 4-grammes pour les phrases de 30 secondes. Il y a 12 modèles de langage par décodeur acoustico-phonétique. Au total, 144 scores sont produits lors de la phase de reconnaissance.

Système basé sur des paramètres articulatoires

Le système est similaire au précédent, excepté le fait que ce ne sont plus des phonèmes qui sont reconnus, mais un ensemble de symboles représentant des mouvements articulatoires. Les mouvements articulatoires sont séparés en grandes classes :
- la mode d'articulation, représentée par 16 symboles,
- le lieu d'articulation, représentée par 18 symboles,
- la position de la langue, représentée par 12 symboles,
- l'arrondissement des lèvres, représentée par 10 symboles.

L'ensemble ainsi déterminé comprend 56 classes ou symboles par langue. Les modèles sont des Modèles de Markov Cachés à trois états, avec 2 lois gaussiennes par état.

Pour chaque décodeur acoustico-phonétique, et pour chaque grande classe articulatoire, des modèles de langage tri-grammes sont appris. Il en résulte un total de 12 décodeurs acoustico-phonétiques * 12 modèles de langage par décodeur * 4 grandes classes articulatoires par modèle de langage, soit une production de 576 scores.

Combinaison des scores

La méthode de combinaison des scores est optimisée sur l'ensemble de développement : les scores des différentes sources sont moyennés si le taux d'identification augmente, sinon seule est gardée la combinaison offrant les meilleures performances.

Décision

Le système présenté ne fonctionne pas en vérification d'une langue, mais en identification, c'est-à-dire qu'il choisit une langue parmi les 12 possibles. Afin de prendre une décision, les scores de vraisemblance sont tout d'abord normalisés. Si la différence entre le score obtenu pour la langue cible et le deuxième meilleur score est supérieure à un seuil, alors la décision « Vrai » est prise pour la langue cible. Les seuils peuvent varier selon la langue cible.

Module de rejet

La décision de rejet est prise à partir d'expériences d'identification des langues effectuées avec des langues absentes du corpus CallFriend et présentes dans le corpus Ogi 22. Le système est testé sur ces langues, et des seuils sur les log-vraisemblances sont appris en fonction des scores obtenus par les segments de parole en langue non reconnue par le système, de manière à rejeter ces langues tout en minimisant les rejets abusifs.

2.5.7 IRIT/DDL

Le laboratoire IRIT et le laboratoire DDL (Dynamique du Langage, Lyon) se sont associés pour cette campagne d'évaluation. Le système proposé se décompose en 2 sous-systèmes :

– un module de modélisation acoustique,
– un module de décodage acoustico-phonétique suivi de modèles de langage (PPRLM).

Les modèles acoustiques sont étudiés plus en détail dans la suite du document. Des descriptions plus complètes pourront être retrouvés dans la section 5.9.1.

Système acoustique

Le modèle acoustique utilise un prétraitement permettant de segmenter automatiquement le signal de parole et de détecter les segments correspondant à des voyelles et à du silence. Les paramètres employés sont des MFCC (8 MFCC + 8 ΔMFCC + durée du segment), ils ne sont calculés que sur les segments vocaliques. Les modèles employés sont des mélanges de lois gaussiennes. Contrairement à la majorité des systèmes proposés

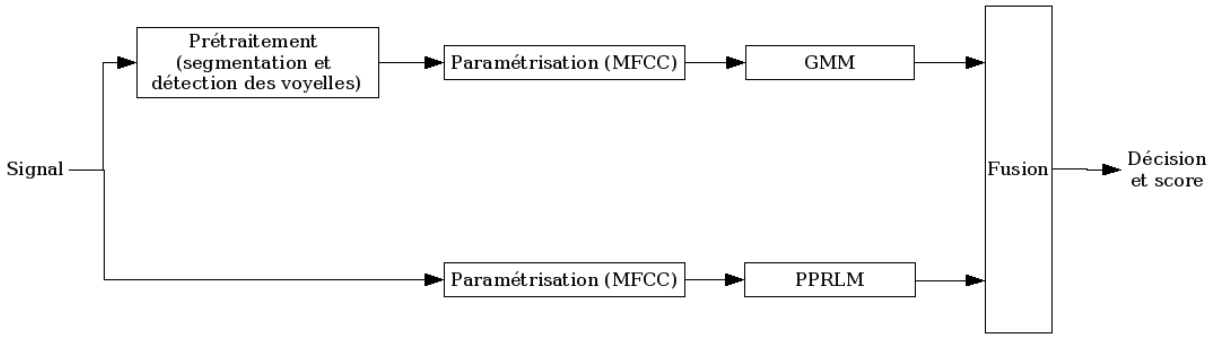


Fig. 2.9 : Vue d'ensemble du système présenté par l'IRIT et le DDL

plus haut (notamment ceux du MIT et d'OGI, §2.5.2 et §2.5.3), les modèles spécifiques pour chaque langue ne sont pas adaptés à partir du modèle du monde. Il y a pour chaque langue cible un modèle appris sur l'ensemble des données de cette langue et un modèle du monde appris sur l'ensemble des données des autres langues, soit 24 modèles différents.

Système phonotactique (PPRLM)

Le modèle employé est similaire à ceux décrits précédemment (figure 2.3). Les paramètres employés sont des PLP, ainsi que l'énergie et la dérivée de l'énergie.

Les phonèmes sont modélisés par des Modèles de Markov Cachés (HMM) avec un nombre d'états variant de 1 à 4 selon la catégorie phonétique du son à modéliser. Il y a 10 lois gaussiennes par état. Les décodeurs acoustico-phonétiques sont appris pour les 6 langues étiquetées phonétiquement du corpus OGI-TS [93]. Ensuite, des modèles de langage sont entraînés après chaque décodeur acoustico-phonétique pour chaque langue disponible dans l'ensemble d'apprentissage de CallFriend.

Il y a 12 modèles de langage par décodeur acoustico-phonétique, ce qui se traduit par 72 scores au total en phase de test.

Fusion

La fusion est réalisée par somme pondérée des scores obtenus avec chaque système.

2.5.8 Résultats

Les résultats sont résumés sur la figure 2.10. Ce graphique reprend les résultats en moyenne sur l'ensemble des douze langues considérées dans le test, hommes et femmes

confondus et pour les segments de test d'une durée de 30 secondes. Les courbes sont obtenues pour les systèmes fusionnés de chaque laboratoire.

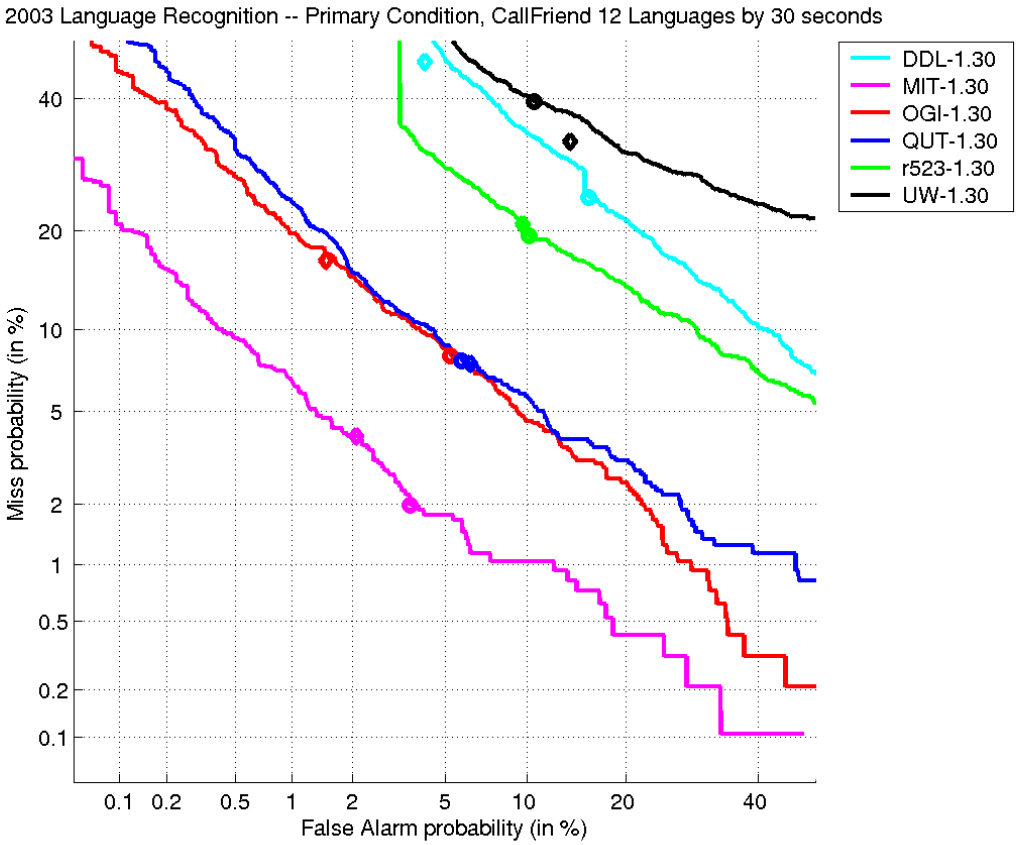


Fig. 2.10 : Résultats pour l'évaluation NIST 2003

Le meilleur résultat est obtenu par le système du MIT (§2.5.2). Ce système emploie les deux approches les plus efficaces combinées avec une approche novatrice (utilisant des SVM) mais qui a depuis fait ses preuves, que ce soit en identification de la langue ou du locuteur [18].

Les systèmes suivants sont ceux du OGI et de QUT (§2.5.3 et §2.5.4). Les résultats obtenus par ces deux systèmes sont très proches. Le système du OGI considéré ici est le système le plus efficace : il ne prend pas en compte la modélisation de la prosodie et ne comporte que la fusion d'un système acoustique et d'un système phonotactique. Il emploie donc les mêmes techniques que le système proposé par le QUT (fusion de 4 systèmes : 2 systèmes basés sur la reconnaissance des phonèmes et 2 systèmes basés sur l'acoustique).

Le quatrième système est le système du R523 (§2.5.5) qui est en quelque sorte une version simplifiée du système acoustique du MIT.

Ensuite vient le système IRIT/DDL (§2.5.7), fusion d'un système phonotactique et d'un système acoustique. Les paramètres employés - qui ne sont pas des *SDC* mais de classiques MFCC - et le fait qu'ils ne soient calculés que sur les segments vocaliques pour l'approche acoustique constituent les différences principales de ce système par rapport aux précédents. Enfin, la méthode de fusion n'est pas la plus optimale.

Le dernier système est le système de l'Université de Washington (§2.5.6). Ce système est différent de tous les précédents puisqu'il n'est pas conçu pour faire de la vérification de la langue mais de l'identification des langues.

Pour conclure, les approches dominantes sont toujours les mêmes (acoustique et phonotactique), et les améliorations sont faites sur la paramétrisation (coefficients *SDC*) et sur les méthodes d'apprentissage des modèles (pour l'approche acoustique : *UBM* et adaptation).

Une approche novatrice et efficace est celle concernant l'emploi de Machine à Vecteurs Support (SVM). Cette technique semble très prometteuse puisqu'elle permet d'obtenir des performances correctes et d'améliorer les performances globales du système lorsqu'elle est fusionnée avec l'approche acoustique MMG et l'approche phonotactique.

Nous noterons toutefois l'absence de systèmes efficaces de modélisation de la prosodie. Malgré le fait que OGI ait proposé un système incluant des paramètres prosodiques, ce système n'a pas été retenu lors de la fusion. Il reste encore de nombreux efforts à faire en paramétrisation et en modélisation pour arriver à inclure un système efficace prenant en compte la prosodie, et ainsi arriver à modéliser le plus d'informations discriminantes possible.

2.6 Le système du LIMSI [49]

L'approche décrite ci-dessous est l'approche la plus efficace pour le moment. Elle est fondée sur une architecture de type PPRLM (figure 2.3).

2.6.1 Cadre théorique

Étant donné un segment X, un jeu de modèles acoustiques Λ associés à un ensemble de phonèmes et un jeu de modèles phonotactiques Φ, le problème d'identification des langues peut se résumer à trouver la langue qui possède la probabilité a posteriori $Pr(L|X, \Lambda, \Phi)$ la plus élevée.

Si l'on suppose que les langues considérées sont équiprobables, le problème peut alors être formulé sous la forme :

$$L^* = \arg\max_L \sum_H f(X|H, L, \Lambda) P(H|L) \tag{2.4}$$

avec L^* la langue identifiée et $f(X|H, L, \Lambda)$ la vraisemblance du segment X sachant la séquence de phonème H et la langue L.

$f(X|H, L, \Lambda)$ est en général estimée avec des modèles de phones (Modèles de Markov Cachés). $P(H|L)$ est estimée avec les modèles N-grammes : $P(H|L) = \prod_i P(h_i|h_{i-N+1}, ..., h_{i-1}, L)$

L'équation 2.4 peut être approximée par :

$$L^* = \arg\max_L \max_H f(X|H, L, \Lambda)P(H|L) \tag{2.5}$$

Dans l'implémentation de type PPRLM, les modèles de phones sont supposés indépendant des langues. On peut alors remplacer $f(X|H, L, \Lambda)$ par $f(X|H, \Lambda)$. Il en résulte :

$$L^* = \arg\max_L P(H^*|L) \tag{2.6}$$

avec H^* la séquence de phonèmes la plus probable :

$$H^* = \arg\max_H f(X|H, \Lambda) \tag{2.7}$$

Une meilleure alternative consiste à maximiser l'espérance de $\log P(H|L)$ sur L, sachant H :

$$L^* = \arg\max_L E_H[\log P(H|L)|X, \Lambda, L] \tag{2.8}$$

Ceci peut être effectué si l'on ne considère plus uniquement la séquence de phonèmes la plus probable mais tout le treillis composé de l'ensemble des séquences possibles. Le treillis est un graphe où les noeuds correspondent aux trames et les arcs correspondent aux hypothétiques phonèmes auxquels sont associées des valeurs de probabilité.

2.6.2 Expériences

Les expériences ont été menées sur le corpus CallFriend pour une tâche de vérification avec les mêmes contraintes que pour la campagne d'évaluation Nist 2003 (voir §2.5).

La structure du système correspond à l'architecture PPRLM décrite plus haut (figure 2.3). Il y a trois décodeurs acoustico-phonétiques (anglais, espagnol et arabe). Chacun de ces décodeurs est entraîné avec des corpus différents. Pour l'espagnol et l'arabe, des conversations extraites du corpus CallHome (parole téléphonique conversationnelle, `http://www.ldc.upenn.edu`) sont utilisées. Pour l'anglais, il s'agit de conversations extraites de SwitchBoard (parole téléphonique conversationnelle, `http://www.ldc.upenn.edu`).

À la sortie de chacun des décodeurs, 12 modèles de langage tri-grammes (correspondant à chacune des langues à reconnaître) sont entraînés sur les données d'apprentissage de

CALLFRIEND. La décision est prise soit en moyennant les probabilités a posteriori obtenues pour chaque décodeur acoustico-phonétique soit en employant un réseau de neurones.

Les résultats sont résumés dans le tableau suivant :

Tab. 2.2 : Résultats (en taux d'erreur EER) du système de référence (PPRLM), du système employant les treillis de phones effectuant ou non la fusion des scores par réseaux de neurones

Méthode	3s	10s	30s
Référence (PPRLM)	23,7	12,6	6,8
PPRLM (Treillis)	18,3	8,3	4,0
PPRLM (Treillis et RN)	18,3	7,9	2,7

Sur ces données, le système du LIMSI est actuellement le plus performant (2,7% d'EER par rapport à 2,8% pour le MIT (§2.5.2)). La principale différence entre ces deux approches est le temps de calcul nécessaire, beaucoup plus faible pour le LIMSI (0,5xRT contre 15*RT pour le MTI).

2.7 Conclusion

Ce panorama des systèmes actuels d'identification automatique des langues montre que soit les caractéristiques acoustiques des langues, soit les caractéristiques phonétiques ou phonotactiques sont privilégiées.

L'ensemble des sources d'information présentées au chapitre 1 n'est pas toujours pris en compte. La dimension prosodique, malgré l'intérêt certain qu'elle présente, n'est que marginalement employée, voire pas du tout lorsque l'objectif premier est la performance.

Lors de l'évaluation NIST précédente de 1996, tous les systèmes présentés n'utilisaient que la modélisation phonotactique. Au cours de la dernière évaluation, nous avons pu voir une évolution, avec des systèmes employant des modélisations acoustiques, qui exploitent une autre source d'information. Les principales améliorations des performances sont dues :
- au nombre de lois gaussiennes employé dans les MMG, qui a nettement augmenté,
- à la création et l'adaptation de modèles UBM
- à l'émergence de nouveaux paramètres, les *Shifted Delta Cepstra*.

Dernièrement, le système du LIMSI montre que l'emploi de treillis de phones permet d'obtenir des estimations des fréquences de n-grammes plus précises, ce qui offre de meilleures performances.

Nous allons voir dans le chapitre suivant quelques modélisations de certains aspects de la prosodie. Les systèmes présentés ne sont plus alors dirigés vers les performances mais sont établis dans une optique de vérifications de théories linguistiques.

Chapitre 3

L'identification automatique des langues : méthodes & approches prosodiques

Sommaire

3.1 Systèmes comparatifs 54
3.1.1 Les travaux de Ramus 54
3.1.2 Les travaux de Grabe 55
3.1.3 Les travaux de Galves 55
3.2 Systèmes descriptifs (intonation) 57
3.2.1 Le système ToBI [119] 58
3.2.2 Le système IViE [51] 60
3.2.3 Modèle Intsint [60] 62
3.2.4 Modèle de Fujisaki [42] 63
3.2.5 Modèle de Gårding [48] 67
3.2.6 Modèle de Mertens [89] 68
3.3 Systèmes applicatifs 71
3.3.1 Modèle de Leavers [76] 71
3.3.2 Modèle d'Itahashi [66] 72
3.3.3 Le système de Cummins [24] 74
3.3.4 Le système de Li [78] 76
3.3.5 Modèle d'Adami [2] 77
3.4 Conclusion 79

Ce chapitre est consacré aux approches prosodiques employées ou employables pour l'identification automatique des langues. Ces systèmes peuvent être classés en trois grandes catégories :

- les systèmes comparatifs, conçus dans le but de vérifier les hypothèses linguistiques de différences entre les langues (classes rythmiques, isochronie),
- les systèmes descriptifs, conçus pour prendre en compte les réalités perceptuelles et de production de parole, qui peuvent permettre de mettre en valeur des différences entre les langues,
- les systèmes applicatifs, conçus dans un but d'amélioration de performances en identification des langues.

Les systèmes comparatifs sont consacrés à l'étude des différences rythmiques entre les langues. Le point faible de ces systèmes est le manque d'automatisation des processus (une segmentation et un étiquetage manuel sont souvent nécessaires), ce qui rend difficile les expériences portant sur des bases de données de taille importante.

Les systèmes applicatifs peuvent aussi bien prendre en compte des paramètres rythmiques qu'intonatifs, voire corrélés à la fois au rythme et à l'intonation. Ces systèmes sont directement appliqués à des tâches d'identification des langues sur des bases de données conséquentes, et permettent ainsi de mesurer les apports de la modélisation de la prosodie. Cependant, le manque de lien avec des théories linguistiques nuit à l'intérêt de tels systèmes.

Les systèmes descriptifs sont uniquement conçus pour la représentation et la compréhension des phénomènes prosodiques. Ils sont souvent appliqués à une seule langue, même si les recherches s'orientent de plus en plus vers des descriptions indépendantes de la langue. L'application de ces systèmes à plusieurs langues permet de visualiser les différences exploitables pour l'identification des langues.

3.1 Systèmes comparatifs

Les principales méthodes de modélisation du rythme en vue de la classification des langues sont décrites ci-dessous. Les auteurs de ces méthodes essaient de trouver des moyens efficaces et simples pour confirmer ou infirmer les théories linguistiques portant sur les regroupements rythmiques des langues.

Le point de départ des travaux récents sur le rythme a été la méthode de description du rythme proposée par Ramus à partir de 1999 [111]. Par la suite, d'autres se sont inspirés de ces travaux et ont proposé différentes méthodes de modélisation du rythme plus complexes (par exemple [50] et [47]).

Le point faible commun à la plupart de ces méthodes est qu'elles n'ont pour le moment été testées qu'après une segmentation manuelle des voyelles dans le signal de parole, ce qui implique souvent des corpus de taille relativement réduite.

3.1.1 Les travaux de Ramus

Dans son article [111], Ramus propose une méthode de classification des langues selon le rythme. Cette approche est basée sur une conception du rythme de parole comme étant la conséquence de propriétés phonologiques liées à l'identité des langues : la complexité des syllabes, la corrélation entre poids syllabique et accent, la présence ou non de réduction vocalique. Ramus propose une analyse de la complexité syllabique d'une langue afin de déterminer sa classe rythmique. La complexité est mesurée à l'aide d'une segmentation manuelle en consonnes et voyelles. Les paramètres sont :

- %V la proportion (en durée) d'intervalles vocaliques dans la phrase,
- ΔV l'écart-type des durées d'intervalles vocaliques par phrase,
- ΔC l'écart-type des durées d'intervalles consonantiques.

Ces paramètres ont été employés sur un corpus composé de huit langues (anglais, néerlandais, polonais, français, espagnol, italien, catalan et japonais). Quatre locutrices sont enregistrées par langue, chacune lisant cinq phrases.

Sur ces données, les paramètres font apparaître clairement des regroupements entre les langues.

- Le plan (%V,ΔC) fait ressortir trois groupes qui correspondent aux classes rythmiques décrites dans la littérature : anglais, néerlandais et polonais pour les langues accentuelles, espagnol, italien et français pour les langues syllabiques, et japonais pour les langues moraïques. Les langues accentuelles admettent plus de syllabes complexes, donc des groupes de consonnes de taille importante. En conséquence, les langues accentuelles ont un faible %V. Les langues admettant les syllabes complexes admettent aussi les syllabes plus simples, donc les groupes consonantiques sont plus variés. Le ΔC des langues accentuelles est donc plus élevé que celui des langues syllabiques.
- En considérant le plan (%V,ΔV), la variable ΔV est moins directement liée aux

classes de rythme. Cependant ΔV apporte une information supplémentaire puisqu'elle suggère que le polonais a des différences importantes avec les autres langues accentuelles.

3.1.2 Les travaux de Grabe

Dans ses articles [50] et [52], Grabe propose une méthode de prise en compte de la durée pour l'identification des langues. Elle propose de mesurer la variabilité de la durée d'intervalles acoustico-phonétiques successifs en employant un paramètre appelé « Pairwise Variability Indice » (PVI). Cette approche est novatrice car le PVI est différent du %V et ΔC de Ramus [111] puisqu'il prend en compte le niveau de variabilité entre les intervalles vocaliques et intervocaliques successifs (normalisé pour les variations de débit).

Le Pairwise Variability Indice (PVI) est défini selon :

$$rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}|/(m-1) \tag{3.1}$$

avec d_k la durée de la voyelle à l'instant k et m le nombre de voyelles dans l'extrait.

Pour s'affranchir des variations liées au débit, une version normalisée de cet indice est définie :

$$nPVI = \sum_{k=1}^{m-1} |\frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2}|/(m-1) \tag{3.2}$$

Enfin, cet indice peut être calculé en prenant en compte les durées des voyelles ou les durées inter-vocaliques, on notera alors respectivement intra-PVI et inter-PVI.

Les expériences sont menées sur un corpus de 18 langues ou dialectes (thai, hollandais, allemand, anglais britannique, tamoul, malais, anglais de Singapour, estonien, roumain, gallois, grec, polonais, français, catalan, japonais, luxembourgeois, espagnol et mandarin), avec un locuteur par langue sauf pour le français et l'espagnol (7 locuteurs).

Les langues accentuelles (anglais, allemand, hollandais) sont bien séparées des langues syllabiques (français, espagnol), elles possèdent des valeurs de intra-nPVI plus importantes. Le Pairwise Variability Indice ne donne cependant pas une séparation des 18 langues en groupes syllabiques et accentuels, mais une répartition suivant un continuum.

3.1.3 Les travaux de Galves

Les résultats donnés dans Ramus [111] résultent d'expériences sur de courts échantillons de quelques langues. L'implémentation de cette méthode sur des données plus

volumineuses est une tâche difficile puisqu'elle dépend d'un étiquetage manuel préalable (coûteux). Cela justifie l'emploi d'une méthode alternative, décrite dans [47]

Cette approche emploie une mesure de la sonorité. En effet, des expériences ont montré que les nouveau-nés sont capables de discriminer entre différentes classes rythmiques avec un signal filtré à 400 Hz [88]. Il est pourtant difficile de distinguer les différents sons après ce filtrage. Galves suggère que la discrimination des classes rythmiques ne repose pas sur une distinction fine entre les voyelles et les consonnes mais sur une distinction grossière entre la perception de la sonorité par opposition à l'obstruence. Il est donc possible d'envisager une classification utilisant une mesure de sonorité.

Galves propose de définir une fonction de mesure de la sonorité. Elle prendra des valeurs comprises entre 0 et 1 : proche de 1 pour des fenêtres montrant des motifs réguliers (caractéristiques des portions sonores du signal), et proche de 0 pour les régions caractérisées par l'obstruence.

Cette fonction est appliquée au spectrogramme du signal. Elle sera appelée $s(t)$, t étant le temps. t appartient à l'ensemble ku, $k = \{1, ..., T\}$, avec u l'unité de base du spectrogramme et T le nombre de points du spectrogramme. Ici, la valeur de u est fixée à 2 ms.

Les valeurs du spectrogramme sont estimées avec une fenêtre gaussienne de 25 ms. Seules les fréquences entre 0 et 800 Hz sont considérées. Soit $c_t(i)$ le coefficient de Fourier pour la fréquence i autour du temps t pour le spectrogramme, le spectre de puissance normalisé est défini par :

$$p_t(i) = \frac{c_t(i)^2}{\sum_f c_t(f)^2} \tag{3.3}$$

ce qui définit une séquence de mesures de probabilités $p_t : t = 1, ..., T$.

Les motifs réguliers (parties sonores) vont correspondre aux séquences dont les mesures de probabilité sont proches en termes d'entropie relative. L'entropie relative pour p_t par rapport à p_{t-1} est donnée par :

$$h(p_t|p_{t-1}) = \sum_i p_t(i) log(\frac{p_t(i)}{p_{t-1}(i)}) \tag{3.4}$$

L'entropie relative est toujours un nombre positif (inégalité de Jensen) et est proche de 0 quand les mesures sont similaires. La fonction de sonorité est alors définie par :

$$s(t) = 1 - min(1, \frac{1}{27} \sum_{u=t-4}^{t+4} \sum_{i=1}^{3} h(p_u|p_{u-i})) \tag{3.5}$$

Cette fonction est caractérisée par :

- La moyenne de $s(t)$: $\overline{S} = \frac{1}{T} \sum_{t=1}^{T} s(t)$ qui jouera le rôle de %V ;
- $\delta S = \frac{1}{T} \sum_{t=1}^{T} |s(t) - s(t-1)|$: les valeurs de $p(t)$, et donc de $s(t)$, montrent de grandes variations quand t appartient à des intervalles de forte obstruence et sont

quasi-constantes pour les zones de grande sonorité. Cet estimateur jouera le rôle de ΔC.

Les calculs de S et δS sont effectués sur les phrases utilisées par Ramus [111]. Ils permettent d'effectuer les mêmes regroupements rythmiques (§3.1.1) sans étiquetage préalable. Les graphiques représentant (%V,$\overline{S}$) et (ΔC,δS) montrent que les statistiques sont corrélées. La distribution de la sonorité montre une plus grande dispersion pour le japonais que pour l'allemand, avec les langues syllabiques au milieu. La probabilité d'avoir une sonorité inférieure à 0.3 augmente du japonais vers l'allemand, avec toujours les langues syllabiques en position intermédiaire.

Galves suggère que les affirmations des psycholinguistes (tels que Mehler [88] : « les enfants se servent des voyelles pour la représentation prosodique, et les enfants peuvent représenter la parole comme une séquence de voyelles ») peuvent être améliorées en remplaçant le terme « voyelle » par « sonorité ». Les mécanismes employés par les enfants pour distinguer les classes rythmiques ne pourraient pas employer des calculs statistiques sur grands échantillons de parole (comme le ΔC) mais devraient reposer sur une procédure simple utilisant des paramètres acoustiques robustes (comparaisons entre les valeurs successives de s(t)).

3.2 Systèmes descriptifs (intonation)

L'intonation, composante primordiale de la prosodie, se rapporte aux mouvements mélodiques de la parole et se caractérise par les mouvements de fréquence fondamentale (§1.1). Actuellement, deux approches formelles existent : l'approche « holistique » et l'approche « autosegmentale » [33, 73].

L'approche autosegmentale décrit l'intonation en terme de séquences de segments tonaux engendrés au moyen d'une grammaire à état finis. Cette approche est souvent associée à la « théorie des deux niveaux » (représentés par des tons bas L pour « LOW », et des tons haut H pour « HIGH »). Cette théorie, du fait de sa simplicité, facilite l'élaboration d'outils fiables pour la transcription et l'étiquetage de l'intonation. Il existe au moins trois systèmes de ce type : TOBI [119], INTSINT ([60]) et IVIE [51]. TOBI présuppose une connaissance approfondie de la phonologie prosodique de la langue, ce qui rend difficile son utilisation en dehors de l'anglais américain. Ce n'est pas le cas des deux autres systèmes qui se révèlent relativement neutres vis-à-vis de la langue.

Dans l'approche holistique, l'intonation est décrite sous forme de patrons mélodiques prototypiques par Vaissière [123, 124], de configurations globales constitutives d'un lexique intonatif par Aubergé [9] ou bien de contours stylisés considérés comme représentatifs des différents « intonèmes » fonctionnellement distincts dans une langue donnée par Delattre [28]. L'approche holistique est parfois abordée dans une logique superpositionnelle, postulant que la construction du contour intonatif résulte de la superposition de domaines. Par exemple Fujisaki [42] considère la superposition d'une composante accentuelle et d'une

composante syntagmatique.

Quelques systèmes de description de l'intonation sont détaillés ci-dessous. Ces systèmes tentent de décrire la courbe de fréquence fondamentale au moyen de règles liées aux mécanismes de production ou de perception de la parole.

3.2.1 Le système ToBI [119]

Le système ToBI (*Tones and Break Indices*) est certainement le système de transcription de la prosodie le plus connu. L'objectif est d'avoir à long terme un seul système de transcription de l'intonation (pour l'anglais américain dans un premier temps) qui serait utilisé par la communauté scientifique et qui permettrait de comparer les résultats obtenus.

Description

En addition à la représentation graphique du contour de la fréquence fondamentale, le système possède quatre niveaux de transcription parallèles (appelés *tiers*). Les quatre niveaux d'étiquetage symbolique sont :

- un niveau orthographique, pour indiquer les mots prononcés au cours de l'énoncé,
- un niveau tonal, pour indiquer les éléments constrastifs présents dans le contour de fréquence fondamentale (*Tones*),
- un niveau d'indice de rupture (*Break Indices*), pour indiquer la force de la cohérence et de la disjonction entre des mots adjacents,
- un niveau additionnel qui indique des effets induits par la variabilité de la parole spontanée tels que les rires ou les hésitations, qui sont nécessaires pour interpréter les éléments des autres niveaux de transcription.

Chaque niveau est constitué d'un jeu de symboles représentant des évènements prosodiques, associés à l'instant auquel ils se produisent.

L'analyse du niveau tonal suppose une hiérarchie des phrases intonatives contenant une ou plusieurs phrases intermédiaires. Le niveau tonal permet de transcrire l'intonation et la structure prosodique. L'intonation est transcrite par une séquence linéaire d'évènements distribués sur la phrase. Ceux-ci sont basés sur la phonologie intonative de Pierrehumbert [106]. Trois types d'évènement tonaux sont annotés :

- Les frontières de phrases intonatives sont marquées par H% ou L%. Une étiquette optionnelle %H peut indiquer le début.
- Un accent tonal noté H ou L doit se situer après le dernier accent annoté dans la phrase intermédiaire. Ce ton est alors considéré jusqu'à la fin de la phrase ou jusqu'au ton de la frontière suivante.
- Les accents sont associés avec les syllabes accentuées des mots proéminents. Il doit y avoir au moins un accent sur le mot qui est le plus proéminent dans la phrase.

Les cinq types d'accent sont notés ci-dessous, avec des exemples de contours pour lesquels ils apparaissent habituellement (en anglais américain) :

- H* : ton haut (déclarative),
- L* : ton bas (questions oui/non),
- L+H* : montée du ton bas vers le ton haut (contraste),
- L*+H : montée tardive (incertitude pragmatique),
- H+ !H* : descente sur une syllabe accentuée (inférence pragmatique).

Le niveau d'indice de rupture indique le degré de disjonction entre les mots par un chiffre variant entre 0 et 4. Le codage employé correspond à ·

- 0 : entre des mots qui sont regroupés phonétiquemment (en débit rapide),
- 1 : entre deux mots prosodiques différents,
- 2 : une disjonction forte marquée par une pause mais sans marque tonale, ou une disjonction qui est plus faible que celle attendue lors d'une frontière de phrase intermédiaire ou intonative,
- 3 : pour une frontière de phrase intermédiaire,
- 4 : pour une frontière de phrase intonative.

La figure 3.1 illustre un étiquetage avec le système ToBI.

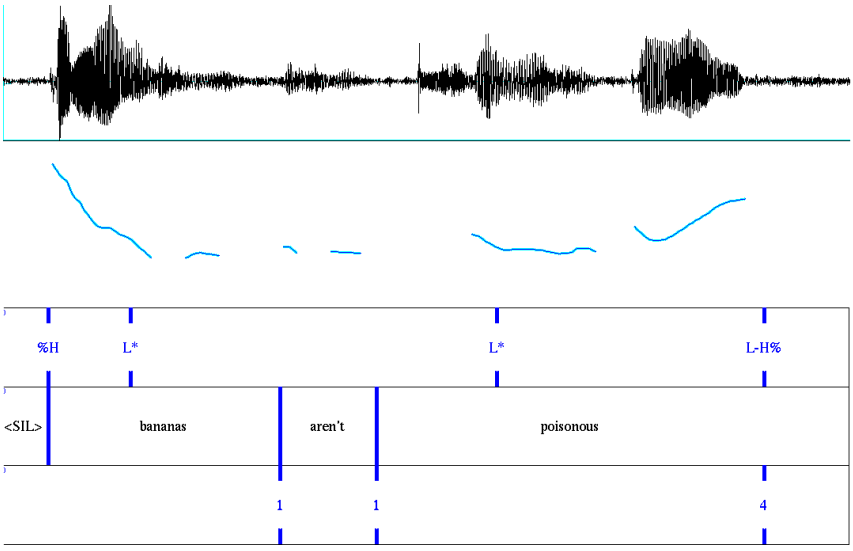


Fig. 3.1 : Exemple d'étiquetage de la prosodie avec ToBI sur la phrase *Bananas aren't poisonous.*

Critiques

Le système ToBI n'a été développé que sur l'anglais américain standard, mais avec la possibilité de l'étendre à d'autres langues. La volonté de faire de ToBI un standard universel pour les études sur la prosodie de l'anglais américain a motivé l'adaptation de ce système à d'autres langues. Ainsi, malgré la difficulté de la tâche, Grice [54] a développé un ToBI allemand avec toutes les modifications que cela suppose. Les contours prosodiques allemands sont représentés par un jeu de symboles marquant le type de ton et les marques

de frontières.

Le système ToBI a été largement critiqué ces dernières années, notamment par Wightman [126]. Un nombre croissant de chercheurs rejettent la nature descriptive du système ToBI en faveur de systèmes qui prennent plus en compte les aspects liés à la perception. Martin [86] suggère que le système ToBI ne peut pas prétendre à l'universalité, et que certaines langues nécessitent d'autres outils. Nolan et Grabe [97] jugent que ce système hésite entre une représentation phonologique et une représentation phonétique, ce qui entraîne des difficultés pour adapter ToBI aux autres langues, et même à d'autres variantes de l'anglais. Grabe et ses collègues [51] proposent d'ailleurs une alternative avec IViE (pour *Intonational Variation in English*), dont le principe s'inspire du système ToBI mais qui est plus adapté à l'anglais britannique.

3.2.2 Le système IViE [51]

IViE signifie *Intonational Variation in English*. IViE s'inspire de ToBI, le standard actuel pour l'annotation prosodique de l'intonation en anglais américain [119], mais IViE permet l'obtention de transcriptions comparables de plusieurs variétés d'anglais par un seul système d'étiquetage. De plus, les transcriptions IViE capturent les différences rhythmiques entre les variétés et les différences dans la réalisation phonétique.

Dans le système IViE, la prosodie est transcrite sur trois niveaux :

1. structure rythmique,
2. structure acoustico-phonétique,
3. structure phonologique.

Les trois niveaux permettent de transcrire les variations rythmiques, les variations de réalisation des accents et la variation dans la structure intonative :

1. Tout d'abord, les syllabes rythmiquement proéminentes sont repérées. Ces syllabes peuvent être accentuées ou non et sont étiquetées « P » (proéminentes). Cette étiquette est généralement placée au milieu d'une voyelle. Deux symboles supplémentaires sont également disponibles : « % » et « # ». « % » indique la localisation d'une frontière rythmique, et « # » transcrit la localisation des hésitations ou des interruptions.
2. Ensuite, les mouvements de fréquence fondamentale autour des syllabes proéminentes sont étiquetés. La réalisation des accents est transcrite dans des Domaines d'Implémentation (*Implementation Domains ou IDs*). Une ID contient : la syllabe pré-accentuelle, la syllabe accentuée, toutes les syllabes inaccentuées (s'il y en a) jusqu'à la prochaine syllabe accentuée.

   Trois niveaux de fréquence fondamentale sont disponibles pour la transcription : h(igh), m(mid) et l(ow). Ces niveaux sont relatifs. Lorsque ces niveaux sont transcrits en majuscule, ils indiquent le niveau sur la syllabe accentuée, les autres syllabes sont transcrites en minuscules. Le niveau atteint à la fin d'un ID est précédé par un tiret.

Ainsi, l'étiquette `lL-h` signifie que la valeur de fréquence fondamentale sur la syllabe préaccentuelle est basse, elle reste basse sur la syllabe accentuée et elle augmente sur la syllabe finale.

S'il n'y a pas de syllabe pré- ou post-accentuelle, l'étiquette comporte un symbole minuscule de moins. Si la direction de la fréquence fondamentale change au cours d'une syllabe, une quatrième étiquette peut être ajoutée.

3. Le niveau phonologique est utilisé pour transcrire l'intonation au niveau phonologique. Les étiquettes disponibles pour ce niveau ne sont pas dépendantes d'une variété particulière d'anglais. Elles font partie d'un ensemble parmi lequel sont choisies les étiquettes qui correspondent à chaque variété dialectale. Les étiquettes disponibles sont les suivantes :

   (a) étiquettes d'intonation :

   - H*+L,
   - L*+H,
   - H*,
   - L*,
   - L*H+L,
   - H*L+H

   (b) modificateurs :

   - ˆ : montée
   - ! : descente
   - _ : déplacement d'un accent vers la droite (e.g. H*+_L) : modifie la localisation d'un accent dans le domaine temporel.

   (c) spécifications des frontières :

début de phrase	fin de phrase
%H	H%
%0	0%
%L	L%

   Comme ToBI, IViE possède des étiquettes de frontières de phrase : H% et L%. Additionellement, IViE propose 0%, qui permet d'annoter les frontières pour lesquelles il n'y a pas de mouvements de F_0.

Un exemple d'étiquetage par le système de transcription IViE est illustré sur la figure 3.2.

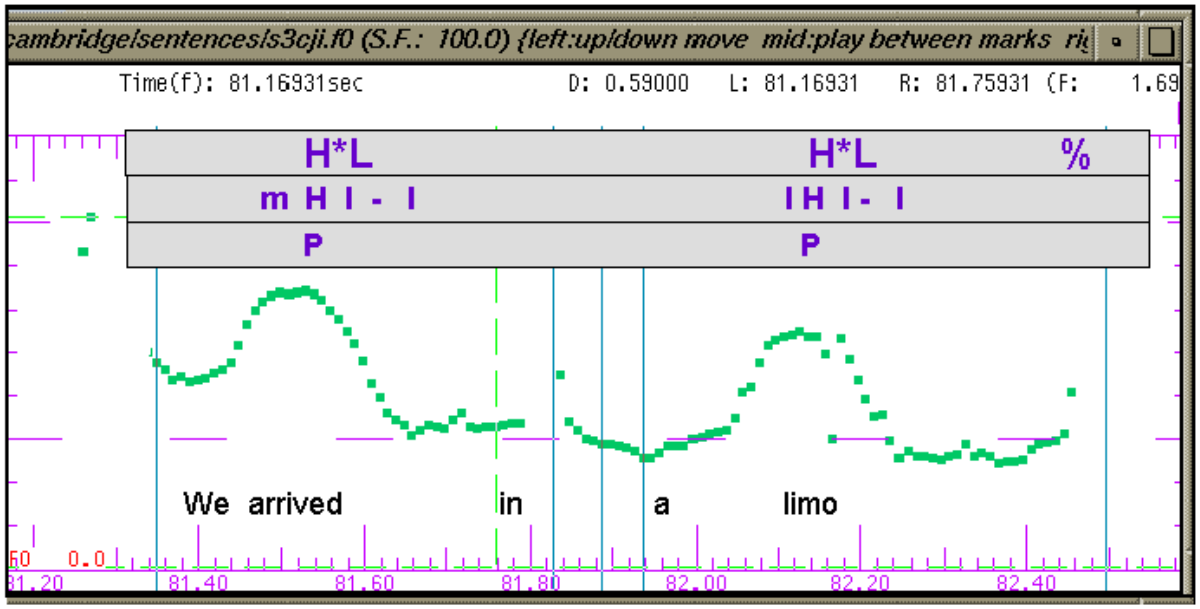


Fig. 3.2 : Exemple d'étiquetage par le système de transcription IViE

3.2.3 Modèle Intsint [60]

Différents modèles phonétiques décrivant les motifs de fréquence fondamentale ont été développés pour essayer de prendre en compte les aspects physiologiques de la production de contours intonatifs ([42] par exemple). D'autres chercheurs ont concentré leurs efforts sur la modélisation des aspects perceptuels des motifs intonatifs ([89] par exemple). La conviction de Hirst est qu'aucune de ces approches ne donne une image complète de la réalité : tous les aspects (perception et production) doivent être pris en compte dans un modèle plus général.

Plusieurs modèles phonétiques/phonologiques ont été proposés afin de générer une courbe intonative à partir d'une entrée symbolique. Dans le cadre de l'analyse de la parole, c'est le problème inverse qui est rencontré. Le modèle proposé ici a pour principal objectif d'être inversible, c'est-à-dire qu'il ne s'agit pas uniquement de passer de la courbe de fréquence fondamentale à une représentation symbolique, mais que l'inverse doit aussi être possible.

Le codage Intsint est basé sur une modélisation de la courbe de fréquence fondamentale appelée Momel. La procédure employée est la suivante :

– Momel :
L'algorithme Momel permet une stylisation de la courbe de fréquence fondamentale. Des points cibles sont automatiquement déterminés. Ils sont ensuite reliés (y compris au niveau des parties non voisées) au moyen d'une fonction spline quadra-

tique.

– Intsint : Chaque point cible est codé par un symbole, représentant soit un ton absolu (T, B, M) dépendant du locuteur, soit des tons relatifs (H,S,L,U,D) dépendants du point-cible précédent. Pour ces tons relatifs, une distinction est faite entre les tons non-itératifs (H, S, L) et les tons itératifs (U, D). Le codage des tons est résumé dans le tableau suivant :

Tab. 3.1 : Codage des tons Intsint

tons		montant	neutre	descendant
Absolus		T	M	B
Relatifs	Non-itératifs	H	S	L
	Itératifs	U	-	D

Un exemple d'étiquetage des contours de la fréquence fondamentale est donné sur la figure 3.3.

