ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAITRISE EN GÉNIE ÉLECTRIQUE
M.Ing

PAR
FILIPE VELHO

LA RECONNAISSANCE DU LOCUTEUR À L'AIDE DE LA TRANSFORMÉE EN ONDELETTES CONTINUE

MONTRÉAL, LE 8 MARS 2006

CE MÉMOIRE A ÉTÉ ÉVALUÉ
PAR UN JURY COMPOSÉ DE :

M. Gheorghe Marcel Gabrea, directeur de mémoire
Département de Génie Électrique à l'École de technologie supérieure

M. Christian Gargour, codirecteur de mémoire
Département de Génie Électrique à l'École de technologie supérieure

Mme Rita Noumeir, président du jury
Département de Génie Électrique à l'École de technologie supérieure

M. Jean-Marc Lina, membre du jury
Département de Génie Électrique à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC
LE 8 FÉVRIER 2006
À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

LA RECONNAISSANCE DU LOCUTEUR À L'AIDE DE LA TRANSFORMÉE EN ONDELETTES CONTINUE

Filipe Velho

SOMMAIRE

La vaste majorité des méthodes de traitement de la parole basées sur la transformée en ondelettes rapportées dans la littérature fait usage de bancs de filtres discrets, d'arbres dyadiques et de paquets d'ondelettes. Très peu d'auteurs seulement se sont penchés sur l'utilisation de la transformée en ondelettes continue (TOC) pour le traitement de la parole avec des résultats souvent intéressants. Les raisons de cette mise à l'écart concernent son caractère fortement redondant. Il est très difficile d'en extraire les données pertinentes en choisissant les bons coefficients.

Nos travaux présentent un système d'identification du locuteur, fonctionnant en mode indépendant du texte, dans un environnement non bruité et combinant la TOC et les MFCC pour l'extraction des vecteurs de caractéristiques, le tout basé sur une modélisation par GMM.

Notre système de reconnaissance exploite l'hypothèse selon laquelle les différentes échelles de la TOC peuvent servir à mettre en évidence la variabilité inter-locuteurs d'une population; ceci, en faisant ressortir d'autres caractéristiques qui apparaissent à certaines échelles et qui ne s'expriment pas de la même manière dans le signal de parole d'origine. La méthode que nous utiliserons pour sélectionner les coefficients de la TOC consiste à les sélectionner tous sur une même échelle. Ceci élimine substantiellement la redondance de celle-ci et nous retrouvons des temps de calcul comparables à ceux du système de référence. Nous avons également montré que les bonnes performances de notre système sont liées à la distribution des coefficients cepstraux calculés avec les points retenus pour cette échelle. De plus, l'utilisation de la TOC permet au système d'utiliser moins de parole pour les phases d'entraînement et de reconnaissance.

Notre système est testé sur des populations de 30 et de 50 locuteurs avec des échantillons de parole (de qualité téléphonique) tirés de la base de données publique YOHO. Les taux de reconnaissance sont respectivement de 96 % et de 96.7 %, soit une amélioration de plus de 6 points par rapport au même système sans TOC.

SPEAKER RECOGNITION USING THE CONTINUOUS WAVELET TRANSFORM

Filipe Velho

ABSTRACT

Few authors only have studied speech signal processing with the CWT which performances are often good in other domains. The reasons why it has been set apart concern the high redundancy which is obtained with the CWT. It is very difficult to extract the useful data by choosing the good coefficients.

This work presents high performance speaker identification text-independent systems combining the CWT and the MFCC for feature extraction, and based on Gaussian mixture models.

This speaker identification strategy combines three elements of novelty. First, it exploits the fact that the different scales of the CWT can be used to increase the inter-speaker variability of a population and differentiate speakers better just using different characteristics which appears at certain scales and which are not expressed as the same way in the original speech signal in order to create a new signal peculiar to each speaker. Secondly, it has been shown that the best way of picking up the coefficients of the CWT is taking all the coefficients at a fixed specific scale. Moreover, we have shown that these performances are related to the distribution of the cepstral coefficients of these coefficient lines. And finally, the last element deals with the advantages of selecting the coefficients as an entire line for a fixed scale. It sets apart all the drawbacks of the CWT: there is no redundancy in the new signal, and the computational time is reduced to practically the same as standard identification systems. Moreover, the use of CWT may allow both the training of the system and the recognition of speakers based on shorter samples of speech.

The systems are evaluated on a publically available speech database with the constraint of telephone speech quality: YOHO. On a 30 and 50 speaker population the identification accuracy were respectively 96 % and 96.7 %, that is, more than 6 points better comparatively to the same system without using the CWT.

REMERCIEMENTS

Je tenais à remercier mon directeur de recherche, M. Marcel Gabrea, Ph.D. et mon co-directeur M. Christian Gargour, Ph.D. pour m'avoir permis de réaliser ce projet dans de bonnes conditions et pour m'avoir aidé tout au long de mes recherches grâce à leur grande expertise sur le sujet.

Je remercie également mes parents et ma famille pour m'avoir soutenu et encouragé pendant toutes ces années.

TABLE DES MATIÈRES

Page

SOMMAIRE iii

ABSTRACT iiv

REMERCIEMENTS iv

TABLE DES MATIÈRES vi

LISTE DES TABLEAUX ixi

LISTE DES FIGURES xiv

LISTE DES ABRÉVIATIONS ET SIGLES xviii

INTRODUCTION 1

CHAPITRE 1 LE SIGNAL DE PAROLE 5

1.1 Introduction 5

1.2 Les mécanismes de la parole 6

1.2.1 L'appareil phonatoire 7

1.2.2 Production de la parole 8

1.3 L'information vocale 9

1.3.1 Traits acoustiques du signal de parole 9

1.3.1.1 La fréquence fondamentale 10

1.3.1.2 Le spectre du signal de parole 11

1.4 La perception de la parole 13

1.4.1 Le système auditif 13

1.4.2 Analyse fréquentielle 14

1.4.3 Aire d'audition 14

1.4.4 Échelles de modélisation 15

1.4.4.1 L'échelle de Bark 16

1.4.4.2 L'échelle de Mel 17

1.5 Conclusion 20

CHAPITRE 2 SYSTÈME DE RÉFÉRENCE POUR LA RECONNAISSANCE DU LOCUTEUR 21

2.1 Introduction 21

2.2 Principe de la reconnaissance du locuteur 22

2.3 Module de pré-traitement et de paramétrisation 25

2.3.1 Pré-traitement 25

2.3.2 Extraction des points caractéristiques – Méthode des MFCC 26

2.3.2.1 Blocage des trames 28

2.3.2.2 Fenêtrage et mise en trames 29

2.3.2.3 Transformée de Fourier Rapide (FFT) 33
2.3.2.4 Application de l'échelle de perception de Mel 34
2.3.2.5 Coefficients MFCC 35
2.3.2.6 Dérivées temporelles des coefficients MFCC 36
2.3.2.7 Motivation de l'utilisation des coeffcients cepstraux MFCC 37
2.4 Module de modélisation et d'apprentissage 38
2.4.1 Modélisation du locuteur par mélange de Gaussiennes (GMM) 40
2.4.2 Modèle du locuteur avec les GMM 40
2.4.2.1 Définition 40
2.4.2.2 Type de modèle 43
2.4.3 Estimation à maximum de vraisemblance des paramètres des GMM - phase d'entraînement 44
2.4.3.1 Maximisation directe 46
2.4.3.2 Maximisation à l'aide de l'algorithme EM 47
2.5 Module de reconnaissance - phase de test 54
2.5.1 Étape préliminaire 55
2.5.2 Identification du locuteur 55
2.6 Derniers réglages de perfectionnement pour la mise au point du système de référence pour la reconnaissance du locuteur 57
2.6.1 Choix du nombre de Gaussiennes pour les GMM 57
2.6.2 Initialisation des paramètres du modèle pour l'algorithme EM 60
2.6.2.1 Méthode d'initialisation des poids 61
2.6.2.2 Méthode d'initialisation des centres 61
2.6.2.3 Méthode d'initialisation des covariances 63
2.6.3 Nombre d'itérations pour l'algorithme EM 63
2.7 Expérimentation du système d'identification de référence 65
2.7.1 Paramètres optimaux du système de référence 66
2.7.1.1 Normalisation du signal de parole 67
2.7.1.2 Extraction des vecteurs de caractéristiques avec les MFCC 67
2.7.1.3 Phase d'apprentissage 68
2.7.1.4 Phase de test 68
2.7.2 Résultats expérimentaux trouvés pour le système de référence 68
2.8 Conclusions 71

CHAPITRE 3 LES ONDELETTES ET SES APPLICATIONS 72
3.1 Introduction 72
3.1.1 Position du problème 73
3.1.2 Introduction aux ondelettes 76
3.2 Étude de la méthode des ondelettes 78
3.2.1 Qu'appelle-t-on une ondelette ? 78
3.2.1.1 Présentation générale 78
3.2.1.2 Condition d'admissibilité 79
3.2.1.3 Condition de régularité 80
3.2.1.4 Compression et dilatation d'une ondelette 82

3.2.2 Exemples classiques d'ondelettes continues 1D 83
3.2.2.1 Présentation de l'ondelette de Morlet 84
3.3 Étude de la transformée en ondelettes continue 86
3.3.1 La transformée en ondelettes continue 87
3.3.2 Transformée de Fourier d'une ondelette ; analyse temps-fréquence 88
3.3.3 Avantages de la TOC 88
3.3.4 La transformée en ondelettes continue inverse 89
3.3.5 Qu'y a-t-il de continu dans la TOC ? 90
3.4 Passage en revue des autres transformées en ondelettes 91
3.4.1 Décomposition discrète en série d'ondelettes 92
3.4.2 Transformée en ondelettes à temps discret 93
3.4.3 Transformée en ondelettes discrète 94
3.5 Conclusion 94