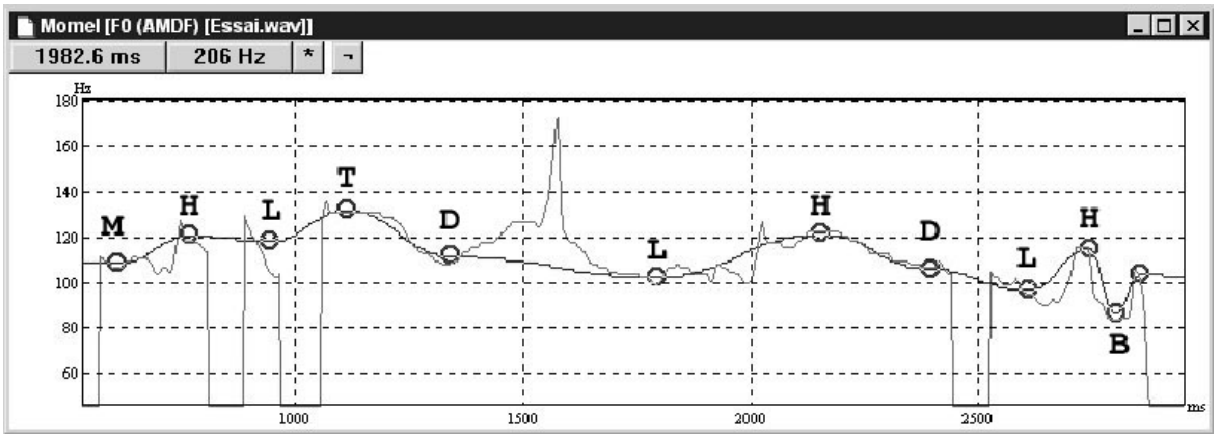


Fig. 3.3 : Exemple d'étiquetage des contours de la fréquence fondamentale (stylisation MOMEL et codage INTSINT)

3.2.4 Modèle de Fujisaki [42]

Les informations exprimées par la parole peuvent être vues comme faisant partie de trois catégories : linguistiques, paralinguistiques et non-linguistiques. Un locuteur organise les différentes unités linguistiques dans une phrase expressive et sensée en général grâce à trois moyens : l'accentuation, le phrasing, et les pauses.

– L'accentuation est définie comme le fait de changer (en général augmenter) la proéminence relative d'une syllabe dans un mot ou un groupe de mots. Cet aspect est lié à la fréquence fondamentale, la durée et l'intensité, même si les langues diffèrent dans la manière d'utiliser ces paramètres.

– Le « *Phrasing* » correspond à un regroupement de mots en un constituant perceptuellement cohérent. Cela s'effectue grâce à la fréquence fondamentale et au débit local.
– Le « *Pausing* » signifie mettre une pause après un élément pour indiquer que les éléments aux deux extrémités de la pause doivent être traités séparément.

Dans le but d'inclure dans le modèle l'organisation sous-jacente aux observations caractéristiques de la parole, il est logique de penser à ces deux étapes :

1. inclure les commandes des caractéristiques de la parole ;
2. inclure les unités et les structures de la prosodie à partir de ces commandes.

Puisque l'étape 1 est l'opération inverse de la production de parole, elle est plus certainement et objectivement conduite si nous disposons d'un modèle quantitatif du processus de production. Un tel modèle a été présenté pour les contours de fréquence fondamentale du japonais, et a été utilisé avec succès pour l'analyse et la synthèse, c'est le modèle de Fujisaki [44].

Un modèle quantitatif pour générer les contours de fréquence fondamentale des mots et phrases du japonais

Après observation, Fujisaki suggère qu'un contour typique de fréquence fondamentale sur une phrase peut être considéré comme constitué de deux éléments. Le premier est une fonction avec peu de variation qui peut éventuellement présenter une légère montée initiale et qui diminue ensuite graduellement vers une asymptote. Elle peut être renforcée à certaines frontières syntaxiques, au moins dans le cas du japonais. Le deuxième est composé de "bosses" (pics ou plateaux) qui correspondent aux motifs accentuels des mots constituant la phrase. Les hypothèses suivantes sont formulées :

1. Les commandes de phrase sont un ensemble d'impulsions et les constituants de phrase sont la réponse d'un système linéaire du second ordre à ces commandes.
2. Les commandes accentuelles sont un ensemble de fonctions échelon et les constituants accentuels sont les réponses d'un autre système linéaire du second ordre à ces commandes.
3. Les constituants de phrase et d'accent sont superposés et produisent un changement proportionnel dans le logarithme de fréquence fondamentale.

A partir de ces hypothèses, un modèle a été construit pour générer les contours de fréquence fondamentale sur les phrases. Dans ce modèle, le contour de la fréquence fondamentale est exprimé par :

$$\log_e F_0(t) = \log_e F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \qquad (3.6)$$

Avec :

$$G_p(t) = \begin{cases} \alpha^2 t e^{(-at)} & \text{pour } t \geq 0 \\ 0 & \text{pour } t < 0 \end{cases} \qquad (3.7)$$

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma] & \text{pour } t \geq 0 \\ 0 & \text{pour } t < 0 \end{cases} \tag{3.8}$$

où $G_p(t)$ représente la réponse impulsionnelle du mécanisme de contrôle de phrase et $G_a(t)$ est la réponse à un échelon du mécanisme de contrôle de l'accentuation.

Nous avons :
F_b valeur initiale de la fréquence fondamentale, I nombre de phrases de commande, J nombre de commandes accentuelles, A_{pi} amplitude de la ième commande de phrase, A_{ai} amplitude de la jème commande d'accent, T_{0i} durée de la ième commande de phrase, T_{1j} début de la jème commande d'accent, T_{2j} fin de la jème commande d'accent, α fréquence angulaire naturelle du mécanisme de contrôle de phrase, β fréquence angulaire naturelle du mécanisme de contrôle d'accent, γ niveau relatif de plafond des constituants accentuels.

Un exemple de génération de contours de fréquence fondamentale est représenté sur la figure 3.4.

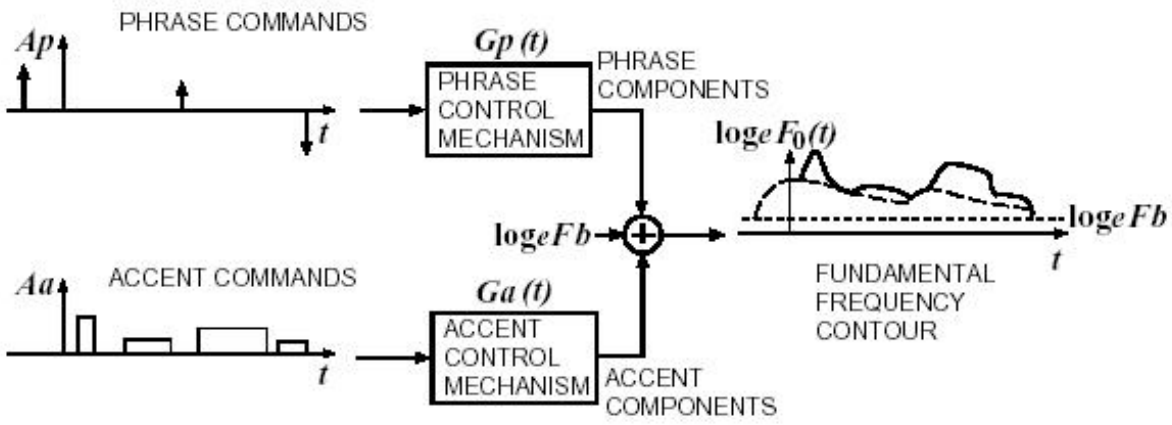


Fig. 3.4 : Exemple de génération d'un contour de fréquence fondamentale par le modèle de Fujisaki (image extraite de [42]).

L'hypothèse est faite que les paramètres α et β sont constants au moins sur la phrase, tandis que le paramètre γ est fixé à 0,9. Une descente rapide de fréquence fondamentale, parfois observée à la fin d'une phrase et occasionnellement sur une frontière, peut être vue comme la réponse du mécanisme de contrôle de phrase à une impulsion négative. Grâce à la technique « analyse-par-synthèse », il est possible de décomposer un contour de fréquence fondamentale donné en ses constituants et, ainsi, d'estimer l'amplitude et la durée de leurs commandes sous-jacentes par déconvolution. Le contour de fréquence fondamentale ainsi généré est si proche du contour mesuré qu'ils sont perceptiblement indistinguables en parole resynthétisée. Le modèle peut prédire et générer le contour entier à partir d'un

ensemble de commandes. De plus, l'accord entre la sortie du modèle et la mesure atteste de la validité du modèle.

Les occurrences temporelles des commandes semblent proches du contenu linguistique de la phrase. Les commandes d'accent débutent 40 à 50 ms avant le déclenchement d'une voyelle accentuée et finissent 40 à 50 ms avant la fin du segment moraïque. Les commandes de phrase sont localisées approximativement 200 ms avant le déclenchement de la phrase et également devant une frontière syntaxique majeure. En général, les commandes de phrase sont plus importantes en position initiale, et plus faibles en milieu de phrase. Ainsi, le contour global de fréquence fondamentale montre une déclinaison au long de la phrase.

L'analyse montre aussi que les variations des valeurs de α et β sont aussi faibles d'une phrase à l'autre que d'un individu à l'autre.

Ce modèle peut donc permettre de séparer les facteurs qui sont proches des informations linguistiques et paralinguistiques comme l'amplitude et le timbre de ces commandes, des facteurs qui sont proches des mécanismes physiques et physiologiques de contrôle phonatoire comme les réponses caractéristiques, c'est-à-dire les formes de constituants de phrase et d'accent.

Application du modèle aux contours de fréquence fondamentale de différentes autres langues

Le succès du modèle sur les contours mélodiques du japonais suggère son applicabilité aux autres langues, en considérant le fait que le modèle capture les caractéristiques essentielles du larynx humain, qui est similaire (sinon identique) parmi les locuteurs des différentes langues. De nombreuses expériences ont été menées dans ce sens (voir [43, 45, 46, 87] par exemple). Les résultats montrent des différences majeures entre deux types de langues :

- langues avec uniquement des commandes locales positives : les contours de fréquence fondamentale de plusieurs phrases de différentes langues (anglais, allemand, grec, coréen, polonais et espagnol) ont été analysés. Les résultats montrent que les polarités des commandes en entrée et les mécanismes de réponse sont essentiellement similaires à ceux du japonais courant, ce qui prouve que le modèle peut être appliqué aux motifs de fréquence fondamentale de toutes ces langues.
- langues avec à la fois des commandes locales positives et négatives : l'analyse de contours de différentes langues incluant le chinois standard, le thai et le suédois indique que les constituants locaux (associés aux tons dans le cas du chinois) ne sont pas toujours positives mais peuvent être positives et négatives.

Conclusions

Ce modèle est capable de générer de bonnes approximations d'un grand nombre de contours de fréquence fondamentale à partir d'un nombre limité de paramètres. Les mé-

canismes physiologiques et physiques sous-jacents ont été élucidés. Finalement, l'applicabilité du modèle à la génération de contours de fréquence fondamentale pour différentes langues a été démontrée, indiquant l'utilité du modèle pour la synthèse vocale multilingue. Cependant, cette méthode est difficile à mettre en œuvre pour l'analyse automatique, la technique d'« analyse-par-synthèse » étant coûteuse en termes de temps de calcul.

3.2.5 Modèle de Gårding [48]

Hypothèses

Le modèle de Gårding a été tout d'abord conçu sur le suédois, puis appliqué à d'autres langues telles que le français et le grec. Ce modèle est basé sur une analyse qui sépare la prosodie lexicale de la prosodie de la phrase. L'entrée du modèle est une phrase, pour laquelle les tons ou accents lexicaux sont marqués, ainsi que les accents au niveau de la phrase, les frontières morphologiques et de phrase, et le mode d'intonation de phrase. L'hypothèse est que tous ces facteurs sont combinés et interagissent pour produire le motif temporel et tonal du signal de parole. Le modèle simule ce procédé en un nombre fini d'étapes. Ces étapes ne sont pas censées simuler le procédé de production de la parole. Ce modèle doit être vu comme étant un essai de systématisation de la description prosodique d'une langue.

Le modèle

Le modèle se décompose en plusieurs étapes :

1. Règles de structures syllabiques
2. Règles de durées syllabiques
3. Règles phonologiques intermédiaires
4. Représentation intermédiaire de l'intonation
5. Algorithme de génération de fréquence fondamentale

Ici, nous allons considérer que les deux premières étapes sont réalisées. Les règles phonologiques intermédiaires (étape 3) sont nécessaires pour prendre en compte les cas où les marqueurs sur la phrase d'entrée ont un effet uniquement sur la durée, et non pas sur la fréquence fondamentale.

Les règles de la quatrième étape, la représentation intermédiaire de l'intonation, permettent de convertir les étiquettes d'entrée (abstraites) en symboles plus concrets. Les caractéristiques globales, qui concernent l'intonation de phrase, sont exprimées en termes de montées, descentes, ou stabilités, ou une combinaison de ces éléments. Les caractéristiques locales, concernant les syllabes ou les mots, sont exprimées par une combinaison de hauts et de bas. La spécification de haut et de bas peut varier selon les langues, et même selon les différents dialectes de chaque langue. Cette différence, combinée avec la règle de

délétion de l'accent décrite ci-dessus, peut faire émerger différents motifs intonatifs (un motif en « chapeau » dans une langue, contre un motif de « trou » dans une autre).

La dernière étape du modèle est l'algorithme de génération de la fréquence fondamentale. Il consiste en sept règles :

1. Intonation de phrase (détermination de la grille tonale en fonction du type de phrase et des frontières majeures de phrase (cadre global pour l'intonation de phrase dans lequel les mouvements locaux se développent))
2. Frontières de phrase (insertion des « hauts » et « bas » sur la grille d'après la langue ou le dialecte)
3. Accents de phrase
4. Accents de mots
5. Accents contrastifs de mots
6. Règles de contexte (ajustement des « hauts » et les « bas » d'après le contexte segmental et prosodique)
7. Concaténation (raccordement des points générés dans les segments voisés par une courbe)

Différentes applications à ce modèle ont été proposées dans [48]. Le modèle est employé pour comparer la prosodie de différents dialectes en suédois, puis pour comparer la prosodie de différentes langues (suédois, français et grec). Ensuite, le modèle est appliqué à la modélisation et la comparaison de contours intonatifs de phrases déclaratives ou interrogatives.

3.2.6 Modèle de Mertens [89]

Ce modèle, appelé *prosogramme*, est un modèle destiné à la représentation de la prosodie, contrairement aux deux précédents ([42] et [48]) qui sont plus axés sur la génération. Les principes de ce système de transcription de la prosodie sont :

1. la représentation de la prosodie se doit d'être objective, robuste, et facile à interpréter. Cette transcription doit représenter l'intonation perçue, et devrait permettre de distinguer les évènements prosodiques audibles de ceux qui sont inaudibles, que ce soit pour une syllabe ou une série de syllabes.
2. la transcription devrait montrer l'évolution de la fréquence fondamentale sur toute une phrase, afin de permettre d'identifier des phénomènes tels que la déclinaison et le registre, par exemple.
3. la courbe de fréquence fondamentale affichée doit être quantifiée pour permettre l'estimation d'intervalles mélodiques au niveau local ou bien global.
4. l'organisation temporelle du signal de parole doit être préservée afin d'identifier les pauses et les hésitations, pour déterminer le débit et étudier le rythme.
5. la transcription doit être automatique ou semi-automatique.

6. la transcription doit être neutre, c'est-à-dire indépendante d'un modèle, pour que des personnes avec des connaissances différentes puissent l'utiliser.
7. la transcription doit fournir des étiquettes phonétiques et orthographiques alignées, dans un souci de lisibilité et pour la recherche par le contenu.
8. La représentation quantifiée de l'intonation et du temps doit autoriser les manipulations, pour la synthèse et la resynthèse, ce qui permet d'évaluer la transcription.

La méthode employée pour ce modèle utilise une technique de stylisation de la fréquence fondamentale décrite dans [26]. Sa particularité réside dans le fait qu'elle est basée sur une simulation de la perception de l'intonation et que l'unité de base est le noyau syllabique. Le prosogramme est constitué de deux parties :

– le contour de fréquence fondamentale (stylisé),
– une ou plusieurs transcriptions alignées temporellement.

La transcription phonétique est nécessaire pour le calcul de la stylisation. Les autres types d'annotations (orthographiques, unités prosodiques, ...) sont optionnelles.

Différentes variantes sont disponibles pour la partie concernant les données de fréquence fondamentale. Le prosogramme basique n'affiche que la fréquence fondamentale perçue, tandis que la version « riche » montre également la valeur extraite de fréquence fondamentale et l'intensité.

Perception

Plusieurs phénomènes ont été observés pour la perception de la fréquence fondamentale dans la parole.

1. Pour être audible, une variation de fréquence fondamentale doit avoir une taille minimale. Cette taille varie en fonction de la fréquence de départ et de la durée du stimulus. Ce seuil de *glissando* a été mesuré pour des variations linéaires de la fréquence pour les tons purs, la parole de synthèse et la parole resynthétisée. Le seuil standard déterminé par des expériences psychoacoustiques pour les voyelles est : $G = 0.16/T^2$ en demi-tons par seconde, T étant la durée de la variation.
2. Les variations de fréquence fondamentale sont bien évidemment rarement linéaires pour la parole naturelle. Il faut alors tenir compte des changements de pente dans les variations. Les changements de pente sont comparés à un seuil de *glissando* différentiel. Quand le changement de pente est inférieur à ce seuil, le mouvement de chaque partie est remplacé par une simple variation linéaire. Il y a eu peu de recherche sur ce seuil, la valeur utilisée est de 20 demi-tons par seconde.
3. Ici, seules les variations de fréquence fondamentale pour des sons isolés ont été mentionnées. Cependant, dans la parole, la concaténation de sons entraîne des changements en intensité et en voisement, ainsi que des changements importants dans le spectre. Dans la plupart des cas, l'alternance de voyelles et de consonnes (ou de groupes de consonnes) a pour conséquence l'apparition d'un pic de sonorité et d'intensité pendant la voyelle, caractérisé par une relative stabilité spectrale. La voyelle constitue alors le noyau syllabique.

4. En parole continue, les sons et les syllabes se suivent rapidement et l'information de fréquence fondamentale doit être traitée en temps réel avant qu'elle ne soit masquée par les autres sons. La perception tonale est optimale pour des voyelles isolées, le cas le plus difficile étant pour de la parole continue à débit rapide : le seuil de glissando est plus élevé pour la parole continue. Les variations de fréquence fondamentale sont plus facilement perçues quand elles sont suivies par une pause. En d'autres termes, la présence d'une pause après la variation fait baisser le seuil de glissando.

Procédure de stylisation

À cause de l'absence de segmentation automatique de la parole en noyaux syllabiques, la solution adoptée ici est de ne considérer que les voyelles. L'information sur l'identité des voyelles est donnée par la segmentation phonétique.

Pour chaque voyelle, le noyau vocalique est déterminé. Il est défini comme étant la partie voisée autour du pic d'intensité, et est délimité par les points localisés à -3 dB (pour la droite) et -9 dB (pour la gauche) du pic. La valeur pour la frontière droite élimine la plupart des perturbations microprosodiques au début de la syllabe, de même que les phénomènes microprosodiques pour les consonnes voisées au niveau de la frontière syllabique. La valeur pour la frontière gauche préserve les variations tardives de fréquence fondamentale pour les voyelles accentuées. Les résultats dépendent largement de la qualité de l'alignement phonétique disponible.

Afin de déterminer le seuil de glissando optimal pour la parole continue, des stylisations obtenues pour différents seuils ont été comparées à des transcriptions manuelles.

En utilisant un seuil de $G = 0.16/T^2$, la stylisation rend compte de nombreux glissements intrasyllabiques qui ne sont pas présents dans les transcriptions manuelles. En d'autres termes, la stylisation avec ce seuil surestime les capacités auditives moyennes.

En employant un seuil de $G = 0.32/T^2$, le double du seuil standard, la stylisation est très proche des notations manuelles. Pour les variations de fréquence à long terme (plusieurs secondes), la transcription semi-automatique semble plus fiable que la transcription manuelle.

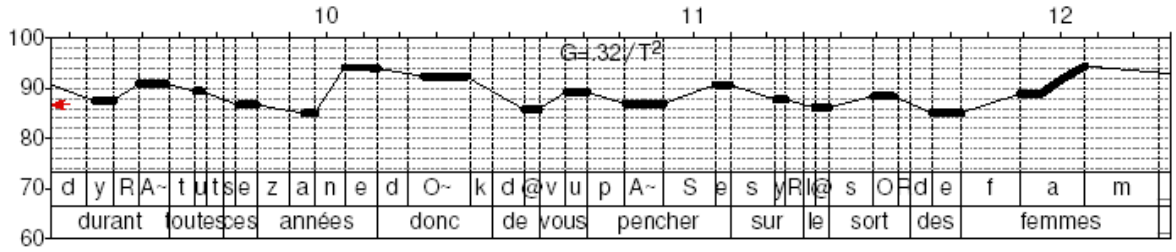


Fig. 3.5 : Exemple de stylisation de la fréquence fondamentale (seuil $G = 0.32/T^2$)

3.3 Systèmes applicatifs

3.3.1 Modèle de Leavers [76]

Le but de cette étude est d'imiter les stratégies cognitives employées par les humains pour identifier les langues : informations spectrales, temporelles et suprasegmentales (rythme, intonation, accent), segmentales (phonétiques), lexicales (mots). L'étude s'appuie sur des travaux en psycholinguistique, en production de la parole, et en reconnaissance des formes. Les paramètres linguistiques sont classés selon une stratégie similaire à celle éventuellement employée par les humains : un arbre de décision. La base de données se compose de 14 locuteurs (tous des hommes) : 4 chinois (locuteurs du mandarin), 5 anglais, 2 espagnols, 2 portugais, qui lisent dans un premier enregistrement le même passage du même livre et par la suite différents passages de ce même livre.

Ce système est basé sur les hypothèses suivantes :

1. intonation : il existe différents groupes de langues (tonales et autres). Pour les langues tonales (mandarin par exemple) les changements de F_0 se font au niveau du mot. Pour les autres (anglais par exemple) les mots sont accentués mais l'intonation porte sur la phrase.
2. paramètres acoustiques locaux : une comparaison des espaces vocaliques montre la présence de différences entre le portugais et l'espagnol. Ces différences sont caractérisées par la distribution spectrale de F_0 et la durée des voyelles.

Les paramètres correspondants seront employés :

1. suprasegmental : le but est d'étudier la fréquence fondamentale sur les phrases entières (différences pour les langues tonales), c'est à dire prendre en compte les dépendances temporelles de la fréquence fondamentale sur la phrase. Les langues tonales ne devraient pas avoir de dépendances fortes. Pour cela l'exponentielle de Hurst (caractérisant des similarités à long terme) sera employée (voir Annexe A).
2. segmental : il faut caractériser les différences entre les systèmes vocaliques (par exemple portugais / espagnol) avec la durée des voyelles (normalisée). Les plages sonores couvertes par les voyelles différant selon les langues, les distributions seraient asymétriques, d'où l'emploi du skewness ou coefficient d'asymétrie.

Les résultats obtenus sont les suivants :

1. Exponentielle de Hurst : les résultats sont donnés sous forme graphique, avec en abscisse le nombre d'échantillons de F_0 et en ordonnée les valeurs correspondante de l'exponentielle. Si la courbe est linéaire, cela caractérise un effet de mémoire. La pente est la valeur de l'exponentielle de Hurst. Les ruptures de linéarité caractérisent une perte de l'effet mémoire. Ici, on a une pente supérieure à 0.5 pour l'anglais, l'espagnol et le portugais (l'intonation varie sur la phrase : mémoire) ; alors que pour le mandarin, la pente est inférieure à 0.5 (l'intonation varie sur les mots, pas d'intonation globale de phrase).

2. Skewness et durée des voyelles : ces paramètres permettent la séparation entre l'espagnol et le portugais, mais ne donnent pas de résultats concluants sur les autres langues.

En conclusion, il n'est pas approprié d'employer la prosodie seule, d'où la nécessité d'informations segmentales. Un système hiérarchique avec une classification par arbre et l'utilisation de stratégies cognitives est possible. Ici, deux séparations majeures sont opérées : les langues tonales sont distinguées des autres, puis deux langues de la même famille (espagnol et portugais) sont séparées.

3.3.2 Modèle d'Itahashi [66]

Le système décrit dans [66] utilise des paramètres extraits de la fréquence fondamentale et des coefficients cepstraux. Les tests sont réalisés en prenant en compte soit uniquement les paramètres extraits de la fréquence fondamentale, soit les coefficients cepstraux, soit l'ensemble des paramètres.

Fréquence fondamentale

Les contours de la fréquence fondamentale sont extraits grâce à la méthode AMDF [58] avec :
- une fréquence d'échantillonnage de 8 kHz,
- une quantification sur 16 bits,
- une fenêtre d'analyse de 30 ms, intervalles de 10 ms.

Des erreurs dites d'harmoniques peuvent se produire lors de l'extraction de la fréquence fondamentale par la méthode AMDF. La fréquence estimée peut être le double ou la moitié de la fréquence réelle.

Une méthode de correction de ces erreurs grossières est proposée par [11]. Le principe est de calculer l'autocorrélation de toutes les valeurs candidates de la période fondamentale à chaque instant. La fréquence fondamentale estimée correspond au candidat qui possède la plus forte valeur d'autocorrélation.

Le motif décrit par la fréquence fondamentale est alors modélisé soit par des lignes polygonales soit par des fonctions exponentielles. Au total, on extrait 7 paramètres du contour de la fréquence fondamentale et 9 paramètres des lignes approximées ou des fonctions exponentielles. Les 7 paramètres extraits du contour de fréquence fondamentale sont :
- 1,2,3 : écart-type, skewness et kurtosis de la distribution des valeurs de fréquence fondamentale sur la phrase,
- 4,5,6 : écart-type, skewness et kurtosis de la distribution des valeurs de l'énergie sur la phrase,
- 7 : coefficient de corrélation entre la fréquence fondamentale et l'énergie.

Pour les deux modélisations, on a :

Lignes polygonales : $y_k(t) = a_k(t_k - t_{k-1}) + b_k$, $k = 1, 2, ..., K$.

On détermine a_k et b_k de façon à minimiser l'erreur des moindres carrés. Le nombre de lignes est choisi de sorte que l'erreur d'approximation soit inférieure à un seuil. Les paramètres extraits sont :

- 1 : rapport de durée entre la pente positive et la pente négative,
- 2,3 : nombre de lignes par unité de durée (pour les pentes positives et négatives),
- 4,5 : pente moyenne (positive et négative),
- 6,7 : écart-type moyen des pentes (positives et négatives),
- 8,9 : fréquence de départ relative des pentes positives et négatives.

Fonctions exponentielles : $y(t) = a(\frac{e}{\tau})e^{(-\frac{t}{\tau})} + bt + c$

Les paramètres extraits sont :

- 1 : durée moyenne de l'intervalle d'approximation,
- 2 : nombre de fonctions par unité de durée,
- 3,4 : moyenne et écart-type de l'amplitude a,
- 5,6 : moyenne et écart-type de la pente b,
- 7,8 : moyenne et écart-type de la constante de temps τ,
- 9 : fréquence de départ relative.

Modèles de Markov Cachés (HMM) pour les coefficients cepstraux

Les coefficients cepstraux (suivant l'échelle de Mel) sont extraits du signal acoustique en utilisant :

- une fréquence d'échantillonnage de 8 kHz,
- une fenêtre d'analyse (Hamming) de 5 ms, à intervalles de 10 ms.

On extrait alors 12 coefficients cepstraux, les 12 dérivées de ces coefficients, et la dérivée de l'énergie. On utilise un modèle de Markov par langue, en faisant l'apprentissage sur 30 locuteurs. Le nombre d'états de chaque modèle est variable, des tests sont effectués pour 4, 8, 16, 32 et 64 états. Cette modélisation est de nature acoustico-phonétique.

Expérimentations

Les expériences sont réalisées sur le corpus Multilingual Telephone Speech Corpus de l'Oregon Graduate Institute (OGI-MLTS) (Annexe G) [93]. Itahashi utilise des données de 45 secondes de parole spontanée par locuteur, pour 50 locuteurs dans 10 langues (anglais, français, espagnol, farsi (perse), chinois, coréen, japonais, tamoul et vietnamien). L'apprentissage se fait sur 30 locuteurs, les 20 locuteurs restants (différents de ceux employés à l'apprentissage) sont utilisés pour les tests.

Résultats

Aux résultats obtenus par chacune des approches viennent s'ajouter ceux obtenus en combinant les deux approches initiales.

Modèles de Markov cachés : Les taux d'identification ont étés calculés en utilisant uniquement les modèles de Markov pour cinq nombres différents d'états (4, 8, 16, 32, 64). Les meilleurs résultats sont de 56% d'identifications correctes. Ils sont obtenus pour 32 ou 64 états.

Fréquence fondamentale : Une analyse discriminante des paramètres dérivés de la fréquence fondamentale donne les résultats suivants :
- modèle lignes polygonales : 25,5% d'identifications correctes,
- modèle fonctions exponentielles : 28,0% d'identifications correctes.

Méthode combinée : Afin de fusionner les résultats obtenus par les deux méthodes, les scores sont normalisés de façon à avoir une moyenne nulle et une variance de 1. Le résultat obtenu par les HMM est pondéré avec un poids w. En combinant les deux méthodes, le meilleur taux de reconnaissance (60%) est obtenu pour des HMM avec 32 états, le modèle de lignes polygonales et un poids w=2,4.

Les résultats montrent que l'apport de la fréquence fondamentale augmente de 5% le score obtenu par les modèles de Markov seuls, dans le cas d'une modélisation par lignes polygonales, ou 2%, pour les fonctions exponentielles. La prise en compte de la fréquence fondamentale peut apporter une information sur la langue à identifier, mais l'apport réalisé par comparaison aux résultats donnés par les modèles de Markov cachés seuls est faible.

3.3.3 Le système de Cummins [24]

Le système [24] utilise une décision par réseaux de neurones, en fusionnant les résultats obtenus par estimation de la fréquence fondamentale et de l'enveloppe d'amplitude.

Estimation de la fréquence fondamentale

La fréquence fondamentale est estimée à intervalles de 1 ms. La dérivée est calculée, puis sous-échantillonnée à 100 Hz (fenêtre glissante rectangulaire). Ensuite on effectue un lissage (avec une fenêtre rectangulaire de 15 points) et un changement d'échelle est opéré pour avoir des valeurs comprises entre -1 et 1. Le paramètre obtenu ainsi est noté ΔF_0.

Estimation de l'enveloppe d'amplitude

Le signal est filtré au moyen d'un filtre de Butterworth passe-bande centré sur 1000 Hz, de largeur de bande 500 Hz. La valeur absolue de ce signal est alors calculée, puis fournie en entrée d'un autre filtre Butterworth passe-bas (fréquence de coupure 10 Hz)

qui effectue un lissage. La dérivée est évaluée, puis sous-échantillonnée à 100 Hz. Enfin, on fait un lissage et un changement d'échelle pour se ramener à des valeurs comprises entre 0 et 1. Le paramètre ainsi obtenu est noté ΔEnv.

Réseau de neurones

Les neurones utilisés sont dits à « Long Short Term Memory » [62]. Dans un réseau LSTM, les unités cachées conventionnelles sont remplacées par des *blocs* mémoire contenant une ou plusieurs cellules (voir figure 3.6).

Une cellule est une unité linéaire avec une connexion récurrente de poids 1. Ce poids permet à la cellule de rester activée en l'absence d'entrée. Le flux d'activation entrant (net_c) passe par une grille d'entrée (*input gating*) et une fonction sigmoïde. L'entrée (net_{in}) pour chaque cellule est multipliée par l'activité de la grille d'entrée, autorisant la grille d'entrée à décider à quelle information la cellule est exposée. La sortie se comporte de la même façon. L'apprentissage est une combinaison de « Back Propagation Through Time » et de « Real Time Recurrent Learning ».

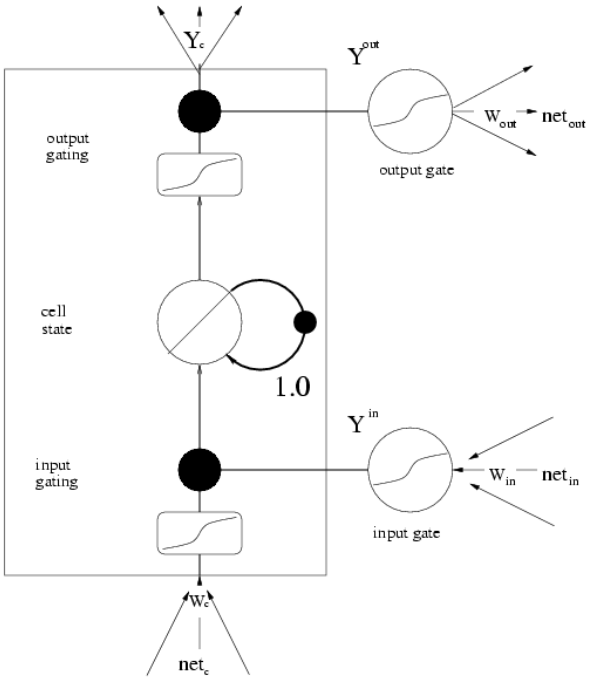


Fig. 3.6 : Bloc LTSM contenant une seule cellule [24]

Expérimentations

Le corpus utilisé est le corpus OGI-MLTS [93]. 5 langues ont été considérées (anglais, japonais, espagnol, mandarin et allemand). L'apprentissage se fait sur 50 locuteurs par langue sur des fichiers de parole spontanée courts. Les tests sont effectués sur 20 locuteurs par langue (différents de ceux employés à l'apprentissage) sur des enregistrements de plus grande durée de parole spontanée.

Résultats

Les expérimentations sont faites en utilisant un seul critère à la fois : ΔF_0 ou ΔEnv. Lors des tests, les tâches d'identification consistent à choisir parmi deux langues. Les tests sont donc effectués sur chaque paire de langues (anglais/allemand, anglais/espagnol, ...).

- En utilisant uniquement ΔEnv on obtient un taux d'identifications correctes variant selon les paires de langues entre 50 et 63 %.
- Avec ΔF_0, on obtient des taux d'identifications correctes légèrement supérieurs, entre 50 et 69 %.

Ces résultats concordent avec la littérature [121] où la fréquence fondamentale est une variable plus discriminante que la modulation de l'amplitude. Cependant, les expériences suggèrent que la modulation de l'enveloppe soit plus exploitée.

3.3.4 Le système de Li [78]

Li [78] a développé un système d'identification automatique des langues basé sur la reconnaissance du locuteur. Une vue générale du système est représentée sur la figure 3.7. Son idée est de classer un signal en mesurant la similarité entre son locuteur et les locuteurs les plus proches dans chaque langue.

Durant l'apprentissage, un réseau de neurones est utilisé pour extraire tous les noyaux syllabiques. Des coefficients spectraux sont extraits à différents endroits des noyaux et sauvegardés.

Durant la phase de reconnaissance, les noyaux syllabiques sont extraits de la même façon et les coefficients spectraux sont comparés à tous ceux mémorisés pour chaque locuteur. La plus petite différence entre chaque noyau du fichier à examiner et les noyaux des autres locuteurs est alors calculée. La somme des différences est considérée comme la différence entre le locuteur et chacune des références. La différence moyenne entre les locuteurs les plus semblables de chaque langue constitue la distance entre la langue du fichier de test et les langues cibles : la langue pour laquelle la différence est la plus petite est sélectionnée.

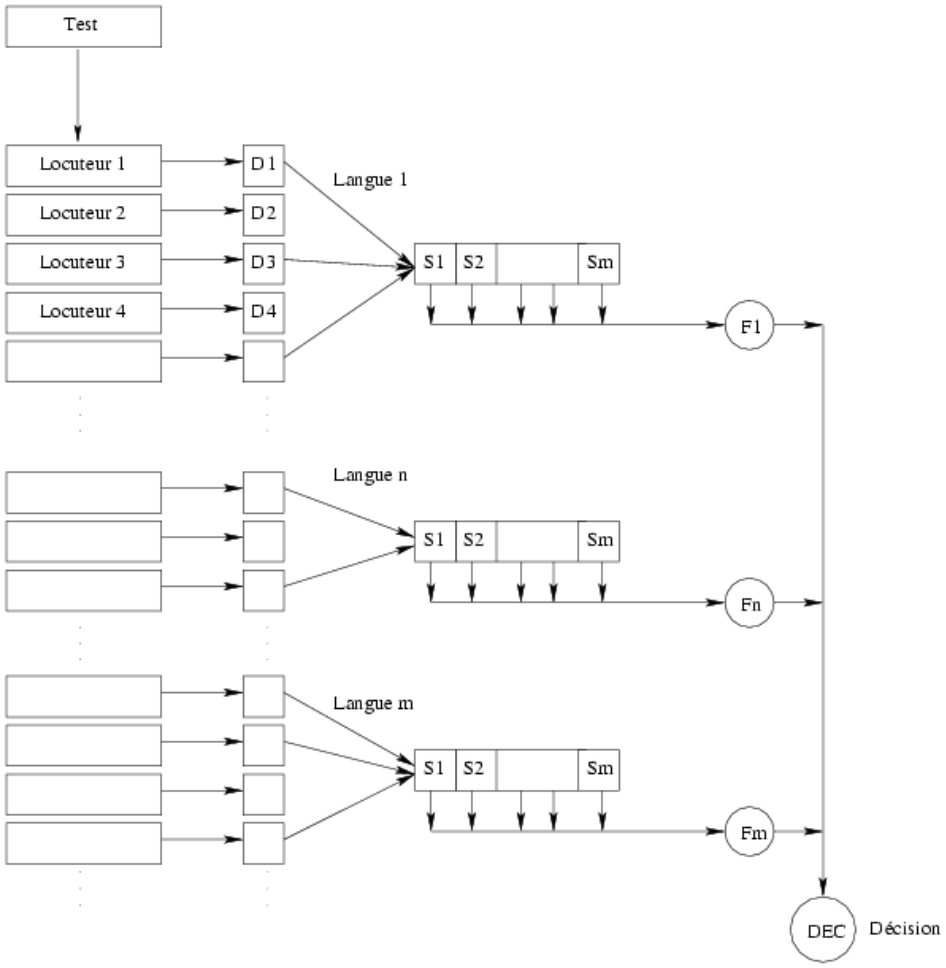


Fig. 3.7 : Schéma descriptif du système de K.P. Li [78]

Le système a été évalué sur le corpus téléphonique OGI-MLTS [93], en utilisant 449 locuteurs répartis sur 10 langues. Les résultats sont d'environ 78% et 58% pour 10 langues en utilisant des séquences de 45 s. et 10 s. respectivement. Si l'on ne considère que les identifications par paire de langues, les taux moyens sont de 91% et 82% pour des fichiers de 45 s. et 10 s. respectivement.

3.3.5 Modèle d'Adami [2]

Ce modèle est dédié à la modélisation de la fréquence fondamentale et de l'énergie dans un but d'identification des langues. Cette méthode est également employée en vérification du locuteur. Elle essaie de rendre compte de l'aspect dynamique de la prosodie au moyen d'une segmentation selon des évènements prosodiques associés à un étiquetage. Les enchaînements des étiquettes sont ensuite modélisés grâce à des modèles n-grammes.

Méthode

Il s'agit tout d'abord d'effectuer une segmentation à partir des courbes de fréquence fondamentale et d'énergie.

- Traitement de la fréquence fondamentale :
  La première étape consiste à prendre les débuts et les fins de chaque partie voisée de la phrase considérée. Sur ces parties voisées, les points d'inflexion (maximum et minimum locaux) sont détectés. Dans l'article d'Adami, ils sont détectés en calculant la dérivée sur 5 trames (50 ms).
- Traitement de la courbe d'énergie :
  Les traitements effectués sur la courbe d'énergie sont de même nature, c'est-à-dire que l'on repère les points d'inflexion par la méthode décrite ci-dessus. La combinaison des points d'inflexion sur les deux courbes donne la segmentation finale (voir la figure 3.8).
- Étiquetage :
  Ensuite, des étiquettes sont déterminées sur chaque segment. Les étiquettes sont choisies comme suit :
  1. montée de F_0 et montée de l'énergie,
  2. montée de F_0 et descente de l'énergie,
  3. descente de F_0 et montée de l'énergie,
  4. descente de F_0 et descente de l'énergie,
  5. segment non voisé.

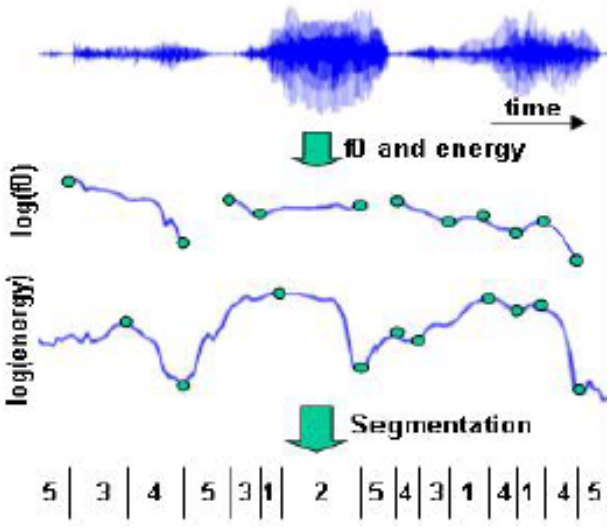


Fig. 3.8 : Illustration de la segmentation et de l'étiquetage (image extraite de [2]).

À ces premières étiquettes en sont ajoutées d'autres qualifiant la durée des segments :
- pour les parties voisées, les segments d'une durée inférieure à 8 trames (80 ms) sont étiquetés « courts », les autres sont dits « longs »,

– pour les parties non voisées, les segments d'une durée inférieure à 14 trames (140 ms) sont étiquetés « courts », les autres sont dits « longs ».

– Modélisation :
Les séquences d'étiquettes les plus fréquentes sont déterminées et représentées par des modèles tri-grammes.

– Résultats :
Ce système a été testé dans [2] sur le corpus CALLFRIEND utilisé lors de la campagne NIST 2004 décrite plus haut (§2.5). La tâche était de vérifier si la langue parlée dans un enregistrement correspondait à la langue hypothèse ou non, comme pour la campagne. Seules trois langues - l'anglais, le mandarin et l'allemand - ont été considérées ici, uniquement pour des durées de 30 secondes.
Le système permet d'obtenir les résultats suivants, comparés et fusionnés avec une approche acoustico-phonétique « classique » (reconnaissance des phonèmes de l'anglais uniquement : 39 symboles) :

Tab. 3.2 : Résultats du système prosodique d'Adami et comparaison avec un système phonotactique (% EER (Equal Error Rate))

Langue	Système prosodique	Système phonotactique	Fusion
Anglais	27,5 %	17,5 %	15,0 %
Mandarin	23,8 %	26,3 %	22,5 %
Allemand	21,3 %	20,0 %	17,5 %

Le système prosodique permet d'obtenir des résultats intéressants comparativement au nombre de symboles employés (10 symboles uniquement pour le système prosodique contre 39 pour le système acoustico-phonétique). Les résultats sont même légèrement supérieurs à ceux obtenus avec l'approche « classique » pour le mandarin. Enfin, la fusion entre les deux approches permet d'obtenir de meilleures performances dans tous les cas.

3.4 Conclusion

Nous avons vu dans ce chapitre quelques systèmes utilisant la prosodie, que ce soit dans un but de caractérisation ou d'identification des langues, ou simplement de description de la prosodie.

Les résultats obtenus par les systèmes applicatifs montrent l'intérêt d'utiliser la prosodie pour l'identification des langues. Le système de Leavers (§3.3.1) montre que des séparations de groupes de langues sont possibles en employant uniquement la prosodie et que l'emploi de techniques acoustiques permettrait de distinguer les langues dans chaque groupe. Le système d'Itahashi (§3.3.2) met l'accent sur l'apport de l'emploi de techniques de modélisation de la prosodie par rapport aux méthodes "classiques" (ici acoustico-phonétique), même si le gain en termes de performance n'est pas très conséquent. Le système de Cummins (§3.3.3) confirme que l'on peut distinguer deux langues uniquement

par la prosodie, pour peu qu'elles n'appartiennent pas au même groupe linguistique. Le système de Li (§3.3.4) prouve que la modélisation de la prosodie permet d'améliorer les performances. Seul le système d'Adami (§3.3.5) essaie de modéliser les enchaînements d'évènements prosodiques sur une phrase et démontre l'efficacité d'une telle démarche.

Le nombre et la diversité des modèles de description de l'intonation prouvent la difficulté de trouver un formalisme multilingue. Chaque méthode de description possède ses avantages propres, qu'elle soit basée sur les mécanismes de production ou de perception de la parole. L'utilisation de tels modèles pour l'identification des langues est donc difficile, puisque très peu de travaux ont porté sur des différences éventuelles entre les langues observables à l'aide de ces méthodes de description.

Les systèmes de caractérisation du rythme (§3.1) emploient des méthodes similaires. Ils essaient de décrire le rythme au travers de mesures de durées et de dispersion des durées, que ce soit des intervalles vocaliques, intervocaliques ou consonantiques, voire qui tendent à s'en rapprocher (sonorité). Nous avons également vu que la faiblesse de ces méthodes est qu'elles n'ont pas été testés sur des corpus suffisamment conséquents, ce qui est principalement dû à la nécessité d'employer un étiquetage manuel préalable au traitement. Dans la section suivante, nous allons voir comment tirer parti des outils disponibles à l'IRIT afin de valider ces méthodes sur un corpus plus conséquent.

Chapitre 4

Extraction automatique et caractérisation d'unités prosodiques

Sommaire

4.1 Introduction 83
4.2 Extraction automatique d'informations prosodiques 85
4.2.1 Segmentation de la parole 85
4.2.2 Détection de l'activité vocale 86
4.2.3 Localisation des voyelles 87
4.2.4 Conclusion 88
4.3 Cadre expérimental 89
4.3.1 Corpus 89
4.3.2 Protocole expérimental 91
4.3.3 Modélisation : cadre statistique 92
4.4 Adaptation de quelques approches présentées au chapitre précedent 93
4.4.1 Évaluation automatique des paramètres proposés par Ramus (§3.1.1) 93
4.4.2 Évaluation automatique des paramètres proposés par Grabe (§3.1.2) 96
4.4.3 Discussion 98
4.5 Caractérisation d'une unité rythmique : la pseudo-syllabe . . 99
4.5.1 Localisation de la pseudo-syllabe 99
4.5.2 Modélisation des pseudo-syllabes 100
4.5.3 Modélisation des caractéristiques temporelles des pseudo-syllabes 100
4.5.4 Modélisation des caratéristiques intonatives des pseudo-syllabes 104
4.5.5 Fusion des modèles de durée et d'intonation pseudo-syllabiques 108
4.6 Conclusion 109

4.1 Introduction

AVANT de pouvoir modéliser automatiquement la prosodie des langues, nous devons décrire les moyens de traitement automatique permettant l'extraction d'informations de nature prosodique. Ensuite, nous proposons une évaluation de quelques paramétrisations décrites dans le chapitre 3 (celles de Ramus et de Grabe) sur le corpus MULTEXT [20]. Nous nous appuyons sur ces expériences pour définir une unité prosodique, proche de la syllabe. Nous testons cette unité au moyen de modélisations de durée et d'intonation afin de montrer son efficacité.

4.2 Extraction automatique d'informations prosodiques

Les méthodes que nous décrivons comme préalable à toute étude ont pour but d'atteindre des informations de bas niveau. Trois traitements de base conduisent à la localisation de frontières pertinentes quand à la notion consonne/voyelle :

- la segmentation en zones quasi-stationnaires,
- la détection d'activité vocale,
- la localisation des voyelles.

4.2.1 Segmentation de la parole

La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) [5] qui est fondé sur une étude statistique du signal dans le domaine temporel. Le signal de parole est hypothétiquement décrit par une suite de zones quasi-stationnaires. Chacune de ces zones est caractérisée par un modèle statistique, le modèle autorégressif gaussien :

$$\begin{cases} y_n = \sum a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases} \tag{4.1}$$

où (y_n) est le signal de parole et (e_n) est un bruit blanc gaussien.

La méthode consiste à détecter les changements de modèles autorégressifs au travers des erreurs de prédiction calculées sur deux fenêtres d'analyse (figure 4.1). La distance entre ces deux modèles est obtenue à partir de l'entropie mutuelle des deux lois conditionnelles correspondantes.

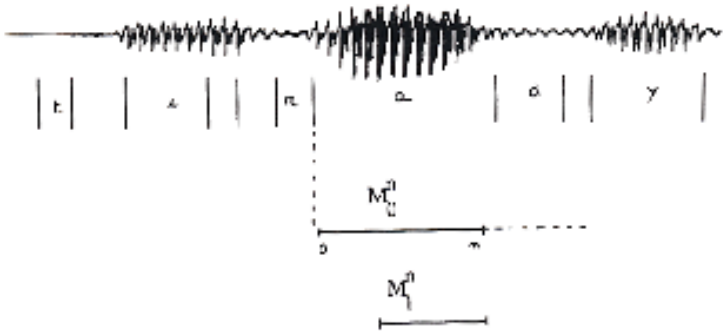


Fig. 4.1 : Localisation des fenêtres d'estimation des modèles M_0^n et M_1^n à l'instant n ; l'instant « 0 » correspond à la dernière frontière validée. La phrase prononcée est : « il se garan**tira du**... ».

La statistique est définie comme une somme cumulée : $W_n = \sum_{k=1}^{n} w_k$ avec w_k l'entropie mutuelle entre les deux modèles dans le cas gaussien :

$$w_k = \frac{1}{2}\left\{2\frac{e_k^0 e_k^1}{\sigma_1^2} - [1 + \frac{\sigma_0^2}{\sigma_1^2}]\frac{e_k^0}{\sigma_0^2} + [1 - \frac{\sigma_0^2}{\sigma_1^2}]\right\} \tag{4.2}$$

et l'erreur de prédiction pour chaque modèle à l'instant k :

$$e_k^i = y_k - \sum_{j=1}^{p} a_j^i y_{k-j}, i = 0, 1 \tag{4.3}$$

Cette méthode a été comparée à de nombreuses autres méthodes de segmentation [6]. Elle a déjà fourni des résultats intéressants pour la reconnaissance automatique de la parole : des expériences ont montré que la durée des segments est porteuse d'une information pertinente [7].

Elle permet d'atteindre, notamment pour la parole, une segmentation infra-phonétique où 3 types de segments se distinguent :

- les segments quasi-stationnaires qui correspondent à la partie stable des phonèmes lorsqu'elle existe,
- les segments transitoires,
- les segments courts (environ 20 ms).

Leur longueur varie entre 20 et 100 ms pour la parole (figure 4.2).

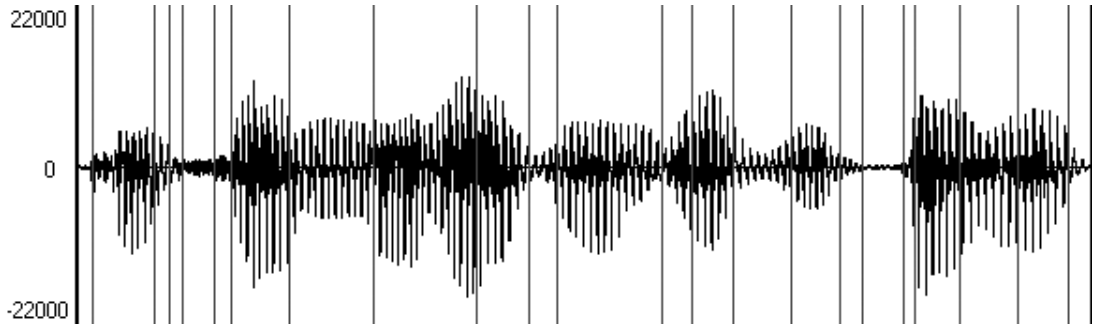


Fig. 4.2 : Résultat de la segmentation sur environ 1 seconde de parole. La phrase prononcée est : « Confirmez le rendez-vous par écrit ».

4.2.2 Détection de l'activité vocale

Nous avons implanté un détecteur d'activité vocale basé sur une analyse statistique du premier ordre du signal temporel [99]. On note N le nombre de segments issus de la segmentation automatique et $\{S_1, ..., S_N\}$ la suite de ces segments. Le seuil d'activité σ_a est défini par :

$$\sigma_a = \alpha \min_i \left(var\left(S_i\right)\right) \tag{4.4}$$

où le coefficient α a été expérimentalement fixé à 2,5. Les segments ayant une variance inférieure à S_a sont étiquetés comme étant des silences. On distingue les silences signalant une absence d'activité de parole (segments longs), des silences se produisant en cours de locution (les silences des occlusives, les pauses courtes). Un post-traitement permet de regrouper les segments de non activité en cas de sur-segmentation. Si plusieurs segments courts sont étiquetés comme étant des silences et ont une durée totale supérieure à 150 ms, on considère qu'il s'agit d'une zone de non activité.

4.2.3 Localisation des voyelles

L'algorithme de détection automatique des voyelles est basé sur une analyse spectrale [99]. Un critère appelé Rec (*Reduced Energy Cumulating*) est calculé pour chaque trame du signal.

Ce critère est calculé à partir de 24 coefficients d'énergie calculés selon l'échelle Mel :

$$Rec(t) = \frac{E_{BF}(t)}{E(t)} \sum_{i=1}^{24} \alpha_i \left(E_i(t) - \overline{E}(t)\right)^+ \tag{4.5}$$

où t est le numéro de la trame du signal, $E_i(t)$ est l'énergie de la trame t dans le i^{eme} filtre selon l'échelle Mel, $E(t)$ est l'énergie totale de la trame t, $E_{BF}(t)$ est l'énergie de la trame t dans la bande de fréquences inférieures à 1 kHz, $\overline{E}(t)$ est la moyenne de l'énergie pour la trame t sur les 24 bandes spectrales et α_i est le poids affecté au $i^{ième}$ filtre.

Le critère Rec met en évidence la présence ou l'absence de structure formantique sur la trame considérée. Les voyelles sont caractérisées par des valeurs élevées et nous considérons que les pics de $Rec(t)$ localisent des trames vocaliques. Un segment issu de la segmentation automatique (§4.2.1) sera étiqueté vocalique s'il contient une trame vocalique.

Le fait que ni données étiquetées ni apprentissage supervisé ne soient nécessaires constitue le principal avantage de cet algorithme. Il a été évalué pour cinq langues - le français, le japonais, le coréen, l'espagnol et le vietnamien - sur la partie phonétiquement étiquetée du corpus OGI MLTS (parole téléphonique, 8 KHz).

Le tableau 4.1 montre les performances de l'algorithme, comparé à d'autres systèmes de détection de voyelles issus de la littérature. Le taux d'erreur est défini par :

$$\text{taux d'erreur} = 100 * \frac{\text{fausses détections} + \text{fausses alarmes}}{\text{nombre de voyelles}}$$

Tab. 4.1 : Comparaison de différents algorithmes de détection automatique de voyelles.

Référence	Corpus	Langue	Taux d'erreur
Pfitzinger & al., 1996[4] [104]	PhonDatII (parole lue)	Allemand	12,9%
	Verbmobil (parole spontanée)	Allemand	21,0%
Fakotakis & al., 1997 [36]	TIMIT (parole lue)	Anglais	32,0%
Pfau & Ruske, 1998 [103]	Verbmobil (parole spontanée)	Allemand	22,7%
Howitt, 2000 [63]	TIMIT (parole lue)	Anglais	29,5%
Pellegrino & André-Obrecht, 1998 [98]	OGI MLTS (parole spontanée)	Coréen	28,5%
		Espagnol	19,2%
		Français	19,5%
		Japonais	16,3%
		Vietnamien	21,1%
		Moyenne	22,9%

Les performances de l'algorithme de détection de voyelles sont correctes, comparées à celles des autres systèmes. Le fait qu'aucun apprentissage ne soit nécessaire permet d'utiliser l'algorithme sur différentes langues, mais cela signifie également qu'il n'est optimal pour aucune des langues.

4.2.4 Conclusion

À la suite de ces prétraitements automatiques et indépendants des langues, nous obtenons une segmentation du signal en segments vocaliques, consonantiques et de silences. Des étiquettes « V », « C » ou « # » sont utilisées pour qualifier chaque segment (figure 4.3).

[4]dans cette étude, le taux d'erreur est calculé en tenant compte des noyaux syllabiques et non pas directement des voyelles

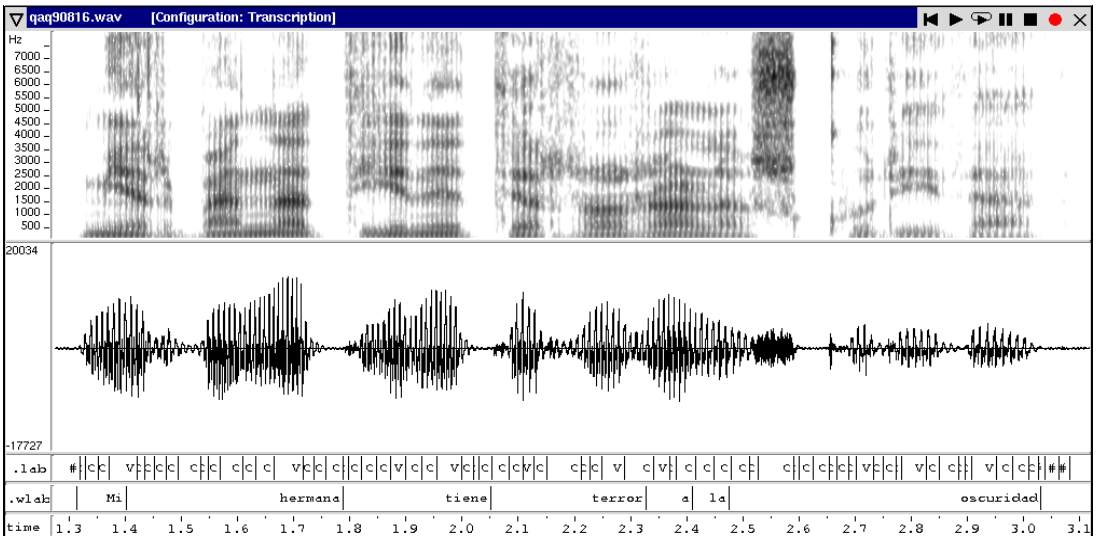


Fig. 4.3 : Résultat de la segmentation en segments consonantiques, vocalique et de silence sur un enregistrement d'espagnol du corpus MULTEXT

À cause des propriétés intrinsèques de l'algorithme (et en particulier le fait que les parties voisées et non voisées d un même phonème peuvent être séparées), il est quelque part incorrect de considérer cette segmentation comme étant une exacte dichotomie entre les consonnes et les voyelles.

Toutefois, elle est indéniablement corrélée à la structure rythmique de la parole. Nous étudions l'hypothèse que cette corrélation peut permettre à un modèle statistique de discriminer les langues suivant leur structure rythmique.

4.3 Cadre expérimental

La détection automatique des voyelles permet l'évaluation des différents algorithmes présentés dans le chapitre 3, dans une version automatisée et unifiée, sur un corpus plus conséquent : « MULTEXT » [20].

Toutes les expériences menées par la suite seront effectuées sur ce même corpus et avec le même protocole expérimental quels que soient les paramètres étudiés.

4.3.1 Corpus

Les enregistrements sont extraits du corpus de parole EUROM1, réalisé à l'occasion du projet ESPRIT 2589 « Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation » [22]. Les enregistrements audio sont de haute qualité (échantillonnage à 20 KHz, 16 bits) et effectués en chambre anéchoïque. Ils ont été contrôlés durant l'ac-

quisition de manière à rejeter toute donnée bruitée ou toute erreur de lecture. MULTEXT reprend cinq des huit langues de EUROM1 (allemand, anglais, espagnol, français et italien). Les données correspondent au jeu de locuteurs « FEW TALKER SET » comprenant dix locuteurs par langue - cinq femmes et cinq hommes - et à des passages lus de cinq phrases connectées par une structure sémantique cohérente. Il est demandé à chaque locuteur de lire un extrait du passage et d'essayer d'avoir l'intonation la plus naturelle possible. La durée de chaque passage est d'environ 20 s et la durée des enregistrements par langue est de 45 minutes en moyenne.

Un ensemble de phrases en japonais a été rajouté à ce corpus par Kitazawa [67]. Ces phrases sont enregistrées dans des conditions similaires à celles du corpus original. Un autre ensemble de phrase en mandarin a également été enregistré dans les mêmes conditions [70]. De même que pour le japonais, ces ensembles sont ajoutés au corpus initial.

Au final, le corpus se compose de sept langues : anglais, français, allemand, italien, japonais, mandarin et espagnol.

Nous avons choisi d'utiliser le maximum de locuteurs pour l'apprentissage, c'est-à-dire quatre hommes et quatre femmes pour toutes les langues sauf le japonais pour lequel on dispose de deux hommes et deux femmes. Les tests seront effectués avec les deux locuteurs restants pour chaque langue, un homme et une femme.

Notons qu'une même phrase peut être prononcée par deux ou trois locuteurs et que cela entraîne une dépendance possible au texte dans les modélisations. Pour pallier à ce défaut, nous avons divisé le corpus en ensembles disjoints de test et d'apprentissage, tant au niveau des locuteurs qu'au niveau des textes prononcés par ces mêmes locuteurs.

Trois jeux de données sont ainsi déterminés, en changeant les locuteurs de test et d'apprentissage. L'ensemble d'apprentissage est aussi l'ensemble de développement à cause du manque de données. Le premier jeu est décrit ci-dessous (tableaux 4.2 et 4.3). Les deux autres jeux de données sont décrits dans l'annexe B.

Les expériences rapportées dans le corps de ce manuscrit correspondent à celles effectuées sur le jeu n°1. Le cas échéant, les expériences sur les autres jeux de données seront reportées en annexe.

Tab. 4.2 : Description de l'ensemble d'apprentissage du jeu1 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Durée
Anglais	8	80	24 mn
Français	8	80	29 mn
Allemand	8	80	29 mn
Italien	8	80	30 mn
Japonais	4	80	39 mn
Mandarin	8	80	26 mn
Espagnol	8	80	27 mn
Total	52	560	204 mn

Tab. 4.3 : Description de l'ensemble de test du jeu1 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Durée
Anglais	2	20	6 mn
Français	2	19	7 mn
Allemand	2	20	7 mn
Italien	2	20	7 mn
Japonais	2	20	11 mn
Mandarin	2	20	6 mn
Espagnol	2	20	8 mn
Total	14	139	52 mn

4.3.2 Protocole expérimental

Le protocole expérimental se décompose en trois étapes :

1. Prétraitement : le signal de parole est étiqueté automatiquement en segments vocaliques, consonantiques et de silence.
2. Les distributions des différents paramètres sont représentées graphiquement afin de déterminer leur pouvoir discriminant. Étant donné le nombre de points à représenter, nous avons décidé par souci de lisibilité de nous limiter à visualiser la moyenne des paramètres pour chaque langue autour de laquelle nous avons dessiné une barre d'erreur ayant pour longueur l'écart-type.
3. Des expériences en identification des langues sont menées, avec l'emploi de modèles de mélange de lois gaussiennes.
   (a) Les modèles, des Modèles de Mélanges de lois Gaussiennes (MMG), sont estimés à partir des données de l'ensemble d'apprentissage. Ces modèles sont appris pour différents nombres de composantes gaussiennes (2, 4, 8, 16, 32, 64) dans le MMG.

(b) Pour chaque dimension des MMG, des expériences sont effectuées en utilisant l'ensemble d'apprentissage comme ensemble de développement. Cela permet de déterminer le nombre de lois gaussiennes optimal, mais ne réduit pas les risques de sur-apprentissage.

(c) Une fois le nombre de lois gaussiennes déterminé, les expériences d'identification sont effectuées sur l'ensemble de test et les matrices de confusion correspondantes sont données. Des regroupements à l'intérieur des matrices de confusion permettent de visualiser les différents groupes rythmiques et d'interpréter les résultats.

4.3.3 Modélisation : cadre statistique

Chaque observation est définie par un vecteur de paramètres de dimension d : $\psi = (x_1, x_2, x_3, ..x_d)$. L'ensemble des observations composant une phrase est traité. On note $\Psi = \{\psi_1, \psi_2, ..., \psi_{n_p}\}$ la suite des n_p vecteurs d'observations de la phrase.

Pour chaque langue, les paramètres d'un Modèle de Mélange de lois Gaussiennes (MMG) sont appris à partir des observations, en utilisant l'algorithme LBG [79] suivi de l'algorithme EM (annexe F).

La probabilité d'observer ψ_k sachant que la langue L_i est utilisée s'exprime sous la forme suivante :

$$p(\psi_k|L_i) = \sum_{j=1}^{Q_i} \frac{\alpha_j^i}{(2\pi)^{d/2}\sqrt{\left|\Sigma_j^i\right|}} \exp\left(-\frac{1}{2}\left(\psi_k - \mu_j^i\right)^t \left(\Sigma_j^i\right)^{-1}\left(\psi_k - \mu_j^i\right)\right) \tag{4.6}$$

où Q_i est le nombre de composantes du mélange de lois gaussiennes, d est la dimension du vecteur ψ_k, et (μ_j, Σ_j) représente les paramètres de la loi Gaussienne j.

En faisant l'hypothèse que les observations sont indépendantes, nous obtenons :

$$p(\Psi|L_i) = \prod_{k=1}^{n_p} p(\psi_k|L_i) \tag{4.7}$$

Dans le cadre de l'approche bayésienne classique, la langue la plus probable L^* est définie par l'équation suivante :

$$L^* = arg \max_{1 \leq i \leq N_L} p(L_i|\Psi) \tag{4.8}$$

4.4 Adaptation de quelques approches présentées au chapitre précedent

Nous allons expérimenter les paramètres décrits par Ramus et par Grabe (voir §3.1) afin de vérifier s'il est possible d'automatiser ces méthodes et ainsi de les appliquer à des corpus de taille plus conséquente. Par voie de conséquence, si la réponse est positive, l'étude de grandes bases de données doit permettre de confirmer la théorie. Les paramètres sont calculés à partir de la segmentation automatique en consonnes/voyelles. Les expériences sont faites sur le corpus MULTEXT.

4.4.1 Évaluation automatique des paramètres proposés par Ramus (§3.1.1)

Pour calculer les paramètres proposés par Ramus, nous appliquons tout d'abord la segmentation automatique en segments consonantiques, vocaliques et de silence décrite plus haut (§4.2 et figure 4.3).

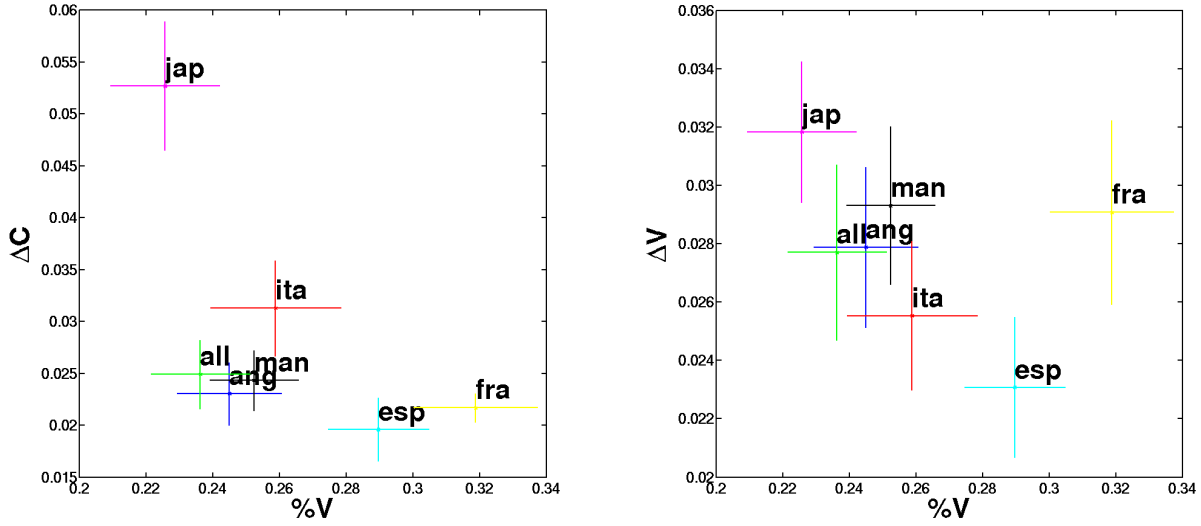
Les paramètres suivants sont calculés sur chaque phrase :
- %V la proportion (en durée) des segments vocaliques,
- ΔV l'écart-type des durées des intervalles entre les segments vocaliques,
- ΔC l'écart-type des durées des intervalles entre les segments consonantiques.

Représentations graphiques

Les distributions de ces paramètres sont étudiées au travers de leurs projections dans les plans (%V, ΔC) et (%V, ΔV) (figure 4.4). Dans le plan (%V, ΔC), on voit apparaître des différences entre certains groupes de langues : le japonais se dissocie de toutes les autres langues, avec un ΔC très élevé. Les autres langues considérées ici ont un ΔC assez proche. En revanche, des regroupements apparaissent entre l'allemand, l'anglais et le mandarin au niveau du paramètre %V. En conclusion, on peut distinguer trois groupes : un correspondant uniquement au japonais (langue moraïque), un autre regroupant l'anglais, l'allemand, le mandarin[4] (langues accentuelles) et l'italien (langue syllabique) et un groupe français-espagnol (langues syllabiques).

Le plan (%V, ΔV) ne fait pas apparaître aussi clairement de tels regroupements. On notera toutefois que le paramètre ΔV permet de séparer le français de l'espagnol.

[4]Nous avons considéré le mandarin comme étant une langue accentuelle, de même que Komatsu [70], mais cette catégorisation n'est pas admise unanimement



(a) (b)

Fig. 4.4 : Paramètres de Ramus
(a) Paramètres (%V, ΔC)
(b) Paramètres (%V, ΔV)

Ces résultats sont concordants avec ceux trouvés par Ramus dans ses expériences. Les groupes de langues sont retrouvés dans le plan (%V, ΔC). Le plan (%V,ΔV) ne donne pas d'informations liées aux groupes rythmiques, mais il permet de différencier certaines langues à l'intérieur de ces groupes rythmiques.

Expériences en identification des langues

Les vecteurs d'observations sont constitués des trois paramètres (%V, ΔV, ΔC). Les expériences ont montré que le meilleur taux d'identification sur les données de l'ensemble d'apprentissage est de 50,2 %, soit 281 identifications correctes sur 560 fichiers. Ce résultat est obtenu avec 4 gaussiennes par MMG. Ce nombre est le même quelque soient les langues La faible dimension des mélanges de lois gaussiennes peut s'expliquer par le faible nombre de données d'apprentissage. En effet, les paramètres sont calculés sur chaque phrase, il n'y a que 80 vecteurs d'observations par langue pour l'apprentissage. La matrice de confusion correspondante est représentée sur le tableau 4.4.

Tab. 4.4 : Paramètres de Ramus (%V, ΔV, ΔC)
Expériences sur l'ensemble d'apprentissage : correct : 50,2 ± 4.1 % (281/560)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	17	24	24	-	3	12	-
Allemand	6	39	30	-	3	-	2
Mandarin	1	25	41	-	3	10	-
Français	-	3	11	48	-	18	-
Italien	3	22	15	2	17	13	8
Espagnol	-	6	4	16	3	51	-
Japonais	-	7	1	-	4	-	68

En reconnaissance, l'expérience est menée avec les modèles donnant le meilleur résultat sur l'ensemble d'apprentissage, c'est-à-dire des MMG avec 4 gaussiennes. Le taux d'identification correcte est de 43,9 % (61 identifications correctes sur 139 fichiers). La matrice de confusion est représentée sur le tableau 4.5.

Tab. 4.5 : Paramètres de Ramus (%V, ΔV, ΔC)
Expériences sur l'ensemble de test : correct : 43,9 ± 8,3 % (61/139))

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	1	9	8	-	1	1	-
Allemand	8	5	7	-	-	-	-
Mandarin	-	8	4	1	-	6	1
Français	-	-	1	13	-	5	-
Italien	-	3	6	-	7	-	4
Espagnol	-	-	1	5	1	13	-
Japonais	-	2	-	-	-	-	18

Les résultats en identification des langues sont corrects, tout en reflétant les regroupements que nous avons pu mettre en évidence sur les graphiques :

- les langues accentuelles sont bien séparées des autres groupes de langues, mais elles sont confondues entre elles (l'anglais est totalement confondu avec l'allemand et le mandarin). L'italien est également en partie confondu avec ces langues,
- les langues syllabiques sont également bien identifiées (sauf l'italien). On notera aussi des confusions entre le français et l'espagnol,
- le japonais, seule langue moraïque du corpus, est la langue la mieux reconnue.

Ces résultats sont illustrés dans le tableau 4.6, où les regroupements sont effectués en fonction des groupes linguistiques. En faisant ces regroupements, le taux d'identifications correctes est de 80,5 %.

Tab. 4.6 : Paramètres de Ramus (%V, ΔV, ΔC)
Expériences sur l'ensemble de test, Regroupement en classes rythmiques : correct : 80,5 ± 6,0 % (112/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	50	9	1
L. Syllab.	11	44	4
L. Mora.	2	-	18

4.4.2 Évaluation automatique des paramètres proposés par Grabe (§3.1.2)

Pour calculer les paramètres de Grabe, la segmentation automatique en segments consonantiques, vocaliques et de silence (§4.2) est appliquée à chaque phrase. Les intervalles nécessaires au calcul des indices PVI sont directement mesurables à partir des segments étiquetés automatiquement. Les paramètres *Pairwise Variability Indices* sont calculés sur chaque phrase. Ces paramètres sont étudiés pour les écarts inter- et intra-vocaliques, dans leurs versions « brutes » (équation 3.1) et normalisées (équation 3.2).

Représentations graphiques

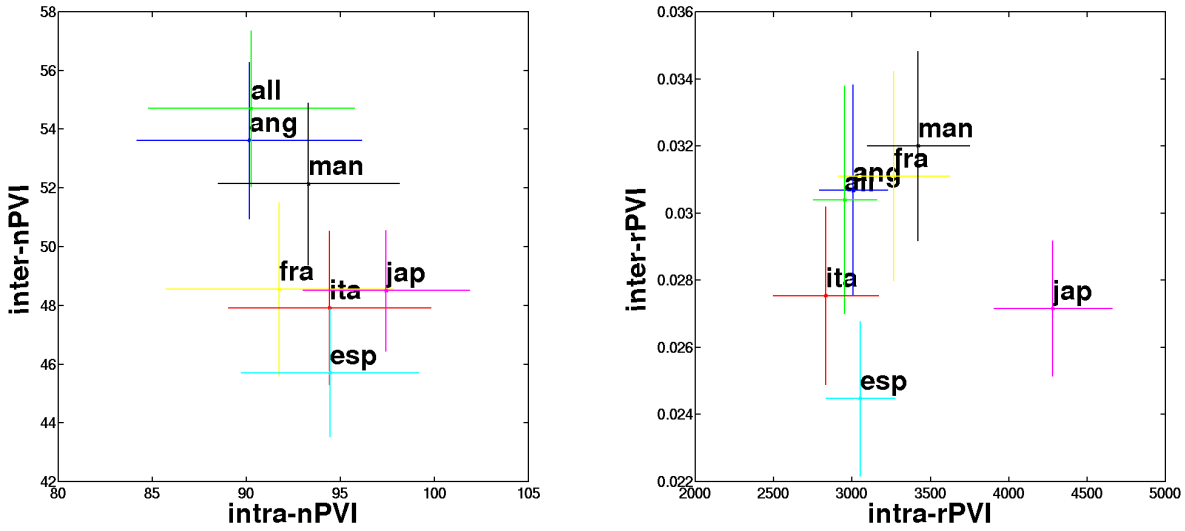
Les distributions de ces paramètres sont représentées sur la figure 4.5.

Dans le plan (Intra-nPVI, Inter-nPVI), on remarque une séparation de l'anglais et de l'allemand par rapport aux autres langues, qui sont par contre bien regroupées, avec le mandarin en position intermédiaire.

Les paramètres (Intra-rPVI, Inter-rPVI) permettent de distinguer le japonais des autres langues, plus particulièrement le paramètre intra-rPVI.

Ces expériences montrent des résultats moins clairs que ceux obtenus avec le système de Ramus. Les groupes de langues ne sont pas aussi bien séparés les uns des autres que dans les expériences réalisées dans [50].

Cela peut être dû à un manque de robustesse vis-à-vis de la variabilité des paramètres employés : les expériences proposées dans [50] n'emploient qu'un seul locuteur par langue. De plus, l'étiquetage automatique fournit une segmentation différente d'une segmentation phonétique. La variabilité en durée des segments est plus importante que celle des phonèmes. Il y a donc peu de segments pertinents, ce qui peut expliquer ces résultats.



(a) (b)

Fig. 4.5 : Paramètres de Grabe
(a) Paramètres *PVI* normalisés
(b) Paramètres *PVI* non normalisés

Expériences en Identification des Langues

Chaque vecteur d'observation est composé des quatres paramètres (Intra-nPVI, Inter-nPVI, intra-rPVI, inter-rPVI). Ici aussi, les paramètres sont calculés sur chaque phrase, il n'y a qu'un vecteur d'observation par phrase. Les données d'apprentissage étant peu nombreuses, le nombre de lois gaussiennes dans les modèles est faible. Les expériences ont montré que le meilleur taux d'identification sur les données de l'ensemble d'apprentissage est de 67,0 %, soit 375 identifications correctes sur 560 fichiers. Ce résultat est obtenu avec 8 gaussiennes par MMG. La matrice de confusion correspondante est représentée dans le tableau 4.7.

Tab. 4.7 : Paramètres de Grabe (Intra-nPVI, Inter-nPVI, intra-rPVI, inter-rPVI)
Expériences sur l'ensemble d'apprentissage : correct : 67.0 ± 3.9 % (375/560)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	54	12	2	4	3	4	1
Allemand	9	57	2	3	4	5	-
Mandarin	10	6	43	6	9	2	3
Français	12	3	10	45	2	2	6
Italien	5	3	3	2	57	10	-
Espagnol	7	2	1	1	3	65	1
Japonais	4	3	6	6	3	4	54

En reconnaissance, le taux d'identifications correctes est de 36,7 % (51 identifications correctes sur 139 fichiers). La matrice de confusion est représentée sur le tableau 4.8.

Tab. 4.8 : Paramètres de Grabe (Intra-nPVI, Inter-nPVI, intra-rPVI, inter-rPVI) Expériences sur l'ensemble test : correct : 36,7 ± 8,0 % (51/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	5	3	2	4	2	-	4
Allemand	6	5	-	1	3	-	1
Mandarin	3	1	2	5	1	3	5
Français	1	1	3	10	1	-	3
Italien	4	3	-	-	6	4	3
Espagnol	1	1	1	3	5	5	4
Japonais	-	-	1	-	1	-	18

L'étude des distributions laisse entrevoir que les paramètres *PVI* ne sont pas suffisamment robustes pour discriminer les langues sur des corpus conséquents. Cela est confirmé par les expériences en identification des langues qui ne permettent pas de distinguer quelque regroupement que ce soit. On notera toutefois la bonne performance en identification du système pour le japonais.

Tab. 4.9 : Paramètres de Grabe (Intra-nPVI, Inter-nPVI, intra-rPVI, inter-rPVI) Expériences sur l'ensemble de test, Regroupement en classes rythmiques : correct 56,8 +- 8,2 % (79/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	27	19	10
L. Syllab.	15	34	10
L. Mora.	1	1	18

4.4.3 Discussion

L'adaptation automatique des paramètres de Grabe [50] n'est pas très efficace. Ce système n'est performant que pour isoler le japonais des autres langues.

Les paramètres proposés par Ramus [111] sont plus concluants. Les graphiques permettent de faire apparaître des différences entre les langues et les expériences d'identification permettent de retrouver les groupes rythmiques linguistiques.

Ces résultats montrent qu'il n'est pas aisé de trouver une modélisation automatique adéquate pour rendre compte des différences rythmiques pouvant exister entre les langues. En s'appuyant sur ces expériences, nous allons déterminer une unité prosodique et extraire des caractéristiques rythmiques et intonatives sur cette unité afin de vérifier sa pertinence.

4.5 Caractérisation d'une unité rythmique : la pseudo-syllabe

Le système que nous proposons est basé sur une unité de type syllabique [37].

La syllabe est une unité privilégiée pour la modélisation du rythme. Cependant, l'extraction automatique des syllabes (et plus particulièrement la détection des frontières entre syllabes) est une opération qui se révèle peu robuste : la qualité de la prononciation et la vitesse d'élocution sont des facteurs qui influent directement sur le découpage en syllabes [71]. Segmenter le signal de parole en syllabes est une tâche spécifique à chaque langue [104], aucun algorithme indépendant de la langue ne peut donc être appliqué aisément.

Pour cette raison, nous avons introduit la notion de pseudo-syllabe [38]. L'idée de base consiste à articuler l'unité prosodique autour des éléments centraux des syllabes - les voyelles - puis de rassembler autour de ces noyaux les consonnes voisines. Nous avons choisi de ne rattacher que les consonnes précédant chaque voyelle. Ce choix s'explique, d'une part, par le fait que la détection des limites des syllabes n'est pas une tâche aisée dans un cadre multilingue sans connaissances a priori et, d'autre part, par le fait que les syllabes les plus fréquentes correspondent à la structure consonne/voyelle [27].

Le processus de segmentation et de paramétrisation du signal de parole en pseudo-syllabes est décrit ci-dessous.

4.5.1 Localisation de la pseudo-syllabe

A l'issue du processus de localisation automatique des noyaux vocaliques (§4.2), les segments identifiés comme vocaliques sont étiquetés V. Tous les autres segments de parole sont conservés et étiquetés C. Le signal de parole est désormais une suite de segments étiquetés consonne (C) ou voyelle (V), qui se traduit en une suite de segments de type $[C^n.V]^+$ dont les frontières sont déterminées automatiquement. La figure 4.6 illustre le résultat de cette localisation.

Nous appelerons « pseudo-syllabe » une structure de type C^nV (où n est un entier qui peut être nul). Par exemple, si la séquence (CCVCCCCCVCCCCCCCCVCCCCV-VCCC) est obtenue après le découpage et la détection de segments vocaliques, elle sera partitionnée en 6 pseudo-syllabes : (CCV.CCCCCV.CCCCCCCCV.CCCCV.V.CCC).

Une fois cette étape réalisée, il reste à trouver une paramétrisation adaptée à la pseudo-syllabe pour en faire une unité acoustique, prosodique et rythmique, et trouver les modélisations correspondantes.

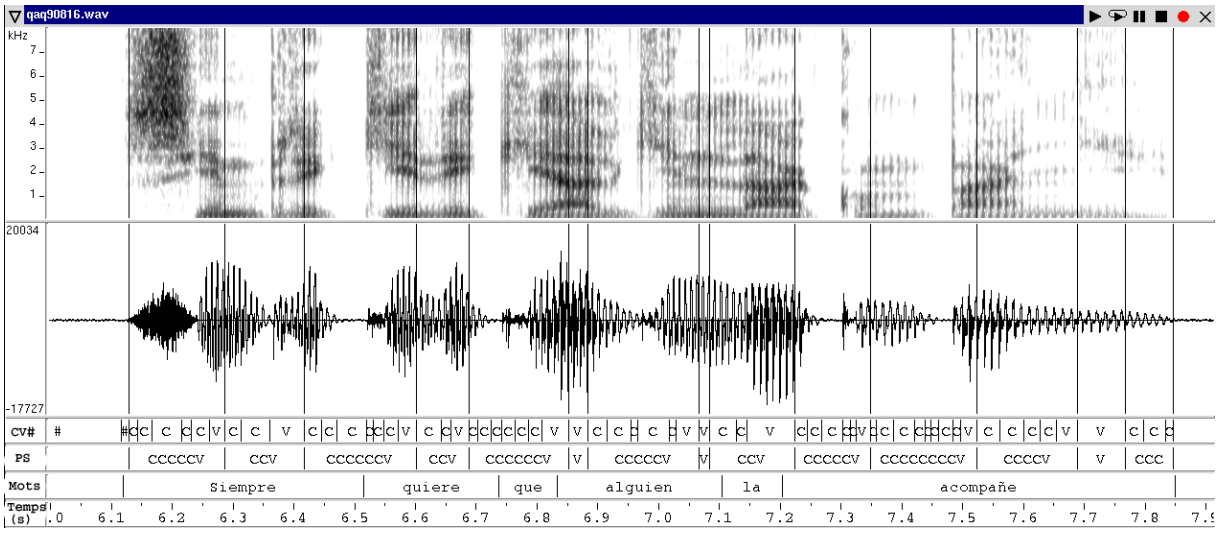


Fig. 4.6 : Exemple d'extraction de pseudo-syllabes sur la phrase « *Siempre quiere que alguien la acompañe* ». La première ligne de transcription correspond à l'étiquetage automatique en Consonnes, Voyelles et Silences. La deuxième ligne est le résultat du regroupement en pseudo-syllabes. La troisième ligne est l'étiquetage en mots fourni avec le corpus.

4.5.2 Modélisation des pseudo-syllabes

L'étape suivante consiste à caractériser les langues au travers des propriétés de leurs pseudo-syllabes. Des informations sont extraites de chaque pseudo-syllabe afin de former un vecteur de paramètres, qui sera considéré comme une observation en vue d'effectuer la reconnaissance de la langue.

4.5.3 Modélisation des caractéristiques temporelles des pseudo-syllabes

Caractérisation de la durée d'une pseudo-syllabe

Pour chaque pseudo-syllabe $C^n.V$, trois paramètres sont calculés, correspondant respectivement à la durée totale des n segments consonantiques D_c, à la durée totale du segment vocalique D_v et à la complexité de la pseudo-syllabe qui s'exprime en termes de nombre de segments consonantiques $N_c = n$ (voir figure 4.7). Les durées sont exprimées en millisecondes.

$$\text{On note : } \psi = (D_c, D_v, N_c) \tag{4.9}$$

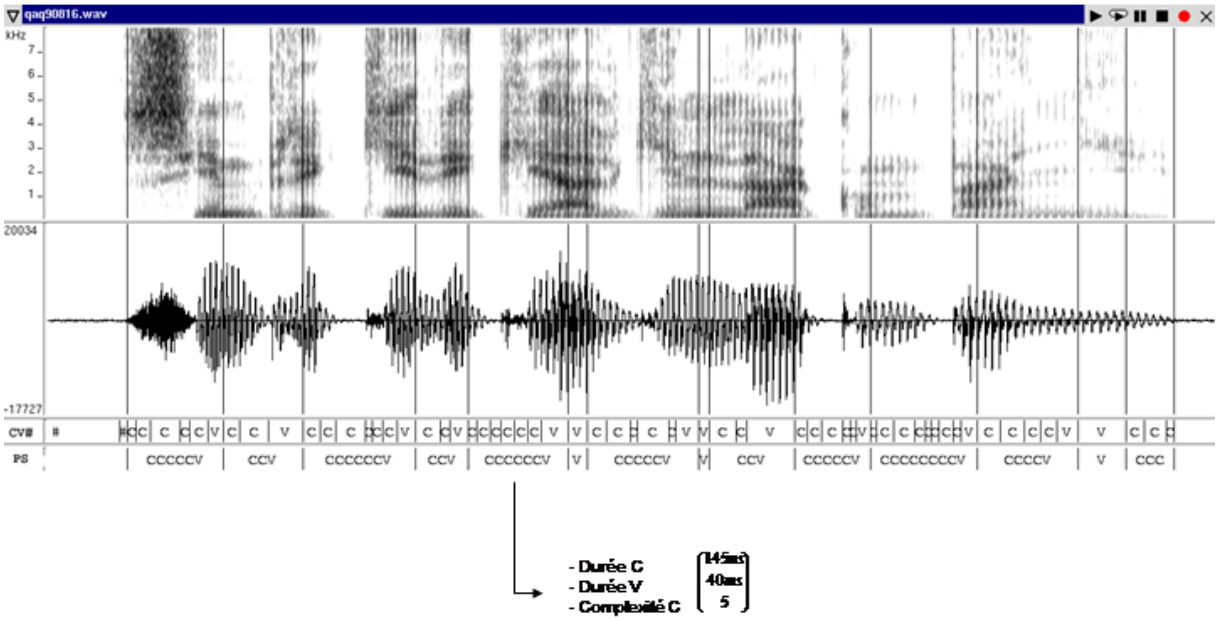


Fig. 4.7 : Extraction de paramètres de durée à partir de la segmentation en pseudo-syllabes sur la phrase « *Siempre quiere que alguien la acompañe* ».