CHAPITRE 4 MODIFICATION DU SYSTÈME DE RECONNAISSANCE DE RÉFÉRENCE À L'AIDE DE LA TOC : UTILISATION D'UNE GRILLE DE SÉLECTION DES COEFFICIENTS DE LA TOC 96
4.1 Introduction 96
4.2 Présentation des bases de données utilisées dans la phase de tests 96
4.2.1 Base de données YOHO 97
4.2.2 Bases de données dérivées de YOHO 99
4.3 Modification du système de reconnaissance de référence avec la TOC 100
4.3.1 Mise en place de la TOC pour essayer d'améliorer les performances du système 101
4.3.1.1 Empreinte graphique du locuteur 102
4.3.1.2 Proposition d'utilisation de la TOC pour faire de la reconnaissance de mots isolés 105
4.3.1.3 Méthode proposée pour exploiter l'empreinte du locuteur 107
4.3.2 Présentation des systèmes de reconnaissance hybrides proposés 109
4.3.2.1 Procédure 109
4.3.2.2 Systèmes de reconnaissance hybrides utilisant une grille pour la sélection des coefficients de la TOC : systèmes hybrides G 111
4.3.2.3 Recombinaison des coefficients de la TOC en un nouveau signal 1D .. 112
4.4 Phase expérimentale 115
4.4.1 Paramètres optimaux additionnels pour les systèmes hybrides 116
4.4.1.1 Ondelette analysante 116
4.4.1.2 Échelle d'analyse 117
4.4.2 Essais expérimentaux pour les systèmes hybrides G 117
4.4.2.1 Tests des systèmes d'identification hybrides G1 117
4.4.2.2 Tests des systèmes d'identification hybrides G2 120
4.5 Conclusions 122

CHAPITRE 5 MODIFICATION ET AMÉLIORATION DU SYSTÈME DE RECONNAISSANCE HYBRIDE PROPOSÉ : UTILISATION DE LIGNES POUR SÉLECTIONNER LES COEFFICIENTS DE LA TOC 125
5.1 Introduction 125
5.2 Présentation du système de reconnaissance hybride amélioré 125
5.2.1 Systèmes de reconnaissance hybrides utilisant une combinaison de lignes pour la sélection des coefficients de la TOC : systèmes hybrides C 125
5.2.2 Tests du système d'identification hybride C 127
5.2.3 Conclusions sur le système hybride C 131
5.3 Présentation d'un nouveau système hybride amélioré 132
5.3.1 Systèmes de reconnaissance hybrides utilisant une seule ligne pour la sélection des coefficients de la TOC : systèmes hybrides L 132
5.3.2 Tests du système d'identification L 134
5.3.3 Conclusions sur le système hybride L 138
5.4 Amélioration des performances du système hybride L par un raffinement d'échelles 140
5.4.1 Sélection des coefficients de la TOC : premier raffinement des échelles 140
5.4.2 Performances 141
5.4.3 Conclusions sur le premier raffinement des échelles 143
5.4.4 Nouvelle sélection des coefficients de la TOC : second raffinement des échelles 145
5.4.5 Performances 145
5.4.6 Conclusions sur le second raffinement des échelles 148
5.5 Vérification des tests 148
5.6 Conclusions 150
CHAPITRE 6 CHOIX DE L'ÉCHELLE D'ANALYSE DE LA TOC POUR LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR 152
6.1 Introduction 152
6.2 Étude des lignes de coefficients de la TOC 152
6.2.1 Analyses statistiques du premier ordre 152
6.2.2 Analyses énergétiques 154
6.2.3 Analyse du taux de passage par zéro 156
6.2.4 Conclusions sur l'étude des lignes de coefficients de la TOC 157
6.3 Étude des coefficients MFCC extraits à partir des lignes de la TOC 157
6.3.1 Analyses statistiques du premier ordre 158
6.3.2 Analyses graphiques des histogrammes 160
6.3.3 Hypothèse 165
6.3.3.1 Analyse de l'énergie 166
6.3.3.2 Analyse de l'entropie 169

6.3.4 Conclusions sur l'étude des coefficients MFCC extraits à partir des lignes de la TOC 173
6.4 Proposition d'une technique pour la reconnaissance automatique du locuteur utilisant la TOC 175
6.4.1 Méthodologie 175
6.4.2 Essais expérimentaux 177
6.4.3 Conclusions sur la méthode pour la reconnaissance automatique du locuteur 179
6.5 Conclusions 180
CONCLUSION 182
ANNEXE 1 REPRÉSENTATION DES ÉCHELLES DE MEL ET DE BARK PAR UN BANC DE FILTRES 185
ANNEXE 2 L'OPTIMISATION DE LAGRANGE 187
ANNEXE 3 L'ALGORITHME DES K-MEANS 189
ANNEXE 4 BASE DE DONNÉES DE YOHO 192
ANNEXE 5 SOUS BASES DE DONNÉES DE YOHO 201
ANNEXE 6 RÉSULTATS COMPLETS DES TESTS DU SYSTÈME DE RECONNAISSANCE HYBRIDE L 204
ANNEXE 7 RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (1ère partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L.. 207
ANNEXE 8 RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (2ème partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L. 212
ANNEXE 9 RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (3ème partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L. 216
ANNEXE 10 L'ESTIMATEUR DE DENSITÉ DE KERNEL 220
BIBLIOGRAPHIE 222

LISTE DES TABLEAUX

Page

Tableau I Bande passante de la fenêtre de Hamming en fonction de son nombre d'échantillons [62] 32

Tableau II Taux de reconnaissance et IER du système de référence en fonction du nombre de Gaussiennes du modèle GMM pour une base de données de 30 locuteurs 69

Tableau III Taux de reconnaissance et IER du système de référence en fonction du nombre de Gaussiennes du modèle GMM pour une base de données de 50 locuteurs 70

Tableau IV Caractéristiques des principales bases de données utilisées en traitement de la parole [48] (PSTN = Public Switched Telephone Network) 98

Tableau V Composition des bases de données dérivées de la base YOHO 99

Tableau VI Performances (en %) des systèmes hybrides G pour différentes tailles de grilles et pour une recombinaison verticale (1) des coefficients de la TOC 118

Tableau VII Temps de traitement (en secondes) des systèmes d'identifications hybrides G pour les phases d'entraînement et de test, en fonction de la taille de la grille et pour une recombinaison verticale (1) des coefficients de la TOC 119

Tableau VIII Performances (en %) des systèmes hybrides G pour différentes tailles de grilles et pour une recombinaison horizontale (2) des coefficients de la TOC 120

Tableau IX Temps de traitement (en secondes) des systèmes d'identifications hybrides G pour les phases d'entraînement et de test, en fonction de la taille de la grille et pour une recombinaison horizontale (2) des coefficients de la TOC 122

Tableau X Performances des systèmes hybrides C pour différentes bandes de lignes consécutives à caractère particulier 128

Tableau XI Performances des systèmes hybrides C pour différentes bandes de lignes consécutives sans signification particulière 128

Tableau XII Performances des systèmes hybrides C pour différentes combinaisons aléatoires de plusieurs lignes de coefficients de la TOC. 129

Tableau XIII Performances des systèmes hybrides C pour différentes combinaisons aléatoires de peu de lignes de coefficients de la TOC 130

Tableau XIV Temps de traitement des systèmes d'identifications hybrides C pour les phases d'entraînement et de test, pour une combinaison formée par une seule ligne de coefficients de la TOC 131

Tableau XV Performances des systèmes hybrides L pour différentes lignes de coefficients de la TOC 134

Tableau XVI IER des systèmes hybrides L pour différentes lignes de coefficients de la TOC 135

Tableau XVII Performances des systèmes hybrides L pour une même ligne de coefficients de la TOC, mais avec un pas temporel variable (population de 30 locuteurs)..... 137

Tableau XVIII Performances des systèmes hybrides L améliorés, pour différentes lignes (à 1 chiffre significatif après la virgule) de coefficients de la TOC..... 141

Tableau XIX Performances des systèmes hybrides L améliorés, pour différentes lignes (à 2 chiffres significatifs après la virgule) de coefficients de la TOC dans la zone d'analyse définie pour 30 locuteurs..... 146

Tableau XX Performances des systèmes hybrides L améliorés, pour différentes lignes (à 2 chiffres significatifs après la virgule) de coefficients de la TOC dans la zone d'analyse définie pour 50 locuteurs..... 147

Tableau XXI Vérification du système L avec les données d'entraînement 149

Tableau XXII Statistiques moyennes pour diverses lignes de coefficients pour une population de 30 locuteurs 153

Tableau XXIII Énergie moyenne pour diverses lignes de coefficients pour une population de 30 locuteurs 155

Tableau XXIV Taux de passage par zéro moyen pour diverses lignes de coefficients pour une population de 30 locuteurs 156

Tableau XXV Statistiques moyennes pour les coefficients MFCC extraits à partir de diverses lignes pour une population de 30 locuteurs 159

Tableau XXVI Valeurs du banc de filtres pour la modélisation de l'échelle des fréquences de Mel et de Bark [99] 186

Tableau XXVII Base de données de 50 locuteurs formée par des sous ensembles de 10, 20, et 30 locuteurs 203

Tableau XXVIII Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à valeur entière, et pour des populations de 10, 20, 30, 40, et 50 locuteurs 206

Tableau XXIX Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à un chiffre significatif après la virgule, et pour des populations de 30 et 50 locuteurs 211

Tableau XXX Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à deux chiffres significatifs après la virgule, et pour une population de 30 locuteurs 215

Tableau XXXI Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à deux chiffres significatifs après la virgule, et pour une population de 50 locuteurs 219

LISTE DES FIGURES

Page

Figure 1 Physiologie de la production de la parole [62] 6

Figure 2 Représentation schématique de la production de la parole [5]. 7

Figure 3 Comparaison d'un son voisé et d'un son non-voisé 8

Figure 4 Évolution de la fréquence de vibration des cordes vocales dans la phrase « je me suis enrhumé » 10

Figure 5 Observation spectrale du conduit vocal [4] 11

Figure 6 Évolution temporelle (en bas) et transformée de Fourier (en haut) d'un signal de parole voisé. 12

Figure 7 Système auditif [106] 13

Figure 8 Réponse en fréquence d'une cellule ciliée [15] 14

Figure 9 Diagramme de Fletcher [62] 15

Figure 10 Tracé de l'échelle de Bark en fonction de la fréquence (en Hz) [18] 17

Figure 11 L'échelle des Mels [62] 18

Figure 12 Approximation de l'échelle de Mel : Banc de filtres triangulaires [22] .. 19

Figure 13 Structure de base des systèmes de vérification du locuteur 22

Figure 14 Structure de base des systèmes de d'identification du locuteur 23

Figure 15 Principales étapes du processus de génération des coefficients MFCC... 28

Figure 16 Signal de parole continu (a) et zoom sur 30ms de ce signal (b) 28

Figure 17 Fenêtre rectangulaire [64] 30

Figure 18 Fenêtre de Hamming [64] 30

Figure 19 Spectre d'un signal de parole fenêtré par Hamming [62] 31

Figure 20 Fenêtrage du signal de parole pour l'obtention de vecteurs paramétriques 33

Figure 21 Processus général de production des coefficients MFCC 36

Figure 22 La fonction de modélisation de la distribution est ici créée à partir de la somme de deux fonctions Gaussiennes 39

Figure 23 Composition d'une suite **X** de vecteurs de caractéristiques extraite d'un locuteur 41
Figure 24 Schématisation de la modélisation GMM d'un locuteur 43
Figure 25 Phase d'entraînement du système 45
Figure 26 Variation de la vraisemblance des données 47
Figure 27 Schéma récapitulatif de l'algorithme EM 53
Figure 28 Modélisation GMM des données L d'un locuteur 54
Figure 29 Schéma bloc de la phase de test du locuteur 57
Figure 30 Modélisation GMM des vecteurs caractéristiques extraits à partir de la parole d'un locuteur (Distribution des coefficients cepstraux), avec (a) 2 Gaussienne, (b) 4 Gaussiennes, (c) 18 Gaussiennes 59
Figure 31 Entrainement et généralisation des GMM [21] 60
Figure 32 Division de l'espace de données en régions d'intérêt. Pour chaque région on calcule son centre géométrique [59] 62
Figure 33 Estimation du GMM d'un locuteur pour 2, 4, 10, et 40 itérations 64
Figure 34 Variation de la vraisemblance des données pour la modélisation du locuteur 1, 2, 3, et 4 65
Figure 35 Schéma descriptif du système de reconnaissance de référence 66
Figure 36 Analyse à fenêtre glissante dans la TFFG [71] 75
Figure 37 Représentation d'ondelettes continues usuelles en 1D [Matlab] 83
Figure 38 Représentation de la partie réelle (tracé continu) et imaginaire (tracé en pointillés) de l'ondelette de Morlet 85
Figure 39 Plan temps-fréquence (a) TFFG (b) TO [33] 89
Figure 40 Décalage de l'ondelette analysante [Matlab] 91
Figure 41 Analyse par l'ondelette de Morlet sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 128 103
Figure 42 Analyse par l'ondelette de Morlet sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 64 103
Figure 43 Spectrogramme du signal de parole « zéro » (YOHO) calculé sur WaveSurfer 104
Figure 44 Spectrogramme du signal de parole « zéro » (YOHO) calculé sur WaveSurfer, avec détermination des quatre premières formantes 105

Figure 45 Analyse par l'ondelette de Haar sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 256 106

Figure 46 Analyse par l'ondelette de Haar sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 128 107

Figure 47 Principales étapes constituant un système de reconnaissance hybride .. 110

Figure 48 Sélection des coefficients de la TOC à l'aide d'une grille 112

Figure 49 Recombinaison verticale des points extraits de la grille (recombinaison 1) 113

Figure 50 Recombinaison horizontale des points extraits de la grille (recombinaison 2) 114

Figure 51 Sélection des coefficients de la TOC à l'aide d'une combinaison de plusieurs lignes de coefficients 126

Figure 52 Sélection des coefficients de la TOC à l'aide d'une seule ligne de coefficients 133

Figure 53 Performances des systèmes hybrides L en fonction des lignes de coefficients de la TOC et suivant la taille de la population 136

Figure 54 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 1 de 2 à 61, et pour une population de 10, 20, 30, 40, et 50 locuteurs... 138

Figure 55 Performances des systèmes hybrides L améliorés, en fonction des lignes (à 1 chiffre significatif après la virgule) de coefficients de la TOC et suivant la taille de la population 142

Figure 56 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.1 de 25 à 36, et pour une population de 30, et 50 locuteurs 143

Figure 57 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.01 de 34.5 à 35.5, et pour une population de 30 locuteurs 146

Figure 58 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.01 de 32.1 à 32.8, et pour une population de 50 locuteurs 147

Figure 59 Histogrammes simple et cumulatif des lignes de coefficients de la TOC 154

Figure 60 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 35.4 160

Figure 61 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 35.5 161

Figure 62 Histogrammes de la répartition des coefficients MFCC extraits à partir du signal de parole d'origine (celui du système de référence) 162

Figure 63 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 16 163

Figure 64 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 96 163

Figure 65 Représentations des densités de Kernel des coefficients MFCC extraits à partir de différentes lignes de coefficients de la TOC 165

Figure 66 Effets de l'énergie des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybrid L 168

Figure 67 Effets de l'entropie de Shannon des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybride L 171

Figure 68 Pour les échelles 15 à 40, variation du taux de reconnaissance, et la somme des énergies et des entropies de la population en fonction de l'échelle de la ligne de coefficients de la TOC 174

Figure 69 Effets des écarts du maximum de vraisemblance des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybride L 178

LISTE DES ABRÉVIATIONS ET SIGLES

F_0	Fréquence fondamentale
B	Unité de mesure [Barks]
B'	Barks modifié
M	Unité de mesure [Mels]
LPC	Linear Prediction Coefficients
$MFCC$	Mel Frequency Cepstral Coefficients
$LPCC$	Linear Prediction Cepstral Coefficients
E	Nombre d'échantillons des trames
C	Nombre d'échantillons de non-chevauchement des trames
FFT	Fast Fourier Transform
$\Gamma(n)$	Fenêtre
N_Γ	Nombre d'échantillons de la fenêtre
$x_l(n)$	Trame
L	Nombre de vecteurs paramétriques
$\mathbf{x}_l$	Vecteur paramétrique
$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_L\}$	Séquence de L vecteurs paramétriques
T_s	Nombre d'échantillons du signal de parole
T_{ch}	Nombre d'échantillons de chevauchement des trames
TFR	Transformée de Fourier Rapide
TFD	Transformée de Fourier Discrète
DFT	Discrete Fourier Transform
$\mathbf{x}_l = \{x_1, x_2, ..., x_E\}$	Vecteur paramétrique
$\boldsymbol{\chi}_l = \{\chi_1, \chi_2, ..., \chi_E\}$	DFT du vecteur paramétrique
TCD	Transformée en Cosinus Discrète
DCT	Discrete Cosine Transform

$\tilde{\chi}_{\log} = \left\{ \tilde{\chi}_{\log_1}, \tilde{\chi}_{\log_2}, ..., \tilde{\chi}_{\log_E} \right\}$	Logarithme du vecteur énergie du signal obtenu en sortie du banc de filtres de Mel
N_F	Nombre de filtres du banc de Mel
D	Nombre de coefficients cepstraux
$\tilde{C}_d$	Coefficients cepstraux
$\tilde{C}_0$	Premier élément cepstral : coefficient d'énergie
$\Delta \tilde{C}_{d_t}$	Coefficients delta-cepstraux
$\Delta\Delta \tilde{C}_{d_t}$	Coefficients delta-delta-cepstraux
GMM	Gaussian Mixture Model (Modélisation par Mélange de Gaussiennes)
HMM	Hidden Markov Modeling (Chaînes de Markov Cachées)
VQ	Vector Quantization
$\underline{\mathbf{x}}_l = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_D\}$	Vecteur paramétrique aléatoire composé de D variables aléatoires
$\underline{\mathbf{x}}_l$	Vecteur paramétrique aléatoire ; ce dernier est obtenu à partir de la $l^{ième}$ trame
λ	Modèle GMM
N	Nombre de Gaussiennes du GMM
g_n	Densité de probabilité unimodale suivant une loi Gaussienne
$p(\underline{x}_l \mid \lambda)$	Densité de probabilité du GMM λ à partir d'un vecteur paramétrique $\underline{\mathbf{x}}_l$
$\underline{\boldsymbol{\mu}}_n$	Le GMM est paramétré par son vecteur de moyennes
Σ_n	Le GMM est paramétré par sa matrice de covariances
$E[\ ]$	Espérance d'un vecteur
Π_n	Poids du mélange de Gaussiennes

r	Locuteur
λ_r	Modèle du locuteur r
MLE	Maximum Likelihood Estimation (Estimation du Maximum de Vraisemblance)
ML	Maximum Likelihood (Maximum de Vraisemblance)
$p(\mathbf{X}\|\lambda) = \prod_{l=1}^{L} p(\underline{\mathbf{x}}_l\|\lambda)$	Densité de probabilité du GMM λ à partir d'une séquence de L vecteurs paramétriques
EM	Expectation-Maximization (Espectative-Maximisation)
$\underline{x}_d$	Variable aléatoire
$\underline{z}_{l;d}$	Label de la variable aléatoire $\underline{x}_d$
j	Numéro de l'itération
λ^j	Modèle pour l'itération j
κ	Multiplieur de Lagrange
F'	Dérivée première de la fonction F
$\hat{\lambda}$	Estimation du modèle λ
ε	Erreur
R	Nombre de locuteurs
p	Probabilité
S_j	Sous-ensembles disjoints
L_j	Nombre de données contenues dans S_j
J	Critère de la somme des carrés
μ_j	Centre géométrique de S_j
TOC	Transformée en Ondelettes Continue
CWT	Continuous Wavelet Transform
TF	Transformée de Fourier
TFFG	Transformée de Fourier à fenêtre glissante
TO	Transformée en ondelettes

TOTD	Transformée en ondelettes à temps discret
TOD	Transformée en ondelettes discrète
TOR	Transformée en ondelettes rapide
ASR	Automatic Speaker Recognition
DDSO	Décomposition Discrète en Série d'Ondelettes
DPWT	Discrete Parameter Wavelet Transform
DTWT	Discrete Time Wavelet Transform
DWT	Discrete Wavelet Transform
FWT	Fast Wavelet Transform
IER	Identification Error Rate

INTRODUCTION

L'utilisation de la parole est considérée de nos jours comme une des formes les plus simples de la technologie biométrique, car elle n'est pas intrusive, et ne demande aucun contact physique avec le locuteur à identifier. Certes, la voix n'est pas aussi efficace et fiable que peuvent l'être les empreintes digitales, l'iris, ou encore l'ADN. Cependant, elle est dans certains cas le seul moyen d'identification d'un individu (identification via une liaison téléphonique, ...). La parole offre ainsi un grand nombre d'application, notamment en matière de télécommunications (accès privé à des bases de données, ...), de sécurité (serrures vocales, distributeurs automatiques, ...), ou encore pour le domaine judiciaire (identification de suspects, preuves juridiques, ...) [97].