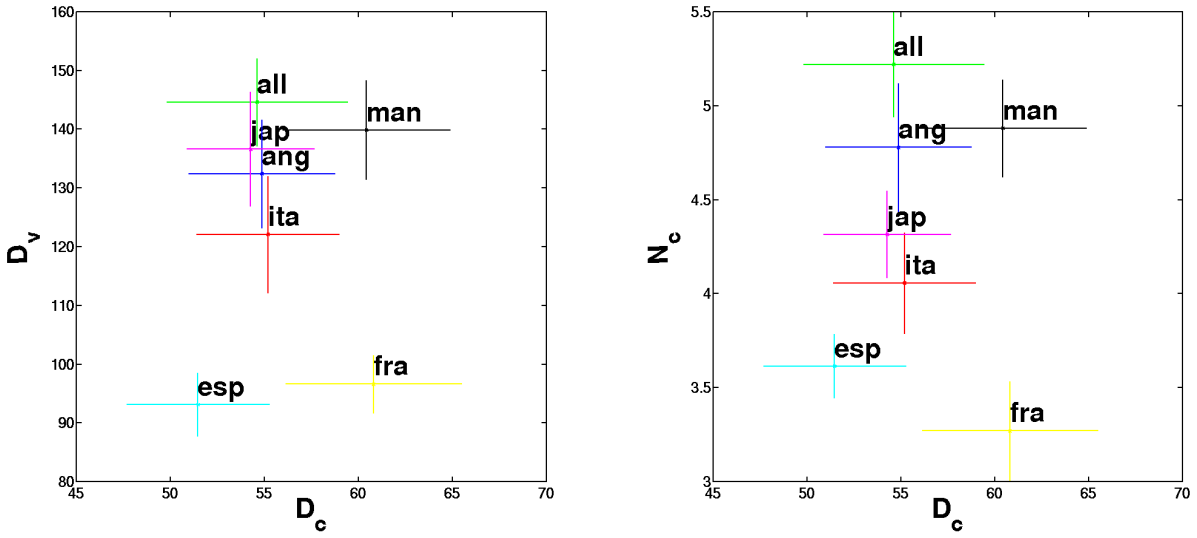
Étude des distributions

Comme pour les systèmes de Ramus et Grabe, les distributions des paramètres sont observées au travers des données du jeu n°1 du corpus MULTEXT afin d'évaluer le pouvoir discriminant des paramètres employés. Afin de pouvoir comparer les graphiques, les paramètres pseudo-syllabiques sont moyennés sur chaque phrase pour donner une seule observation par phrase.

Les projections des distributions des différents paramètres pour chaque langue sont représentées sur la figure 4.8.

Dans le plan (D_c, D_v), on peut remarquer que le paramètre D_v permet de séparer le groupe français-espagnol d'un groupe formé par l'ensemble des autres langues. Le paramètre D_c permet la distinction entre le français et l'espagnol dans le premier groupe, ainsi que celle du mandarin dans le deuxième groupe.

Dans le plan (D_c, N_c), nous pouvons effectuer un regroupement en classes rythmiques, avec un groupe de langues accentuelles (anglais, allemand et mandarin), un groupe de langues syllabiques (français, espagnol) et un groupe intermédiaire (japonais, italien).



(a) (b)

Fig. 4.8 : Paramètres extraits des pseudo-syllabes
(a) Paramètres (D_c, D_v),
(b) Paramètres (D_c, N_c).

Ces regroupements effectués diffèrent de ceux obtenus avec les systèmes de Ramus et de Grabe (§4.4.1 et §4.4.2). Les langues sont mieux séparées qu'avec le système de Grabe, et le paramètre N_c permet une bonne distinction des langues. La complexité syllabique augmente graduellement du français à l'allemand, en passant successivement par l'espagnol puis l'italien, le japonais, l'anglais et le mandarin.

Nous pouvons supposer que ces paramètres seront plus efficaces en identification des langues, car ils permettent d'effectuer des regroupements similaires aux groupes rythmiques linguistiques (comme le système de Ramus), tout en conservant des différences suffisantes entre les langues d'un même groupe.

Expériences en identification des langues

Le protocole expérimental suivi est le même que celui employé pour tester les approches de Ramus et Grabe (§4.3.2). Chaque pseudo-syllabe constitue un vecteur d'observation de dimension 3 (D_c, D_v, N_c).

Les expériences ont montré que le meilleur taux d'identification sur les données de l'ensemble d'apprentissage est de 68,7 %, soit 385 identifications correctes sur 560 fichiers. Ce résultat est obtenu pour 8 gaussiennes. La matrice de confusion correspondante est représentée sur le tableau 4.10.

Tab. 4.10 : Modèle de durées pseudo-syllabiques
Expériences sur l'ensemble d'apprentissage : correct : 68,75 ± 3,8 % (385/560)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	65	9	2	-	-	2	2
Allemand	22	51	6	-	-	1	-
Mandarin	18	10	46	1	4	-	1
Français	4	-	-	68	-	8	-
Italien	22	-	-	10	31	10	7
Espagnol	2	-	2	14	2	60	-
Japonais	9	-	-	1	3	3	64

En reconnaissance, l'expérience est menée pour les paramètres donnant le meilleur résultat sur l'ensemble d'apprentissage, soit 8 gaussiennes par MMG. Le taux d'identification correcte est de 66,9 % (93 identifications correctes sur 139 fichiers). La matrice de confusion est représentée sur le tableau 4.11.

Tab. 4.11 : Modèle de durées pseudo-syllabiques
Expériences sur l'ensemble de test : correct : 66,9 ± 7,8 % (93/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	16	1	1	-	1	1	-
Allemand	5	14	1	-	-	-	-
Mandarin	4	3	11	-	1	-	1
Français	-	-	-	19	-	-	-
Italien	6	1	1	-	11	-	1
Espagnol	-	-	-	8	2	6	4
Japonais	2	-	-	-	2	-	16

Ces résultats confirment la pertinence de la modélisation du rythme par les pseudo-syllabes. Toutes les langues sont correctement reconnues, sauf l'espagnol qui est principalement confondu avec une autre langue syllabique, le français. L'italien, l'allemand et le mandarin sont légèrement confondus avec l'anglais, une autre langue accentuelle.

Lorsque l'on effectue un regroupement des résultats selon les classes rythmiques, le taux d'identification correcte est de 85,6 %. La matrice de confusion est donnée en fonction des groupes rythmiques sur le tableau 4.12.

Tab. 4.12 : Modèle de durées pseudo-syllabiques : Regroupement selon les classes rythmiques
Expériences sur l'ensemble de test : correct : 85,6 ± 7,0 % (119/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	56	3	1
L. Syllab.	8	47	5
L. Mora.	2	2	16

Les confusions sont peu nombreuses entre les différents groupes de langues. Cette expérience montre que les caractéristiques rythmiques extraites des pseudo-syllabes sont efficaces pour l'identification des langues ou du moins des groupes de langues.

4.5.4 Modélisation des caratéristiques intonatives des pseudo-syllabes

Nous allons utiliser la segmentation en pseudo-syllabes pour caractériser des évènements prosodiques. Des paramètres liés à la forme et à l'ampleur des variations de la courbe intonative sont extraits sur chaque pseudo-syllabe. Des expériences sont menées en identification des langues et permettent de juger de la pertinence de l'unité pseudo-syllabe dans un cadre de segmentation de la parole en unités prosodiques.

Les difficultés rencontrées lors de la modélisation de l'intonation sont similaires à celles rencontrées pour la modélisation du rythme : le problème principal est la définition d'une unité prosodique intonative qui permettrait de prendre en compte les caractéristiques propres à l'intonation.

Dans la section précédente, nous avons évalué la pseudo-syllabe, unité que nous avons décrite (§4.5) et utilisée avec succès pour la modélisation du rythme. Afin de confirmer la validité prosodique de cette unité, nous avons procédé à des expériences de caractérisation de l'intonation à partir de la segmentation en pseudo-syllabes.

Caractérisation de l'intonation d'une pseudo-syllabe

De manière semblable à celle utilisée pour modéliser le rythme (§4.5.3), un ensemble de paramètres est extrait sur chaque pseudo-syllabe. Les paramètres sont calculés à partir de la segmentation en pseudo-syllabes et des valeurs de la fréquence fondamentale extraites toutes les 10 ms. Les paramètres sont :

- Le skewness ou coefficient d'aplatissement de la distribution des valeurs de fréquence fondamentale sur chaque pseudo-syllabe, que nous noterons $f_{0_{Skew}}$;
- Le kurtosis ou coefficient d'asymétrie de la distribution des valeurs de fréquence fondamentale sur chaque pseudo-syllabe, que nous noterons $f_{0_{Kurt}}$;

- L'écart entre la position de la valeur maximale de la fréquence fondamentale sur la pseudo-syllabe et le début du segment vocalique de la pseudo-syllabe, noté δ_{max} ;
- L'écart entre la position de la valeur minimale de la fréquence fondamentale sur la pseudo-syllabe et le début du segment vocalique de la pseudo-syllabe, noté δ_{min} ;
- La bande passante normalisée ou écart entre la valeur maximale et la valeur minimale de fréquence fondamentale sur chaque pseudo-syllabe, normalisé par la moyenne des valeurs de fréquence fondamentale sur la pseudo-syllabe, noté $\Delta f_0 N$.

On obtient ainsi un vecteur d'observation de dimension 5 pour chaque pseudo-syllabe ·

$$\psi = (f_{0_{Skew}}, f_{0_{Kurt}}, \delta_{max}, \delta_{min}, \Delta f_0 N) \tag{4.10}$$

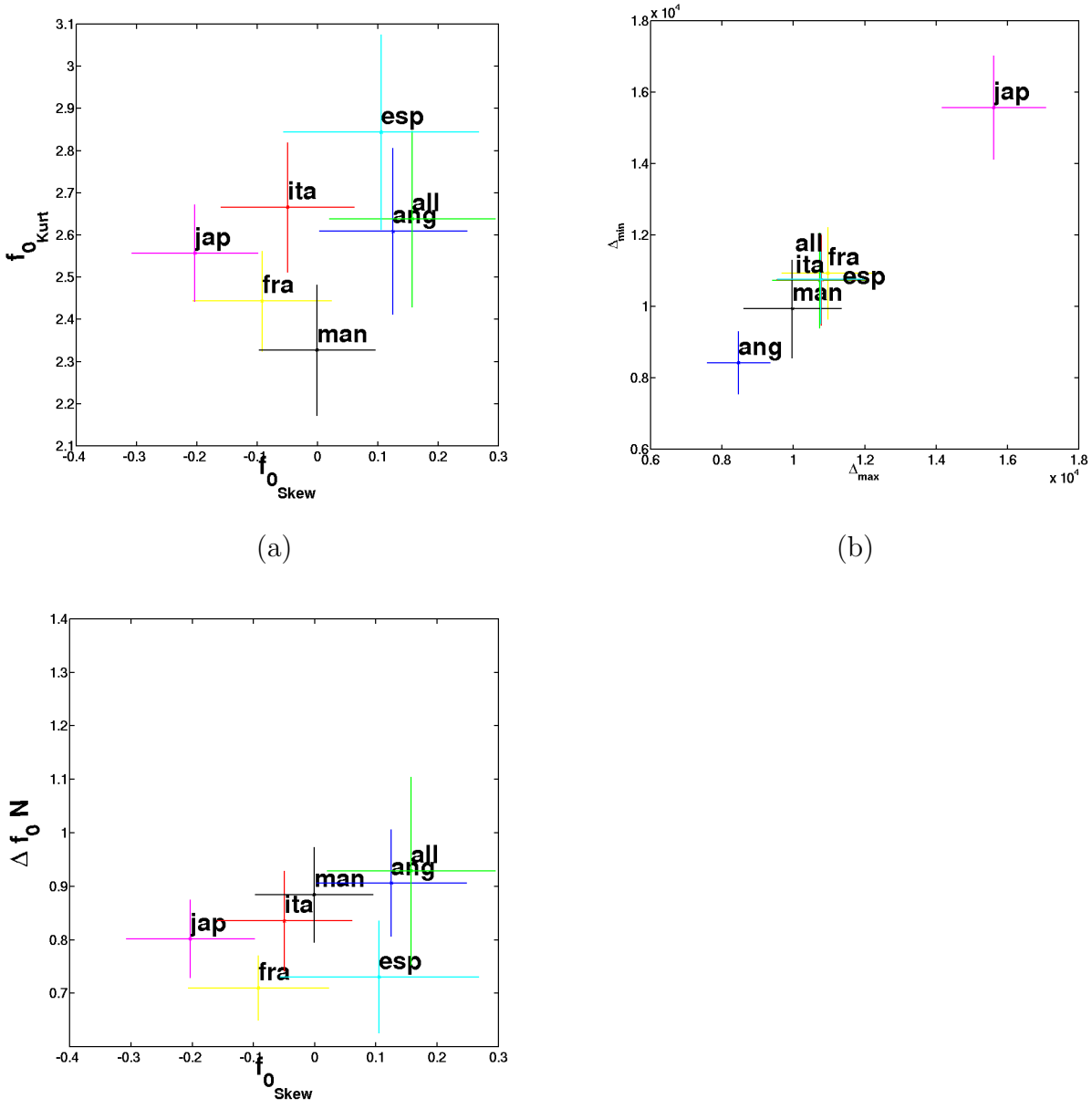
Représentations graphiques

Comme lors de l'étude des systèmes précédents, des représentations graphiques permettent de voir la répartition des différents paramètres selon les langues (figure 4.9).

Le plan ($f_{0_{Skew}}$, $f_{0_{Kurt}}$) n'apporte pas d'informations pertinentes. On notera que le japonais possède le plus faible coefficient d'aplatissement, tandis que le mandarin possède le plus faible coefficient d'asymétrie.

Le plan (δ_{max}, δ_{min}) montre que ces deux variables sont parfaitement corrélées. Nous remarquons toutefois que le japonais se situe très à l'écart des autres langues.

Le plan ($f_{0_{Skew}}$, $\Delta f_0 N$) montre que le français et l'espagnol ont une dynamique de fréquence fondamentale faible par rapport à toutes les autres langues. Ce ne sont pas les langues asiatiques qui possèdent la plus forte dynamique, mais les langues accentuelles.



(a) (b)

(c)

Fig. 4.9 : Paramètres extraits des pseudo-syllabes
(a) Paramètres ($f_{0_{Skew}}$, $f_{0_{Kurt}}$),
(b) Paramètres (δ_{max}, δ_{min}),
(c) Paramètres ($f_{0_{Skew}}$, $\Delta f_0 N$).

Expériences en identification des langues

Les expériences sur les données d'apprentissage ont montré que le meilleur taux d'identification correcte (74,1 ± 3,6 %) est obtenu pour des modèles MMG à 8 composantes.

La matrice de confusion correspondante est représentée sur le tableau 4.13.

Tab. 4.13 : Modèle intonatif pseudo-syllabique
Expériences sur l'ensemble d'apprentissage : correct : 74,1 ± 3,6 % (415/560)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	59	10	3	3	2	2	1
Allemand	6	62	7	-	3	2	-
Mandarin	11	10	48	1	3	6	1
Français	1	-	5	71	2	-	1
Italien	9	1	5	8	51	2	4
Espagnol	6	1	1	6	5	58	3
Japonais	1	-	2	7	4	-	66

En reconnaissance, les expériences d'identification des langues sont menées pour les modèles MMG à 8 composantes. Le taux d'identification correcte est de 52,5 ± 8,3 %, soit 73 fichiers correctement identifiés sur 139. La matrice de confusion est représentée sur le tableau 4.14.

Tab. 4.14 : Modèle intonatif pseudo-syllabique
Expériences sur l'ensemble de test : correct : 52,5 ± 8,3 % (73/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	10	-	2	4	2	2	-
Allemand	5	4	6	1	4	-	1
Mandarin	5	-	13	1	1	-	-
Français	-	-	2	16	-	-	1
Italien	7	-	-	2	10	-	1
Espagnol	1	-	-	11	3	1	4
Japonais	-	-	-	1	-	-	19

La langue la mieux reconnue est le japonais, ce qui est concordant avec les résultats graphiques. Les langues syllabiques sont principalement confondues entre elles. Pour les langues accentuelles, le mandarin est confondu avec l'anglais tandis que l'allemand est confondu avec la plupart des autres langues.

Tab. 4.15 : Modèle intonatif pseudo-syllabique : Expériences sur l'ensemble de test
Regroupement en classes : correct : 77,0 ± 7,0 % (107/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	45	15	-
L. Syllab.	10	43	6
L. Mora.	-	1	19

En effectuant des regroupements selon les classes rythmiques, le taux d'identification correcte devient 77,0 ± 7,0 %. Ces résultats montrent que la segmentation en pseudo-syllabe est également pertinente pour la modélisation de l'intonation.

4.5.5 Fusion des modèles de durée et d'intonation pseudo-syllabiques

Une expérience de fusion des deux approches précédentes est menée. La décision est prise à partir d'une addition pondérée des log-vraisemblances. Les poids sont appris sur les données de l'ensemble d'apprentissage.

Les poids optimaux sont :
- 0,4 pour le modèle rythmique,
- 0,6 pour le modèle intonatif.

Remarquons que le poids associé au modèle intonatif est plus important que celui associé au modèle rythmique alors que les performances de ce dernier sont meilleures. La taille limitée du corpus - notamment l'absence d'un ensemble de développement - ne permet pas un apprentissage correct de ces poids.

La matrice de confusion correspondante est donnée sur le tableau 4.16.

Tab. 4.16 : Fusion des modélisations rythmiques et intonatives
Expériences sur l'ensemble d'apprentissage : correct : 83,2 ± 3,1 % (455/560)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	69	7	2	-	-	2	-
Allemand	8	70	1	-	1	-	-
Mandarin	11	10	55	-	2	2	-
Français	3	-	-	74	-	2	1
Italien	8	-	-	8	59	2	3
Espagnol	2	1	-	8	3	66	-
Japonais	1	-	-	2	3	1	73

Une fois les poids déterminés, nous les avons employés pour les expériences d'identification des langues sur les données de test. Les résultats sont indiqués sur le tableau 4.17.

Tab. 4.17 : Fusion des modélisations rythmiques et intonatives
Expériences sur l'ensemble de test : correct : 66,2 ± 7,9 % (92/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	17	-	2	1	-	-	-
Allemand	9	8	3	-	-	-	-
Mandarin	8	-	12	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	6	-	-	-	13	-	1
Espagnol	-	-	-	11	2	4	3
Japonais	-	-	-	-	1	-	19

Nous voyons que les résultats sont similaires à ceux obtenus en employant uniquement le modèle de rythme. Cependant, la dégradation est assez forte entre les expériences sur les données d'apprentissage et celles sur les données de test, ce qui n'était pas le cas pour le modèle rythmique. Il est possible qu'il y ait sur-apprentissage du modèle intonatif. Il n'y a pas d'amélioration du taux d'identification global.

Tab. 4.18 : Fusion des modélisations rythmiques et intonatives
Regroupement en classes : correct : 91,3± 4,7 % (127/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	59	1	-
L. Syllab.	6	49	4
L. Mora.	-	1	19

Comme on peut le voir sur le tableau 4.18, les confusions se font très souvent entre langues appartenant à la même famille. Lorsque l'on effectue les regroupements selon les typologies rythmiques, on s'apperçoit que les confusions sont très peu nombreuses entre les groupes linguistiques. Les performances deviennent significativement supérieures à celles obtenues avec la modélisation des durées des pseudo-syllabes (tableau 4.19).

4.6 Conclusion

Au cours de ce chapitre, nous avons étudié différentes approches de modélisation du rythme des langues. Les approches de Ramus (§3.1.1) et de Grabe (§3.1.2) prennent pour point de départ des mesures de durées vocaliques et intervocaliques. Notre méthode consiste à déterminer une unité de type syllabique qui permet d'effectuer des mesures similaires à celles de Ramus et Grabe.

Nous avons ensuite testé cette unité au travers de mesures de durées et de mesures de paramètres liés à l'intonation. Les résultats obtenus en identification des langues (voir

tableau 4.19) sont encourageants, et ont montré que la pseudo-syllabe est une unité prosodique pertinente. Rappelons toutefois que la taille limitée de corpus, particulièrement l'absence d'un ensemble de développement, constitue une limitation. Disposer d'un corpus possédant un ensemble de développement devrait permettre d'améliorer les performances.

Cependant, afin de modéliser la prosodie, nous devons tenir compte des enchaînements de ces unités pour prendre en compte la dynamique de la prosodie selon une échelle suprasegmentale, ce sera l'objet du chapitre suivant.

Tab. 4.19 : Récapitulatif des expériences du chapitre

Paramètres	% correct	% correct groupes
Ramus	43,9	80,5
Grabe	36,7	56,8
Pseudo-syllabes (rythme)	66,9	85,6
Pseudo-syllabes (intonation)	52,5	77,0
Pseudo-syllabes (rythme et intonation)	66,2	91,3

Chapitre 5

Un système d'identification automatique des langues par la prosodie

Sommaire

5.1 Cadre expérimental 114
5.1.1 Création de classes de segments 114
5.1.2 Modélisation par multigrammes 114
5.1.3 Règle de décision 115
5.2 Modélisation du rythme : prise en compte temporelle 115
5.2.1 Regroupement des pseudo-syllabes en classes 115
5.2.2 Expériences en identification des langues 116
5.3 Modélisation de l'intonation à long terme 118
5.3.1 Traitement de la courbe de fréquence fondamentale 118
5.3.2 Traitement de la courbe d'énergie 125
5.3.3 Identification des langues 126
5.4 Modélisation de la prosodie (rythme et intonation) 127
5.4.1 Ajout d'étiquettes au modèle rythmique 128
5.4.2 Ajout d'étiquettes au modèle intonatif 129
5.4.3 Conclusion 131
5.5 Expériences sur le système d'Adami 131
5.6 Expériences en modélisation de la prosodie à court terme . . 134
5.7 Fusion : modèle dynamique et modèle statique 136
5.8 Fusion : modèle accentuel et modèle intonatif 137
5.9 Comparaison avec d'autres systèmes d'identification 138
5.9.1 Méthode acoustique 138
5.9.2 Méthode phonotactique (PPRLM) 140
5.9.3 Conclusion 141
5.10 Conclusion 141

Les études précédentes montrent que la pseudo-syllabe peut être proposée comme une unité de base prosodique, facilitant l'extraction de caractéristiques prosodiques. Nous avons cherché à exploiter l'enchaînement de ces unités, tout comme les modèles phonotactiques étudient l'enchaînement des phonèmes.

Dans un premier temps, les pseudo-syllabes sont regroupées en classes selon leurs caractéristiques rythmiques (§5.2) et intonatives (§5.3) ou une combinaison des deux (§5.4). À chaque classe correspond naturellement une étiquette. Les séquences de ces étiquettes sont ensuite modélisées au moyen de modèles multigrammes (§5.1.2). Des expériences en identification des langues, toujours sur le corpus Multext, sont menées et commentées.

Nous avons également étudié un système similaire à celui d'Adami (§3.3.5, [2]), que nous testons sur le corpus Multext à des fins de comparaison.

Enfin, une approche alternative est proposée, décrite et expérimentée. La notion de classes rythmiques étant au cœur de la problématique, elle est développée au cours de la discussion.

5.1 Cadre expérimental

Le protocole expérimental suivi pour valider nos propositions ultérieures se décompose en trois phases :

- une définition de classes de pseudo-syllabes et un étiquetage des pseudo-syllabes en ces classes,
- une modélisation statistique des séquences d'étiquettes selon chaque langue considérée,
- une règle de décision.

5.1.1 Création de classes de segments

Les segments issus de la segmentation en pseudo-syllabes (§4.5) sont étiquetés selon la procédure suivante :

1. des caractéristiques sont calculées sur chaque pseudo-syllabe (durées, pente de la fréquence fondamentale, ...),
2. des règles empiriques sont ensuite utilisées pour créer des classes, par exemple en mesurant la distance des caractéristiques par rapport à la moyenne.
3. L'ensemble des segments du corpus est étiqueté d'après ces règles.

5.1.2 Modélisation par multigrammes

Les enchaînements des étiquettes sur les phrases de l'ensemble d'apprentissage sont modélisés par des modèles multigrammes. Les séquences les plus fréquentes sont alors identifiées pour chaque langue et des probabilités leur sont associées.

Un modèle de langage multigramme est un modèle statistique qui probabilise chaque motif de suites d'unités [16, 17, 29, 30]. La modélisation multigrammes et la modélisation n-grammes permettent de capturer des informations sur l'enchaînement et la fréquence des unités. Dans le cas classique d'une modélisation de séquences d'unités acoustiques, ces modélisations traduisent les contraintes phonologiques du système de production de la parole et font émerger les unités les plus fréquentes dans la langue utilisée. Les modèles multigrammes permettent de s'affranchir de la longueur de l'unité, alors que le modèle n-grammes la fixe. Cette particularité permet de modéliser des informations linguistiques de plus haut niveau, notamment les informations lexicales.

Un modèle de langage multigramme est constitué d'un dictionnaire, contenant des « mots », c'est-à-dire des séquences d'unités z_i associées à leur probabilité d'occurrence dans le corpus $P(z_i)$. La longueur de ces « mots » est variable, mais on peut fixer la longueur maximale à n : on parle alors d'un modèle n-multigramme.

5.1.3 Règle de décision

Lors de la phase de reconnaissance, la vraisemblance de chaque phrase est calculée par rapport à chacun des modèles multigrammes appris. La langue pour laquelle cette vraisemblance est la plus forte est la langue identifiée.

Les expériences sont toujours conduites sur l'ensemble de test du corpus MULTEXT (§4.3.1).

5.2 Modélisation du rythme : prise en compte temporelle

Les paramètres utilisés précédemment pour caractériser les durées des pseudo-syllabes sont repris pour les regrouper en classes. Ces classes sont créées statistiquement à partir des histogrammes de répartition des valeurs des différents paramètres.

5.2.1 Regroupement des pseudo-syllabes en classes

Pour classer les pseudo-syllabes, nous avons cherché à modéliser statistiquement chacun des différents paramètres extraits des pseudo-syllabes, à savoir :
- la durée des consonnes,
- la durée de la voyelle,
- la complexité (le nombre de consonnes),
- l'énergie.

Nous nous sommes limités à la recherche de modèles statistiques paramètriques et compte tenu des histogrammes que nous avons obtenus (figure 5.1), les lois probabilistes choisies sont :
- loi inverse gaussienne (loi de Wald) pour la durée des consonnes,
- loi inverse gaussienne (loi de Wald) pour la durée des voyelles,
- loi de poisson pour la complexité de la pseudo-syllabe (voir §4.5.2),
- loi normale pour l'énergie.

Les histogrammes correspondants aux pseudo-syllabes de l'ensemble d'apprentissage ainsi que les lois statistiques associées (en rouge) à chaque paramètre sont représentés sur la figure 5.1.

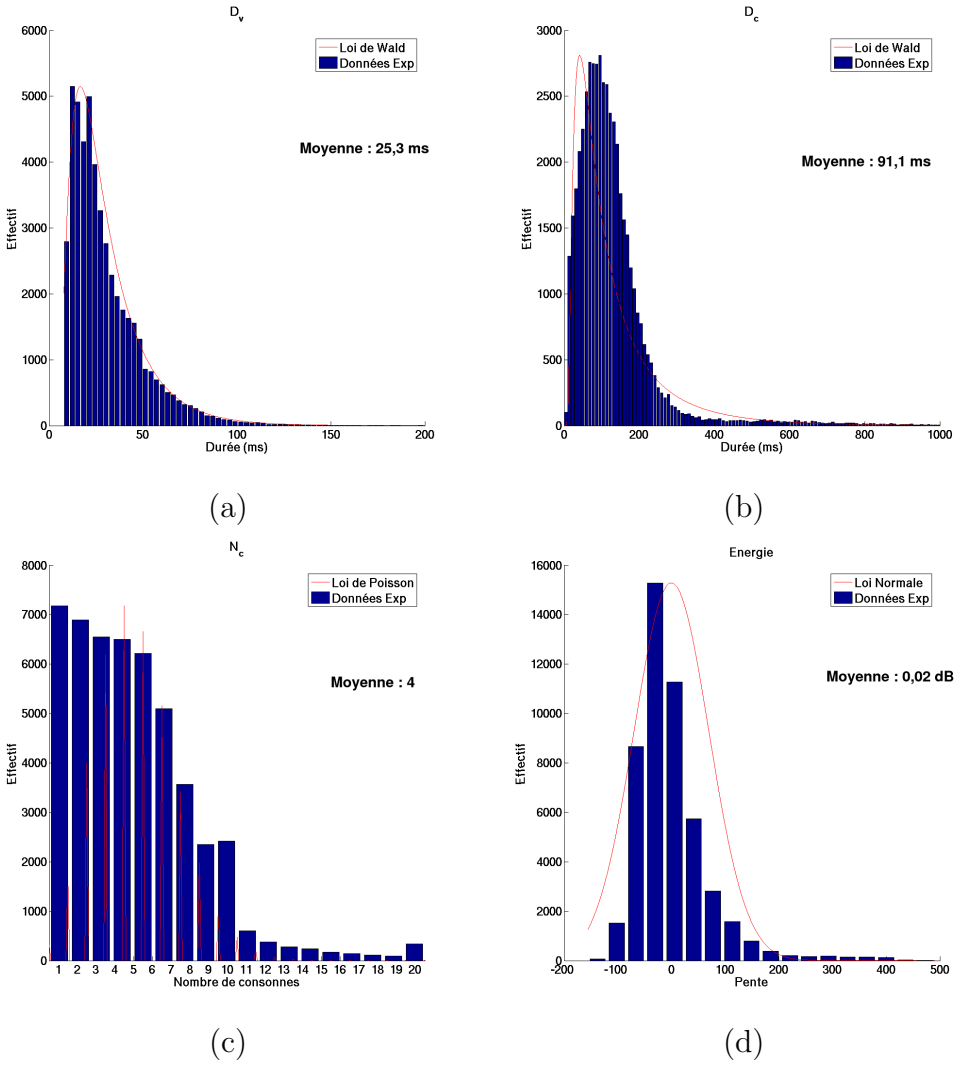


Fig. 5.1 : Histogrammes de répartition des paramètres de pseudo-syllabes sur l'ensemble des langues,
(a) Durées des consonnes, loi de Wald (λ=71, μ=30),
(b) Durées des voyelles, loi de Wald (λ=138, μ=133),
(c) Complexité, loi de Poisson (λ=4.6),
(d) Energie, loi Normale (μ=0, σ=70).

À l'issue de cet étalonnage, deux classes sont définies pour chaque type de paramètre, selon que ce paramètre est supérieur ou inférieur à la moyenne de la loi statistique associée. Il en résulte 16 classes possibles. Ce jeu de symboles est commun à toutes les langues.

5.2.2 Expériences en identification des langues

Les expériences sont toujours menées sur le jeu n°1 du corpus MULTEXT. Les modèles multigrammes sont entraînés pour chaque langue avec les pseudo-syllabes de l'ensemble

d'apprentissage, étiquetées grâce aux règles précédentes.

Les expériences en identification des langues montrent que le taux d'identification correcte est de 42,4 ± 8,2 % (59 identifications correctes sur 139 fichiers). Ce taux est obtenu pour des modèles 2-multigrammes. La matrice de confusion est représentée sur le tableau 5.1.

Tab. 5.1 : Modèle rythmique 2-multigrammes
Expériences sur l'ensemble de test : correct : 42,4 ± 8,2 %(59/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	3	1	5	3	5	-	3
Allemand	-	17	-	-	-	-	3
Mandarin	3	5	5	-	1	1	5
Français	-	-	-	8	7	4	-
Italien	-	3	-	2	10	1	4
Espagnol	-	-	-	4	5	6	5
Japonais	3	-	3	-	2	2	10

Les résultats montrent un faible taux d'identification global. Les langues accentuelles sont principalement confondues entre elles (sauf l'allemand, bien reconnu). La même observation peut être faite pour les langues syllabiques (français, italien et espagnol). Le japonais est également mal reconnu, alors que c'était la langue la mieux reconnue par les systèmes précédents.

Le japonais est une langue où l'on retrouve des oppositions fortes de durées entre les voyelles courtes et les voyelles longues. La faible performance du système sur le japonais est peut être dûe à un problème de durée, la moyenne n'étant pas nécessairement un bon critère puisque les paramètres de durée suivent des lois inverse gaussiennes.

Tab. 5.2 : Modèle rythmique 2-multigrammes : Expériences sur l'ensemble de test ;
Regroupement en classes rythmiques : correct 69,0 ± 7,7 % (96/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	39	10	11
L. Syllab.	3	47	9
L. Mora.	6	4	10

Après regroupement des langues (figure 5.2), les résultats montrent que le modèle rythmique permet de séparer les langues accentuelles et les langues syllabiques. Par contre, la performance est moindre sur la langue moraïque, ce qui peut être dû à un regroupement non optimal des pseudo-syllabes en classes. Ce modèle peut également être amélioré en considérant également des informations sur l'intonation.

5.3 Modélisation de l'intonation à long terme

Pour modéliser l'intonation, il est important de définir une unité de base, pour ensuite en gérer les enchaînements.

Doit-on employer une unité intonative, c'est-à-dire longue, ou une unité accentuelle de taille plus réduite ?

Notre première hypothèse est que les variations prosodiques caractéristiques des langues sont plus visibles à long terme. En effet, certaines langues possèdent des schémas intonatifs portant sur l'ensemble d'un énoncé (ou d'une phrase) tandis que d'autres langues possèdent moins de contraintes liées à la totalité de l'énoncé.

La définition de l'unité intonative est un concept important pour le traitement de la parole. Cependant, la segmentation d'un discours d'une langue donnée en unités intonatives n'est pas une tâche aisée (voir [96] pour le russe par exemple). De plus, la recherche d'une unité intonative multilingue rend le travail encore plus complexe.

C'est pour ces raisons que nous avons décidé dans un premier temps d'employer l'unité définie plus haut pour la modélisation du rythme, la pseudo-syllabe. La caractérisation des enchaînements de pseudo-syllabes devrait nous permettre d'estimer la régularité des mouvements mélodiques sur un ou plusieurs mots.

Par la suite, et en suivant l'idée d'Adami (§3.3.5 et [2]), nous considérons des unités de taille plus réduite, que l'on pourrait qualifier d'unités infra-accentuelles. De la même façon que précédemment, nous allons caractériser les enchaînements de ces unités par des modèles multigrammes, qui permettront alors de modéliser les variations sur une ou plusieurs syllabes.

5.3.1 Traitement de la courbe de fréquence fondamentale

En s'inspirant des travaux de Fujisaki [42], nous considérons que la courbe mélodique résulte de deux contributions :

- l'accentuation « de phrase » qui décrit les variations macroprosodiques sur une phrase,
- l'accentuation « locale » qui permet de prendre en compte des accentuations portées sur des unités plus courtes, comme les syllabes (ou les phonèmes).

Accentuation de phrase

Afin de calculer automatiquement l'accentuation de phrase, un prétraitement est effectué. Le signal audio est segmenté en phrases séparées par des silences. À l'intérieur de ces phrases, la ligne de base [125], caractérisant les mouvements à long terme de la fréquence

fondamentale, est calculée.

1. Segmentation en phrases :

   La segmentation en phrases se base sur l'analyse automatique du signal déjà utilisée. Le signal est segmenté avec l'algorithme DFB [5]. Les segments de silence (non activité) sont identifiés grâce au détecteur d'activité vocale. Les segments de silence adjacents sont ensuite regroupés (voir chapitre 4).
   Il est classique de supposer que les phrases sont entrecoupées de silences de taille significative, correspondant à des respirations. Notre but étant d'étudier la prosodie, il faut prendre garde de ne repérer que les réelles pauses discursives. C'est pourquoi nous avons sélectionné les « longs » silences. Les phrases seront situées entre ces silences. Le seuil utilisé pour déterminer si le silence est « long » est empiriquement fixé à 250 ms (figure 5.2).
   Pour les expériences qui suivent, l'ensemble des graphiques est donné pour un seul signal audio, contenant 5 phrases de parole lue par un locuteur mâle en espagnol. Ce fichier fait partie de la base de données MULTEXT.

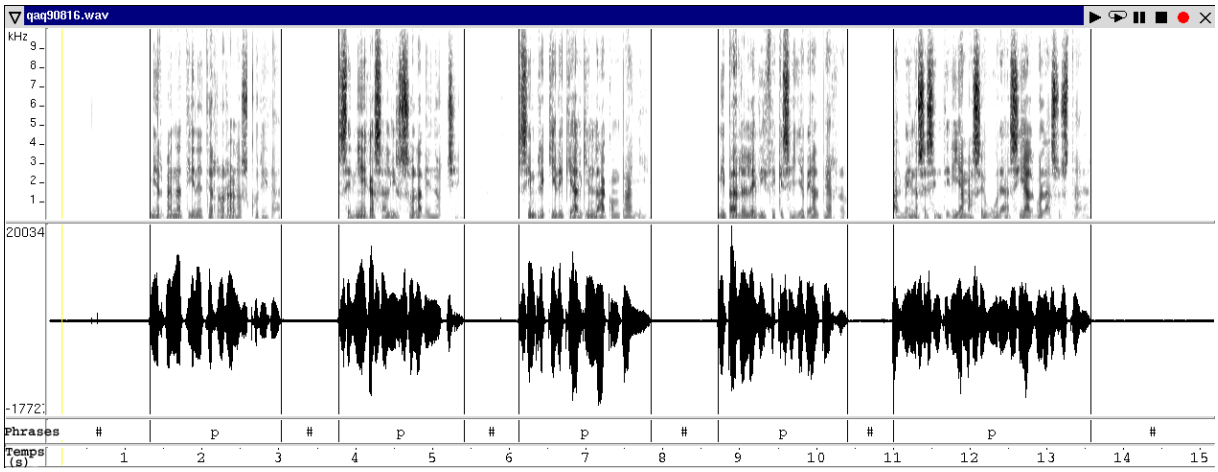


Fig. 5.2 : Exemple de segmentation en phrase, obtenue sur le signal audio spanish/qa/qaq90816. Le locuteur lit : « *Mi hermana tiene terror a la oscuridad. Se niega a salir sola de noche. Siempre quiere que alguien la acompañe. Mi padre me dice que se lleve al perro. Al menos se sentiria màs segura si algo la asustara* ».

2. Approximation de la ligne de base (accentuation de phrase) :

   Afin de déterminer la ligne de base [125], il est supposé qu'elle passe par les minima locaux de F_0, de telle sorte qu'aucun point ne se situe au-dessous d'elle. Pour chaque phrase, le même traitement est appliqué.

   (a) L'algorithme d'extraction de fréquence fondamentale peut donner par erreur des valeurs non nulles dans les zones non voisées. C'est pourquoi on considère

comme nulle toute valeur détectée sur les segments étiquetés « silence ».

(b) Les valeurs de la fréquence fondamentale en Hertz sont converties en demi-tons. L'échelle en demi tons a pour valeur minimale la hauteur du La0 (55 Hz). La quantification en demi tons permet à la fois de se reporter sur une échelle logarithmique, plus proche de l'échelle de la perception humaine, et de lisser la courbe mélodique. La formule de conversion des Hertz en demi-tons est :

$$F(demi-tons) = 12 * \log_n \frac{F(Hz)}{F_{ref}(Hz)} / \log_n 2 \qquad (5.1)$$

avec $F_{ref} = 55$ Hz. Un exemple de conversion en demi-tons est montré sur la figure 5.3.

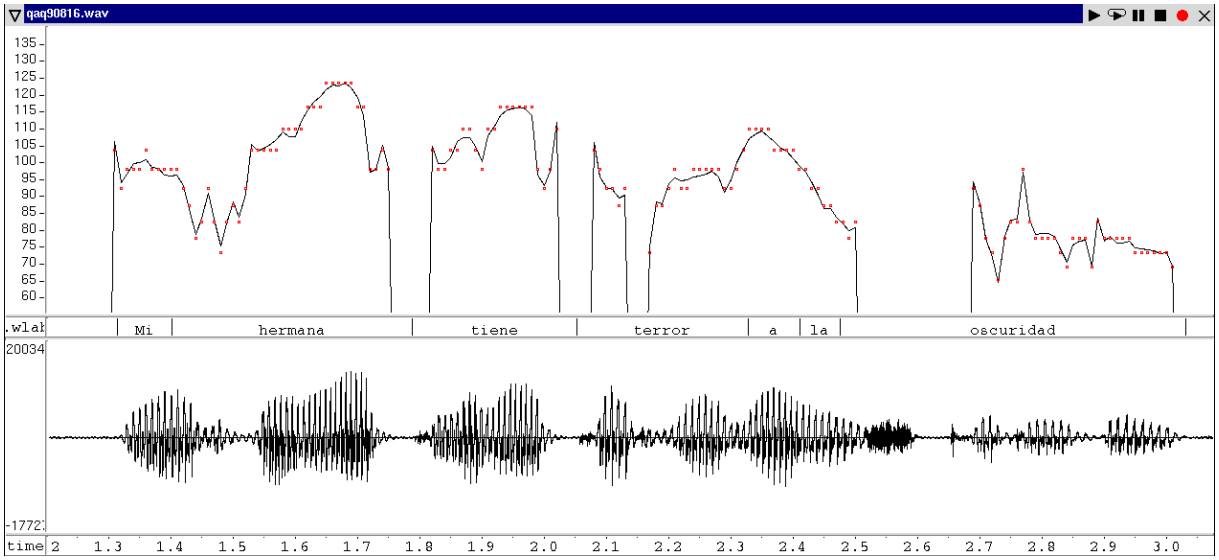


Fig. 5.3 : Exemple de quantification en demi tons, obtenue sur une partie d'un enregistrement d'espagnol (spanish/qa/qaq90816)
- ligne continue : valeurs de f0 en Hz
- ligne pointillée : résultat de la quantification en demi-tons

(c) La droite de régression linéaire est estimée sur l'ensemble des parties voisées de chaque phrase (ligne bleue continue sur la figure 5.4).

(d) Le minimum est alors repéré sur chaque partie voisée située en dessous de cette droite. Par exemple, sur la première partie voisée de la première phrase de l'énoncé de la figure 5.4, deux portions de la courbe de fréquence fondamentale se retrouvent en dessous de la droite de régression. Deux minima sont alors trouvés.

(e) L'accentuation de phrase ou ligne de base est la droite qui rejoint les minima de la phrase. La figure 5.4 montre les lignes de base obtenues pour les 4 premières phrases d'un enregistrement d'espagnol.

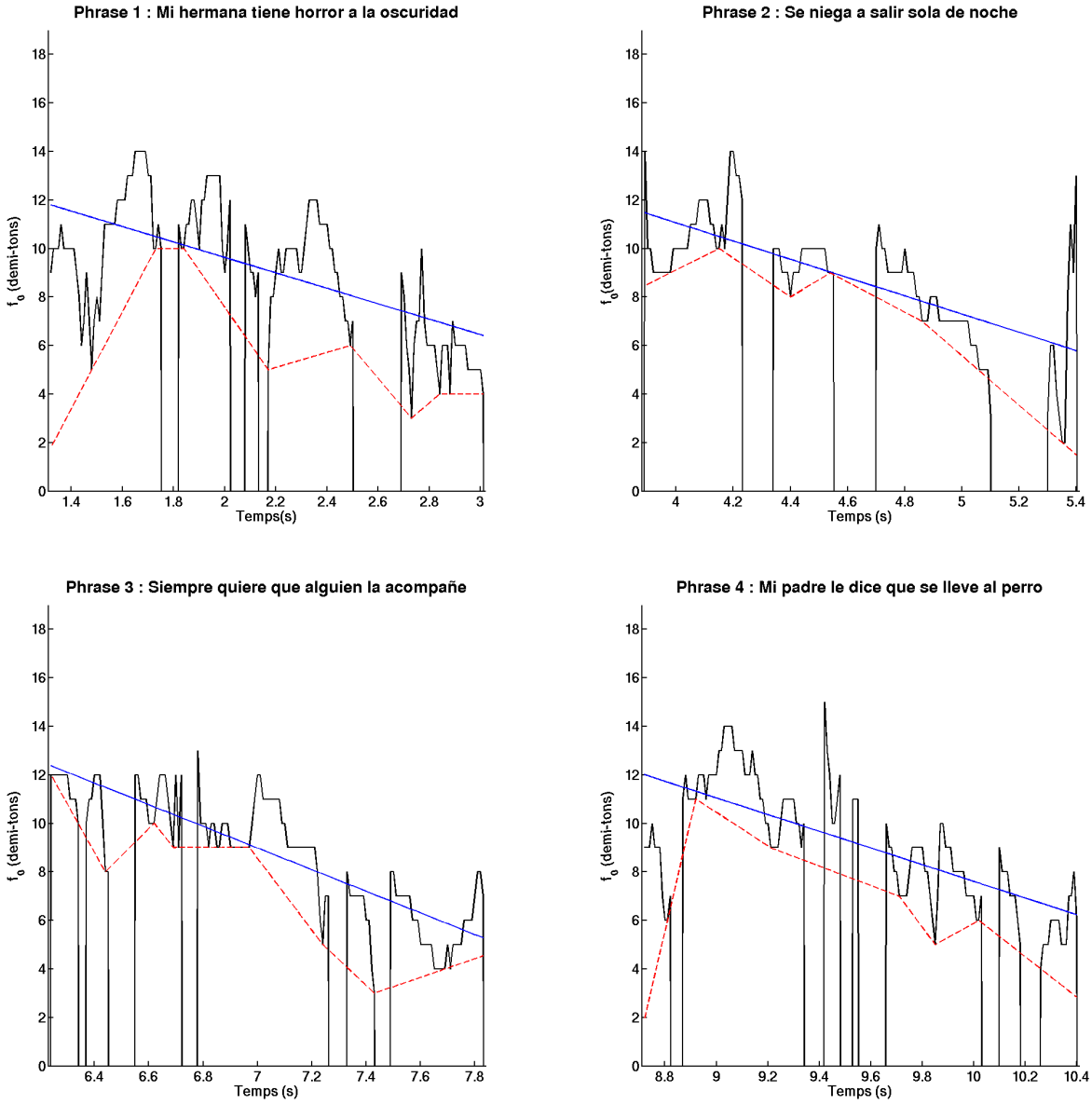


Fig. 5.4 : Lignes de base pour les phrases d'un enregistrement d'espagnol (spanish/qa/qaq90816).
- Ligne continue : approximation linéaire
- ligne pointillée : ligne de base

Ce traitement permet d'obtenir l'accentuation de phrase, en respectant le fait qu'un minimum de points doivent se trouver sous la courbe intonative originale. Pour le fichier montré en exemple (figure 5.4), le nombre de points situés en dessous de la courbe d'accentuation de phrase est 3 sur 694 soit 0.43 %.

Une fois obtenue la ligne de base, nous pouvons estimer l'accentuation locale, qui correspond à des variations prosodiques à court terme.

Accentuation locale

L'approximation de l'accentuation locale s'effectue en deux étapes :

1. La différence entre les valeurs instantanées de fréquence fondamentale et la ligne de base (accentuation de phrase) est appelée « résidu ». Il correspond aux mouvements de fréquence fondamentale sur des évènements locaux (figure 5.5).

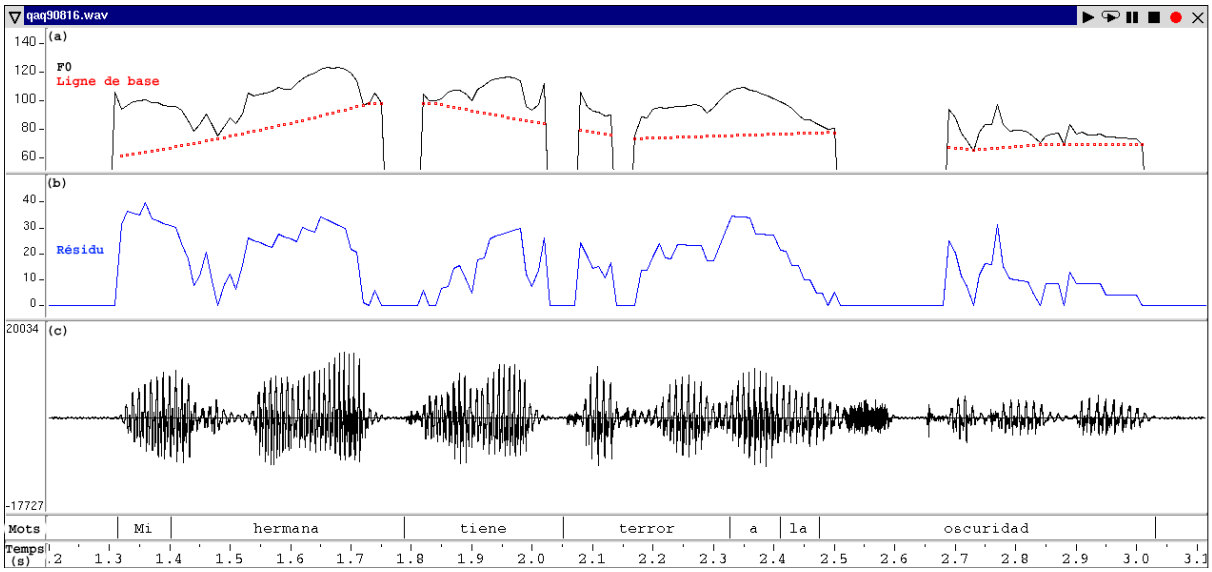


Fig. 5.5 : Première phrase du signal audio spanish/qa/qaq90816 (Multext, homme). Le locuteur dit « *Mi hermana tiene terror a la oscuridad* ». La ligne de transcription correspond à l'étiquetage en mots fourni avec le corpus. Les graphiques correspondent à :
(a) Fréquence fondamentale (ligne) et ligne de base trouvée (points)
(b) Résidu
(c) Représentation du signal

2. Obtention de l'accentuation locale :
L'accentuation locale est liée à des phénomènes prosodiques courts. Une approximation linéaire est calculée sur la partie voisée de chaque pseudo-syllabe. Si la pseudo-syllabe contient plusieurs parties voisées, on ne considère que la dernière. Un exemple est montré sur la figure 5.6.

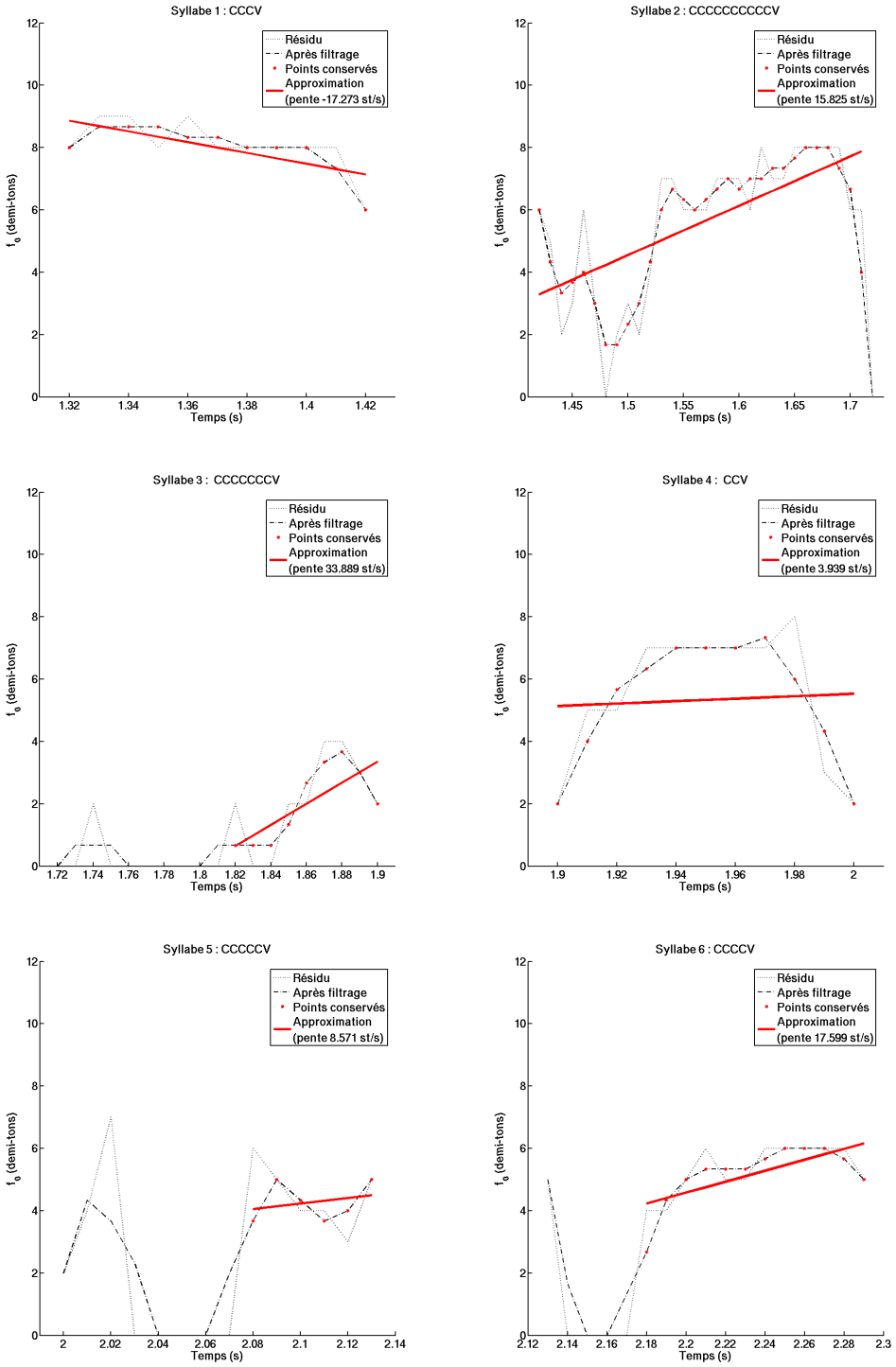


Fig. 5.6 : Six premières pseudo-syllabes de la première phrase du signal audio spanish/qa/qaq90816 (Multext, homme)

3. Reconstruction :
   Une représentation graphique permet de visualiser le résultat de l'approxima-

tion de la fréquence fondamentale. Le graphique 5.7 permet de comparer l'approximation avec la courbe originale. On notera que la huitième pseudo-syllabe ("CCCCCCCCCCCCCV") a une partie voisée considérée trop courte pour faire une approximation.

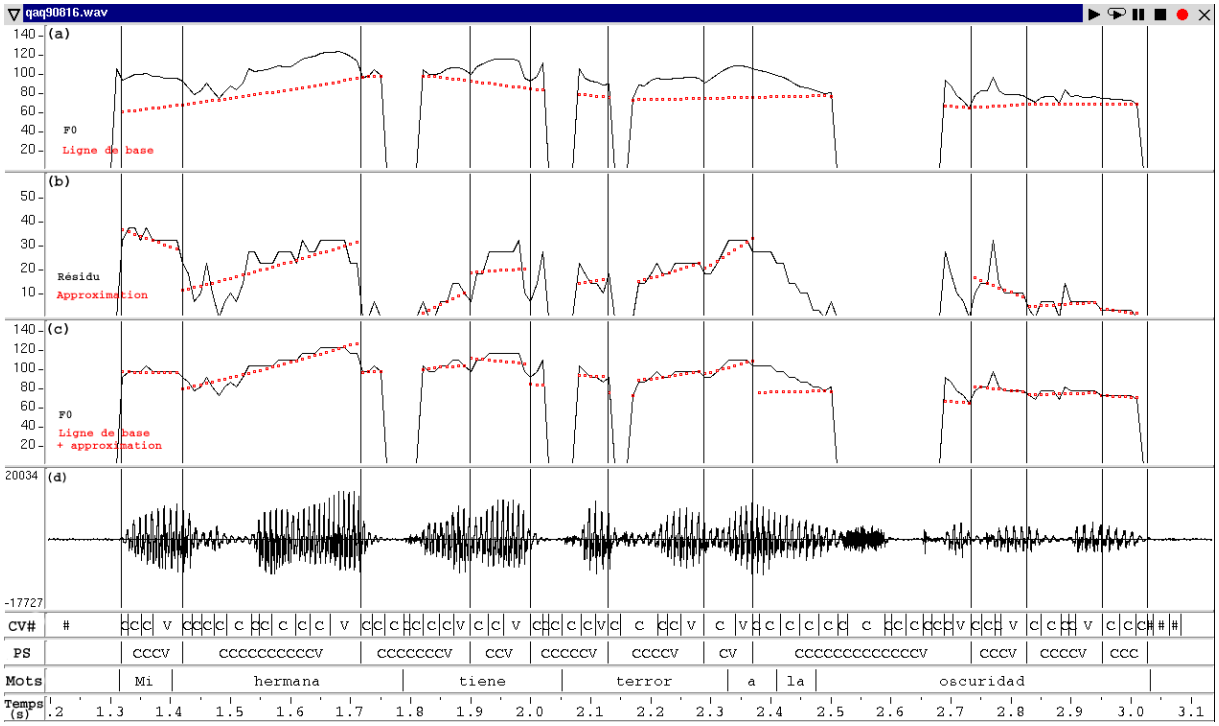


Fig. 5.7 : Première phrase du signal audio spanish/qa/qaq90816 (Multext, homme). Le locuteur dit « *Mi hermana tiene terror a la oscuridad* »
(a) fréquence fondamentale (ligne) et ligne de base trouvée (points)
(b) résidu (ligne) et approximations linéaires sur les pseudo-syllabes (points)
(c) fréquence fondamentale (ligne) et addition des approximations de la ligne de base et du résidu (points)
(d) représentation du signal
Les lignes de transcription correspondent à :
- la segmentation en Consonnes, Voyelles et Silences,
- la segmentation en pseudo-syllabes,
- l'étiquetage en mots fourni avec le corpus.

Codage discret

Les courbes de résidu sont étiquetées selon le sens de la variation de la courbe d'accentuation locale sur chaque pseudo-syllabe. Il y a une étiquette par pseudo-syllabe, « U » (montant), « D » (descendant) ou « # » pour les parties non voisées.

La ligne de base a également été étiquetée suivant le même protocole. Lorsqu'elle est

considérée, il y a alors deux étiquettes par pseudo-syllabe, la première tenant compte des variations mélodiques sur la phrase et la deuxième des variations sur la pseudo-syllabe. Un exemple d'un tel étiquetage est donné sur la figure 5.8.

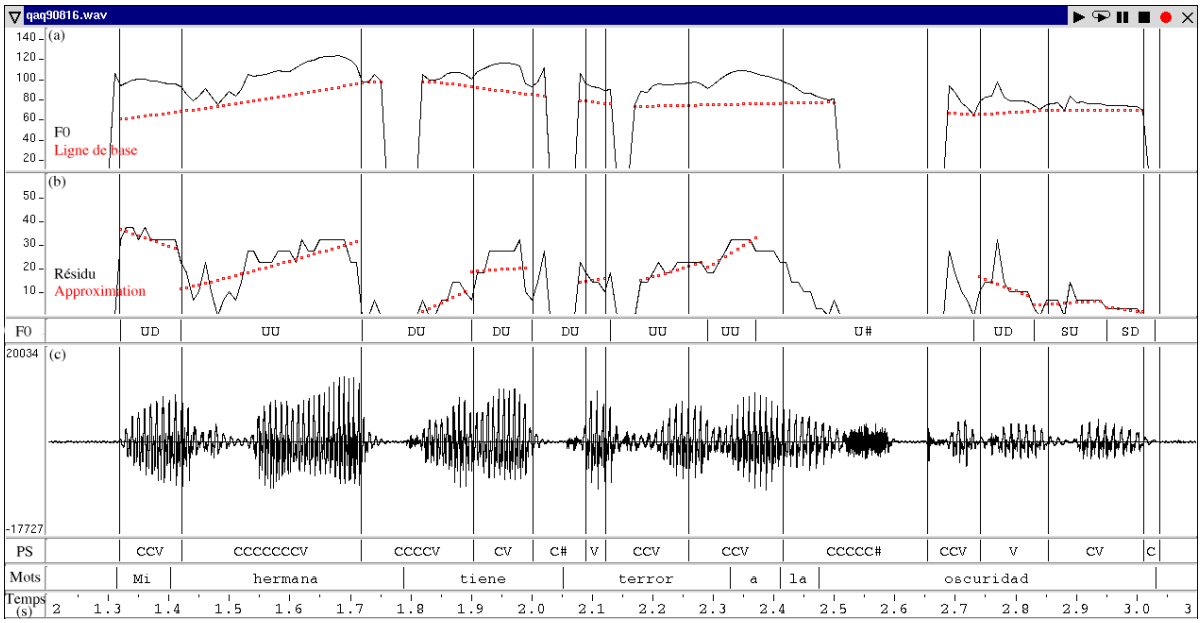


Fig. 5.8 : Exemple d'étiquetage de la fréquence fondamentale sur le signal audio /Spanish/qa/qaq90816.
(a) fréquence fondamentale et ligne de base
(b) résidu et approximation
(c) signal
les lignes de transcription correspondent à :
- l'étiquetage des variations de f0 (ligne de base et résidu),
- l'étiquetage en pseudo-syllabes,
- l'étiquetage en mots

5.3.2 Traitement de la courbe d'énergie

Le traitement de la courbe d'énergie est effectué de manière similaire : une approximation linéaire est effectuée sur chaque pseudo-syllabe. Une étiquette « U », « D » ou « # » est ainsi apposée à chaque pseudo-syllabe. Un exemple est montré sur la figure 5.9.

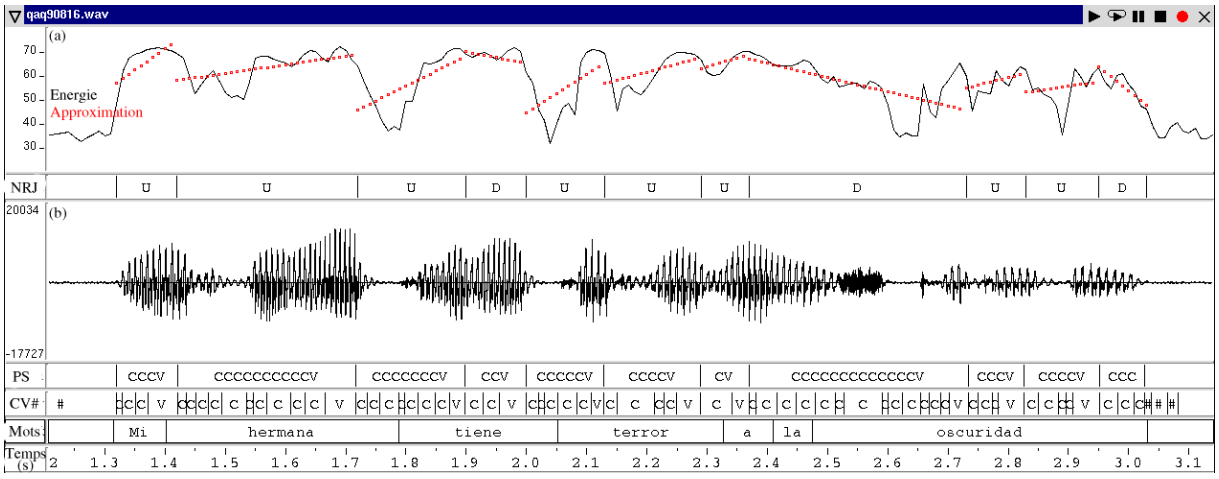


Fig. 5.9 : Exemple d'approximation et d'étiquetage de la courbe d'énergie
(a) Energie et approximation
(b) Signal
Les lignes de transcription correspondent à :
- l'étiquetage de l'énergie
- la segmentation en pseudo-syllabes
- la segmentation en Consonnes/Voyelles
- l'étiquetage en mots.

5.3.3 Identification des langues

Les expériences sont menées avec différents jeux d'étiquettes.
– seules les étiquettes concernant le résidu sont prises en compte,
– les étiquettes du résidu et de la ligne de base sont couplées,
– les étiquettes du résidu, de la ligne de base et de l'énergie sont exploitées.
Les modèles multigrammes sont entraînés avec ces différents jeux d'étiquettes. Les résultats pour chacun de ces modèles sont résumés dans le tableau 5.4.

Tab. 5.4 : Expériences d'identification des langues avec le modèle intonatif

Paramètres étiquetés	Nb d'étiquettes	Modèle	% correct
Ligne de Base (LB)	3	3-multigrammes	31,6 ± 7,7
Résidu	3	3-multigrammes	29,4 ± 7,6
F0 (Résidu + LB)	9	3-multigrammes	36,7 ± 8,0
Énergie	3	3-multigrammes	26,6 ± 7,4
F0 + Énergie	27	3-multigrammes	33,1 ± 7,8

La prise en compte de l'énergie n'améliore pas les performances du système. Le modèle permettant d'obtenir les meilleurs résultats prend en compte les étiquetages de la ligne

de base, du résidu et de l'énergie. La matrice de confusion obtenue pour ce modèle est représentée sur le tableau 5.5.

Tab. 5.5 : Modèle intonatif 3-multigramme
Expériences sur l'ensemble de test : correct : 36,7 ± 8,0 % (51/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	4	2	1	6	2	2	3
Allemand	-	14	-	1	3	2	-
Mandarin	-	3	4	3	3	4	3
Français	2	-	2	9	-	5	1
Italien	4	2	1	5	3	4	1
Espagnol	3	-	-	1	4	9	3
Japonais	2	0	3	3	3	1	8

Avec ce modèle, seul l'allemand est relativement correctement identifié. Nous avons toutefois tenté un regroupement selon les groupes rythmiques.

Tab. 5.6 : Modèle intonatif 3-multigrammes : Expériences sur l'ensemble de test ;
Regroupement en classes rythmiques : correct 54,7 ± 8,3 % (76/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	28	26	6
L. Syllab.	14	40	5
L. Mora.	5	7	8

Avec le tableau 5.6, nous remarquons que les langues syllabiques sont les mieux identifiées.

Cependant, la modélisation proposée ici n'est clairement pas optimale. Dans la section suivante, nous prenons en compte les informations rythmiques et intonatives afin de se rapprocher d'un modèle prosodique plus complet.

5.4 Modélisation de la prosodie (rythme et intonation)

Nous combinons les informations issues de l'étude du rythme et de l'intonation. Les systèmes proposés sont similaires à ceux décrits dans les paragraphes 5.2 et 5.3. Les étiquetages déterminés lors de la modélisation du rythme et de l'intonation sont combinés, la pseudo-syllabe étant conservée comme unité de base.

5.4.1 Ajout d'étiquettes au modèle rythmique

Afin de former les classes de pseudo-syllabes, aux 4 paramètres déterminés pour le modèle du rythme (D_c, D_v, N_c, énergie) ont été rajoutés deux autres paramètres liés à la fréquence fondamentale, à savoir : le signe du coefficient directeur de la régression linéaire sur la courbe de fréquence fondamentale et la valeur absolue de la pente de la même régression.

En créant deux classes par paramètre, il y a au total 36 classes possibles.

En reconnaissance, les expériences d'identification des langues montrent que le taux d'identification correcte est de 51,8 ± 8,3 % (72 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau 5.7).

Tab. 5.7 : Modèle rythmique/intonatif 3-multigrammes
Expériences sur l'ensemble de test : correct : 51,8 ± 8,3 % (72/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	4	1	4	3	4	4	-
Allemand	-	18	1	-	-	-	1
Mandarin	6	2	7	2	-	1	2
Français	-	-	-	17	-	-	2
Italien	5	3	2	2	4	3	1
Espagnol	1	-	-	8	-	7	4
Japonais	-	-	-	2	-	3	15

Les résultats obtenus montrent une amélioration par rapport à ceux obtenus avec les modèles rythmiques ou intonatifs seuls. Malgré le faible taux d'identification global, nous pouvons remarquer que quelques langues sont correctement identifiées : l'allemand, le français et le japonais.

Tab. 5.8 : Modèle rythmique/intonatif 3-multigrammes
Regroupement en classes rythmiques : correct 71,2 ± 8,2 % (79/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	43	14	3
L. Syllab.	11	41	7
L. Mora.	-	5	15

En effectuant des regroupements selon les typologies rythmiques des langues (tableau 5.8), le taux d'identification correcte passe à 71,2 %. Cependant, les confusions sont encore nombreuses.

5.4.2 Ajout d'étiquettes au modèle intonatif

Les durées des pseudo-syllabes sont prises en compte au travers de 4 étiquettes. Les étiquettes adjointes ici sont liées à la durée des groupes consonantiques et à la durée de la voyelle de chaque pseudo-syllabe.

En reprenant les notations précédentes où D_c est la durée des consonnes composant la pseudo-syllabe, Dv est la durée de la voyelle, L est le nombre de langues et P_l est le nombre de pseudo-syllabes par langue étudié dans l'ensemble d'apprentissage, nous introduisons les valeurs moyennes :

$$\begin{aligned} \bar{D_c} &= \frac{1}{L}\sum_{l=1}^{L}\Big[\frac{1}{P_l}\sum_{p=1}^{P_l}\Big(D_c(p,l)\Big)\Big] \\ \bar{D_v} &= \frac{1}{L}\sum_{l=1}^{L}\Big[\frac{1}{P_l}\sum_{p=1}^{P_l}\Big(D_c(p,l)\Big)\Big] \end{aligned} \tag{5.2}$$

Quatre classes sont alors définies :

1. $D_c > \bar{D_c}$ et $D_v > \bar{D_v}$ (les durées à la fois des consonnes et de la voyelle sont importantes),

2. $D_c < \bar{D_c}$ et $D_v < \bar{D_v}$ (les durées à la fois des consonnes et de la voyelle sont faibles),

3. $D_c < \bar{D_c}$ et $D_v > \bar{D_v}$ (les durées des consonnes sont faibles et la durée de la voyelle importante),

4. $D_c > \bar{D_c}$ et $D_v < \bar{D_v}$ (les durées des consonnes sont importantes et la durée de la voyelle faible).

Le tableau 5.9 montre les résultats en fonction des paramètres employés pour l'étiquetage. Les systèmes emploie respectivement :
- les étiquettes de durée définies ci-dessus,
- les étiquettes de la ligne de base,
- les étiquettes du résidu,
- les étiquettes de la ligne de base et de la durée,
- les étiquettes du résidu et de la durée,
- les étiquettes du résidu et de la ligne de base,
- les étiquettes de l'énergie,
- les étiquettes de la ligne de base et de l'énergie,
- les étiquettes de la ligne de base, de la durée et de l'énergie,
- les étiquettes du résidu, de la ligne de base, de l'énergie et de la durée.

Tab. 5.9 : Expériences avec les multigrammes intonation/durée

Paramètre(s)	Nb d'étiquettes	Modèle	% correct
Durée	4	3-multigrammes	23,7 ± 7,1
Ligne de Base (LB)	3	3-multigrammes	31,7 ± 7,8
Résidu	3	3-multigrammes	29,4 ± 7,6
LB + Durée	12	3-multigrammes	25,1 +- 7,2
Résidu + Durée	12	3-multigrammes	28,1 ± 7,5
F0 (Résidu + LB)	9	3-multigrammes	36,7 ± 8,0
Énergie	3	3-multigrammes	26,6 ± 7,4
LB + Énergie	9	3-multigrammes	29,4 +- 7,6
LB + Énergie + Durée	36	3-multigrammes	32,4 +- 7,8
F0 + Énergie + Durée	108	3-multigrammes	41,0 ± 8,2

En reconnaissance, les expériences d'identification des langues montrent que le taux d'identification correcte est de 41,0 ± 8,2 % (57 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau 5.10).

Tab. 5.10 : Modèle intonation/durée 3-multigrammes
Expériences sur l'ensemble de test : Correct : 41,0 ± 8,2 % (57/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	7	-	3	1	4	-	5
Allemand	5	12	2	-	1	-	-
Mandarin	3	3	6	-	5	-	3
Français	1	-	-	12	4	-	2
Italien	5	-	2	4	3	2	2
Espagnol	6	-	3	1	3	4	3
Japonais	3	-	2	2	4	2	7

Les résultats montrent que les seules langues relativement correctement identifiées sont l'allemand et le français. Ce modèle souffre de l'inadéquation entre le nombre d'étiquettes possibles (108 !) et la taille du corpus.

Le tableau 5.11 montre les résultats pour les regroupements selon les typologies rythmiques.

Tab. 5.11 : Modèle intonation/durée 3-multigrammes
Regroupement en classes rythmiques : correct 58,3 ± 8,2 % (81/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	41	11	8
L. Syllab.	17	33	7
L. Mora.	5	8	7

Les langues accentuelles sont les mieux identifiées. Cependant, les résultats sont inférieurs à ceux obtenus par le modèle précédent (tableau 5.8).

5.4.3 Conclusion

Les différentes expériences menées au cours de ce chapitre n'ont pas donné de résultats très satisfaisants. La prise en compte d'enchaînements d'unités telles que les pseudo-syllabes n'est peut être pas adéquate. Cependant, nous restons persuadés que la modélisation d'enchaînements d'évènements prosodiques est une source d'information non négligeable.

Le système d'Adami, décrit au paragraphe 3.3.5, est fondé sur le même principe de modélisation des enchaînements et permet a priori d'obtenir des performances correctes sur le corpus CALLFRIEND. C'est pour cette raison que nous avons décidé d'implémenter ce système afin de tester ses performances sur nos données.

5.5 Expériences sur le système d'Adami

À la suite d'une segmentation du signal (ici effectuée à partir des courbes de fréquence fondamentale et d'énergie), des étiquettes correspondant au sens des mouvements de fréquence fondamentale et d'énergie sont associées à chaque segment.

Un exemple de segmentation et d'étiquetage des mouvements prosodiques est donné sur la figure 5.10.

Les expériences décrites ici ont pour but d'évaluer le système sur la base de données MULTEXT afin de pouvoir le comparer au nôtre. Il y a cependant une différence entre le système original tel que décrit dans [2] et celui que nous avons implémenté pour ces expériences : les modèles utilisés pour les séquences d'étiquettes ne sont pas des modèles n-grammes mais des n-multigrammes (§5.1.2).

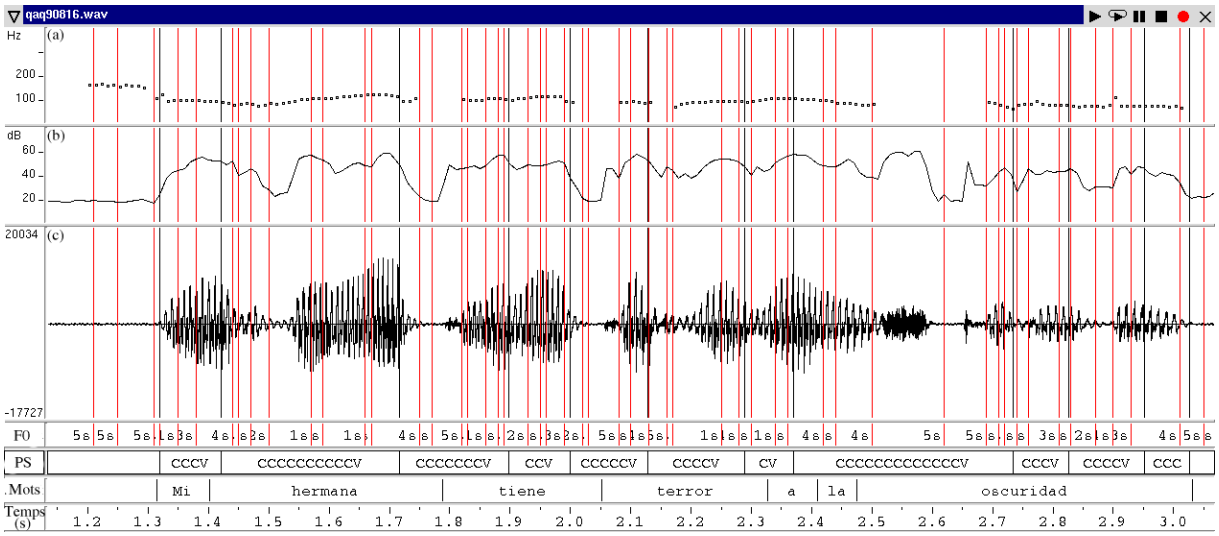


Fig. 5.10 : Exemple d'étiquetage des mouvements prosodiques (fichier spanish/qa/qaq90816)
(a) fréquence fondamentale
(b) énergie
(c) signal
Les lignes de transcription correspondent à :
- l'étiquetage des mouvements prosodiques,
- l'étiquetage en pseudo-syllabes,
- l'étiquetage en mots.

Les expériences sont effectuées en faisant varier le nombre de paramètres prosodiques pris en compte, ce qui influe également sur la segmentation.

Tab. 5.12 : Expériences avec le modèle d'Adami

Paramètre(s)	Nb d'étiquettes	Modèles	% correct
F0	3	3-multigrammes	33,1 ± 7,8
F0 + Durée	6	3-multigrammes	48,2 ± 8,3
Énergie	3	3-multigrammes	44,6 ± 8,3
Énergie + Durée	6	3-multigrammes	48,2 ± 8,3
F0 + Énergie	5	3-multigrammes	55,4 ± 8,3
F0 + Énergie + Durée	10	3-multigrammes	64,7 ± 8,0
		4-multigrammes	65,5 ± 8,0
		5-multigrammes	69,7 ± 7,7

Le meilleur résultat est obtenu lorsque l'ensemble des paramètres est pris en compte. Pour ces paramètres, la segmentation donne des segments voisés de 68,4 ms en moyenne et

des segments non voisés de 103.2 ms en moyenne. La matrice de confusion correspondante est donnée sur le tableau 5.13.

Tab. 5.13 : Modèle d'Adami 5-multigrammes
Expériences sur l'ensemble de test : correct : 69,8 ± 7,7 % (97/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	11	1	4	2	1	-	1
Allemand	2	17	1	-	-	-	-
Mandarin	1	1	18	-	-	-	-
Français	1	-	-	14	4	-	-
Italien	-	2	1	2	8	3	4
Espagnol	1	-	1	1	5	9	3
Japonais	-	-	-	-	-	-	20

Les langues accentuelles sont toutes correctement reconnues, seul l'anglais est confondu avec des langues appartenant aux autres groupes rythmiques. L'italien et l'espagnol sont les langues les moins bien reconnues. Sur ces données, ce système ne commet aucune erreur d'identification pour le japonais.

Le tableau 5.14 montre les résultats en fonction des groupes de langues.

Tab. 5.14 : Modèle d'Adami 5-multigramme : Expériences sur l'ensemble de test
Regroupement en classes rythmiques : correct 87,8 ± 5,5 % (122/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	56	3	1
L. Syllab.	6	46	7
L. Mora.	-	-	20

Les performances de ce système sont supérieures à celles des systèmes présentés ci-dessus.

L'unité de base est la principale différence entre cette approche et la nôtre. D'après les résultats, nous pouvons supposer que la pseudo-syllabe n'est pas nécessairement une unité adéquate pour la modélisation dynamique de la prosodie dans le cadre de l'identification des langues.

La durée de l'unité est la principale différence entre la méthode d'Adami et la nôtre. Un segment voisé issu de sa segmentation dure en moyenne 68,4 ms, alors que la durée moyenne d'une pseudo-syllabe est de 187 ms. La taille de l'unité d'Adami est de l'ordre de celle d'un segment « V » (55.0 ms en moyenne) et non d'une pseudo-syllabe entière.

5.6 Expériences en modélisation de la prosodie à court terme

De même que pour les pseudo-syllabes, des étiquettes correspondant à la direction des mouvements de fréquence fondamentale et de l'énergie sont apposées sur chaque segment.

Le prétraitement de la fréquence fondamentale est le même que précédemment (§5.3.1), avec l'approximation et l'étiquetage de l'accentuation de phrase et de l'accentuation locale (figure 5.11).

Les étiquettes de durée sont fonction de la nature du segment. Nous avons mesuré les durées moyennes sur l'ensemble du corpus d'apprentissage :

– pour les consonnes, la durée moyenne d'un segment est de 29,8 ms,
– pour les voyelles, la durée moyenne d'un segment est de 55,0 ms,
– pour les silences, la durée moyenne d'un segment est de 161,3 ms.

Chaque segment est alors qualifié de « court » ou « long » s'il dépasse ou non la durée moyenne des segments de même nature.

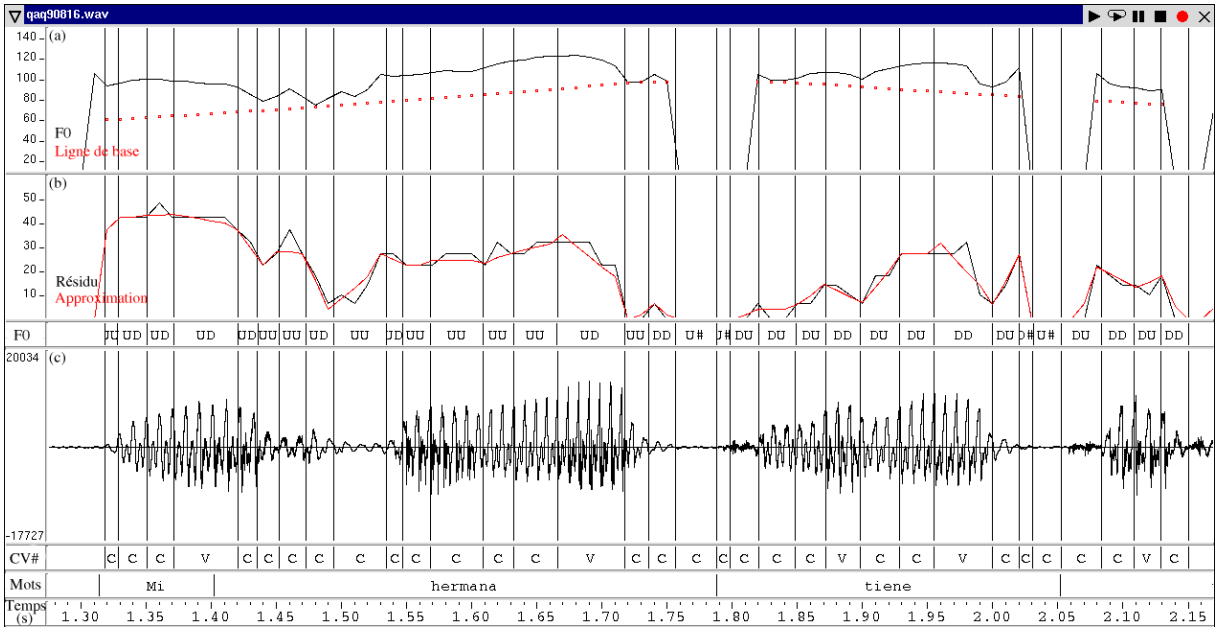


Fig. 5.11 : Exemple d'étiquetage de la fréquence fondamentale sur le fichier /Spanish/qa/qaq90816.
(a) fréquence fondamentale et ligne de base
(b) résidu et approximation
(c) signal
les lignes de transcription correspondent à :
- l'étiquetage des variations de f0,
- l'étiquetage en pseudo-syllabes,
- l'étiquetage en mots

De même que pour le système d'Adami, différentes expériences sont réalisées avec les différents paramètres. Ces expériences sont résumées dans le tableau 5.15.

Tab. 5.15 : Expériences avec le système segmental durée/intonation

Paramètre(s)	Nb d'étiquettes	Modèle	% correct
Durée	2	3-multigrammes	34,5 ± 7,9
Ligne de Base	3	3-multigrammes	38,1 ± 8,1
Résidu	3	3-multigrammes	41,0 ± 8,2
F0 (Résidu + LB)	9	3-multigrammes	49,6 ± 8,3
Ligne de base + Durée	6	3-multigrammes	46,0 ± 8,3
Résidu + Durée	6	3-multigrammes	48,9 ± 8,3
F0 + Durée	18	3-multigrammes	51,1 ± 8,3
Énergie	3	3-multigrammes	41,0 ± 8,2
Ligne de base + Energie + Durée	18	3-multigrammes	53,9 ± 8,3
Résidu + Énergie + Durée	18	3-multigrammes	63,3 ± 8,0
F0 + Énergie + Durée	54	3-multigrammes	61,1 ± 8,1

Le meilleur résultat est obtenu avec les étiquettes calculées à partir des paramètres sur le résidu, l'énergie et la durée des segments. La matrice de confusion correspondante est représentée sur le tableau 5.16.

Tab. 5.16 : Modèle 3-multigrammes
Expériences sur l'ensemble de test : correct : 63,3 ± 8,0 % (88/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	6	-	5	1	5	2	1
Allemand	1	18	1	-	-	-	-
Mandarin	4	2	12	-	2	-	-
Français	-	-	-	16	1	2	-
Italien	1	-	1	2	13	2	1
Espagnol	2	-	-	1	7	9	1
Japonais	-	-	1	-	5	-	14

Les confusions se font principalement à l'intérieur des groupes rythmiques - le mandarin pour les langues accentuelles et l'espagnol pour les langues syllabiques.

Cependant, lorsque l'on effectue les regroupements en fonction des groupes rythmiques linguistiques, le taux d'identification correcte devient 83,4 %.

Tab. 5.17 : Modèle 3-multigrammes : Expériences sur l'ensemble de test
Regroupement en classes rythmiques : correct 83,4 ± 6,1 % (128/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	49	10	1
L. Syllab.	4	53	2
L. Mora.	1	5	14

Les résultats obtenus en regroupant les langues selon les typologies rythmiques sont proches de ceux obtenus avec le système d'Adami. Le système que nous avons développé est légèrement plus efficace sur l'identification des langues syllabiques, mais un peu moins pour les langues accentuelles et moraïques.

5.7 Fusion : modèle dynamique et modèle statique

Une fusion pondérée est opérée avec le meilleur système dynamique (modèle d'Adami) et le meilleur système statique (modélisation des durées des pseudo-syllabes, §4.5.3). La fusion est comme au paragraphe §4.5.5 une addition pondérée des log-vraisemblances, pour laquelle nous avons estimé les poids optimaux sur l'ensemble d'apprentissage. Les poids obtenus sont :

– 0,2 pour le module « statique »
– 0,8 pour le module « dynamique »

La matrice de confusion est représentée sur le tableau 5.18.

Tab. 5.18 : Fusion des approches prosodiques statique et dynamique
Expériences sur l'ensemble de test : correct : 75,5 ± 7,2 % (105/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	12	-	1	1	2	-	4
Allemand	2	17	-	-	-	-	1
Mandarin	-	-	20	-	-	-	-
Français	-	-	-	17	2	-	-
Italien	-	1	3	1	11	2	2
Espagnol	1	-	2	1	5	11	-
Japonais	1	2	-	-	-	-	17

Le modèle statique obtient 66,9 % d'identifications correctes, et le modèle d'Adami 69,7 %. La fusion permet d'améliorer les résultats, le taux d'identifications correctes est de 75,5 %.

Tab. 5.19 : Fusion des approches prosodiques statique et dynamique : Expériences sur l'ensemble de test
Regroupement en classes rythmiques : correct 85,6 ± 5,2 % (119/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	52	2	5
L. Syllab.	7	50	2
L. Mora.	3	-	17

Lorsque l'on considère les regroupements en fonction des classes rythmiques, la fusion n'améliore pas les résultats (tableau 5.19). Le taux d'identifications correctes est le même que lorsque l'on ne considère que le modèle statique (85,6 %), alors que le modèle d'Adami obtenait 87,8 %.

5.8 Fusion : modèle accentuel et modèle intonatif

Nous fusionnons deux systèmes permettant de prendre en compte d'une part les mouvements prosodiques à l'échelle de l'accent et d'autre part les mouvements à l'échelle de la phrase. La fusion est opérée entre le système complet d'Adami (avec les étiquettes de fréquence fondamentale, d'énergie et de durée) et un système n'utilisant que les étiquettes de la ligne de base (§5.6, 38,1 % d'identifications correctes).

La fusion est une addition pondérée des log-vraisemblances (comme au §4.5.5), pour laquelle nous avons estimé les poids optimaux sur l'ensemble d'apprentissage. Les poids obtenus sont :

– 0,1 pour le module « ligne de base »,
– 0,9 pour le module d'Adami.

La matrice de confusion est représentée sur le tableau 5.20.

Tab. 5.20 : Fusion des approches prosodiques à court terme et à long terme
Expériences sur l'ensemble de test : correct : 71,2 ± 7,5 % (99/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	12	-	4	2	1	-	1
Allemand	2	17	1	-	-	-	-
Mandarin	1	1	18	-	-	-	-
Français	1	-	-	14	4	-	-
Italien	-	1	1	2	10	3	3
Espagnol	1	-	1	1	6	9	2
Japonais	-	-	-	-	1	-	19

La fusion entre ces deux modèles n'améliore pas les résultats en identification des langues.