La voix est une biométrie difficile à falsifier étant donné les nombreuses caractéristiques propres au locuteur qui la composent : fréquence de vibration des cordes vocales, forme de l'appareil phonatoire, ... [1], [16], [23], [46]. La variabilité inter-locuteurs (qui démontre les différences du signal de parole en fonction du locuteur) doit être grande pour bien différencier les locuteurs. Cependant, elle également mélangée avec la variabilité intra-locuteurs (conditionnée par l'état physique et émotionnel du locuteur), et la variabilité des conditions d'enregistrement du signal de parole (type de microphone, bande passante du canal de transmission, environnement bruité) qui tendent à réduire la variabilité inter-locuteur, et à rendre ainsi la reconnaissance du locuteur plus difficile [6].

Pour rendre les systèmes de reconnaissance de plus en plus fiables, les scientifiques se sont penchés vers le problème de la reconnaissance du locuteur lors des trente dernières décennies. Citons notamment les chercheurs de Texas Instruments (qui ont été les précurseurs), AT&T, IDIAP, Sandia National Laboratories, et de nombreuses universités françaises, et anglaises [5]. De nos jours, les systèmes offrent des performances de plus en plus élevées (de l'ordre de 98 % de reconnaissance), et sont de plus en plus efficaces

contre toute forme d'intrusion. Au cours des dernières années, un nouvel outil mathématique est venu « révolutionner » les domaines du traitement du signal et du biomédical : les ondelettes [83], [93], [100], [102]. Les ondelettes sont des fonctions mathématiques qui permettent de décomposer un signal selon différentes échelles, avec des résolutions qui varient en fonction de l'échelle d'analyse. Offrir une résolution temporelle et fréquentielle variable est l'un de leur grand avantage. Les ondelettes permettent ainsi une meilleure localisation des signaux en temps et en fréquence. La vaste majorité des méthodes de traitement de la parole basées sur la transformée en ondelettes rapportées dans la littérature fait usage de bancs de filtres discrets, d'arbres dyadiques et de paquets d'ondelettes. Très peu d'auteurs seulement se sont penchés sur l'utilisation de la transformée en ondelettes continue (TOC) pour le traitement de la parole avec des résultats souvent intéressants, notamment dans le domaine du biomédical [36], [45], [73]. La cause principale de cet « abandon » est que cette transformée présente une importante redondance au niveau de ses coefficients. Il est par conséquent très difficile d'extraire l'information pertinente, afin de réduire les temps de calculs requis par les systèmes. Cette transformée apparaît donc aux yeux des scientifiques comme inutile et longue d'utilisation comparée aux autres transformées plus souples à utiliser.

Le but de notre projet est de vérifier si la TOC peut servir à la reconnaissance du locuteur sachant que ceci n'a jamais été référé dans la littérature auparavant. Nous voulons dans un premier temps expérimenter les performances en matière de reconnaissance du locuteur, et ce, indépendamment des temps de calculs qui seront engendrés. Puis si les résultats s'avèrent intéressants, nous pourrons y apporter quelques modifications afin d'améliorer les temps de calculs.

Pour cela, nous devons tout d'abord concevoir un système de reconnaissance de référence, basé sur des techniques courantes et performantes (paramétrisation MFCC,

modélisation GMM, et classificateur à maximum de vraisemblance). Puis, par la suite, essayer de le modifier à l'aide de la TOC dans le but d'en améliorer ses performances.

Nous avons fixé pour objectifs de mettre au point un système de reconnaissance permettant d'identifier l'identité du locuteur dans un environnement non bruité, et ce, en mode indépendant du texte (le locuteur sera libre de dire ce qu'il veut pendant la phase de test du système). Pour ne pas engendrer des temps de calculs longs et inutiles en cas de mauvais résultats, nous voulons pouvoir le tester sur une base de données formée d'une population mixte d'environ 30 locuteurs.

Ce mémoire est organisé selon six chapitres. Dans un premier chapitre, nous vous présenterons comment on peut caractériser un individu à l'aide de sa voix. Nous verrons comment se forme le signal de parole et quelles sont les différentes caractéristiques qu'il véhicule. Ceci nous permettra de comprendre ce qui sera important pour différentier les locuteurs et mettre en évidence la variabilité au sein de la population. Dans la seconde partie nous verrons comment concevoir un système de reconnaissance capable d'extraire ces caractéristiques. Nous passerons en revue différentes techniques d'extraction, de modélisation, et de reconnaissance, en choisissant les plus performantes pour former un système de référence. Le troisième chapitre sera consacré aux ondelettes et à ses applications. Il aura pour but de définir ce nouvel outil mathématique, ses avantages et ses modes d'utilisation. Ceci permettra au lecteur de se familiariser avec les ondelettes et de comprendre comment l'intégrer au sein du système de référence. Dans le quatrième chapitre, nous verrons comment nous avons utilisé la TOC pour faire de la reconnaissance du locuteur. Nous décrirons les systèmes hybrides conçus avec la TOC, puis, nous testerons leurs performances sur différentes populations. Le chapitre 5 se consacré aux différentes modifications que nous avons apportées sur le choix des coefficients de la TOC pendant la phase expérimentale pour essayer d'améliorer les nouveaux résultats obtenus avec les systèmes hybrides. Enfin, dans le dernier chapitre, nous chercherons quels sont les paramètres de la TOC qui jouent un rôle important dans

la reconnaissance. Nous verrons ainsi comment choisir l'échelle d'analyse de la TOC et nous proposerons pour finir une méthode simple permettant d'aboutir aux meilleures performances avec les systèmes de reconnaissance hybrides.

CHAPITRE 1

LE SIGNAL DE PAROLE

1.1 Introduction

Depuis le début des années 70, les étudiants et chercheurs de AT&T, BBN, CMU, IBM, Lincoln Labs, MIT, et SRI ont largement contribué dans la recherche et la compréhension du langage parlé [1], [2].

Basiquement, la parole n'est qu'une séquence de segments sonores discrets, reliés les uns aux autres dans le temps. Ces segments, appelés phonèmes, ont par définition des caractéristiques articulatoires et acoustiques uniques. Bien que l'appareil phonatoire humain puisse produire une infinité de mouvements articulatoires, le nombre de phonèmes quant à lui reste limité [1]. Chaque phonème a des caractéristiques acoustiques distinctes et, en se combinant avec d'autres phonèmes, ils permettront de former des entités plus grandes telles que des syllabes ou des mots. La connaissance des différences acoustiques des sons produits permettra alors de distinguer un mot d'un autre et donc de faire de la reconnaissance de la parole.

Lorsque les sons sont connectés pour former des unités linguistiques encore plus grandes (phrases, texte,...), les propriétés acoustiques d'un phonème donné vont changer en fonction de l'environnement phonétique ; ceci est du aux interactions des diverses structures anatomiques (la langue, les lèvres, les cordes vocales) qui composent l'appareil phonatoire humain et à leur degré de lenteur [1]. Il en résulte alors un chevauchement de l'information phonémique d'un segment à un autre. Cet effet connu sous le nom de coarticulation, peut survenir dans un mot ou à la fin de celui-ci [1]. Par conséquent, nous voyons que lors de la production de la parole, de nombreux paramètres spécifiques à chaque individu vont venir marquer les sons prononcés. L'action de toutes

les structures anatomiques créera une empreinte vocale spécifique à chaque individu, qui sera contenue dans tous les messages vocaux, et qui pourra être exploitée dans les systèmes de reconnaissance du locuteur. C'est sur ces structures que nous allons nous pencher dans ce chapitre afin de comprendre comment est produite la parole. Ainsi, nous serons en mesure de localiser et d'identifier les données pertinentes dans le signal vocal, puis, dans le chapitre suivant, nous verrons comment les extraire pour caractériser et identifier chaque locuteur.

1.2 Les mécanismes de la parole

L'appareil phonatoire humain est formé de différentes parties qui peuvent nous sembler complexes (fig.1). Cependant, il peut facilement être assimilé, et même souvent représenté comme un système composé simplement d'une source vibrante et d'un filtre (résultant du conduit vocal qui est formé d'une cavité résonante complexe) (fig. 2) [3], [4], [62].