Tab. 5.21 : Fusion des approches prosodiques à court terme et à long terme : Expériences sur l'ensemble de test
Regroupement en classes rythmiques : correct 89,1 ± 5,0 % (124/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	56	3	1
L. Syllab.	5	49	5
L. Mora.	-	1	19

En considérant les regroupements rythmiques, les résultats sont meilleurs (tableau 5.21). La prise en compte de la ligne de base, c'est-à-dire de la prosodie globale de la phrase, permet d'améliorer la reconnaissance des langues syllabiques.

5.9 Comparaison avec d'autres systèmes d'identification

Les recherches menées à l'IRIT sur l'identification des langues ne concernent pas que la prosodie (voir §3). Nous avons démontré au début de ce document qu'il est crucial de prendre en compte le maximum de sources d'information disponibles. Pour cela, nous avons continué le travail inauguré par François Pellegrino et continué par Jérôme Farinas concernant les modèles acoustiques et phonotactiques.

Le système acoustique développé à l'IRIT s'appuie sur une différenciation des espaces acoustiques. Une modélisation des espaces vocaliques a été présentée dans [98]. Nos recherches se font dans la continuité de cette approche, en essayant de modéliser les espaces consonantiques.

Le système phonotactique développé est une approche « classique » de type PPRLM (§ 2.3), utilisant 6 décodeurs acoustico-phonétiques.

5.9.1 Méthode acoustique

La modélisation acoustique proposée est fondée sur l'extraction de paramètres cepstraux (MFCC). Ces paramètres sont extraits soit de manière centiseconde (toutes les 10 ms), soit de manière segmentale, à partir de la segmentation automatique décrite plus haut (§4.2.1). Dans ce dernier cas, une seule observation est conservée pour chaque segment.

Grâce à l'algorithme de détection automatique des voyelles (§4.2.2) et à une mesure de voisement, sept modèles sont proposés prenant en compte différents espaces acoustiques :

1. modélisation de l'ensemble des sons sans distinction de classe,
2. modélisation des voyelles,

3. modélisation des consonnes,
4. modélisation des consonnes voisées,
5. modélisation des consonnes non voisées,
6. fusion des modélisations 2 et 3,
7. fusion des modélisations 2, 4 et 5.

L'architecture d'un système acoustique d'identification des langues avec modélisation différenciée des consonnes et des voyelles est présentée sur la figure 5.12.

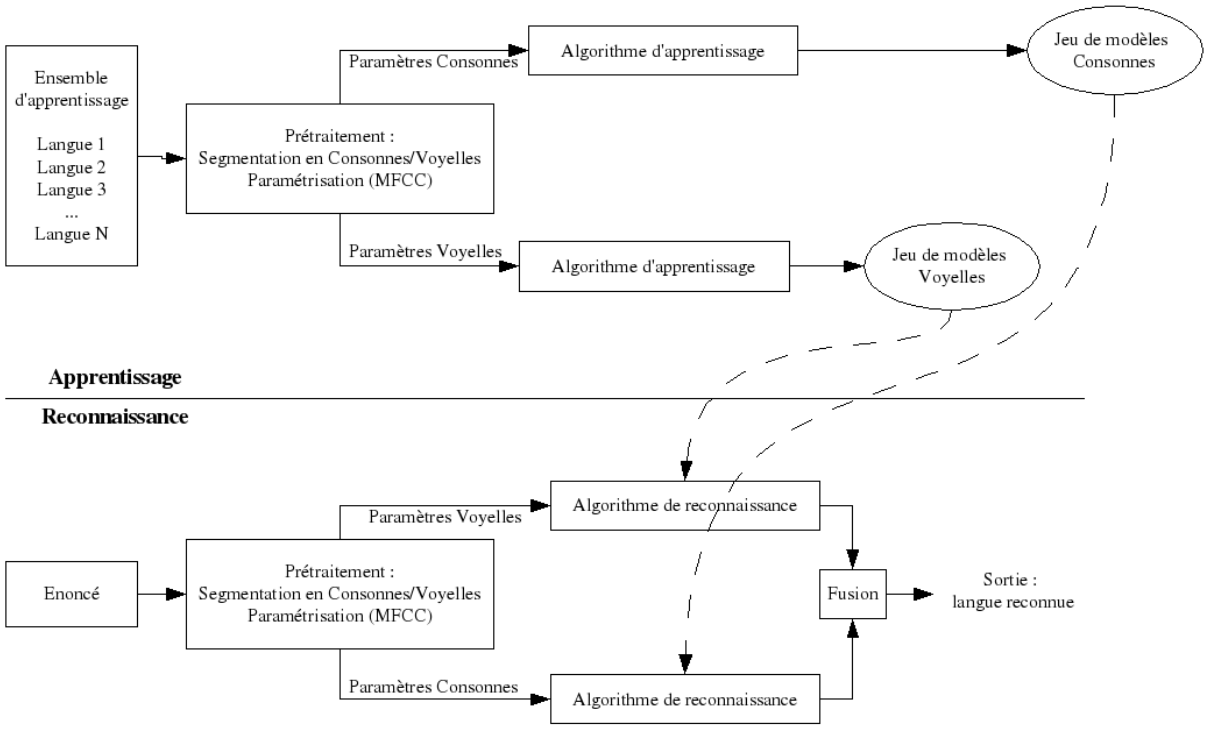


Fig. 5.12 : Système d'identification des langues par modélisation différenciée des consonnes et des voyelles

Le protocole expérimental est le même que celui employé lors des différentes modélisations de la prosodie (§4.3.2).

Les résultats obtenus avec les différents modèles sont résumés ci-dessous. Pour plus de détails, se reporter à l'annexe D.

Modèles centiseconde

Les paramètres MFCC sont extraits toutes les 10 ms, sur des fenêtres de 20 ms avec des recouvrements de 10 ms. 12 MFCC sont calculés sur chaque trame, avec leurs dérivées premières et secondes. Pour chaque observation, nous obtenons un vecteur de dimension 36.

Espaces acoustiques considérés	% correct
Espace global	76,2 ± 7,1 %
Espace vocalique	49,6 ± 8,3 %
Espace consonantique	66,2 ± 7,9 %
Espaces consonantique + vocalique	54,7 ± 10,3 %
Espaces consonantique voisé + consonantique non voisé + vocalique	59,7 ± 8,2 %

Il est anormal d'avoir des résultats aussi faible avec ce système. Nous pensons que le problème peut venir de la paramétrisation.

Modèles segmentaux

Espaces acoustiques considérés	% correct
Espace global	87,8 ± 5,5 %
Espace vocalique	69,8 ± 7,6 %
Espace consonantique	99,3 ± 1,4 %
Espaces consonantique + vocalique	99,3 ± 1,41 %
Espaces consonantique voisé + consonantique non voisé + vocalique	98,5 ± 2,0 %

5.9.2 Méthode phonotactique (PPRLM)

Le système phonotactique d'identification des langues est un système classique de type PPRLM (Parallel Phone Recognition followed by Language Modeling, figure 2.3, §2.3).

Ce système possède 6 décodeurs acoustico-phonétiques, il est similaire au système employé par l'IRIT pour la campagne NIST (voir §2.5.7).

Le taux d'identification correcte est de 85,6 % (tableau 5.22).

Tab. 5.22 : Modèle PPRLM
Expériences sur l'ensemble de test (correct : 85,6 %)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	17	2	-	-	-	-	1
Allemand	-	20	-	-	-	-	-
Mandarin	-	4	16	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	19	-	1
Espagnol	-	-	-	1	10	9	-
Japonais	-	-	-	-	1	-	19

Le modèle phonotactique permet d'obtenir de bons résultats. Cependant, il ne fournit pas d'aussi bonnes performances que les meilleurs modèles acoustiques.

Tab. 5.23 : Modèle PPRLM : Expériences sur l'ensemble de test
Regroupement en classes rythmiques : correct 97,8 ± 2,4 % (136/139)

	L. Accent.	L. Syllab.	L. Mora.
L. Accent.	59	-	1
L. Syllab.	-	58	1
L. Mora.	-	1	19

5.9.3 Conclusion

L'approche la plus efficace est l'approche acoustique. Ces résultats sont cohérents avec les conclusions de la campagne d'évalution NIST (§2.5). Nous remarquons également que l'approche phonotactique permet d'obtenir de bonnes performances. Sur le corpus MULTEXT, ces deux approches fournissent des résultats supérieurs à ceux obtenus avec la modélisation de la prosodie.

5.10 Conclusion

L'unité pseudo-syllabe, présentée dans le chapitre précédent, semblait prometteuse lors des expériences préliminaires. Cependant, la modélisation des enchaînements de ces unités ne s'est pas révélée aussi efficace que ce que nous attendions. Les expériences menées avec le système d'Adami nous ont amené à considérer des enchaînements d'unités plus courtes (infra-phonémiques).

De plus, les mouvements de F_0 ou d'énergie soit monotones sur les unités d'Adami alors que nos segments « C » ou « V » ne sont pas définies prosodiquement même si

elles sont également infra-phonémiques. Comme les mouvements prosodiques ne sont pas monotones sur nos unités, l'étiquetage provoque une perte d'information.

Les différences entre les langues ne seraient donc pas liées à des évènements macro-prosodiques comme nous l'espérions, mais à des événements micro-prosodiques. Le fait que les séquences les plus discriminantes soient des séquences de 3 ou 5 segments montre que ce sont les mouvements de fréquence fondamentale, d'énergie et les variations de durée au cours d'une syllabe qui sont caractéristiques.

Les regroupements que nous avons pu effectuer en fonction des groupes rythmiques linguistiques (langues accentuelles, langues syllabiques et langues moraïques) montrent que la modélisation de la prosodie permet de différencier des ensembles de langues partageant les mêmes caractéristiques.

La technique de classification des durées « courtes » ou « longues » n'est clairement pas optimale. Une amélioration pourra être envisagée à ce niveau.

Ces expériences exposent le problème de relations entre les unités phonémiques et prosodiques pour la définition d'unités accentuelles.

Tab. 5.24 : Récapitulatif des expériences du chapitre

Paramètres	% correct	% correct groupes
Durée	42,4	69,0
Intonation	36,7	54,7
Durée/intonation	51,8	71,2
Intonation/Durée	41,0	58,3
Adami	69,8	87,8
Segments	63,3	83,4
Fusion (pseudo-syllabe rythme + Adami)	75,5	85,6
Fusion (Ligne de base + Adami)	71,2	89,1
Acoustique	99,3	100
Phonotactique	85,6	97,8

Les meilleurs résultats sont obtenus avec les approches « classiques », particulièrement l'approche acoustique.

Le problème de fusion, avec la qualification de l'apport du modèle prosodique, ne peut cependant pas être évoqué ici au vu des trop bonnes performances des autres méthodes sur nos données. Il serait intéressant de tester l'ensemble de nos systèmes sur des corpus plus « difficiles », par exemple en considérant de la parole téléphonique spontanée.

Chapitre 6

Quelques pistes pour la parole spontanée

Sommaire

6.1 Expériences de discrimination des langues sur OGI 146
6.1.1 Corpus 146
6.1.2 Expériences 146
6.1.3 Conclusion 148
6.2 Mesure du débit 150
6.2.1 Définitions et méthodes d'estimation 150
6.2.2 Analyse des données 151
6.2.3 Evaluation des algorithmes comme estimateurs du débit 154
6.2.4 Conclusion et Perspectives 155
6.3 Conclusion 156

Afin d'étudier la robustesse des théories et des modèles précédemment présentés, nous avons appliqué le système mis au point sur le corpus Multext à un corpus de parole spontanée téléphonique, le corpus Ogi Mlts [93]. Les premières expériences menées dans l'optique de faire une distinction entre deux langues ont donné dans l'ensemble de bons résultats, mais avec beaucoup de variabilité selon les langues [117]. Dès que l'on considère un nombre de langues plus important, les taux de réussite diminuent fortement. Nous supposons que ces résultats sont dûs principalement à la variabilité plus importante de la parole spontanée. Afin de tenir compte de ces phénomènes, nous avons mis au point une méthode automatique de mesure du débit de parole que nous avons évalué sur la parole spontanée. Ces expériences ouvrent des perspectives certaines aux approches proposées.

6.1 Expériences de discrimination des langues sur OGI

6.1.1 Corpus

Le corpus OGI MLTS (Multilingual Telephone Speech Corpus) [93] fait figure de référence en termes de base de données pour les applications en identification automatique des langues. Il a été utilisé lors des campagnes d'évaluation NIST en identification des langues jusqu'en 1996. Il s'agit de parole téléphonique échantillonnée à 8 kHz, enregistrée le plus souvent dans des conditions bruitées.

Un résumé du protocole expérimental et quelques statistiques sur ce corpus sont donnés dans l'annexe G. Cette base de données est intéressante puisqu'elle permet de se rapprocher des conditions réelles d'utilisation d'un système d'identification des langues, car elle contient notamment des énoncés de parole spontanée.

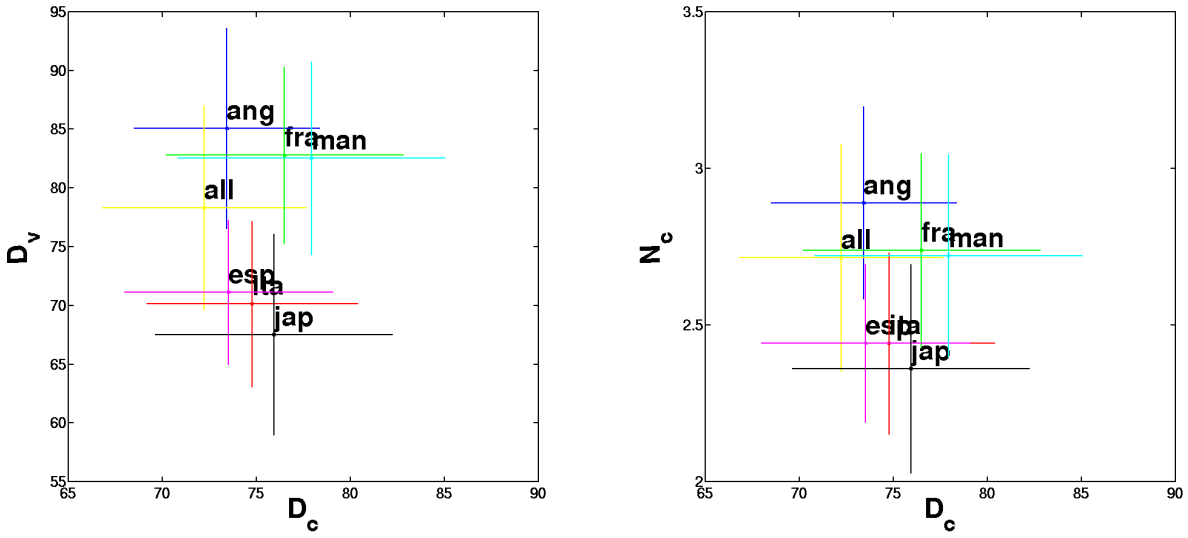
Cependant, certains aspects peuvent se révéler gênants : les locuteurs de français ont parfois un fort accent québécois. Il s'agirait donc plus d'un corpus francophone que français, ce qui est handicapant lorsque l'on considère l'étude de la prosodie. Comme il est vraisemblable que cet aspect se retrouve pour les autres langues du corpus (notamment en anglais et en espagnol), les résultats obtenus avec notre système prosodique sont à prendre avec précaution.

Pour les expériences suivantes, nous avons considéré 10 langues extraites de ce corpus. Ces langues sont l'anglais, le farsi, le français, l'allemand, le japonais, le coréen, le mandarin, l'espagnol, le tamoul et le vietnamien. Les tests sont effectués en utilisant les fichiers de 45 secondes de parole spontanée.

6.1.2 Expériences

Un premier essai a consisté à appliques directement le système prosodique dynamique utilisant les étiquettes de fréquence fondamentale, d'énergie et de durée (décrit à la fin du chapitre précédent). Cependant, les résultats d'identification sur les 10 langues du corpus OGI sont loin d'être satisfaisant. En effet, le taux d'identification correcte moyen est de l'ordre de 20%.

Les projections des paramètres D_c, D_v et N_c sur la figure 6.1 illustrent la difficulté de la tâche. Il est probable que les variations de débit entre les locuteurs d'une même langue influent sur les valeurs des paramètres de manière importante. Une illustration de tels phénomènes sur les paramètres de Ramus est donnée par Dellwo dans [31].



(a) (b)

Fig. 6.1 : Paramètres extraits des pseudo-syllabes
(a) Paramètres (D_c, D_v),
(b) Paramètres (D_c, N_c).

Nous avons décidé de ne considérer que la tâche de discrimination de paires de langues, qui permet d'évaluer les capacités du système à distinguer deux langues, qu'elles appartiennent à la même famille ou non.

Les expériences sont menées avec notre système prosodique « statique » donnant les meilleurs résultats. Il s'agit de la fusion des systèmes employant les paramètres caractérisant les durées et l'intonation des pseudo-syllabes. Chaque jeu de paramètres (durée, intonation) est modélisé par des modèles de mélange de lois gaussiennes (§4.5.5).

Le protocole expérimental employé est exactement le même que pour les expériences sur le corpus MULTEXT (§4.3.2), à la différence que le système ne doit plus trouver une langue parmi sept, mais distinguer une langue d'une autre.

Le corpus OGI est divisé en trois ensembles : un ensemble d'apprentissage, un ensemble de développement et un ensemble de test. Les modèles MMG sont entraînés à partir des données de parole spontanée de l'ensemble d'apprentissage. Les tests sont effectués sur l'ensemble de test, composé d'enregistrements de 45 secondes de parole spontanée. Les résultats sont donnés dans le tableau 6.1.

Sur ces données spontanées, la discrimination est plus aisée pour les langues qui n'appartiennent pas à la même famille rythmique et intonative.

– L'anglais et l'allemand, langues accentuelles qui n'utilisent pas l'intonation comme marqueur lexical, sont bien discriminés contre le japonais (langue moraïque), le mandarin et le vietnamien (qui utilisent l'intonation comme marqueur lexical), ainsi

que contre le coréen, le tamoul et le farsi. Mais discerner l'anglais de l'allemand est difficile.
– De la même façon, nous pouvons observer que le mandarin est bien discerné des autres langues, à part le japonais et le vietnamien.

Nous comparons nos résultats à ceux obtenus par Cummins dans [25], où il est également question de discrimination de paires de langues sur le même corpus. Comme nous l'avons décrit précédemment (§3.3.3), les paramètres utilisés sont différents des nôtres : il s'agit de la modulation de l'enveloppe d'amplitude et d'un paramètre lié à la fréquence fondamentale. Nos résultats semblent meilleurs que ceux obtenus par Cummins et al. Nous pouvons supposer que nos paramètres nous permettent de prendre en compte plus d'informations, plus spécialement lorsqu'il s'agit de distinguer des langues appartenant à des classes rythmiques et intonatives différentes (comme la distinction entre l'anglais et le japonais). Toutefois, notre méthode est relativement moins efficace lorsqu'il s'agit de distinguer deux langues appartenant à la même classe intonative.

6.1.3 Conclusion

Il est très difficile de transposer directement notre approche, mise au point sur de la parole lue, sur de la parole spontanée. La principale raison de ces performances médiocres peut être la variabilité intrinsèque des données due au style de parole. Cette variabilité peut être liée à la grande variation du débit d'élocution rencontrée en parole spontanée. Afin d'envisager une prise en compte dans nos modèles de ce facteur, nous proposons une méthode de mesure automatique du débit.

Tab. 6.1 : Taux d'identifications correctes dans la tâche de discrimination entre paires de langues sur 10 langues du corpus OGI-MLTS

	Anglais	Allemand	Français	Espagnol	Mandarin	Vietnamien	Japonais	Coréen	Tamoul	Farsi
Anglais	-	59.5	51.5	67.7	**75.0**	67.7	67.6	**79.4**	**77.4**	**76.3**
Allemand	59.5	-	55.9	59.4	62.2	65.7	65.8	**71.4**	69.7	**71.8**
Français	51.5	55.9	-	64.3	60.6	58.1	55.9	54.8	60.1	68.6
Espagnol	67.7	59.4	64.3	-	**80.6**	62.1	62.5	**75.9**	65.4	66.7
Mandarin	**75.0**	62.2	60.6	**80.6**	-	50.0	50.0	**73.5**	**74.2**	**76.3**
Vietnamien	67.7	65.7	58.1	62.1	50.0	-	68.6	56.2	**71.4**	66.7
Japonais	67.6	65.8	55.9	62.5	54.1	68.6	-	65.7	59.4	66.7
Coréen	**79.4**	**71.4**	54.8	**75.9**	**73.5**	56.2	65.7	-	62.1	**75.0**
Tamoul	**77.4**	69.7	60.1	65.4	**74.2**	**71.4**	59.4	62.1	-	69.7
Farsi	**76.3**	**71.8**	68.6	66.7	**76.3**	66.7	66.7	**75.0**	69.7	-

6.2 Mesure du débit

6.2.1 Définitions et méthodes d'estimation

Définition(s) du débit de parole

La notion de débit de parole (DP) est liée à la notion de rythme et génère des problèmes de définition similaires, puisqu'ils font tous les deux intervenir le comptage d'unités par seconde. Le choix de l'unité demeure crucial : syllabes et phonèmes en constituent les meilleurs candidats. Pfitzinger a montré [105] que le débit de parole perçu est plus corrélé au débit syllabique qu'au débit phonétique (respectivement $R = 0.81$ contre $R = 0.73$). Dans une perspective de typologie rythmique ou d'identification des langues, il semble évident que les DP calculés en termes de phonèmes par seconde ou de syllabes par seconde apportent des informations complémentaires sur la structure rythmique et l'organisation phonotactique des langues. Par ailleurs, des expériences ont montré que les DP calculés en termes de syllabes ou de phonèmes sont corrélés (pour l'allemand : $R = 0.6$ [105]), tout du moins pour un débit de parole voisin de la norme. Le niveau de corrélation est probablement plus élevé pour les langues ayant une structure syllabique simple en CV que pour les langues qui autorisent une plus grande complexité syllabique en termes de nombre de segments consonantiques consécutifs. À des débits de parole élevés, des stratégies de dépendance par rapport à la langue peuvent aussi intervenir (voir [31] pour une étude de l'impact du débit sur l'organisation temporelle de la parole en termes de quantité de voyelles et de variance de durées des segments consonantiques).

Par conséquent, les débits observés résultent des intéractions entre les facteurs dépendants des locuteurs et/ou dépendants des langues. De même que Ramus [108], nous considérons que l'étude de grands corpus va conduire à une meilleure compréhension de la contribution respective de chaque facteur. Nous proposons d'étudier les DP en termes de phonèmes et de syllabes par seconde dans une perspective multilingue et d'évaluer notre algorithme de segmentation automatique et de détection automatique de voyelles comme estimateur des DP.

Estimation du débit de parole

L'algorithme de segmentation et de détection de voyelles utilisé est celui décrit précedemment (voir section §4.2). Il est basé sur une segmentation statistique combinée à une analyse spectrale du signal de parole. Il est appliqué de manière indépendante de la langue et du locuteur, sans aucune phase d'adaptation. La segmentation est intrinsèquement infra-phonémique puisqu'elle est basée sur une détection de ruptures et que les parties transitoires des phonèmes sont dissociées des parties stables. Nous testons cependant si le nombre de segments par seconde se révèle corrélé de manière forte au nombre de phonèmes par seconde. La détection des voyelles fournit quant à elle un estimateur a priori fiable du nombre de syllabes par seconde. Les erreurs de détection les plus fréquentes sont

des erreurs d'omission de voyelles de faible énergie ou dévoisées et des fausses détections de liquides.

6.2.2 Analyse des données

Corpus

Les expériences sont menées sur un sous ensemble du corpus « OGI MLTS » pour lequel des transcriptions phonétiques manuelles sont fournies. Le tableau 6.2 donne les caractéristiques de cette sous base de données. Pour chaque locuteur, un enregistrement d'environ 40 secondes est étiqueté phonétiquement et qualifié de « spontané » ou « lu ». Cette distinction n'a pas été faite pour l'hindi. Pour les autres langues, la plupart des enregistrements sont considérés « spontanés » et la taille du corpus varie de 64 fichiers pour le japonais à 144 pour l'anglais. Ce corpus nous permet de calculer le débit à partir de l'étiquetage manuel et à partir de l'algorithme de détection des voyelles.

Tab. 6.2 : Description du corpus, nombre total de locuteurs et nombre de locuteurs considérés comme « spontanés », durée moyenne des fichiers (et écart-type)

Langue	Nombre de locuteurs (spontanés)	Durée moyenne par locuteur (écart-type)
Anglais	144 (111)	47,1 (3,4)
Allemand	98 (89)	42,7 (8,4)
Hindi	68 (n.c.)	46,5 (5,9)
Japonais	64 (55)	46,1 (5,1)
Mandarin	69 (69)	39,9 (10,8)
Espagnol	108 (106)	45,6 (5,6)

Conventions et calcul du débit

Les conventions d'étiquetage développées au CSLU [75] sont fondées sur des règles indépendantes des langues et adaptées à chaque langue suivant la liste des phonèmes. Les frontières phonémiques sont déterminées avec une précision de l'ordre d'une milliseconde. Par convention, les diphtongues sont considérées comme une seule voyelle dans le calcul du débit. Puisque les évenements ne correspondant pas à de la parole (pauses silencieuses, respirations, etc.) sont également étiquetés, il est possible de les écarter pour le calcul du débit.

Soit u la phrase sur laquelle on calcule le débit. Soit $N_v(u)$ le nombre de d'étiquettes « voyelle » dans cette phrase et $D(u)$ la durée de la phrase. Le débit moyen en termes de syllabes par seconde mesuré sur la phrase ($DP(u)$) à partir de l'étiquetage manuel est

alors défini par :

$$DP(u) = \frac{N_v(u)}{D(u)} \tag{6.1}$$

En considérant la durée totale des évenements ne correspondant pas à de la parole $D_{np}(u)$, le débit moyen non biaisé est alors défini par :

$$DP_{np}(u) = \frac{N_v(u)}{(D(u) - D_{np}(u))} \tag{6.2}$$

Cette mesure globale du débit est évidemment limitée puisqu'elle sous-estime l'impact des variations locales de débit qui ont lieu pendant la production de parole. Elle doit cependant permettre d'évaluer l'impact de la variabilité inter-locuteur et inter-langue sur le DP.

L'algorithme de détection automatique des voyelles fournit une estimation du nombre de voyelles présentes dans le signal (segments vocaliques). Il permet ainsi d'estimer le débit $\widehat{DP}(u)$ à partir d'un traitement automatique :

$$\widehat{DP}(u) = \frac{\widehat{N_v}(u)}{D(u)} \tag{6.3}$$

Le débit de parole peut également être calculé en termes de nombre de phonèmes par seconde. On remplacera alors N_v dans les formules précédentes par N_{phon}, le nombre de phonèmes de la phrase.

Comparaisons inter-langues à partir de l'étiquetage manuel

– Débit syllabique :

Le tableau 6.3 donne les débits moyens (DP et DP_{np}) calculés pour chaque langue du corpus en fonction du nombre de voyelles par seconde. La première constatation est que, même en écartant les pauses, les différences inter-langues sont significatives (ANOVA F(5) = 15 ; p<.0001). Le débit d'information (en termes de syllabes par seconde) est donc dépendant de la langue ce qui confirme indirectement que le débit d'information global résulte non seulement du niveau phonético-phonologique mais également des niveaux morpho-syntaxiques. Si l'on écarte les pauses, le plus faible débit est obtenu pour le mandarin (4,61) tandis que le plus important est obtenu pour l'espagnol (5,71). L'ordre est quasi-identique que l'on considère ou non les pauses. L'anglais et l'allemand montrent des débits DP_{np} très proches qui peuvent être liés au fait que ces deux langues sont très proches rythmiquement.

Tab. 6.3 : Moyenne et écart-type du DP syllabique (étiquetage manuel)

Langue	DP (NbVoy/s) (avec pauses)	DP_{np} (NbVoy/s) (sans pauses)
Anglais	3,80 ±0,11	4,71 ±0,09
Allemand	3,60 ±0,11	4,68 ±0,11
Hindi	3,67 ±0,16	5,40 ±0,14
Japonais	3,89 ±0,20	5,21 ±0,15
Mandarin	3,04 ±0,18	4,61 ±0,16
Espagnol	4,24 ±0,15	5,71 ±0,13

– Débit phonémique :

Le tableau 6.4 donne les débits moyens (DP et DP_{np}) calculés pour chaque langue du corpus en fonction du nombre de phonèmes par seconde. La mise en correspondance avec les résultats du tableau 6.3 donne des indices intéressants sur la structure syllabique des langues présentées. Par exemple, l'allemand présente le débit syllabique le plus faible et le débit phonémique le plus important, révélant ainsi une structure syllabique complexe.

Tab. 6.4 : Moyenne et écart-type du DP phonémique (étiquetage manuel)

Langue	DP (NbPhon/s) (avec pauses)	DP_{np} (NbPhon/s) (sans pauses)
Anglais	11,61 ±0,30	13,73 ±0,25
Allemand	11,44 ±0,33	14,20 ±0,31
Hindi	9,90 ±0,42	13,54 ±0,37
Japonais	11,47 ±0,51	14,63 ±0,38
Mandarin	8,82 ±0,52	12,45 ±0,48
Espagnol	10,96 ±0,34	13,95 ±0,30

– Corrélation entre les deux types de débits proposés :

Le tableau 6.5 montre les corrélations entre les deux estimateurs de débit proposés. Ces résultats montrent que ces deux mesures de débit sont très corrélées, de manière plus importante même que la corrélation indiquée pour l'allemand dans [105]. La pente des régressions linéaires permet d'estimer que la longueur moyenne des syllabes est de 2,8 phonèmes (maximum 3,1 pour l'allemand et minimum 2,4 pour le japonais).

Tab. 6.5 : Corrélation entre les deux types de débits proposés (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,94	0,89	$DP_{syl} = 0{,}35 DP_{pho}$ - 0,23
Allemand	0,93	0,86	$DP_{syl} = 0{,}32 DP_{pho}$ - 0,04
Hindi	0,96	0,92	$DP_{syl} = 0{,}37 DP_{pho} + 0{,}03$
Japonais	0,98	0,95	$DP_{syl} = 0{,}38 DP_{pho}$ - 0,46
Mandarin	0,96	0,91	$DP_{syl} = 0{,}34 DP_{pho} + 0{,}07$
Espagnol	0,96	0,91	$DP_{syl} = 0{,}41 DP_{pho}$ - 0,21

6.2.3 Evaluation des algorithmes comme estimateurs du débit

Estimation du débit syllabique par le nombre de voyelles détectées par seconde

Les résultats sont donnés dans le tableau 6.6 à la fois en termes de coefficients de corrélation (R) et de régression linéaire (figure 6.2). Toutes les corrélations sont très significatives ($p < .001$). La plus mauvaise corrélation est obtenue pour l'espagnol, mais elle demeure élevée ($R = 0,79$). En moyenne, la corrélation obtenue est de $R = 0,86$ ce qui indique que le détecteur de voyelles est un bon à très bon estimateur du débit syllabique. La qualité du détecteur de voyelles est également confirmée par les valeurs des pentes, proches de l'unité (en moyenne 0,89).

Tab. 6.6 : Corrélation entre débits syllabiques réels et estimés par la détection des voyelles (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,84	0,70	$DP_{syl} = 0{,}90\widehat{DP_{syl}} + 0{,}41$
Allemand	0,81	0,65	$DP_{syl} = 0{,}75\widehat{DP_{syl}} + 0{,}85$
Hindi	0,89	0,80	$DP_{syl} = 0{,}91\widehat{DP_{syl}} + 0{,}58$
Japonais	0,92	0,85	$DP_{syl} = 0{,}97\widehat{DP_{syl}} + 0{,}44$
Mandarin	0,90	0,81	$DP_{syl} = 0{,}94\widehat{DP_{syl}} + 0{,}11$
Espagnol	0,79	0,62	$DP_{syl} = 0{,}88\widehat{DP_{syl}} + 1{,}05$

Estimation du débit phonémique par le nombre de segments détectés par seconde

L'utilisation de l'algorithme de segmentation comme estimateur du débit phonémique montre des résultats plus contrastés (tableau 6.7 et figure 6.3). La pente moyenne (0,55) confirme le caractère infra-phonémique de la segmentation. Les coefficients de corrélation s'étendent de 0,51 pour l'allemand à 0,86 pour l'hindi. Il est difficile d'estimer la part de cette variation liée à la structure syllabique des langues et celle liée à un éventuel biais de

la segmentation en fonction des langues.

Tab. 6.7 : Corrélation entre débits phonétiques réels et estimés par la segmentation (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,74	0,55	$DP_{pho} = 0{,}52\widehat{DP_{pho}} + 3{,}91$
Allemand	0,51	0,27	$DP_{pho} = 0{,}37\widehat{DP_{pho}} + 6{,}16$
Hindi	0,86	0,74	$DP_{pho} = 0{,}62\widehat{DP_{pho}} + 1{,}71$
Japonais	0,74	0,55	$DP_{pho} = 0{,}51\widehat{DP_{pho}} + 4{,}80$
Mandarin	0,72	0,51	$DP_{pho} = 0{,}64\widehat{DP_{pho}} + 1{,}47$
Espagnol	0,74	0,55	$DP_{pho} = 0{,}56\widehat{DP_{pho}} + 3{,}48$

Corrélation entre les estimateurs de débits syllabiques et phonémiques

Tab. 6.8 : Corrélation entre les estimateurs de débits syllabiques et phonémiques (avec pauses)

Langue	R	R^2	Régression linéaire
Anglais	0,60	0,35	$\widehat{DP_{syl}} = 0{,}14\widehat{DP_{pho}} + 1{,}66$
Allemand	0,55	0,31	$\widehat{DP_{syl}} = 0{,}15\widehat{DP_{pho}} + 1{,}57$
Hindi	0,82	0,66	$\widehat{DP_{syl}} = 0{,}22\widehat{DP_{pho}} + 0{,}47$
Japonais	0,79	0,62	$\widehat{DP_{syl}} = 0{,}20\widehat{DP_{pho}} + 0{,}94$
Mandarin	0,72	0,52	$\widehat{DP_{syl}} = 0{,}22\widehat{DP_{pho}} + 0{,}62$
Espagnol	0,71	0,51	$\widehat{DP_{syl}} = 0{,}20\widehat{DP_{pho}} + 0{,}89$

Les corrélations entre les estimateurs de débits syllabiques et phonémiques sont similaires à celles obtenues entre les débits syllabiques et phonémiques déduits de l'étiquetage manuel.

6.2.4 Conclusion et Perspectives

Les statistiques présentées montrent que le débit de parole est dépendant de la langue. Elles montrent également que, même pour des langues accentuelles comme l'anglais, pour lesquelles une grande variabilité de la complexité syllabique est avérée ([111]), les débits de parole phonémique et syllabique sont extrêmement corrélés ($R = 0,94$), tout comme dans des langues ayant des structures syllabiques plus simples (espagnol ou mandarin par exemple).

Le détecteur de voyelles fournit un bon estimateur du débit syllabique (en moyenne $R = 0,86$). Par contre, le comportement du nombre de segments automatiques comme

estimateur du débit phonémique semble dépendre de la langue, la corrélation étant plutôt bonne pour l'hindi et faible pour l'allemand. On peut penser que ces différences inter-langues sont liées aux inventaires phonémiques des langues étudiées.

Cependant plusieurs autres paramètres peuvent influencer l'estimation du débit. On peut citer par exemple la variation du débit au cours du temps, qui peut être liée à plusieurs aspects linguistiques (allongement final, etc.), ou encore la qualité de la prise en compte des pauses silencieuses et remplies dans le signal.

6.3 Conclusion

L'identification des langues par la prosodie est une tâche plus difficile lorsque l'on considère la parole spontanée. Les expériences, menées dans l'optique de faire une distinction entre deux langues, ont donné dans l'ensemble de bons résultats mais avec beaucoup de variabilité selon les langues [117].

Par contre, dès que la tâche consiste à identifier une langue parmi un nombre plus conséquent, les résultats sont décevants. Nous supposons que cela est dû à la variabilité bien plus importante de la parole spontanée par rapport à la parole lue, qui s'exprime notamment par des différences importantes de débit entre les locuteurs.

C'est pour cette raison que nous avons mis au point une méthode de mesure automatique du débit qui nous a permis de comparer les différences entre les langues, les locuteurs et les styles de parole (lue et spontanée) [102, 116].

Afin de prendre en compte la mesure de débit dans les modèles, il faut estimer le débit sur les données d'apprentissage. Ce débit peut être catégorisé en trois classes : rapide, normal et lent. Les moyennes des modèles de mélange de lois gaussiennes décrivant les durées des pseudo-syllabes seront adaptés pour chaque catégorie de débit. Pour chaque langue, il y aura trois modèles statistiques caractérisant les durées des pseudo-syllabes en fonction des vitesses d'élocution. Lors de la phase de reconnaissance, le débit sera identifié et le modèle correspondant sera sélectionné.

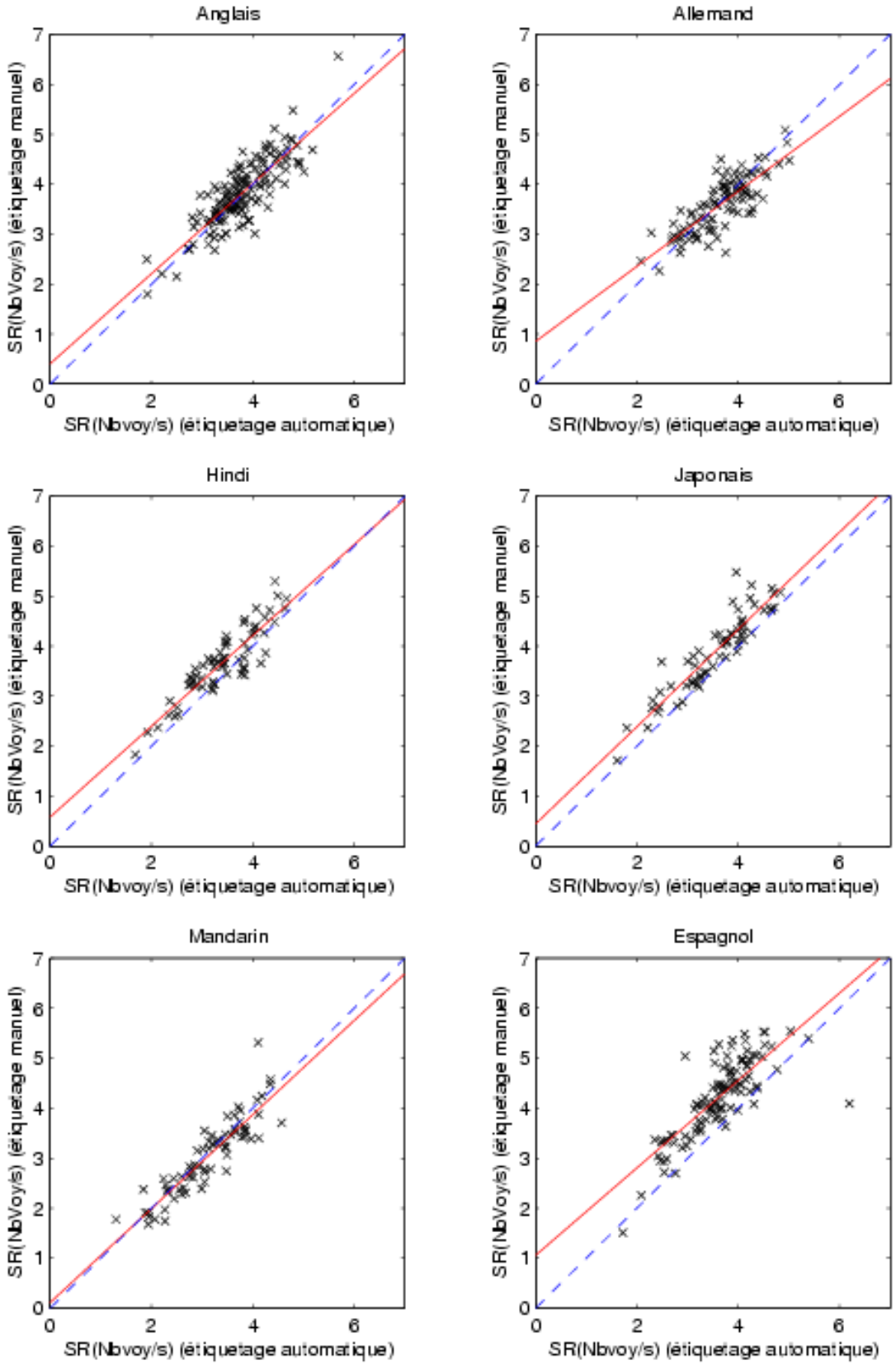


Fig. 6.2 : DP syllabique estimé par le nombre de voyelles par seconde

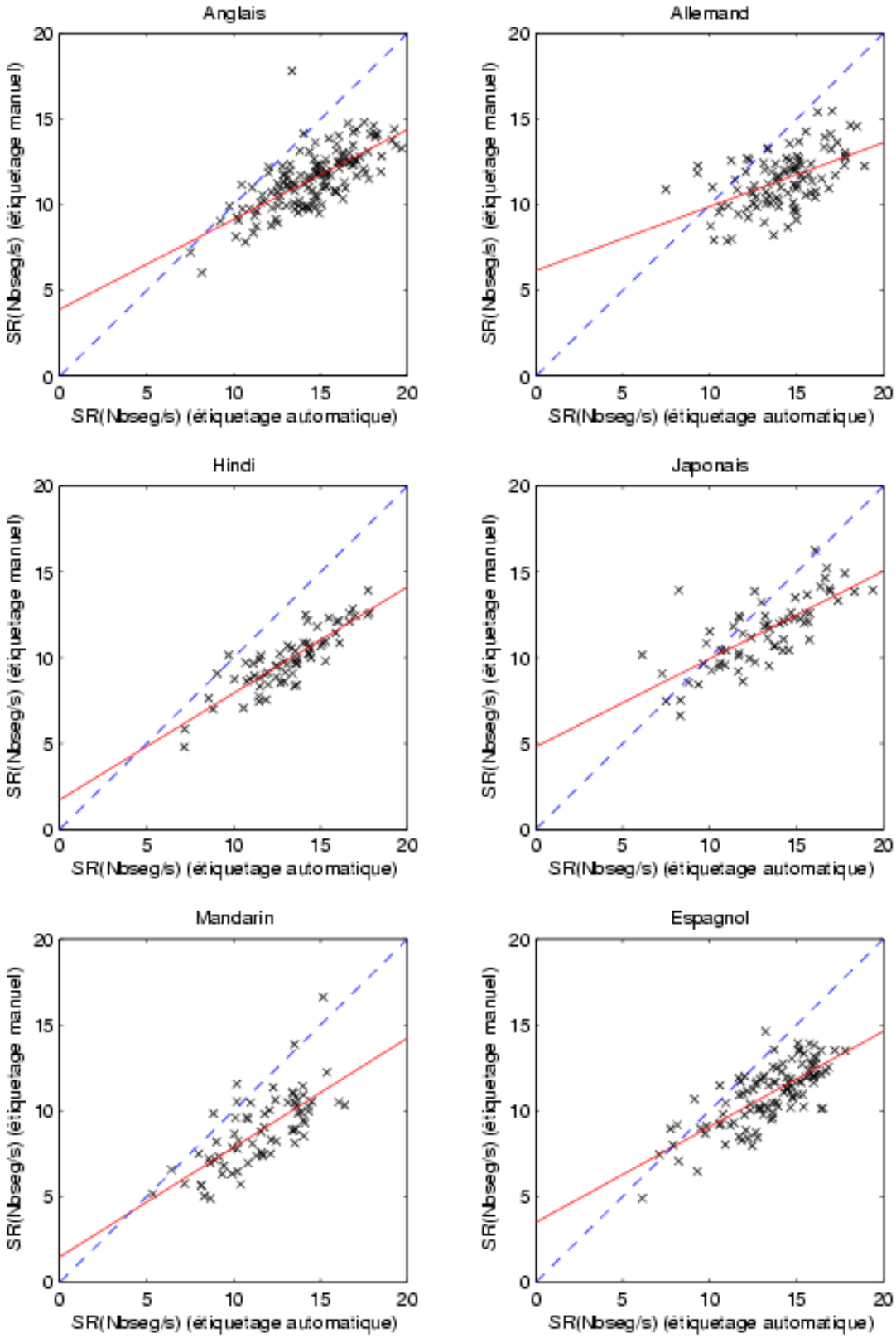


Fig. 6.3 : DP phonémique estimé par le nombre de segments par seconde

Conclusion et perspectives

Il n'est pas aisé de donner une définition de la prosodie. Les principales difficultés rencontrées à la suite de cette définition ont été évoquées, notamment la nature polysémique des informations véhiculées. Parallèlement, nous avons ensuite montré quelles sont les possibilités offertes pour l'extraction automatique de paramètres liés à la prosodie à partir du signal de parole brut. Malgré ces difficultés, nous avons vu quelles sont les potentialités offertes par la prosodie en vue de l'identification automatique des langues, notamment comment l'étude de la prosodie permettrait de confirmer ou d'infirmer les théories proposées par les linguistes et s'inscrire dans le cadre de théories cognitives liées à la compréhension et à l'évolution du langage. Cet espoir est confirmé par les études en perception de parole dégradée, qui permettent d'affirmer que les humains sont capables de distinguer des langues en considérant uniquement des informations prosodiques. La difficulté principale est de savoir quelles sont les informations les plus importantes et comment les modéliser.