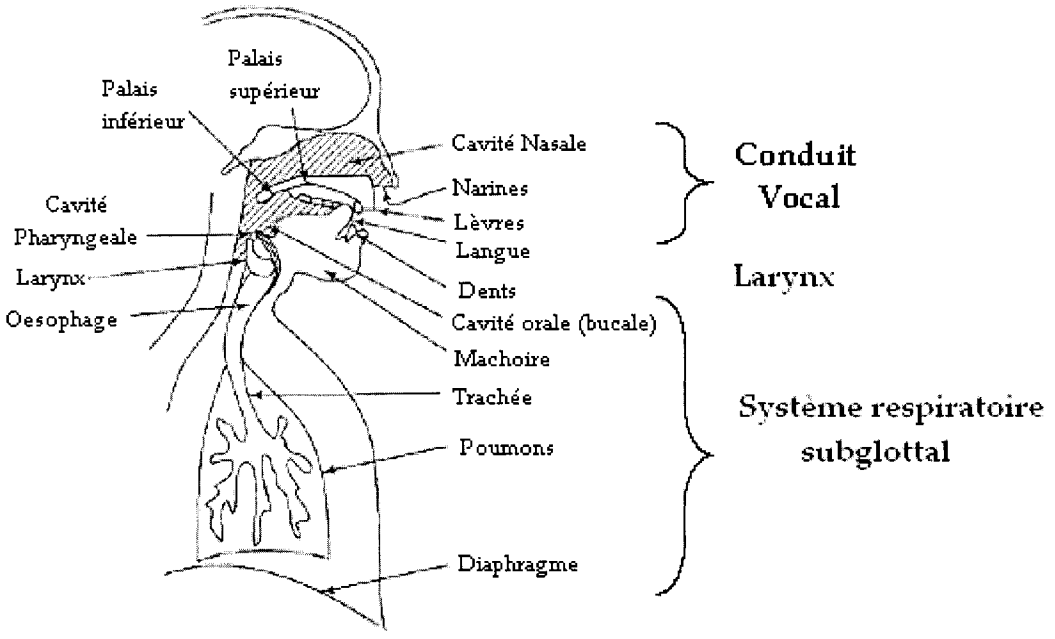


Figure 1 Physiologie de la production de la parole [62].

1.2.1 L'appareil phonatoire

La cavité résonnante (ou résonateur) de l'appareil phonatoire se compose de quatre cavités principales (fig. 2) [5]:

- tout d'abord, nous avons le pharynx ou arrière gorge (1) ;
- puis, les deux cavités buccales (2 et 3) délimitées par la langue (que l'on simplifiera à une seule) ;
- ensuite, nous avons l'ajutage labiale (4) situé entre les dents et les lèvres ;

Ces trois cavités sont placées en « série » à la suite de la source vibrante.

- enfin, la cavité nasale (5), qui vient compléter le résonateur.

Cette dernière cavité quant à elle, est placée en « parallèle sur l'ensemble « série » précédent.

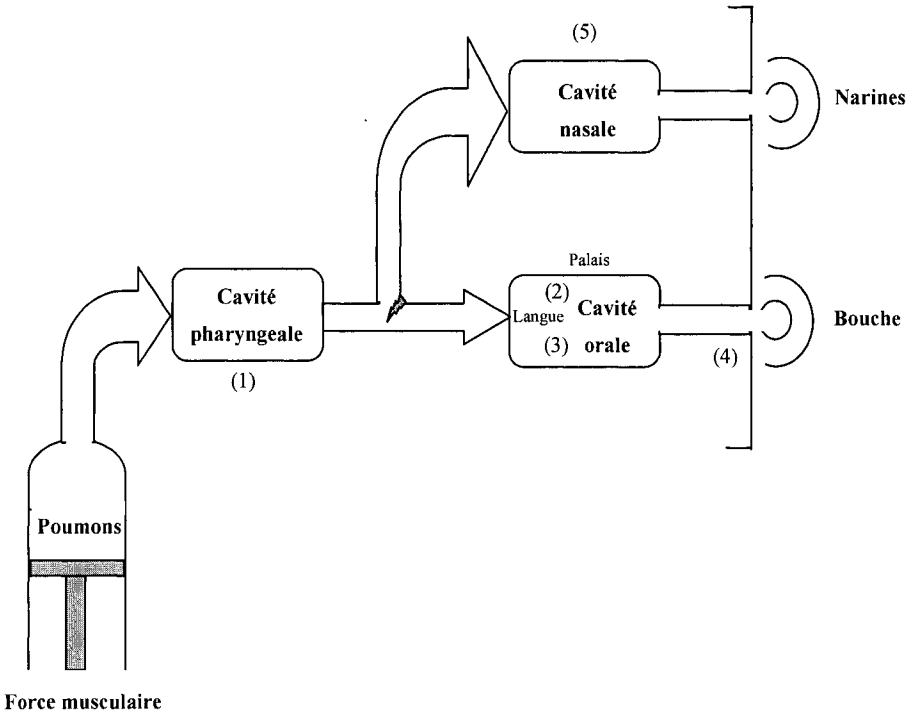


Figure 2 Représentation schématique de la production de la parole [5].

1.2.2 Production de la parole

La parole naît de l'excitation de la cavité résonante. L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant l'air à travers l'appareil phonatoire, vers la source du résonateur [1].

Selon Joseph Campbell, la source du résonateur est en fait décomposable en deux émissions distinctes et d'origines différentes [5]:

- Les cordes vocales, qui possèdent la particularité de produire, en plus de leur fréquence fondamentale, un spectre riche en harmoniques ; elles produisent les sons voisés (fig. 3a).
- Le bruit d'écoulement de l'air en provenance des poumons, dont le spectre est similaire à un bruit blanc ; il crée les sons non-voisés (fig. 3b).

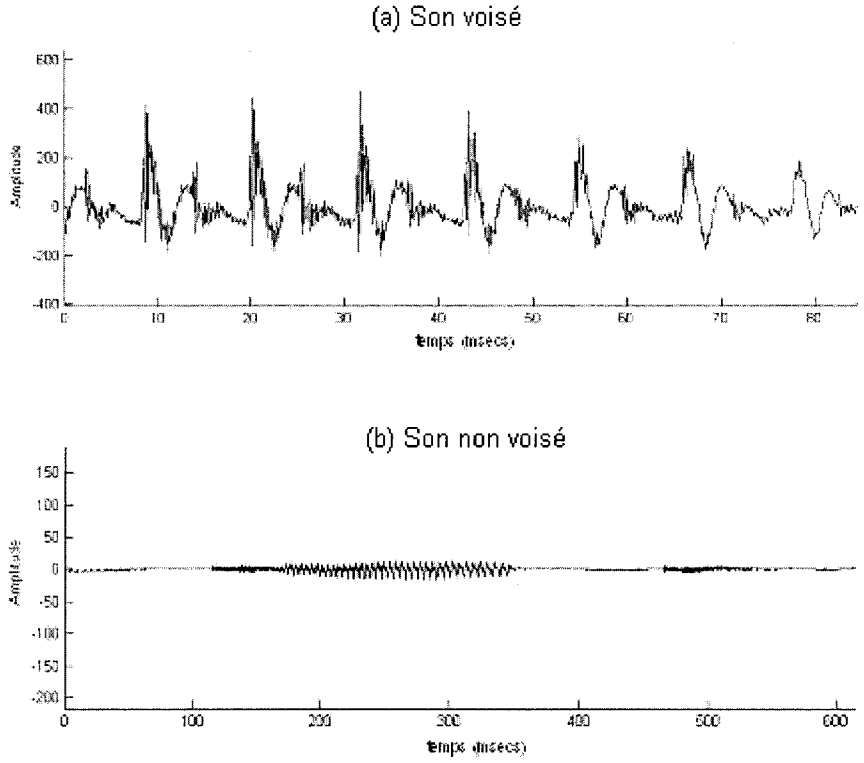


Figure 3 Comparaison d'un son voisé et d'un son non-voisé

Cependant, une source vibrante placée devant une cavité résonante, produira toujours un son dont les fréquences seront filtrées par la bande passante du résonateur.

1.3 L'information vocale

L'information contenue dans le signal de parole peut être analysée de bien des façons. Pour la reconnaissance du locuteur, nous nous limiterons à l'analyse acoustique du signal.

Comme nous venons de le voir plus haut, l'information vocale caractéristique de chaque individu peut être extraite à partir des signaux sortant du résonateur. En effet, chaque locuteur possédant une cavité résonante qui lui est propre, va produire un signal vocal qui aura été altéré par les caractéristiques de cette cavité et qui contiendra par conséquent de l'information sur ce dernier.

1.3.1 Traits acoustiques du signal de parole

Les traits acoustiques du signal de parole sont directement liés à sa production dans l'appareil phonatoire. Tout d'abord, nous avons l'énergie du son [6]; celle-ci est liée à la pression de l'air en amont du larynx. Puis, pour les sons voisés, nous avons la fréquence fondamentale F_0 [7]; cette fréquence correspond à la fréquence du cycle d'ouverture/fermeture des cordes vocales. Enfin, nous avons le spectre du signal de parole [8]; celui-ci résulte du filtrage dynamique du signal en provenance du larynx par le conduit vocal qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses. Chacun de ces traits acoustiques est lui-même intimement lié à une autre grandeur perceptuelle, à savoir l'intensité, le rythme, et le timbre.

1.3.1.1 La fréquence fondamentale

L'évolution temporelle de la fréquence fondamentale (F_0) ou pitch est une information spécifique à chaque locuteur, qui varie en fonction des phonèmes qu'il prononce au cours d'une phrase. Sur la figure 4, nous pouvons observer l'évolution temporelle de la fréquence fondamentale de la phrase « je me suis enrhumé ». Les sons non-voisés sont associés à une fréquence nulle. Ces résultats sont obtenus avec WaveSurfer [13].