Le panorama des systèmes actuels d'identification automatique des langues, notamment ceux utilisés lors de la campagne d'évaluation Nist 2003, montre que sont utilisées uniquement les caractéristiques acoustiques des langues, et les caractéristiques phonétiques ou phonotactiques. La dimension prosodique, malgré l'intérêt certain qu'elle présente, n'est que marginalement employée, voire pas du tout lorsque l'objectif premier est la performance.

Nous avons recherché quelles étaient les théories linguistiques qui pouvaient se prêter à une modélisation automatique et nous en avons décliné plusieurs systèmes

Les résultats obtenus par ces systèmes montrent l'intérêt d'utiliser la prosodie pour l'identification des langues. Avec le système de Leavers (§3.3.1), des séparations de groupes de langues sont possibles en employant uniquement la prosodie et l'emploi de techniques acoustiques permet de distinguer les langues dans chaque groupe. Le système d'Itahashi (§3.3.2) met l'accent sur l'apport de l'emploi de techniques de modélisation de la prosodie par rapport aux méthodes "classiques" (ici acoustico-phonétique), même si le gain en termes de performance n'est pas très conséquent. Le système de Cummins (§3.3.3) confirme que l'on peut distinguer deux langues uniquement par la prosodie, pour peu qu'elles n'appartiennent pas au même groupe linguistique. Le système de Li (§3.3.4) prouve que la modélisation de la prosodie permet d'obtenir de bonnes performances. Toutes ces approches utilisent des modélisations statistiques sur les paramètres prosodiques. Or la prosodie est par nature dynamique, il semble réaliste de prendre cet aspect en compte dans les modèles. Le système d'Adami (§3.3.5) essaie de modéliser les enchaînements d'évènements prosodiques sur une phrase et démontre l'efficacité d'une telle démarche. Le nombre et la diversité des modèles de description de la prosodie prouvent la difficulté de trouver un formalisme multilingue. Chaque méthode de description possède ses avantages propres, qu'elle soit basée sur les mécanismes de production ou de perception de la parole. L'utilisation de tels modèles pour l'identification des langues est donc difficile, puisque très peu de travaux ont porté sur des différences éventuelles entre les langues observables à l'aide de ces méthodes de description. Les systèmes de caractérisation du rythme (§3.1) se ressemblent. Ils essaient de décrire le rythme au travers de mesures de durées et de dispersion des durées que ce soit des intervalles vocaliques, intervocaliques ou consonantiques, voire qui tendent à s'en rapprocher (sonorité). Ces méthodes ont pour faiblesse de n'être testées que sur des corpus peu conséquents, ce qui est principalement dû à la nécessité d'employer un étiquetage manuel préalable au traitement.

Fort de ces constatations, nous avons cherché à tirer parti des outils disponibles à l'IRIT afin de valider ces méthodes sur un corpus plus conséquent. Nous avons étudié différentes approches de modélisation du rythme des langues. Les approches de Ramus (§3.1.1) et de Grabe (§3.1.2) prennent pour point de départ des mesures de durées vocaliques et intervocaliques. Notre méthode a consisté à déterminer une unité extraite automatiquement de type syllabique, la pseudo-syllabe, qui permet d'effectuer des mesures similaires à celles de Ramus et de Grabe.

La pseudo-syllabe est ensuite caractérisée par des paramètres tentant de refléter sa nature prosodique. Afin de caractériser sa constitution temporelle et structurelle, nous avons extrait trois paramètres :

- la durée de la voyelle,
- la durée de l'ensemble des consonnes,
- le nombre de consonnes.

Pour décrire son contour intonatif, nous avons calculé cinq paramètres sur les valeurs de fréquence fondamentale de chaque pseudo-syllabe :

- le *skewness* ou coefficient d'aplatissement,
- le *kurtosis* ou coefficient d'asymétrie,
- l'écart entre la position de la valeur maximale de la fréquence fondamentale et le début du segment vocalique,
- l'écart entre la position de la valeur minimale de la fréquence fondamentale et le début du segment vocalique,
- la bande passante normalisée ou écart entre la valeur maximale et la valeur minimale de fréquence fondamentale sur chaque pseudo-syllabe, normalisé par la moyenne des valeurs de fréquence fondamentale sur la pseudo-syllabe.

Des modèles statistiques, les modèles de mélanges de lois gaussiennes, sont employés pour modéliser les caractéristiques extraites des pseudo-syllabes pour chaque langue.

Avec les paramètres rythmiques, nous obtenons ainsi un taux d'identification correcte de 67% sur les sept langues du corpus MULTEXT (voir [39] pour des résultats avec moins de langues). En utilisant les paramètres relatifs à l'intonation, nous obtenons 50% d'identifications correctes (voir [115] pour des résultats avec moins de langues). Caractériser cette unité avec des paramètres relatifs à l'intonation conjointement aux paramètres de durée, permet d'obtenir 70% d'identifications correctes (voir [40, 114, 118] pour des résultats avec moins de langues). Nous remarquons que lorsqu'il y a des confusions, elles se font principalement à l'intérieur des ensembles de langues évoqués dans les théories linguistiques. En considérant l'identification non plus des langues mais des groupes rythmiques (langues accentuelles, syllabiques et moraiques), la modélisation de la prosodie se revèle efficace (91% d'identifications correctes en fusionnant les modèles rythmiques et intonatifs).

TOUTEFOIS, les modèles statistiques (modèle de mélange de lois gaussiennes) que nous avons employé pour modéliser les caractéristiques des pseudo-syllabes sont intrinsèquement des modèles statiques. En effet, chaque pseudo-syllabe est caractérisée par un unique vecteur de paramètres, ce qui ne correspond pas à la réalité perceptive de la prosodie, qui est par nature continue. Nous devons donc employer des modèles de nature dynamique afin de prendre en compte cet aspect temporel.

En prenant l'exemple du travail décrit dans [2], nous nous servons de paramètres calculés sur chaque pseudo-syllabe pour leur attribuer des étiquettes donnant le sens des mouvements de fréquence fondamentale et d'énergie. Les enchaînements de ces étiquettes sont modélisés par des modèles multigrammes, qui permettent d'identifier les séquences les plus fréquentes pour chaque langue et de leur associer des probabilités. En attribuant ainsi automatiquement une étiquette à chaque pseudo-syllabe, le taux de reconnaissance des langues est de l'ordre de 40% sur le corpus MULTEXT. Ce résultat nous surprenant puisque nous nous attendions à améliorer les résultats en prenant en compte la dynamique dans nos modèles, nous avons réitéré l'expérience en n'étiquetant plus les pseudo-syllabes mais les segments qui les constituent, ce en réadaptant l'idée d'Adami. Le taux d'identification correcte obtenu par cette méthode est de 63%. Ces expériences nous permettent de supposer que les éléments les plus caractéristiques des langues ne seraient pas les enchaînements des pseudo-syllabes, mais les enchaînements des éléments les constituant.

Suite à ces expériences, nous avons tenté d'appliquer le système mis au point sur le corpus MULTEXT à un corpus de parole spontanée téléphonique, le corpus OGI MLTS [93]. Les premières expériences menées dans l'optique de faire une distinction entre deux langues ont donné dans l'ensemble de bons résultats, mais avec beaucoup de variabilité selon les langues [117]. Par contre, dès que la tâche consiste à identifier une langue parmi un nombre plus conséquent, les résultats sont décevants. Nous supposons que cela est dû à la variabilité bien plus importante de la parole spontanée par rapport à la parole lue, qui s'exprime notamment par des différences importantes de débit entre les locuteurs. C'est pour cette raison que nous avons mis au point une méthode de mesure automatique du débit qui nous a permis de comparer les différences entre les langues, les locuteurs et les styles de parole (lue et spontanée) [102, 116].

Ce travail peut bénéficier de plusieurs extensions. Des extensions techniques seront utiles pour l'amélioration des performances de notre système et des extensions liées à l'intérêt de la pseudo-syllabe pour l'étude de la prosodie.

Les extensions techniques concernent les algorithmes de traitement automatiques :

- Amélioration de la détection des voyelles :
  L'algorithme que nous utilisons pour la détection des voyelles possède quelques faiblesse, notamment pour la détection des voyelles réduites dans les langues accentuelles. Une idée pourrait être de créer des modèles HMM sur les étiquettes « C » ou « V » issues de la segmentation automatique, puis d'utiliser ces modèles pour faire une deuxième passe d'étiquetage.
- Amélioration de la définition des pseudo-syllabes :
  Ne considérer que les syllabes de type « CV » est une limitation importante. A l'instar du travail décrit dans [8], nous pourrions employer une méthode de syllabation utilisant un critère de sonorité, ce qui permettrait d'obtenir des syllabes plus pertinentes.
- Amélioration des unités prosodiques :
  Les expériences menées dans le chapitre 5, notamment avec le système d'Adami, montrent l'importance du choix de l'unité prosodique. Nous avons utilisé trois types d'unités différentes : les segments « C » ou « V », les pseudo-syllabes, et l'unité d'Adami. De nombreux travaux restent à effectuer sur la définition d'une unité prosodique qui permettrait d'améliorer les performances du système.
- Amélioration des performances en parole spontanée :
  Pour être efficace sur des données de parole spontanée, il paraît crucial de prendre en compte les variations occasionnées par ce style de parole, particulièrement en ce qui concerne le débit. La prise en compte du débit peut être directement appliquée dans les modèles, ou plus simplement par une normalisation. La difficulté majeure est que les variations de débit influent sur les durées des différents phonèmes. Selon les langues, ce ne sont pas les mêmes phonèmes qui sont compressés ou étendus.

D'autres extensions concernent l'emploi de la modélisation de la prosodie proposée. Il s'agit d'explorer plus avant les possibilités offertes par l'unité prosodique qu'est la pseudo-syllabe et, le cas échéant, d'envisager une adaptation par rapport aux réalités linguistiques.

Cette recherche s'étend naturellement en étudiant les théories linguistiques actuellement développées et en proposant des traitements automatiques adaptés à des besoins de validation. Les expériences ont montré que le système prosodique proposé, construit à partir d'observations linguistiques, se comporte différemment selon le type de parole considéré : parole lue ou parole spontanée. La variabilité liée au style de prononciation influe sur le modèle prosodique, et par voie de conséquence peut remettre en cause certaines affirmations linguistiques.

Plusieurs niveaux de recherche s'offrent alors pour approfondir ce problème :
- La définition très simpliste de l'unité élémentaire pseudo-syllabe qui a démontré ses avantages, peut être affinée en ne se limitant pas à une localisation grossière des segments vocaliques et des segments consonantiques. Très peu d'informations sont actuellement utilisées pour localiser et identifier ces zones ; des modèles acoustiques caractéristiques de sons de base devraient apporter l'information complémentaire.
- Les paramètres issus des pseudo syllabes peuvent dépendre du type de prononciation. À titre d'exemple, il est facile de prédire que ceux liés au débit de parole peuvent devenir totalement inutiles en parole spontanée, voire induire un bruit pour caractériser une langue. Cette dépendance vis à vis du type de prononciation peut être traduite par un indice de confiance dans les modèles.
- Enfin les linguistes eux mêmes doivent se pencher sur ces résultats pour que, dès lors que l'extraction des paramètres et la modélisation automatique ne sont assurément pas les sources de problèmes, le modèle linguistique soit approfondi. Il s'en suit un nouveau cycle de validation.

Ce type de synergie entre linguistique et modélisation automatique doit permettre peu à peu une validation des connaissances sur les langues et bien sûr sur les dialectes.

Annexe A

Exponentielle de Hurst

Sommaire

A.1 Méthode de l'échelle réduite (*Rescaled Range*) 169
A.2 Estimation de l'exponentielle de Hurst 170

L'exponentielle de Hurst quantifie le degré de persistance ou d'anti-persistance d'un processus évoluant au cours du temps. Ce type d'analyse est appelé méthode de l'échelle réduite (*Rescaled Range* ou R/S) [64].

A.1 Méthode de l'échelle réduite (*Rescaled Range*)

Soit $X = \{X_1, ..., X_n, ..., X_N\}$ une série temporelle d'observations. On calcule alors la somme des déviations sur les observations :

$$Y_{n,N} = \sum_{u=1}^{n} [X_u - \bar{X}_N], \text{ avec } \bar{X}_N = \frac{1}{N} \sum_{u=1}^{N} X_u \tag{A.1}$$

$\bar{X}_N$ est la moyenne des observations calculée sur la durée N. $Y_{n,N}$ est une somme cumulée des déviations par rapport à la moyenne sur la durée n.

Le *Range* (R) est la différence entre les valeurs maximales et minimales de déviation :

$$R = \max_{1 \leq n \leq N} (Y_{n,N}) - \min_{1 \leq n \leq N} (Y_{n,N}) \tag{A.2}$$

En divisant R par l'écart type des observations :

$$S = \sqrt{\frac{1}{N} \sum_{u=1}^{N} (X_u - \bar{X}_N)^2}, \tag{A.3}$$

on obtient le *Rescaled Range* (R/S).

Hurst a formulé la relation suivante de manière empirique :

$$\frac{R}{S} = (a * n)^H \tag{A.4}$$

avec R/S le *Rescaled Range* et n le nombre d'observations. a est une constante et H est l'exponentielle de Hurst.

La valeur de H est égale à 0,5 lorsque la série est aléatoire. C'est-à-dire que $R(n)$ devrait augmenter proportionnellement à la racine carrée du temps. Cependant, lors que H est différent de 0,5, les observations ne sont pas indépendantes, mais contiennent une mémoire des évènements précédents. L'effet mémoire caractérisé ici n'est pas une mémoire à court terme ou Markovienne, mais une mémoire à long terme.

Lorsque $0 \leq H < 0,5$, le système est anti-persistant ou ergodique.

Lorsque $0,5 \leq H < 1$, le système est persistant.

La figure A.1 montre une illustration du calcul de R.

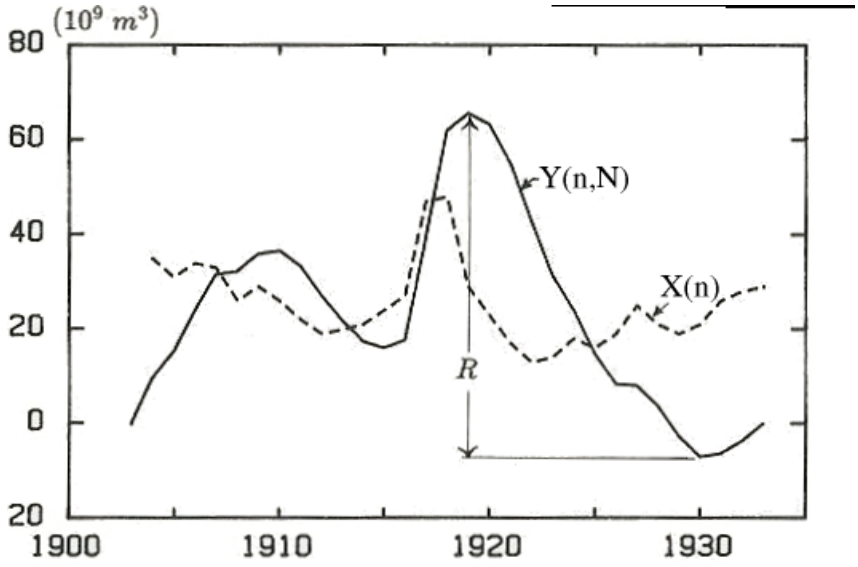


Fig. A.1 : Exemple de calcul de R pour des variations de débit hydraulique sur plusieurs années

A.2 Estimation de l'exponentielle de Hurst

L'exponentielle de Hurst est estimée en calculant les R/S pour de multiples périodes. L'espérance de R/S est alors calculée pour un ensemble de régions. Pour chaque ensemble de régions on a :

$$R/S_{moy} = E\big[\frac{R}{S}(r)\big] = \frac{1}{R}\sum_{u=1}^{R} R/S(u) = Cn_r^H \text{ quand } n \to \infty \tag{A.5}$$

avec C = constante, R le nombre de régions et n la taille des régions..

Le R/S est calculé tout d'abord sur l'ensemble des données. Ensuite, l'ensemble des données est divisé en deux et le R/S est calculé sur chacune des deux parties. Les deux valeurs résultantes sont alors moyennées.

Le processus continue et chacune des deux sections déterminées précédemment est divisée en deux et le R/S est calculé sur chaque nouvelle partie. La valeur moyenne de R/S est ensuite calculée.

Le processus de subdivision s'arrête lorsque les régions sont trop petites. En général, les régions doivent contenir au moins 8 observations.

Un exemple est montré sur la figure A.2.

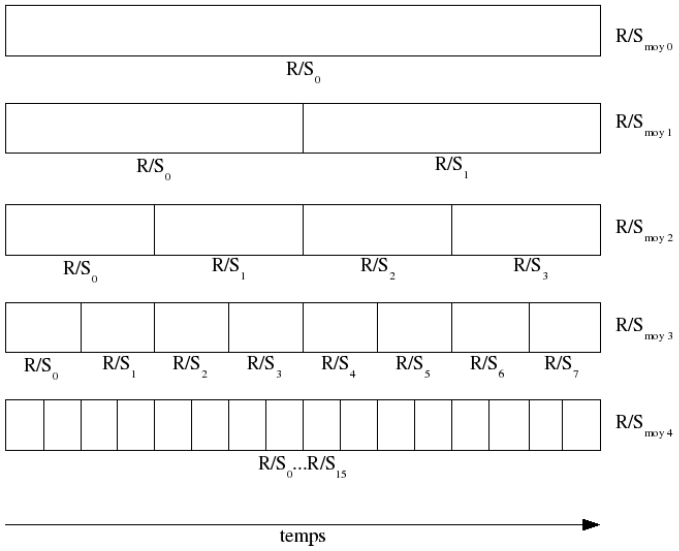


Fig. A.2 : Estimation de l'exponentielle de Hurst : calcul des différents R/S

Pour estimer l'exponentielle de Hurst, un vecteur est alors créé, ayant pour coordonnées le logarithme de la taille de la région considérée et le logarithme de la moyenne de R/S (R/S_{moy}).

L'exponentielle de Hurst est estimée par une régression linéaire sur ces points. La pente de la régression linéaire correspond à l'estimation de l'exponentielle de Hurst.

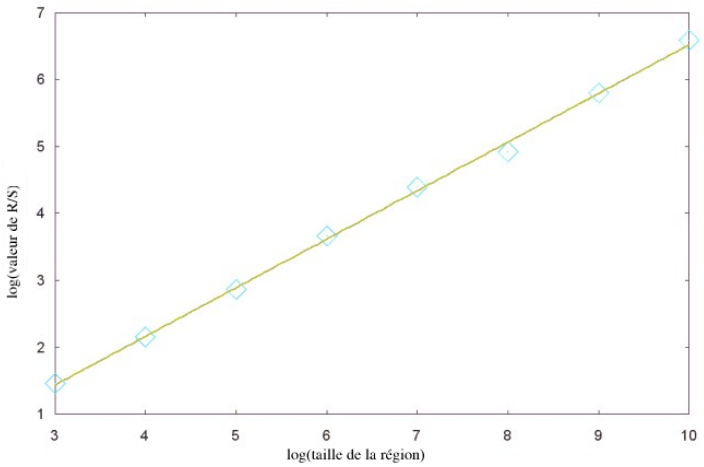


Fig. A.3 : Estimation de l'exponentielle de Hurst : régression linéaire

Cette méthode pour déterminer l'exponentielle de Hurst a été développée et analysée par Benoit Mandelbrot [82].

Annexe B

Jeux de données du corpus Multext

Sommaire

B.1 Description du corpus 175
B.2 Enregistrements audio 175
B.3 Répartition des passages par locuteurs 176
B.3.1 Anglais 176
B.3.2 Allemand 176
B.3.3 Espagnol 177
B.3.4 Français 178
B.3.5 Italien 178
B.4 Jeu de données n°1 180
B.4.1 Données d'apprentissage 180
B.4.2 Données de test 180
B.5 Jeu de données n°2 181
B.5.1 Données d'apprentissage 181
B.5.2 Données de test (2 locuteurs par langue) 181
B.6 Jeu de données n°3 182
B.6.1 Données d'apprentissage 182
B.6.2 Données de test (2 locuteurs par langue) 182

MULTEXT [20] est une base de données prosodiques. Elle est distribuée par l'ELRA[5] à travers l'ELDA[6].

B.1 Description du corpus

Le corpus MULTEXT est proposé pour cinq langues : l'anglais, l'allemand, l'espagnol, le français et l'italien. Le japonais a été rajouté grâce à Kitazawa [67] et le mandarin grâce à Komatsu [70]. Il contient les informations suivantes :

- le signal acoustique ;
- l'alignement sur le signal de la transcription orthographique (au niveau des mots) ;
- la fréquence fondamentale (F_0) extraite toutes les 10 ms ;
- la stylisation de la courbe de F_0 extraite par MOMEL [61], un algorithme qui représente la macro-prosodie d'une phrase par une série de points cibles, puis interpolés par une courbe spline ;
- la re-synthèse de F_0 en utilisant F_0 stylisée ;
- le codage symbolique de la courbe de F_0 ;
- la F_0 résiduelle, issue de la différence entre F_0 et F_0 stylisée ;
- un fichier de description de l'enregistrement (au format SAM).

B.2 Enregistrements audio

Les enregistrements sont extraits du corpus de parole EUROM1, enregistré à l'occasion du projet ESPRIT 2589 « Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation » [22]. Les enregistrements audio sont de haute qualité (échantillonnage à 16 kHz, 16 bits) et enregistrés en chambre anéchoïque. Ils ont été contrôlés durant l'acquisition de manière à rejeter toute donnée bruitée ou toute erreur de lecture. MULTEXT reprend cinq (allemand, anglais, espagnol, français et italien) des huit langues de EUROM1. Les données correspondent au jeu de locuteurs « FEW TALKER SET » (comprenant dix locuteurs par langue, cinq femmes et cinq hommes) sur les passages lus de cinq phrases connectées par une structure sémantique cohérente. Notons qu'une même phrase peut être prononcée en moyenne tous les quatre locuteurs et que cela entraîne une dépendance possible au texte dans les modélisations. Il est demandé à chaque locuteur de lire un extrait du passage et d'essayer d'avoir l'intonation la plus naturelle possible.

[5]European Language Ressources Association, `http://www.elra.fr`
[6]Evaluations and Language ressources Distribution Agency, `http://www.elda.fr/`

B.3 Répartition des passages par locuteurs

B.3.1 Anglais

Les passages de l'anglais sont groupés en 8 ensembles de 5, lus par 10 locuteurs (cf. tableau B.1), de la façon suivante :

- chacun des dix locuteurs prononce 3 ensembles (15 passages) ;
- les six premiers ensembles sont lus par 4 locuteurs ;
- les deux derniers ensembles sont lus par 3 locuteurs.

Au total 150 passages (750 phrases) sont lus, ce qui représente 43,93 minutes de parole. Chaque phrase est lue en moyenne par 3,75 locuteurs.

Tab. B.1 : Répartition des passages par locuteur en anglais sur MULTEXT.

locuteur	*sexe*	o1-o5	o6-o0	p1-p5	p6-p0	q1-q5	q6-q0	r1-r5	r6-r0
fa	M	x	x	x					
fb	M					x	x	x	
fc	M				x	x	x		
fd	M	x	x						x
fe	M	x						x	x
ff	F		x	x	x				
fg	F			x	x	x			
fh	F						x	x	x
fi	F	x	x	x					
fj	F				x	x	x		

B.3.2 Allemand

Les passages de l'allemand sont groupés en 2 ensembles de 20, lus par 10 locuteurs (cf. tableau B.2). Chacun des dix locuteurs ne prononce qu'un ensemble (20 passages).

Au total, 200 passages (1000 phrases) sont lus, ce qui représente 1 heure 22 minutes de parole. Chaque phrase est lue en moyenne par 5 locuteurs.

Tab. B.2 : Répartition des passages par locuteurs en allemand sur MULTEXT.

locuteur	*sexe*	ensemble 1	ensemble 2
bg	M		x
hm	M		x
mj	M	x	
qk	M	x	
sm	M	x	
aj	F	x	
ga	F	x	
jm	F		x
mi	F		x
ss	F		x

B.3.3 Espagnol

Les passages de l'espagnol sont groupés en 8 ensembles de 5, lus par 10 locuteurs (cf. tableau B.3), de la façon suivante :

- chacun des dix locuteurs prononce 3 ensembles (15 passages) ;
- les six premiers ensembles sont lus par 4 locuteurs ;
- les deux derniers ensembles sont lus par 3 locuteurs.

Au total 150 passages (750 phrases) sont lus, ce qui représente 53,87 minutes de parole. Chaque phrase est lue en moyenne par 3,75 locuteurs.

Tab. B.3 : Répartition des passages par locuteur en espagnol sur MULTEXT.

locuteur	*sexe*	o0-o4	o5-o9	p0-p4	p5-p9	q0-q4	q5-q9	r0-r4	r5-r9
na	M					x	x	x	
nb	M	x	x						x
qa	M						x	x	x
ra	M	x	x	x					
ta	M	x						x	x
ba	F			x	x	x			
ca	F				x	x	x		
ea	F		x	x	x				
eb	F	x	x	x					
ha	F				x	x	x		

B.3.4 Français

Les passages du français sont groupés en 4 ensembles de 10, lus par 10 locuteurs (cf. tableau B.4), de la façon suivante :

- chacun des dix locuteurs prononce un ensemble (10 passages) ;
- les deux premiers ensembles sont lus par 3 locuteurs ;
- les deux derniers ensembles sont lus par 2 locuteurs.

Au total 100 passages (500 phrases) sont lus, ce qui représente 36,51 minutes de parole. Chaque phrase est lue en moyenne par 2,5 locuteurs.

Tab. B.4 : Répartition des passages par locuteurs en français sur MULTEXT.

locuteur	*sexe*	o0-o9	p0-p9	q0-q9	r0-r9
bf	M	x			
bo	M		x		
sc	M		x		
sh	M		x		
sl	M			x	
fa	F				x
ja	F	x			
mh	F			x	
ro	F	x			
vi	F				x

B.3.5 Italien

Les passages de l'espagnol sont groupés en 8 ensembles de 5, lus par 10 locuteurs (cf. tableau B.5), de la façon suivante :

- chacun des dix locuteurs prononce 3 ensembles (15 passages) ;
- les six premiers ensembles sont lus par 4 locuteurs ;
- les deux derniers ensembles sont lus par 3 locuteurs.

Au total 150 passages (750 phrases) sont lus, ce qui représente 54,31 minutes de parole. Chaque phrase est lue en moyenne par 3,75 locuteurs.

Tab. B.5 : Répartition des passages par locuteur en italien sur MULTEXT.

locuteur	*sexe*	G1	G2	G3	G4	G5	G6	G7	G8
ag	M					x	x	x	
au	M				x	x	x		
bk	M	x						x	x
b4	M	x	x	x					
b7	M	x	x	x					
a0	F	x	x						x
ba	F			x	x	x			
bf	F				x	x	x		
bl	F		x	x	x				
b6	F						x	x	x

B.4 Jeu de données n°1

B.4.1 Données d'apprentissage

Tab. B.6 : Description de l'ensemble d'apprentissage du jeu1 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	8	80	6609	24 mn
Français	8	80	7428	29 mn
Allemand	8	80	7659	29 mn
Italien	8	80	8041	30 mn
Japonais	4	80	9633	39 mn
Mandarin	8	80	6553	26 mn
Espagnol	8	80	7739	27 mn
Total	52	560	53662	204 mn

B.4.2 Données de test

Tab. B.7 : Description de l'ensemble de test du jeu1 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	2	20	1409	6 mn
Français	2	19	1791	7 mn
Allemand	2	20	1815	7 mn
Italien	2	20	1941	7 mn
Japonais	2	20	2245	11 mn
Mandarin	2	20	1660	6 mn
Espagnol	2	20	2189	8 mn
Total	14	139	13050	52 mn

B.5 Jeu de données n°2

B.5.1 Données d'apprentissage

Tab. B.8 : Description de l'ensemble d'apprentissage du jeu2 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	8	80	6412	24 mn
Français	8	79	7419	29 mn
Allemand	8	80	7508	29 mn
Italien	8	80	7887	28 mn
Japonais	4	80	9523	41 mn
Mandarin	8	80	6836	27 mn
Espagnol	8	80	7264	27 mn
Total	52	559	52849	205 mn

B.5.2 Données de test (2 locuteurs par langue)

Tab. B.9 : Description de l'ensemble de test du jeu2 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	2	20	1556	6 mn
Français	2	20	1800	7 mn
Allemand	2	20	1984	8 mn
Italien	2	20	1969	7 mn
Japonais	2	20	2360	10 mn
Mandarin	2	20	1437	5 mn
Espagnol	2	20	2095	8 mn
Total	14	140	13201	51 mn

B.6 Jeu de données n°3

B.6.1 Données d'apprentissage

Tab. B.10 : Description de l'ensemble d'apprentissage du jeu3 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	5	50	3730	14 mn
Français	5	49	4610	18 mn
Allemand	5	50	5000	19 mn
Italien	5	50	4983	19 mn
Japonais	4	50	5705	26 mn
Mandarin	8	50	4051	16 mn
Espagnol	5	50	5162	18 mn
Total	37	349	33241	130 mn

B.6.2 Données de test (2 locuteurs par langue)

Tab. B.11 : Description de l'ensemble de test du jeu3 (MULTEXT).

Langue	Nombre de locuteurs	Nombre total de fichiers	Nombre total de pseudo-syllabes	Durée
Anglais	2	20	1597	6 mn
Français	2	20	1812	7 mn
Allemand	2	20	1992	8 mn
Italien	2	20	1975	6 mn
Japonais	2	20	2776	11 mn
Mandarin	2	20	1660	6 mn
Espagnol	2	20	1730	6 mn
Total	14	140	13542	50 mn

Annexe C

Expériences complémentaires sur les différents jeux

Sommaire

C.1 Expériences complémentaires avec le modèle de rythme . . . 185
C.1.1 Expériences sur les données du jeu n°2 185
C.1.2 Expériences sur les données du jeu n°3 186
C.2 Expériences complémentaires avec le modèle de l'intonation (statique) 187
C.2.1 Expériences sur les données du jeu n°2 187
C.2.2 Expériences sur les données du jeu n°3 188

C.1 Expériences complémentaires avec le modèle de rythme

Les expériences de §4.5.3 sont reprises sur les deux autres jeux de données.

C.1.1 Expériences sur les données du jeu n°2

Apprentissage

Les expériences ont montré que le meilleur taux d'identification pour l'ensemble d'apprentissage du jeu 2 est de 59,0 ± 4,1 %, soit 330 identifications correctes sur 560 fichiers. Ce résultat est obtenu pour 8 gaussiennes. Ci-dessous, la matrice de confusion correspondante est représentée (tableau C.1).

Tab. C.1 : Résultats du modèle du rythme sur l'ensemble d'apprentissage du jeu 2 (correct : 59,0 ± 4,1 (330/559)).

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	51	2	12	1	-	8	6
Français	-	61	1	2	-	-	15
Allemand	8	-	65	-	-	6	1
Italien	1	9	-	30	-	1	39
Japonais	2	-	7	51	1	8	11
Mandarin	5	-	15	-	-	53	7
Espagnol	1	9	-	-	-	1	69

Reconnaissance

En reconnaissance, l'expérience est menée pour les paramètres donnant le meilleur résultat sur l'ensemble d'apprentissage, c'est-à-dire 8 gaussiennes. Le taux d'idenfication correcte est de 44,3 ± 8,2 % (62 identifications correctes sur 140 fichiers). La matrice de confusion est représentée ci-dessous (Tableau C.2).

Tab. C.2 : Résultats du modèle du rythme sur l'ensemble de test du jeu 2 (correct : 44,3 ± 8,2 (62/140)).

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	5	-	12	-	-	-	3
Français	-	20	-	-	-	-	-
Allemand	5	1	6	-	-	5	3
Italien	3	-	6	3	-	1	7
Japonais	-	-	2	8	-	-	10
Mandarin	2	-	2	4	-	11	1
Espagnol	-	3	-	-	-	-	17

C.1.2 Expériences sur les données du jeu n°3

Apprentissage

Les expériences ont montré que le meilleur taux d'identification pour l'ensemble d'apprentissage du jeu 3 est de 77,4 ± 4,4 %, soit 270 identifications correctes sur 349 fichiers. Ce résultat est obtenu pour 8 gaussiennes. Ci-dessous, la matrice de confusion correspondante est représentée (tableau C.3).

Tab. C.3 : Résultats du modèle du rythme sur l'ensemble d'apprentissage du jeu 3 (correct : 77,3 ± 4,4 (270/349)).

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	34	-	3	5	3	3	2
Français	-	47	-	-	-	-	2
Allemand	2	-	44	-	-	3	1
Italien	3	3	1	25	4	1	13
Japonais	3	-	-	4	37	1	5
Mandarin	5	-	5	1	5	33	1
Espagnol	-	-	-	-	-	-	50

Reconnaissance

En reconnaissance, l'expérience est menée pour les paramètres donnant le meilleur résultat sur l'ensemble d'apprentissage, c'est-à-dire 8 gaussiennes. Le taux d'identification correcte est de 54,3 ± 8,3 % (76 identifications correctes sur 140 fichiers). La matrice de confusion est représentée ci-dessous (Tableau C.4).

Tab. C.4 : Résultats du modèle du rythme sur l'ensemble de test du jeu 3 (correct : 54,3 ± 8,3 (76/140)).

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	6	-	7	-	1	1	5
Français	1	15	-	3	-	-	1
Allemand	4	-	12	-	-	4	-
Italien	-	1	-	6	-	-	13
Japonais	3	-	-	2	15	-	-
Mandarin	1	1	7	-	-	5	6
Espagnol	-	3	-	-	-	-	17

C.2 Expériences complémentaires avec le modèle de l'intonation (statique)

Les expériences de §4.5.4 sont reprises sur les deux autres jeux de données.

C.2.1 Expériences sur les données du jeu n°2

Apprentissage

Tab. C.5 : Résultats du modèle de l'intonation sur l'ensemble d'apprentissage du jeu 2 (correct : 73,3 ± 3,7 (410/559))

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	52	3	13	2	-	6	4
Français	1	73	-	1	2	2	-
Allemand	3	1	62	3	-	9	2
Italien	9	4	2	49	9	2	5
Japonais	2	3	-	3	67	3	2
Mandarin	11	4	10	1	2	51	1
Espagnol	4	8	3	3	1	5	56

Reconnaissance

Tab. C.6 : Résultats du modèle de l'intonation sur l'ensemble de test du jeu 2 (correct : 46,4 ± 8,3 (65/140))

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	7	-	-	-	-	11	2
Français	2	16	-	1	-	-	1
Allemand	2	-	14	3	-	-	1
Italien	-	3	-	4	1	9	3
Japonais	-	2	-	2	15	-	1
Mandarin	11	3	2	-	-	3	1
Espagnol	2	5	5	2	-	-	6

C.2.2 Expériences sur les données du jeu n°3

Apprentissage

Tab. C.7 : Résultats du modèle de l'intonation sur l'ensemble d'apprentissage du jeu 3 (correct : 78,8 ± 4,3 (275/349))

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	43	-	-	1	-	6	-
Français	1	42	-	1	1	4	-
Allemand	1	2	35	5	-	5	2
Italien	3	3	1	41	1	-	1
Japonais	-	3	-	2	44	1	-
Mandarin	5	2	3	1	-	37	2
Espagnol	2	6	-	4	2	3	33

Reconnaissance

Tab. C.8 : Résultats du modèle de l'intonation sur l'ensemble de test du jeu 3 (correct : 52,8 ± 8,3 (74/140))

	Anglais	Français	Allemand	Italien	Japonais	Mandarin	Espagnol
Anglais	6	2	5	2	-	3	2
Français	1	9	-	6	1	3	-
Allemand	1	1	14	-	-	2	2
Italien	1	1	-	16	2	-	-
Japonais	-	-	-	6	13	1	-
Mandarin	5	-	3	1	1	9	1
Espagnol	1	3	-	8	-	1	7

Annexe D

Expériences avec les modèles acoustiques

Sommaire

D.1 Modèles centiseconde 191
D.1.1 Modélisation acoustique globale 191
D.1.2 Modélisation des systèmes vocaliques 192
D.1.3 Modélisation des consonnes 193
D.1.4 Modélisation différenciée consonnes/voyelles 194
D.1.5 Modélisation des consonnes voisées 195
D.1.6 Modélisation des consonnes non voisées 196
D.1.7 Modélisation différenciée consonnes voisées/non voisées/voyelles 197
D.2 Modèles segmentaux 198
D.2.1 Modélisation acoustique globale 198
D.2.2 Modèle consonantique 199
D.2.3 Modélisation des systèmes vocaliques 200
D.2.4 Modélisation différenciée Consonnes/Voyelles 201
D.2.5 Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles 202

La modélisation acoustique proposée se base sur l'extraction de paramètres cepstraux (MFCC). Ces paramètres sont extraits soit de manière centiseconde (toutes les 10 ms), soit de manière segmentale, à partir de la segmentation automatique décrite plus haut (§ 4.2.1). Dans ce dernier cas, une seule observation est conservée pour chaque segment.

Grâce à l'algorithme de détection automatique des voyelles (§ 4.2.2) et à une mesure de voisement, sept modèles sont proposés prenant en compte différents espaces acoustiques :

1. modélisation de l'ensemble des sons sans distinction de classe,
2. modélisation des voyelles,
3. modélisation des consonnes,
4. modélisation des consonnes voisées,
5. modélisation des consonnes non voisées,
6. fusion des modélisations 2 et 3,
7. fusion des modélisations 2, 4 et 5.

Le protocole expérimental est le même que celui employé lors des différentes modélisations de la prosodie :

- Tout d'abord les modèles sont estimés à partir des données de l'ensemble d'apprentissage. Il s'agit ici de Modèles de Mélange de lois Gaussiennes (MMG). Ces modèles sont appris pour différents nombres de lois gaussiennes (2, 4, 8, 16, 32, 64, ...).
- Pour chaque dimension des MMG, des tests sont effectués en utilisant l'ensemble d'apprentissage comme ensemble de test. Cela permet de déterminer le nombre optimal de lois gaussiennes.
- Une fois que le nombre optimal de lois gaussiennes est déterminé, les tests sont effectués sur l'ensemble de test.

D.1 Modèles centiseconde

Les paramètres MFCC sont extraits toutes les 10 ms, sur des fenêtres de 20 ms avec des recouvrements de 10 ms. 12 MFCC sont calculés sur chaque trame, avec leurs dérivées premières et secondes. Pour chaque observation, nous obtenons un vecteur de dimension 36.

D.1.1 Modélisation acoustique globale

Toutes les observations sont prises en compte (sans distinction de classe sonore) pour l'apprentissage et pour le test. Ce système possède l'architecture décrite sur la figure 2.1.

Des tests sont tout d'abord effectués sur l'ensemble d'apprentissage afin de déterminer la meilleure typologie pour les modèles MMG. Les résultats pour différents nombres de lois gaussiennes dans le mélange sont résumés sur le tableau D.1.

Tab. D.1 : Modèle acoustique global centiseconde ;
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128
Taux d'identification correcte	82,6	91,3	94,0	97,5	99,0	100

Lors de la phase de reconnaissance, les expériences d'identification sont menées avec le modèle à 128 composantes. Les résultats sont donnés sur le tableau D.2.

Tab. D.2 : Modèle acoustique global centiseconde
Expériences sur l'ensemble de test : correct : 76,2 ± 7,1 (106/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	12	-	-	-	1	-	-
Allemand	5	14	-	-	-	-	-
Mandarin	1	-	16	3	-	-	-
Français	-	-	-	19	-	-	-
Italien	1	-	-	-	5	-	14
Espagnol	-	-	-	-	-	20	-
Japonais	-	-	-	-	-	-	20

L'italien est la langue la moins bien reconnue, elle est principalement confondue avec le japonais.

D.1.2 Modélisation des systèmes vocaliques

Ici, on ne considère que les coefficients extraits sur les segments étiquetés voyelle.

Les tests effectués sur l'ensemble d'apprentissage montrent que le meilleur taux d'identification (100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 16 gaussiennes. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.3.

Tab. D.3 : Modèle vocalique centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	94,1	98,9	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée pour 16 gaussiennes. Le taux d'identification correcte est de 49,6% (69 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.4).

Tab. D.4 : Modèle vocalique centiseconde
Expériences sur l'ensemble de test : correct : 49,6 ± 8,3 (70/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	7	-	-	-	12	-	1
Allemand	1	17	1	-	-	1	-
Mandarin	-	-	11	9	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	5	-	15
Espagnol	-	-	-	-	20	-	-
Japonais	-	5	-	-	-	5	10

On retrouve ici le même problème d'identification de l'italien, toujours confondu avec le japonais. D'autres confusions sont également remarquables, notamment entre l'anglais et l'italien et entre le mandarin et le français. Le japonais est confondu avec l'espagnol et l'allemand. L'espagnol est totalement confondu avec l'italien.

D.1.3 Modélisation des consonnes

Les tests effectués sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 400 identifications correctes sur 400 fichiers) est obtenu à partir de 8 gaussiennes. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.5.

Tab. D.5 : Modèle consonantique centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

En reconnaissance, l'expérience d'identification des langues est menée avec les modèles à 16 composantes. Le taux d'identification correcte est de 66,2 % (92 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.6).

Tab. D.6 : Modèle consonantique centiseconde
Expériences sur l'ensemble de test : correct : 66,2 ± 7,9 (92/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	4	-	-	-	16	-	-
Allemand	1	19	-	-	-	-	-
Mandarin	-	8	11	1	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	9	-	11
Espagnol	-	-	-	-	10	10	-
Japonais	-	-	-	-	-	-	20

Ici, les principales confusions se font entre l'anglais et l'italien, le mandarin et l'allemand, l'italien et le japonais, et l'italien et l'espagnol.

D.1.4 Modélisation différenciée consonnes/voyelles

Le modèle différencié est la fusion des modèles consonnes et voyelles. Les log-vraisemblances de chaque modèle sont normalisées et additionnées.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 565 identifications correctes sur 565 fichiers) est obtenu à partir de 8 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sut le tableau D.7.

Tab. D.7 : Modélisation différenciée consonnes/voyelles centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	97,2	100	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée avec les modèles à 16 composantes. Le taux d'identification correcte est de 54,7 % (76 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.8).

Tab. D.8 : Modélisation différenciée consonnes/voyelles centiseconde
Expériences sur l'ensemble de test : correct : 54,7 ± 10,3 (76/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	2	-	-	-	18	-	-
Allemand	2	18	-	-	-	-	-
Mandarin	-	2	13	5	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	4	-	16
Espagnol	-	-	-	-	14	6	-
Japonais	-	4	-	-	2	-	14

La combinaison des deux approches précédentes n'amène pas une amélioration des résultats mais une détérioration. Ici, l'anglais est confondu avec l'italien, l'italien avec le japonais, et l'espagnol avec l'italien.

D.1.5 Modélisation des consonnes voisées

Ici, seuls les segments consonantiques voisés sont pris en compte. Les consonnes sont dites voisées si plus de 70% de la durée de la consonne est voisée.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 400 identifications correctes sur 400 fichiers) est obtenu à partir de 16 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.9.

Tab. D.9 : Modélisation des consonnes voisées centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	99	100	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée pour les modèles à 16 composantes. Le taux d'idenfication correcte est de 76,2 %. La matrice de confusion est représentée ci-dessous (tableau D.10).

Tab. D.10 : Modélisation des consonnes voisées centiseconde
Expériences sur l'ensemble de test : correct : 76,2 (106/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	11	-	-	1	8	-	-
Allemand	3	15	-	-	-	-	2
Mandarin	-	1	19	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	12	-	8
Espagnol	-	-	1	-	9	10	-
Japonais	-	-	-	-	-	-	20

La performance s'améliore lorsque l'on ne considère que les consonnes voisées. Les confusions ont toujours lieu entre les mêmes langues, anglais-italien et italien-espagnol.

D.1.6 Modélisation des consonnes non voisées

Ici, seuls les segments consonantiques non voisés sont pris en compte.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 400 identifications correctes sur 400 fichiers) est obtenu à partir de 16 composantes. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.11.

Tab. D.11 : Modélisation des consonnes non voisées centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte
en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	99	100	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée pour les modèles à 16 composantes. Le taux d'idenfication correcte est de 53,9 %. La matrice de confusion est représentée ci-dessous (tableau D.12).

Tab. D.12 : Modélisation des consonnes non voisées centiseconde
Expériences sur l'ensemble de test : correct : 53,9 (75/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	3	-	-	-	17	-	-
Allemand	2	17	-	-	-	1	-
Mandarin	-	-	13	7	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	6	-	14
Espagnol	-	-	-	-	16	4	-
Japonais	-	5	-	-	1	1	13

Les modèles conçus à partir des consonnes non voisées ne permettent pas de grandes performances. Les mêmes confusions que précédemment sont toujours retrouvées (anglais-italien, italien-japonais, espagnol-italien).

D.1.7 Modélisation différenciée consonnes voisées/non voisées/voyelles

Les 3 modules précédents sont fusionnés au moyen d'une somme normalisée des vraisemblances.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 400 identifications correctes sur 400 fichiers) est obtenu à partir de 16 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.13.

Tab. D.13 : Modélisation différenciée consonnes voisées/non voisées/voyelles centiseconde
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	99	100	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée pour les modèles à 16 composantes. Le taux d'idenfication correcte est de 59,7 ± 8,2 %. La matrice de confusion est représentée ci-dessous (tableau D.14).

Tab. D.14 : Modélisation différenciée consonnes voisées/non voisées/voyelles centiseconde
Expériences sur l'ensemble de test : correct : 59,7 ± 8,2 (83/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	3	-	-	-	17	-	-
Allemand	3	17	-	-	-	-	-
Mandarin	-	1	15	4	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	7	-	13
Espagnol	-	-	-	-	13	7	-
Japonais	-	1	-	-	4	-	15

La fusion des approches Consonnes voisées/non voisées et Voyelles ne permet pas d'obtenir de très bonnes performances. Les confusions sont toujours entre l'anglais et l'italien, l'italien et le japonais, et l'espagnol et l'italien

D.2 Modèles segmentaux

Ici, les paramètres MFCC sont extraits sur chaque segment. Ils sont calculés sur une fenêtre centrée au milieu du segment. Nous calculons 8 MFCC par segment, ainsi que leurs dérivées. Pour chaque segment, un vecteur de dimension 16 est ainsi obtenu.

D.2.1 Modélisation acoustique globale

Ici, tous les segments sont pris en compte sans différencier les consonnes des voyelles.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 8 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.15.

Tab. D.15 : Modèle acoustique global segmental
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	100	100	100	100	100	100	100	100

Le taux d'identification correcte est de 87,8 % (122 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.16).

Tab. D.16 : Modèle acoustique global segmental
Expériences sur l'ensemble de test : correct : 87,8 ± 5,5 (122/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	15	-	-	-	5	-	-
Allemand	-	20	-	-	-	-	-
Mandarin	-	-	20	-	-	-	-
Français	-	-	-	17	-	2	-
Italien	2	-	1	-	13	1	2
Espagnol	1	-	-	2	-	17	-
Japonais	-	-	-	-	-	-	20

Le modèle segmental permet d'obtenir de meilleures performances que le modèle centiseconde. Ici, seules de légères confusions sont observables, entre l'anglais et l'italien principalement.

D.2.2 Modèle consonantique

Ici, seuls les segments automatiquement étiquetés "consonne" sont pris en compte.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 16 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.17.

Tab. D.17 : Modèle consonantique segmental
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	99.1	99.8	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée pour les modèles à 32 composantes. Le taux d'identification correcte est de 99,3% (138 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.18).

Tab. D.18 : Modèle consonantique segmental
Expériences sur l'ensemble de test : correct : 99,3 ± 1,4 (138/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	20	-	-	-	-	-	-
Allemand	-	20	-	-	-	-	-
Mandarin	-	-	20	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	20	-	-
Espagnol	-	-	-	-	1	19	-
Japonais	-	-	-	-	-	-	20

Le modèle consonantique permet d'obtenir de très bons résultats, seul un fichier d'espagnol est confondu avec l'italien.

D.2.3 Modélisation des systèmes vocaliques

Ici, on ne considère que les coefficients extraits sur les segments étiquetés "voyelle".

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 32 composantes pour les MMG. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.19.

Tab. D.19 : Modèle vocalique segmental
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	4	8	16	32	64	128	256	512
Taux d'identification correcte	90,9	96,1	99,1	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée avec les modèles à 32 composantes. Le taux d'identification correcte est de 69,8 % (97 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.20).

Tab. D.20 : Modèle vocalique segmental
Expériences sur l'ensemble de test : correct : 69,8 ± 7,6 (97/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	9	1	1	-	7	1	1
Allemand	-	20	-	-	-	-	-
Mandarin	-	6	4	4	5	1	-
Français	-	-	-	19	-	-	-
Italien	-	3	-	3	11	3	-
Espagnol	-	-	-	2	-	18	-
Japonais	-	-	-	1	3	-	16

Le modèle vocalique ne permet pas d'obtenir d'aussi bonnes performances que le modèle consonnantique. On retrouve les confusions entre l'anglais et l'italien. Le mandarin est confondu avec toutes les autres langues sauf l'anglais et le japonais. L'italien est principalement confondu avec le français et l'espagnol.

D.2.4 Modélisation différenciée Consonnes/Voyelles

Le module différencié fait la fusion entre les modules consonnes et voyelles. Les log-vraisemblances normalisées sont additionnées.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 32 gaussiennes. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.21.

Tab. D.21 : Modélisation différenciée Consonnes/Voyelles segmentale
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

Nombre de composantes	2	4	8	16	32	64	128	256	512
Taux d'identification correcte	99	100	100	100	100	100	100	100	100

En reconnaissance, l'expérience d'identification des langues est menée avec les modèles à 32 composantes. Le taux d'idenfication correcte est de 99,3 % (138 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.22).

Tab. D.22 : Modélisation différenciée Consonnes/Voyelles segmentale
Expériences sur l'ensemble de test : correct : 99,3 ± 1,41 (138/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	20	-	-	-	-	-	-
Allemand	-	20	-	-	-	-	-
Mandarin	-	-	20	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	20	-	-
Espagnol	-	-	-	-	1	19	-
Japonais	-	-	-	-	-	-	20

Lorsque l'on effectue la fusion des deux approches précédentes, on retrouve les mêmes résultats que pour le modèle consonantique seul.

D.2.5 Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles

Le module différencié est la fusion entre les modules consonnes voisées, consonnes non voisées et voyelles. Les log-vraisemblances normalisées sont additionnées.

Les expériences sur l'ensemble d'apprentissage montrent que le meilleur résultat (taux d'identification de 100 %, soit 560 identifications correctes sur 560 fichiers) est obtenu à partir de 32 composantes. Les taux d'identification en fonction du nombre de gaussiennes sont donnés sur le tableau D.23.

Tab. D.23 : Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles segmentale
Expériences sur l'ensemble d'apprentissage : taux d'identification correcte en fonction du nombre de composantes gaussiennes

En reconnaissance, l'expérience d'identification des langues est menée avec les modèles à 32 composantes. Le taux d'idenfication correcte est de 98,5 % (137 identifications correctes sur 139 fichiers). La matrice de confusion est représentée ci-dessous (tableau D.24).

Tab. D.24 : Modélisation différenciée Consonnes voisées/Consonnes non voisées/Voyelles segmentale
Expériences sur l'ensemble de test : correct : 98,5 ± 2,0 (137/139)

	Anglais	Allemand	Mandarin	Français	Italien	Espagnol	Japonais
Anglais	19	-	-	-	1	-	-
Allemand	-	20	-	-	-	-	-
Mandarin	-	-	20	-	-	-	-
Français	-	-	-	19	-	-	-
Italien	-	-	-	-	20	-	-
Espagnol	-	-	-	-	1	19	-
Japonais	-	-	-	-	-	-	20

Annexe E

Algorithme VQ (Quantification Vectorielle)

Sommaire

E.1 Objectif 207
E.2 Algorithme des K-means 207
E.3 Algorithme LBG (Linde, Buzo, Gray) 208

E.1 Objectif

La quantification vectorielle consiste à extraire un « dictionnaire » de « prototypes » (ensemble des centroïdes) d'un grand ensemble représentatif de données. Le dictionnaire doit respecter le mieux possible leur répartition dans l'espace.

La première version de l'algorithme de construction du dictionnaire pour la quantification est connue sous le nom de Lloyd [80] et fut utilisée pour la quantification scalaire. Cet algorithme a ensuite été généralisé pour la classification automatique et la reconnaissance des formes sous le nom d'algorithme des « K-means » ou méthode des « nuées dynamiques » [34].

E.2 Algorithme des K-means

$(y_n),\ 0 \leq n \leq N$ représente un nuage de points (observations) de R^d, d est la distance euclidienne et la taille du dictionnaire K est fixée.

1. **Initialisation**
   Soit un dictionnaire D_0 de taille K.

2. **Construction de la partition**
   A la tième itération, le dictionnaire est noté :

$$D_t = \{D_{i,t}\}_{i=1,\cdots,K} \tag{E.1}$$

   La partition qui minimise l'erreur de quantification associée à D_t est composée des classes :

$$C_{i,t} = \{y_n / d(y_n, D_{i,t}) \leq d(y_n, D_{j,t}),\ j \neq i\} \tag{E.2}$$

   L'erreur de quantification vaut :

$$Dis_t = \frac{1}{N} \sum_{n=1}^{N} \left[\min_{i=1}^{K} d(y_n, \mu_{i,t}) \right] \tag{E.3}$$

   où μ_{it} est le centroïde de C_{it}.

3. **Test d'arrêt**
   Si $Dis_{t-1} - Dis_t)/Dis_t < \epsilon$ alors l'algorithme est terminé. Le dictionnaire recherché est D_{t+1} composé des nouveaux centroïdes, soit :

$$D_{i,t+1} = \mu_{i,t} \tag{E.4}$$

   Sinon $t = t + 1$ et l'algorithme est repris à l'étape 2.

Puisque cet algorithme n'est que localement optimal, le choix du dictionnaire de départ est important. Une variante très utilisée de l'algorithme de Lloyd est l'algorithme LBG [79] : il procède hiérarchiquement et réalise une sorte d'initialisation itérative au cours de la construction.

E.3 Algorithme LBG (Linde, Buzo, Gray)

Le but est de construire un dictionnaire de taille K, où $K = 2^p$.

1. **Initialisation**
   Le centre de gravité de l'ensemble d'apprentissage est calculé.
   Soit d_0 ce vecteur. Le dictionnaire est constitué de d_0, $p = 0$.

$$D_0 = \{d_0\}, \quad |D_0| = 2^p \tag{E.5}$$

2. **Éclatement « Splitting »**
   Tous les éléments d en nombre 2^k du dictionnaire sont « éclatés » en deux vecteurs. Ceci se fait par exemple en transformant chaque d en $d + \epsilon$ et $d - \epsilon$, où ϵ est un vecteur aléatoire de variance adaptée aux points du nuage associés à d.

3. **Convergence**
   L'algorithme de Lloyd (cf. section précédente) est appliqué sur le dictionnaire des 2^{k+1} éléments ainsi constitué. Après convergence un dictionnaire « optimal » de 2^{k+1} éléments est obtenu.

4. **Arrêt**
   $k = k + 1$.
   Si $k > k_0$ fixé à l'avance, alors l'algorithme prend fin, sinon le processus est itéré (2).

   Le test d'arrêt peut se faire aussi par rapport à un seuil minimal sur la distorsion des données d'apprentissage par rapport au dictionnaire, comme dans le cas de l'algorithme de Lloyd.

Annexe F

Algorithme EM (Expectation Maximisation)

Sommaire

F.1 Rappels 213
F.2 Algorithme de base 213

F.1 Rappels

L'expression de la vraisemblance d'une observation y de l'ensemble d'apprentissage, supposée réalisation d'un modèle de mélanges de lois gaussiennes, est donnée par :

$$\sum_{k=1}^{N} \nu_k . N(y, \mu_k, \Sigma_k) \tag{F.1}$$

avec :

$$N(y, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \; exp^{\left[-\frac{1}{2}(y-\mu_k)^t \Sigma_k^{-1} (y-\mu_k)\right]} \tag{F.2}$$

et :
N le nombre de composantes du mélange,
ν_k le poids de chaque composante,
μ_k la moyenne de chaque composante,
Σ_k la matrice de covariance associée.

L'algorithme EM est basé sur la vraisemblance de chaque vecteur observé par rapport à chaque composante gaussienne du modèle.

F.2 Algorithme de base

1. **Initialisation (t=0)**

   - Initialisation des moyennes μ_k par N points extraits aléatoirement de l'ensemble des observations X. $X = \{x_1, \cdots, x_N\}$.

   - Initialisation de toutes les matrices de covariance Σ_k à la matrice unité I_p.

   - Initialisation équiprobable des poids des composantes : $\nu_k = 1/N$.

   OU

   - Utilisation de l'algorithme VQ (Quantification Vectorielle) présenté dans l'annexe E pour l'initialisation.

2. **Itération (t)**
   Pour tout $k = 1, \cdots, N$

   – *Phase d'estimation*

   Calcul de la probabilité P_{nk} que le vecteur y_n soit généré par la loi gaussienne k.

$$P_{nk} = \frac{\frac{\nu_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \; exp\left[-\frac{1}{2}(y_n-\mu_k)^t\Sigma_k^{-1}(y_n-\mu_k)\right]}{\sum_{k'=1}^{K} \frac{\nu_{k'}}{(2\pi)^{d/2}|\Sigma_{k'}|^{1/2}} \; exp\left[-\frac{1}{2}(y_n-\mu_{k'})^t\Sigma_{k'}^{-1}(y_n-\mu_{k'})\right]} \tag{F.3}$$

   – *Phase de maximisation*

   Réestimation des paramètres à partir des probabilités P_{nk} :

$$\overline{\nu_k} = \frac{1}{N}\sum_{n=1}^{N} P_{nk} \tag{F.4}$$

$$\overline{\mu_k} = \frac{\sum_{n=1}^{N} P_{nk} y_n}{\sum_{n=1}^{N} P_{nk}} \tag{F.5}$$

$$\overline{\Sigma_k} = \frac{\sum_{n=1}^{N} P_n k (y_n - \overline{\mu_k})(y_n - \overline{\mu_k})^t}{\sum_{n=1}^{N} P_{nk}} \tag{F.6}$$

   – *Incrémentation de t à $t+1$ et retour à la phase d'estimation*

3. **Arrêt de l'algorithme**
   Calcul de la vraisemblance des observations (y_n).
   Si la variation de la vraisemblance descend en dessous d'un seuil fixé, alors l'estimation est terminée, sinon l'estimation est reprise à l'étape 2.

Annexe G

Corpus Ogi-mlts

Sommaire

G.1 Protocole expérimental 216
G.2 Étiquetage phonétique 217

Le corpus *Multi-Language Telephone Speech* [93] a été développé à l'*Oregon Graduate Institute of Science and Technology* par Muthusamy lors de son travail de doctorat.

G.1 Protocole expérimental

La collecte des données est réalisée de manière automatique par enregistrement d'appels téléphoniques. Le signal est échantillonné à 8000 Hz, digitalisé sur 14 bits et compressé en utilisant l'algorithme « shorten ». Le locuteur compose un numéro de téléphone qui le connecte au système.Un premier message d'introduction en langue anglaise lui permet de choisir sa langue en pressant une touche du clavier téléphonique. Le reste des instructions est donné dans la langue choisie. Plusieurs questions sont posées permettant l'enregistrement de différents types d'énoncés. Les questions sont :

1. Quelle est votre langue maternelle ?
2. Quelle est la langue que vous parlez le plus couramment ?
3. Citez les jours de la semaine.
4. Veuillez compter de 0 à 10.
5. Racontez une anecdote sur votre ville.
6. Quel est le climat de votre ville ?
7. Décrivez la pièce d'où vous appelez.
8. Décrivez votre dernier repas.
9. Racontez pendant une minute une histoire de votre choix.

Afin de ne pas couper l'énoncé de la question 9 en plein discours, le locuteur est averti par un signal sonore au bout de 50 s.

Tab. G.1 : Codage des fichiers dans OGI-MLTS.

Numéro de la question	Codage	Type de parole	Durée
1	nlang	mot cible	3
2	clang	mot cible	3
3	dowk	énumération	8
4	nums	énumération	10
5	htl	spontanée	10
6	htc	spontanée	10
7	room	spontanée	12
8	meal	spontanée	10
9 (avant signal)	story-bt	spontanée	50
9 (après signal)	story-at	spontanée	10

G.2 Étiquetage phonétique

Un étiquetage phonétique a été réalisé par des experts phonéticiens pour six langues du corpus OGI-MLTS : l'anglais, l'allemand, l'espagnol, l'hindi, le japonais et le mandarin. Les conventions d'annotation sont disponibles dans le guide du CSLU. Le tableau G.2 liste les locuteurs de chaque langue disposant d'une transcription phonétique. Chaque fichier est de type *story-bt*. Le tableau G.3 indique la durée totale des fichiers audio pour chaque langue.

Tab. G.2 : Liste des fichiers disposant d'une annotation phonétique sur OGI-MLTS.

langue	locuteurs
anglais	5 4 6 8 9 3 11 12 19 18 13 100 105 107 106 109 103 108 112 115 114 117 113 116 118 120 123 125 121 124 127 128 130 133 131 135 134 137 138 139 140 144 146 149 148 147 150 153 152 157 154 151 163 162 164 165 166 167 169 171 170 173 174 175 176 177 178 182 183 184 185 186 181 187 188 20 27 24 28 22 29 202 203 204 205 206 207 200 211 247 30 31 32 38 34 35 37 33 41 42 43 40 45 47 44 48 50 53 51 52 54 57 56 58 59 60 62 63 65 64 66 69 68 61 70 72 77 74 76 73 79 78 83 82 84 86 81 87 88 92 93 94 96 97 98 99 90
allemand	2 3 4 5 6 7 9 1 11 12 14 15 16 18 19 10 101 102 106 109 100 114 116 118 113 123 124 125 127 129 120 134 136 137 138 139 130 141 144 145 148 140 151 152 153 156 157 150 23 24 26 27 28 22 33 34 36 37 38 39 31 41 42 44 45 46 47 40 51 53 56 57 58 59 50 61 63 69 60 72 74 75 77 78 79 70 81 83 85 86 87 88 89 80 91 93 94 95 97 99 90
espagnol	2 3 4 5 6 8 1 12 13 14 15 16 17 18 19 10 102 103 104 105 106 107 108 100 113 115 117 118 119 110 121 122 124 126 127 120 137 130 143 147 149 142 22 23 24 25 26 27 28 29 20 31 32 33 35 36 37 38 30 41 44 46 47 48 49 40 51 52 53 54 55 56 57 50 62 63 64 65 67 68 69 60 71 72 73 76 77 78 79 70 81 82 83 84 85 87 88 89 80 91 93 94 95 96 97 98 99 90
hindi	14 13 104 106 108 109 102 112 126 127 128 125 132 134 136 130 138 142 143 147 140 154 156 159 151 162 164 168 160 174 175 177 170 181 182 180 21 25 20 32 48 52 55 56 57 58 51 67 68 69 63 72 74 75 77 78 79 70 83 84 86 88 80 96 97 98 99 93
japonais	2 3 7 1 15 17 19 13 100 118 121 122 124 126 127 129 120 133 135 136 137 138 139 131 141 140 22 23 24 25 27 29 20 36 35 47 48 40 51 53 54 55 57 58 50 61 62 65 66 67 68 69 60 72 73 75 71 82 83 85 86 88 80 90
mandarin	9 1 13 14 15 16 18 11 101 105 106 107 108 109 100 118 119 113 122 123 124 126 127 129 121 23 24 27 21 31 33 34 35 36 37 39 30 41 42 43 44 46 48 49 40 52 53 55 56 57 58 59 51 65 67 68 69 60 76 77 78 79 73 86 83 92 93 97 98 90

Tab. G.3 : Durée totale des fichiers disposant d'une transciption phonétique sur OGI-MLTS.

langue	durée totale
anglais	2 h 2 mn
allemand	1 h 24 mn
espagnol	1 h 30 mn
hindi	56 mn
japonais	53 mn
mandarin	58 mn

Bibliographie

[1] D. Abercrombie (rédacteur) : *Elements of General Phonetics*, Edinburgh University Press, Edinburgh, 1967

[2] A. Adami, R. Mihaescu, D. A. Reynolds et J. Godfrey : "Modeling Prosodic Dynamics for Speaker Recognition", dans *International Conference on Acoustics, Speech and Signal Processing*, tome 4, p. 788–791, Hong Kong, China, 2003

[3] A. G. Adami et H. Hermansky : "Segmentation of Speech for Speaker and Language Recognition", dans *Eurospeech*, p. 841–844, Geneva, 2003

[4] G. D. Allen : "Speech Rythm : its relation to performance universals and articulatory timing", *Journal of Phonetics*, p. 75–86, 1975

[5] R. André-Obrecht : "A New Statistical Approach For Automatic Speech Segmentation", *IEEE Transanctions on Acoustics, Speech and Signal Processing*, 36(1), p. 29–40, 1988

[6] R. André-Obrecht : *Segmentation et Parole ?*, Habilitation à diriger des recherches, 1993

[7] R. André-Obrecht et B. Jacob : "Direct Identification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition", dans *International Conference on Acoustics, Speech and Signal Processing*, p. 989–992, IEEE, Munich, 1997

[8] F. Antoine, D. Zhu, P. Boula de Mareüil et M. Adda-Decker : "Approches Segmentales multilingues pour l'identification automatique de la langue : phones et syllabes", dans *Journées d'Etude de la Parole*, Fes, Maroc, 2004

[9] V. Aubergé : *La synthèse de la parole : des règles au lexique*, Thèse de doctorat, Université Stendhal, Grenoble, France, 1991

[10] O. Bagou, C. Fougeron et U. Frauenfelder : "Contribution of Prosody to the Segmentation and Storage of "Words" in the Acquisition of a New Mini-Language", dans *Speech Prosody 2002*, p. 159–162, Aix-en-Provence, Avril 2002

[11] P. C. Bagshaw : *Automatic Prosodic Analysis For Computer Aided Pronunciation Teaching*, Thèse de doctorat, Edinburgh University, 1994

[12] P. A. Barbosa : *Caractérisation et génération automatique de la structuration rythmique du français*, Thèse de doctorat, Institut National Polytechnique, Grenoble, France, 1994

URL `http://www.icp.inpg.fr/~bailly/publis/synthese/_pb/these_pb_PHD94.ps`

[13] M. Barkat-Defradas, I. Vasilescu et F. Pellegrino : "Stratégies perceptuelles et identification automatique des langues", *PArole*, 25-26, 2003

[14] M. E. Beckman : "Evidence for speech rhythms across languages", dans *Speech perception, production and linguistic structure* (Y. Tohkura, E. Vatikiotis-Bateson et Y. Sagisaka, rédacteurs), p. 457–463, 1992

[15] L. Benarousse et E. Geoffrois : "Preliminary Experiments On Language Identification Using Broadcast News Recording", dans *Eurospeech*, Aalborg, Denmark, 2001

[16] F. Bimbot, R. Pieraccini, E. Levin et B. Atal : "Modèles de séquences à horizon variable : Multigrammes", dans *XXèmes Journées d'Étude sur la Parole*, Trégastel, France, juin 1994

[17] F. Bimbot, R. Pieraccini, E. Levin et B. Atal : "Variable-length sequence modeling : Multigrams", *IEEE Signal Processing Letters*, 2(6), juin 1995

[18] W. M. Campbell : "A SVM/HMM system for speaker recognition", dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003)*, Hong Kong, China, 2003

[19] W. N. Campbell : *Multi-level timing in speech*, Thèse de doctorat, University of Sussex, UK, 1992

[20] E. Campione et J. Véronis : "A multilingual prosodic database", dans *International Conference on Spoken Language Processing*, Sidney, Australia, http ://www.lpl.univ-aix.fr/projects/multext, 1998

[21] D. Caseiro et I. Trancoso : "Spoken Language Identification Using The SpeechDat Corpus", dans *International Conference on Spoken Language Processing*, Sydney, Australia, December 1998

[22] D. Chan et al. : "EUROM : A Spoken Language Ressource for the E.U.", dans *4th European Conference on Speech Communication and Technology*, Madrid, Espagne, 1995

[23] F. Cummins : "Speech Rhythm and Rhythmic Taxonomy", dans *Speech Prosody*, p. 121–126, Aix-en-Provence, France, 2002

[24] F. Cummins, F. Gers et J. Schimdhuber : "Language Identification From Prosody Without Explicit Features", dans *Eurospeech*, p. 371–374, Budapest, Hungary, 1999

[25] F. Cummins, F. Gers et J. Schmidhuber : "Comparing Prosody Across Many Languages", Technical report idsia-07-99, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, CH, 1999

[26] C. d'Alessandro et P. Mertens : "Automatic pitch contour stylisation using a model of tonal perception", *Computer Speech and Language*, 9, p. 257–288, 1995

[27] R. M. Dauer : "Stress-timing and Syllable-timing Reanalysed", *Journal of Phonetics*, 11, p. 51–62, 1983

[28] P. Delattre : “Les dix intonations de base du français”, *The French review*, 40(1), p. 1–14, 1966

[29] S. Deligne : *Modèles de séquences de longueur variable : application au traitement du langage écrit et de la parole*, Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, France, 1996

[30] S. Deligne et F. Bimbot : “Language modeling by variable length sequences : theorical formulation and evaluation of multigrams”, dans *IEEE 20th International Conference on Acoustics Speech and Signal Processing*, 1995
URL `http://tsi.enst.fr/~speech/Document/Publications/deligne.icassp.95.ps`

[31] V. Dellwo et P. Wagner : “Relations Between Language Rhythm and Speech Rate”, dans *International Congress of the Phonetic Sciences*, p. 471–474, Barcelona, Spain, 2003

[32] A. DiCristo : “Interpréter la prosodie”, dans *Journées d'Etude de la Parole*, Aussois, France, juin 2000

[33] A. DiCristo : “La problématique de la prosodie dans l'étude de la parole dite spontanée”, *revue PArole*, 2000

[34] E. Diday : “Cluster analysis”, dans *Digital Pattern Recognition*, p. 47–94, Springer Verlag

[35] M. Dutat : *Caractérisation de la langue parlée par modèles de séquences d'événements acoustiques*, Thèse de doctorat, École Nationale Supérieure des Télécommunications, Paris, France, décembre 2000

[36] N. Fakotakis, K. Georgila et A. Tsopanoglou : “A Continuous HMM Text-Independent Speaker Recognition System Based on Vowel Spotting”, dans *5th European Conference on Speech Communication and Technology (Eurospeech)*, tome 5, p. 2247–2250, Rhodes, Greece, September 1997

[37] J. Farinas : *Une modélisation automatique du rythme pour l'identification automatique des langues*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 2002

[38] J. Farinas et F. Pellegrino : “Automatic Rhythm Modeling For Language Identification”, dans *Eurospeech*, Aalborg, Denmark, 2001

[39] J. Farinas, F. Pellegrino, J.-L. Rouas et R. André-Obrecht : “Merging Segmental And Rhythmic Features For Automatic Language Identification”, dans *International Conference on Acoutics, Speech and Signal Processing*, tome 1, p. 753–756, Orlando, Florida, 2002

[40] J. Farinas, J.-L. Rouas, F. Pellegrino et R. André-Obrecht : “Extraction automatique de paramètres prosodiques pour l'identification automatique des langues”, *Traitement du Signal*, Soumis

[41] S. Frota, M. Vigario et F. Martins : “Language Discrimination and rhythm classes : evidence from portuguese”, dans *Speech Prosody*, Aix-en-Provence, France, 2002

[42] H. Fujisaki : "Prosody, Information and Modeling - with Emphasis on Tonal Features of Speech", dans *ISCA Workshop on Spoken Language Processing*, Mumbai, India, January 2003

[43] H. Fujisaki, P. Hallé et H. Lei : "Application of F0 contour command-response model to Chinese tones", *Reports of Autumn Meeting, Acoustical Society of Japan*, 1, p. 197–198, 1987

[44] H. Fujisaki et K. Hirose : "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *Journal of the Acoustical Society of Japan*, 5, p. 233–242, 1984

[45] H. Fujisaki et S. Ohno : "Application of the command-response model to the analysis, interpretation, and synthesis of fundamental frequency contours of speech of various languages", *Journal of the Acoustical Society of America*, 1999

[46] H. Fujisaki, S. Ohno et T. Yagi : "Analysis and modeling of fundamental frequency contours of greek utterances", dans *5th European Conference on Speech Communication and Technology (Eurospeech)*, tome 5, Rhodes, Greece, September 1997

[47] A. Galves, J. Garcia, D. Duarte et C. Galves : "Sonority as a Basis for Rhythmic Class Discrimination", dans *Speech Prosody*, Aix en Provence, France, April 2002

[48] E. Garding : "A Generative Model Of Intonation", dans *Prosody : Models And Measurements* (A. Cutler et D. Ladd, rédacteurs), p. 11–25, 1983

[49] J.-L. Gauvain, A. Messaoudi et H. Schwenk : "Language recognition using phone lattices", dans *International Conference on Spoken Language Processing*, Jeju island, Korea, 2004

[50] E. Grabe et E. L. Low : "Durational Variability in Speech and the Rhythm Class Hypothesis", *Papers in Laboratory Phonology 7*, 2002

[51] E. Grabe, F. Nolan et K. Farrar : "IViE - A comparative transcription system for intonational variation in English", dans *International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998

[52] E. Grabe, B. Post, F. Nolan et K. Farrar : "Pitch accent realisation in four varieties of British English", *Journal of Phonetics*, 28, p. 161–185, 2000

[53] S. Greenberg : "Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation", dans *ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, p. 47–56, Kekrade, Pays-Bas, 1998

[54] M. Grice et al. : "Consistency intranscription and labelling of german intonation with GToBI", dans *International Conference on Spoken Language Processing (ICSLP)*, 1995

[55] B. F. Gring (rédacteur) : *Language of the world*, tome 1, 14^{e} édition, Summer Inst of Lnguistics, novembre 2000

[56] C. Gussenhoven : "Discreteness and gradience in intonational contrasts", *Language and Speech*, 42 (2-3), p. 283–305, 1999

[57] H. Hermansky : “Perceptual Linear Predictive (PLP) analysis of speech”, *Journal of Acoustical Society of America*, 87(4), p. 1738–1752, juin 1990

[58] W. Hess : *Pitch Determination Of Speech Signals - Algorithms And Devices*, Springer-Verlag, 1983

[59] J. Hieronymus et S. Kadambe : “Spoken Language Identification Using Large Vocabulary Speech Recognition”, dans *International Conference on Spoken Language Processing*, 1996

[60] D. Hirst, A. Di Cristo et R. Espesser : “Levels of representation and levels of analysis for intonation”, dans *Prosody : Theory and Experiment* (M. Horne, rédacteur), Kluwer, Dordrecht, 2000

[61] D. Hirst, P. Nicolas et R. Espesser : “Coding the F0 of a continuous text in french : an experimental approach”, dans *XIIth International Conference on Phonetic Sciences*, p. 234–237, Aix-en-Provence, 1991

[62] S. Hochreiter et J. Schmidhuber : “Long short-term memory”, *Neural Computation*, 9(8), p. 1735–1780, 1997

[63] A. W. Howitt : “Vowel Landmark Detection”, dans *6th International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000

[64] H. Hurst, R. Black et Y. Simaika : *Long-term storage : an experimental study*, Constable and Co. Ldt., London, 1965

[65] IRIT-ICP-ILPGA-DDL : “Discrmination Multilingue Automatique : Première tentative de classification des langues par le prosodie”, Rapport technique, IRIT-ICP-ILPGA-DDL, 1997

[66] S. Itahashi, K. Kiuchi et M. Yamamoto : “Spoken Language Discrimination Using Speech Fundamental Frequency And Cepstra”, dans *Eurospeech*, Budapest, Hungary, 1999

[67] S. Kitazawa : “Periodicity of japanese accent in continuous speech”, dans *Speech Prosody*, Aix en Provence, France, April 2002

[68] D. H. Klatt : “Linguistic uses of segmental duration in English : acoustic and perception evidence”, *Journal of Acoustical Society of America*, 59, p. 1208–1221, 1976

[69] K. J. Kohler : dans *Invariance and variability in Speech Processes* (J. Perkell et D. H. Klatt, rédacteurs), chapitre Invariability and variability in speech timing : from utterance to segment in German, p. 268–298, Erlbaum Hillsdale, 1987

[70] M. Komatsu, T. Arai et T. Sugawara : “Perceptual discrimination of prosodic types”, dans *Speech Prosody*, p. 725–728, Nara, Japan, 2004

[71] I. Kopecek : “Syllable Based Approach to Automatic Prosody Detection ; Applications for Dialogue Systems”, dans *Proceedings of the Workshop on Dialogue and Prosody*, Eindhoven, Pays-Bas, septembre 1999

[72] A. Lacheret-Dujour et F. Beaugendre : *La prosodie du français*, CNRS Langage, Paris, France, 1999

[73] R. D. Ladd : “”Linear” and ”overlay” descriptions : an autosegmental-metrical middle way”, dans *XIIIth International Congress of Phonetics Sciences*, tome 2, p. 116–123, Stockholm, Suède, août 1995

[74] P. Ladefoged (rédacteur) : *The intonation of American English*, University of Michigan Press, Michigan, USA, 1945

[75] T. Lander et J. L. Hieronymus : “The CSLU labeling guide”, Rapport technique, Center for Spoken Language Understanding - Oregon Graduate Institute, 1997

[76] V.-F. Leavers et C. E. Burley : “The Use Of Cognitive Processing Strategies And Linguistic Cues For Efficient Automatic Language Identification”, *Language Sciences*, 2001

[77] I. Lehiste : “Isochrony reconsidered”, *Journal of Phonetics*, 5, p. 253–263, 1977

[78] K. P. Li : “Automatic language Identification Using Syllabic Spectral Features”, dans *International Conference on Acoustics, Speech and Signal Processing*, p. 297–300, Adelaide, 1994

[79] Y. Linde, A. Buzo et R. M. Gray : “An algorithm for Vector Quantizer design”, *IEEE Transaction on Communications*, 28(1), p. 84–95, 1980

[80] S. Lloyd : “Least Squares Quantization in PCM ?s”, Rapport technique, Bell Telephone Laboratories Papers, 1957

[81] P. MacNeilage : “The frame/content theory of evolution of speech production”, *Behavioral and Brain Sciences*, 21, p. 499–511, 1998
URL `ftp://ftp.princeton.edu/pub/harnad/BBS/.WWW/bbs.macneilage.html`

[82] B. Mandelbrot et J. Wallis : “Some Long-Run Properties of Geophysical Records”, *Water Resources Research*, 5(2), p. 321–340, 1969

[83] S. M. Marcus : “Perceptual centers (P-centers)”, dans *9th International Congress of Phonetic Sciences*, p. 238, Copenhagen, Danemark, 1979

[84] P. Boula de Mareüil, C. Corredor-Ardoy et M. Adda-Decker : “Multi-Lingual Automatic Phoneme Clustering”, dans *International Congress of Phonetic Sciences*, 1999

[85] A. F. Martin et M. A. Przybocki : “NIST 2003 Language Recognition Evaluation”, dans *Eurospeech*, p. 1341–1344, Geneva, 2003

[86] P. Martin : “ToBi : l’illusion scientifique ?”, dans *Journées Prosodie*, Grenoble, France, 2001

[87] B. Möbius, M. Pätzold et W. Hess : “Analysis and synthesis of German F0 contours by means of Fujisaki’s model”, *Speech Communication*, 13(1-2), p. 53–61, 1993

[88] J. Mehler, E. Dupoux, T. Nazzi et D.-L. G. : “Signal to syntax : bootstrapping from speech to grammar in early acquisition”, dans *Coping with linguistic diversity : the infant’s viewpoint* (J. Morgan et K. Demuth, rédacteurs), 1996

[89] P. Mertens : “The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Percption Model”, dans *Speech Prosody 2004*, 2004

[90] P. Monnin et F. Grosjean : "Les structures de performance en français : caractérisation et prédiction", *L'Année Psychologique*, 93, p. 9–30, 1993

[91] Y. Morlec : *Génération multiparamétrique de la prosodie du français par apprentissage automatique*, Thèse de doctorat, ICP, décembre 1997
URL `http://www.icp.inpg.fr/~bailly/publis/synthese/_ym/these_ym_PHD97.ps`

[92] Y. K. Muthusamy, E. Barnard et R. A. Cole : "Automatic language Identification : A Review/Tutorial", *IEEE Signal Processing Magazine*, 1994

[93] Y. K. Muthusamy, R. A. Cole et B. T. Oshika : "The OGI Multilanguage Telephone Speech Corpus", dans *International Conference on Spoken Language Processing*, Alberta, October 1992

[94] T. Nagarajan et H. A. Murthy : "Language Identification Using Parallel Syllable-like Unit Recognition", dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 401–404, Montreal, 2004

[95] M. Nespor : "On the rhythm parameter in phonology", dans *Logical Issues in Language Acquisition* (I. Roca, rédacteur), p. 157–175, Dordrecht : Foris, 1990

[96] I. Nesterenko : "Sur le statut de l'unité intonative dans une tâche métalinguistique : exploitation expérimentale d'un concept linguistique", dans *Journées d'Étude sur la parole*, Fès, Maroc, 2004

[97] F. Nolan et E. Grabe : "Can ToBI transcribe intonational variation in English ?", dans *ESCA Workshop on Intonation : Theory, Models and Applications*, Athens, Greece, 1997

[98] F. Pellegrino : *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 1998
URL `http://www.afcp-parole.org/doc/theses/theseFP98.ps.gz`

[99] F. Pellegrino et R. André-Obrecht : "Vocalic System Modeling : A VQ Approach", dans *IEEE Digital Signal Processing*, Santorini, July 1997

[100] F. Pellegrino et R. André-Obrecht : "An Unsupervised Approach to Language Identification", dans *International Cconference on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, 1999

[101] F. Pellegrino et R. André-Obrecht : "Automatic Language Identification : An Alternative Approach to Phonetic Modeling", *Signal Processing*, 80(7), p. 1231–1244, 2000

[102] F. Pellegrino, J. Farinas et J.-L. Rouas : "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech ", dans *International Conference on Speech Prosody 2004 , Nara, Japon* (B. Bel et I. Marlien, rédacteurs), p. 517–520, ISCA Special Interest Group on Speech Prosody (SproSIG), ISBN 2-9518233-1-2, 23-26 mars 2004

[103] T. Pfau et G. Ruske : "Estimating the speaking rate by vowel detection", dans *International Conference on Audio, Speech and Signal Processing*, Seattle, 1998

[104] H. Pfitzinger, S. Burger et S. Heid : "Syllable Detection in Read and Spontaneous Speech", dans *4th International Conference on Spoken Language Processing*, tome 2, p. 1261–1264, Philadelphia, October 1996

[105] H. R. Pfitzinger : "Local speaking rate as a combinaison of syllable and phone rate", dans *5th International Conference on Spoken Language Processing*, tome 3, p. 1087–1090, décembre 1998
URL `http://www.phonetik.uni-muenchen.de/Publications/Pfitzinger_ICSLP98.ps`

[106] J. Pierrehumbert : *The phonology and phonetics of english intonation*, Thèse de doctorat, Massachussets Institute of Technology, 1980

[107] F. Ramus : *Rythme des langues et acquisition du language*, Thèse de doctorat, EHESS, 1999

[108] F. Ramus : "Acoustic correlates of linguistic rhythm: Perspectives", dans *International Conference on Speech Prosody*, p. 115–120, Aix-en-Provence, France, avril 2002
URL `http://www.ehess.fr/centres/lscp/persons/ramus/ramus_sp02.pdf`

[109] F. Ramus, E. Dupoux et J. Mehler : "The psycological reality of rhythm classes : perceptual studies", dans *International Congress of the Phonetic Sciences*, p. 337–342, Barcelona, Spain, 2003

[110] F. Ramus et J. Mehler : "Language identification with suprasegmental cues : A study based on speech resynthesis", *Journal of the Acoustical Society of America*, 105(1), p. 512–521, 1999

[111] F. Ramus, M. Nespor et J. Mehler : "Correlates of Linguistic Rhythm in the Speech Signal", *Cognition*, 73(3), p. 265–292, 1999

[112] F. Ramus et al. : "Language Discrimination by human Newborns and by Cotton-Top Tamarin Monkeys", *Science*, 288, p. 349–351, 2000

[113] P. Roach : "On the distinction between "stress-timed" and "syllable-timed" languages", *Linguistic Controversies*, p. 73–79, 1982

[114] J.-L. Rouas, J. Farinas et F. Pellegrino : "Merging segmental, rhythmic and fundamental frequency features for automatic language identification", dans *Eusipco*, p. 591–594, vol. III, Eurasip, Toulouse, France, 3-6 septembre 2002

[115] J.-L. Rouas, J. Farinas et F. Pellegrino : "Automatic Modelling of Rhythm and Intonation for Language Identification", dans *15th International Congress of Phonetic Sciences (15th ICPhS) , Barcelona, Spain*, p. 567–570, Causal Productions Pty Ltd, 3-9 août 2003

[116] J.-L. Rouas, J. Farinas et F. Pellegrino : "Evaluation automatique du débit de la parole sur des données multilingues spontanées ", dans *XXVe Journées d'Etude sur la Parole (JEP'2004) , Fès, Maroc*, p. 437–440, LPL-ENS/Fes-AFCP-ISCA, ISBN 2-9518233-3-9, 19-21 avril 2004

[117] J.-L. Rouas, J. Farinas, F. Pellegrino et R. André-Obrecht : "Modeling Prosody for Language Identification on Read and Spontaneous Speech", dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003) , Hong Kong, China*, p. 40–43, Vol. I, IEEE, ISBN 0-7803-7664-1, 6-10 avril 2003

[118] J.-L. Rouas, J. Farinas, F. Pellegrino et R. André-Obrecht : "Rhythmic unit extraction and modelling for automatic language identification", *Speech Communication*, Soumis

[119] K. Silverman et al. : "ToBI : a standard for labeling English prosody", dans *International Conference on Spoken Language Processing (ICSLP)*, tome 2, p. 867–870, 1992

[120] E. Singer et al. : "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification", dans *Eurospeech*, p. 1345–1348, Geneva, 2003

[121] A. Thymé-Gobbel et S. E. Hutchins : "On Using Prosodic Cues In Automatic Language Identification", dans *International Conference on Spoken Language Processing*, Philadelphia, 1996

[122] P. Torres-Carrasquillo, D. A. Reynolds et J. Deller : "Language Identification using Gaussian Mixture Tokenization", dans *International Conference on Acoustics, Speech and Signal Processing*, tome 1, p. 757–760, 2002

[123] J. Vaissière : "On French prosody", Quaterly Progress Report 114, Massachusetts Institute of Technology, 1974

[124] J. Vaissière : "Further note on French prosody", Quaterly Progress Report 115, Massachusetts Institute of Technology, 1975

[125] J. Vaissière : "Language independent prosodic features", dans *Prosody : models and measurements*, Springer series in language and communication, 14, p. 53–66, Cutler, A. and Ladd, D.R. (eds.), Berlin, 1983

[126] C. W. Wightman : "ToBI or no ToBI", dans *Speech Prosody*, Aix-en-provence, France, 2002

[127] I. H. Witten : "A flexible scheme for assigning timing and speech to synthetic speech", *Language and Speech*, (20), p. 240–260, 1977

[128] M. A. Zissman et K. M. Berkling : "Automatic Language Identification", *Speech Communication*, 35(1-2), p. 115–124, 2001