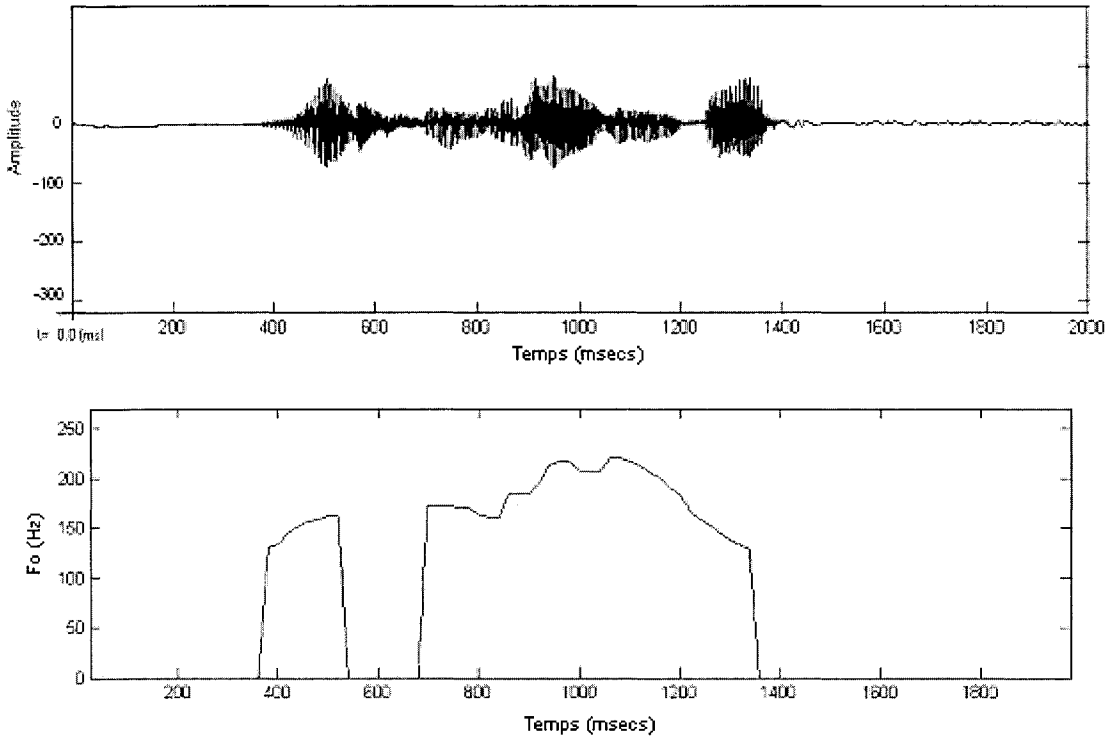


Figure 4 Évolution de la fréquence de vibration des cordes vocales dans la phrase « je me suis enrhumé »

On remarque qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps. D'après toutes les études qui ont été effectuées sur ce sujet, on peut relever les ordres de grandeur des fréquences fondamentales : de 80 Hz à 200 Hz

chez les hommes, de 150 Hz à 350 Hz chez les femmes, et de 200 Hz à 500 Hz pour les enfants [9], [10], [11], [12].

1.3.1.2 Le spectre du signal de parole

Lorsque l'onde acoustique passe à travers la caisse de résonance, son contenu spectral est altéré par les résonances du conduit vocal. L'observation spectrale du conduit vocal (fig. 5) laisse apparaître une enveloppe spectrale formée de pics de résonance, appelés formants, et d'affaiblissements nommés anti-formants (introduits par les sons nasalisés) [5].

Prenons note qu'en général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100 Hz - 5 kHz (300 Hz - 3.4 kHz pour la téléphonie) [63].

	Domaine des Fréquences (Hz)	Domaine des largeurs de bande (Hz)
Formant 1	100 - 1100	45 - 130
Formant 2	1500 - 2500	50 - 190
Formant 3	1500 - 3500	70 - 260

Estimations des largeurs de bande des formants

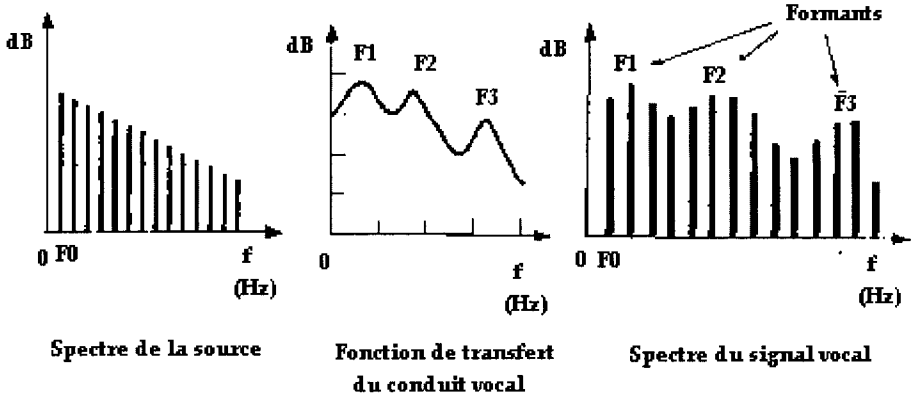


Figure 5 Observation spectrale du conduit vocal [4]

Ainsi, la forme du conduit peut être estimée à partir de la forme du spectre du signal de parole (en localisant les formants et les anti-formants). En effet, la parole est surtout contenue dans les deux premiers formants ; mais l'information proprement dite provient des transitions formantiques [49]. On est par conséquent en mesure de discriminer plusieurs individus selon les caractéristiques du résonateur.

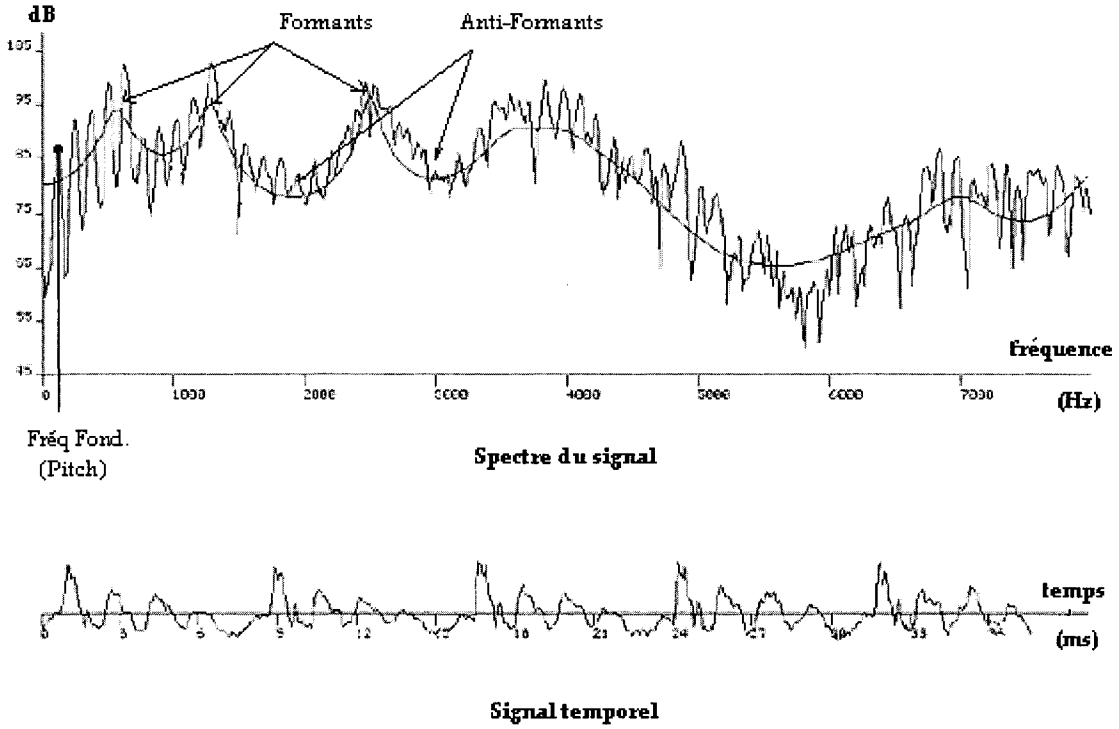


Figure 6 Évolution temporelle (en bas) et transformée de Fourier (en haut) d'un signal de parole voisé.

D'autres aspects de la production de la parole pouvant être utiles pour discriminer les locuteurs sont les caractéristiques de lecture, incluant le taux de parole, les effets prosodiques, et le dialecte [1], [46], [104].

1.4 La perception de la parole

En traitement de la parole, et plus particulièrement en reconnaissance vocale, il est essentiel d'avoir aussi une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille si on désire extraire l'information de façon pertinente. Tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu et par conséquent il y a une partie du signal qu'il est inutile d'analyser. Ceci nous permet alors une réduction des données, et un gain en vitesse de traitement.

1.4.1 Le système auditif

Les vibrations mécaniques du signal sont converties en impulsions nerveuses du nerf auditif par les cellules ciliées au niveau de la cochlée [14].

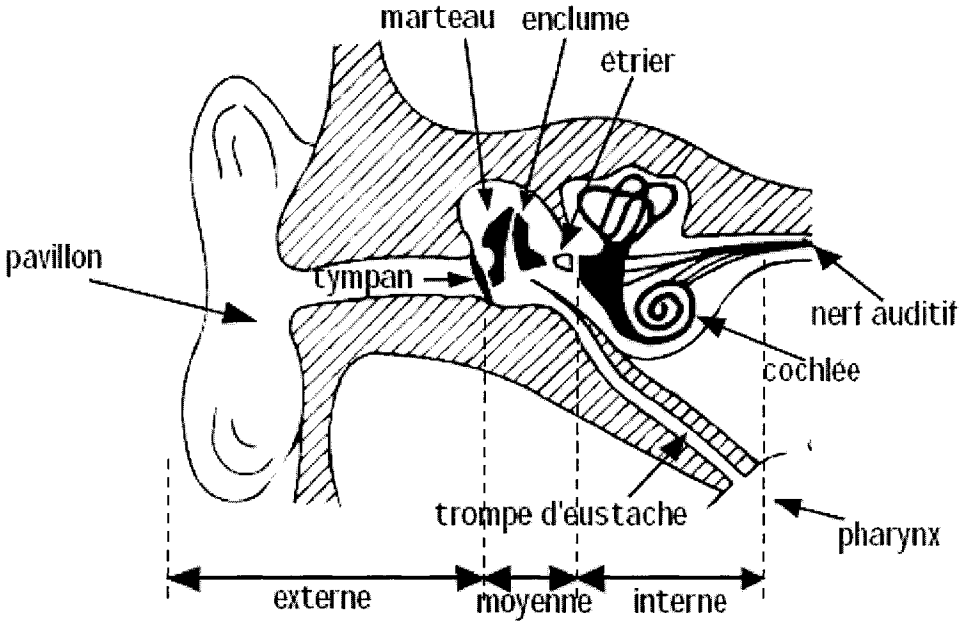


Figure 7 Système auditif [106]

1.4.2 Analyse fréquentielle

Il existe environ 25 000 cellules ciliées qui sont réparties au niveau de la cochlée. Chaque cellule ciliée vibre à une certaine fréquence dite de résonance (fréquence qui dépend de la position de la cellule (fig. 8)). L'oreille effectue donc une sorte d'analyse fréquentielle du signal acoustique. Prenons note que la transformation en impulsion nerveuse est sensible à la fréquence mais est insensible à la phase [15].

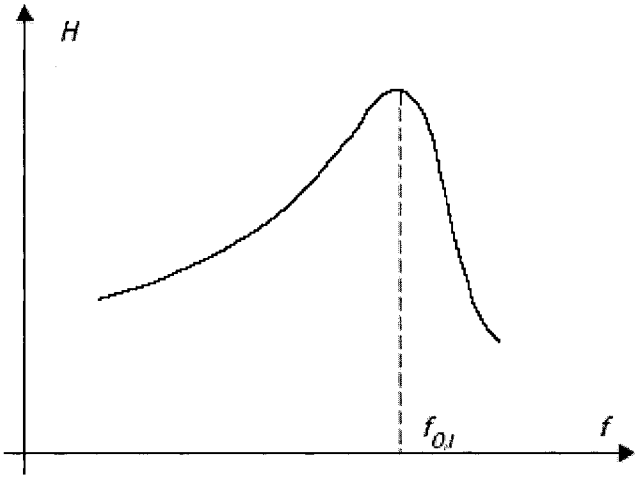


Figure 8 Réponse en fréquence d'une cellule ciliée [15]

1.4.3 Aire d'audition

L'oreille ne répond pas de la même manière à toutes les fréquences. La figure 9 présente le champ auditif humain, délimité par la courbe de seuil de l'audition et celle du seuil de la douleur. Sa limite en fréquence (d'environ 16 kHz à 20 kHz selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (environ 32 kHz).

À l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. On peut voir sur la figure 9, les seuils de perception et

les courbes d'égale sensation de puissance auditive (aussi appelée sonie, exprimée en sones) à différentes fréquences (isotonie). Elles révèlent un maximum de sensibilité dans la plage [500 Hz, 10 kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

Cette aire se caractérise par la pression acoustique, dont la référence à $0\ dB = 10^{-12}\ W.m^{-2}$ est obtenue pour $2.10^5\ Pa$, et par la sonie qui est référencée au $0\ dB$ précédent mais pour 1 kHz [62].

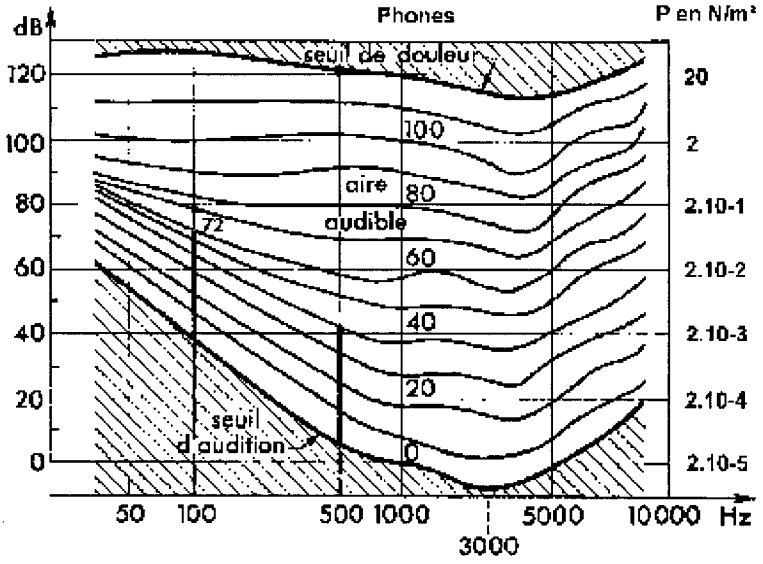


Figure 9 Diagramme de Fletcher [62]

1.4.4 Échelles de modélisation

Les laboratoires de AT&T Bell ont beaucoup contribué dans la recherche auditive, notamment dans la découverte des bandes critiques d'audition, ou encore de l'index d'articulation [16]. Les travaux de Fletcher [17] ont démontré l'existence de bandes critiques dans la réponse de la cochlée. Ces bandes critiques sont d'une importance primordiale pour la compréhension de beaucoup de phénomènes tels que la perception

de l'intensité, le pitch, ou le timbre. Nous venons de voir que le système auditif effectuait une analyse fréquentielle des sons. La cochlée agit comme si elle était constituée par un banc de filtres dont les bandes, appelées "bandes critiques", se chevauchent, et dont les fréquences centrales s'échelonnent continûment. Cette bande critique correspond à l'écartement en fréquence nécessaire pour que deux harmoniques soient discriminées dans un son complexe périodique [9], [63]. Une des modélisations de ces bandes critiques est appelée l'échelle de fréquences de Bark [46]. Une autre échelle de perception des fréquences fut proposée par Mel [19]. Cette dernière est plus efficace pour l'analyse spectrale.

1.4.4.1 L'échelle de Bark

En traitant l'énergie spectrale avec l'échelle de Bark, les chercheurs espéraient qu'on puisse aboutir à un traitement de l'information spectrale aussi proche que celui effectué au niveau de l'oreille. L'échelle de Bark s'étend de 1 à 24 Barks, ce qui correspond aux 24 bandes critiques de l'audition. Notons que les bandes critiques de l'oreille sont continues, et un son de n'importe quelle fréquence audible se trouve toujours centré sur une bande critique. La fréquence de Bark B peut être exprimée en termes de fréquence linéaire (en Hz) par [18]:

$$B(f) = \frac{26.81f}{1960 + f} - 0.53 \quad \text{(exprimée en Barks)} \tag{1.1}$$

Cette approximation s'accompagne de deux fonctions de correction : une pour les basses fréquences, qui assure que la courbe approche au mieux les valeurs standard arrondies ; et l'autre pour les hautes fréquences, qui a pour but de rectifier les mauvaises valeurs obtenues par calcul pour $B(f) > 20.1$:

$$B' = \begin{cases} B + 0.15(2 - B) & \text{si } B(f) < 2 \quad (1.2) \\ B + 0.22(B - 20.1) & \text{si } B(f) > 20.1 \quad (1.3) \end{cases}$$

Sur la figure 10, nous avons la représentation de l'échelle de Barks (approximée par Traunmuller [18]) en fonction de la fréquence exprimée en Hz :

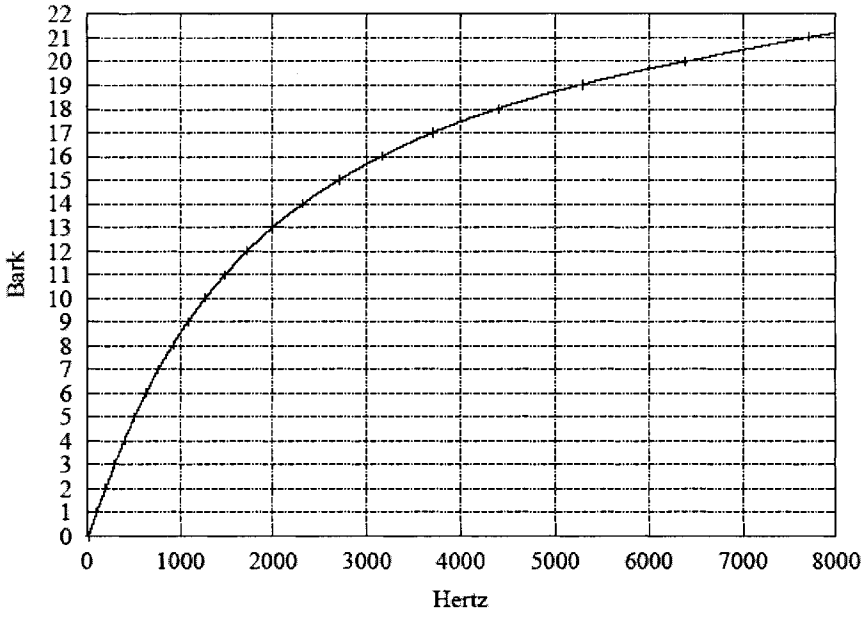


Figure 10 Tracé de l'échelle de Bark en fonction de la fréquence (en Hz) [18]

1.4.4.2 L'échelle de Mel

Après 500 Hz, l'oreille perçoit moins d'une octave pour un doublement de la fréquence. Des expériences psychoacoustiques ont alors permis d'établir la loi qui relie la fréquence et la hauteur perçue : l'échelle des Mels où le « Mel » est une unité représentative de la hauteur perçue d'un son (fig. 11) [62], [63].

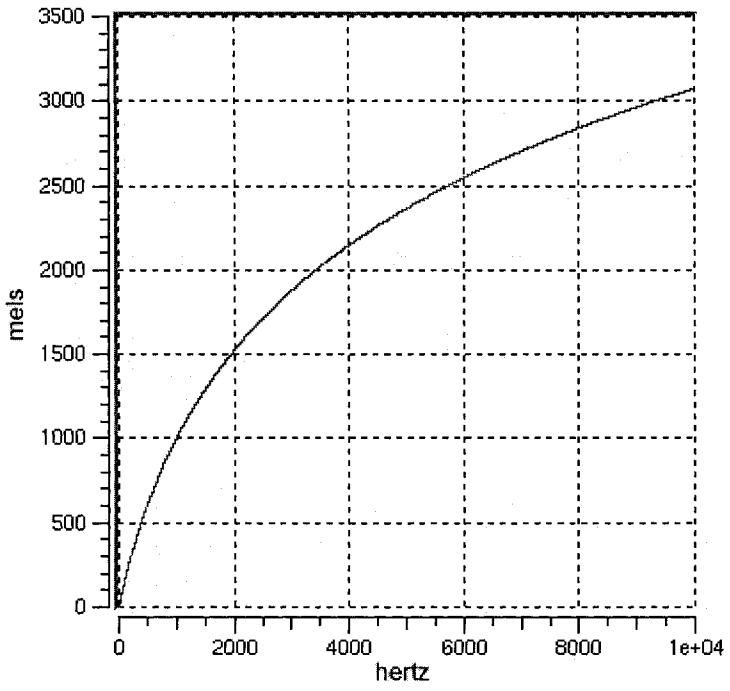


Figure 11 L'échelle des Mels [62]

Cette dernière est linéaire en dessous de 1kHz, et logarithmique pour les fréquences supérieures, avec un nombre égale d'échantillons de part et d'autre de la fréquence 1kHz. L'échelle de Mel[1] (fig. 12) est basée sur des expérimentations avec de simples tons (sinusoïdes) dans lesquels les personnes devaient servir à diviser des bandes de fréquences données en quatre intervalles de perception de même taille, ou bien à ajuster la fréquence d'un ton de stimulation afin que celui ci soit à la mi hauteur du ton de référence [21], [22]. Un Mel est défini comme étant égal à $\frac{1}{1000}ème$ du pitch d'un ton de fréquence 1kHz. Comme pour toutes les échelles de perception, l'échelle de Mel avait pour but de mieux modéliser la sensibilité des segments de l'oreille humaine. L'analyse par l'échelle des fréquences de Mel a été largement utilisée dans les systèmes récents de

[1] Les échelles Mel ou de Bark sont approchées par un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1KHz, puis espacés logarithmiquement jusqu'aux fréquences maximum [9], [10].

reconnaissance de la parole [52], [64]. Elle peut être approximée mathématiquement par [63]:

$$M(f) = 2595\log_{10}(1+\frac{f}{700}) = 1125\ln(1+\frac{f}{700}) \quad \text{(exprimée en Mels)} \tag{1.4}$$

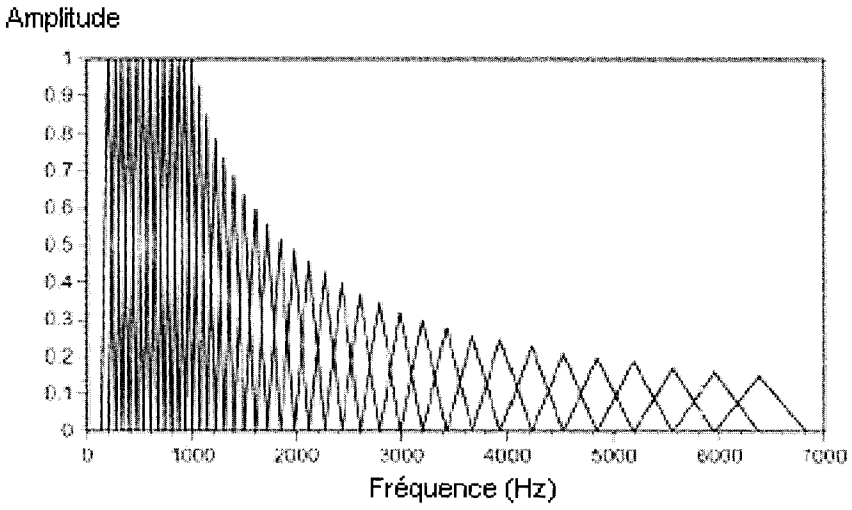


Figure 12 Approximation de l'échelle de Mel : Banc de filtres triangulaires [22]

Un nombre de techniques incalculables employées en traitement de la parole a bénéficié des recherches sur la perception de l'oreille pour rendre les systèmes d'autant plus précis et fidèles; à titre d'exemple, nous pouvons citer l'analyse cepstrale que nous verrons dans la suite du mémoire, et qui servira à extraire les paramètres spécifiques de chaque locuteur.

1.5 Conclusion

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le ton et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application.

Comme nous avons pu le voir dans ce chapitre, l'évolution temporelle d'un signal de parole ne fournit pas directement les traits acoustiques du signal. Pour cela, il est nécessaire d'utiliser des outils mathématiques tels que les transformées, afin de passer dans le domaine fréquentiel et faire apparaître ses caractéristiques. C'est ce que nous verrons dans la suite de ce mémoire, notamment au chapitre 3.

Maintenant que nous avons vu comment s'exprimaient les caractéristiques spécifiques de chaque individu dans le signal de parole, nous pouvons désormais nous pencher sur l'extraction de ces caractéristiques et sur la conception d'un système de reconnaissance de référence basé sur des méthodes qui ont déjà fait leurs preuves dans le passé.

CHAPITRE 2

SYSTÈME DE RÉFÉRENCE POUR LA RECONNAISSANCE DU LOCUTEUR

2.1 Introduction

La reconnaissance du locuteur est un processus qui consiste à reconnaître automatiquement qui a parlé à partir de l'information qui est contenue dans le signal de parole de chaque individu. Grâce à cette technique, il est possible d'utiliser la voix des personnes pour vérifier leur identité et contrôler l'accès à des services tels que les achats par téléphone, les messageries, le démarrage automatique d'une voiture sécurisé par l'empreinte vocale du propriétaire [23-p293], ou encore, pour résoudre des investigations criminelles en trouvant lequel des suspects correspond aux enregistrements trouvés sur la scène du crime par exemple [24], [25].

Les systèmes de reconnaissance du locuteur sont constitués de trois parties principales [63], [64]. Tout d'abord, le module de prétraitement et de caractérisation qui permet d'effectuer quelques conditionnements du signal (tels que la normalisation de l'amplitude, la suppression du bias,...) pour le rendre plus propice à la phase suivante, l'extraction des paramètres acoustiques de chaque locuteur. Ensuite, nous avons le module de modélisation qui va créer un modèle de chaque personne à partir des vecteurs de caractéristiques obtenus précédemment. Et enfin, on trouve le module de reconnaissance qui va tester les nouveaux sujets (pendant la phase de test du système) sur les modèles existant et stockés dans une base de donnée (lors de la phase d'entraînement du système), afin de vérifier ou d'identifier le locuteur qui a parlé.

Dans ce chapitre nous allons étudier les différentes étapes de la conception d'un tel système de reconnaissance. Pour chaque module principal, nous établirons la meilleure méthode en choisissant le meilleur algorithme pour les conditions de fonctionnement

que nous nous sommes fixé. Ainsi, nous aboutirons à un système d'identification de référence comparable à ceux qui peuvent sont décris dans la littérature.

2.2 Principe de la reconnaissance du locuteur

Dépendamment de l'application, le domaine général de la reconnaissance du locuteur se divise en deux taches distinctes [26], [27], [28], [29], [30], [31]:

- la vérification, dont le but est de confirmer ou non si un individu est bien celui qu'il prétend être [32], [33]. La figure 13 nous montre la structure de base des systèmes de vérification du locuteur :

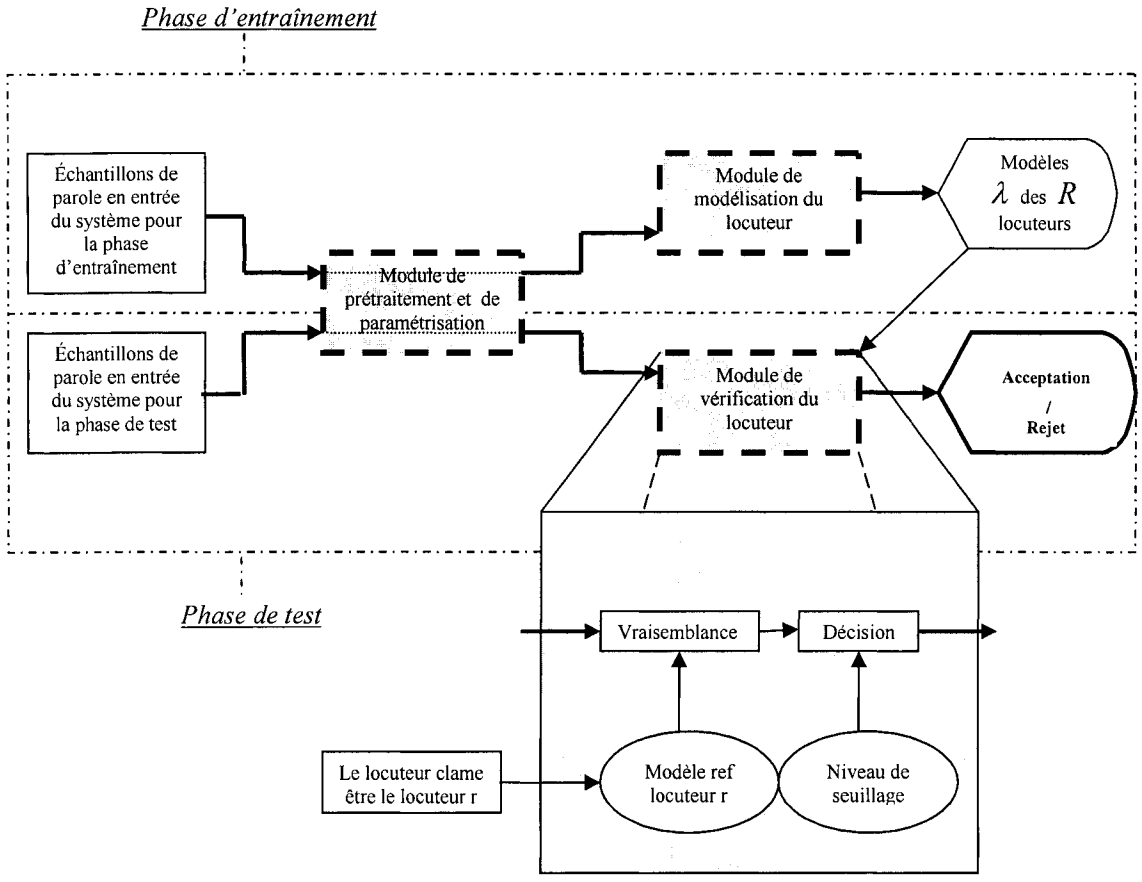


Figure 13 Structure de base des systèmes de vérification du locuteur

- et l'identification qui consiste à indiquer qui est le locuteur. Dans notre projet, nous allons nous intéresser à cette dernière. La figure 14 nous montre la structure de base des systèmes d'identification du locuteur :

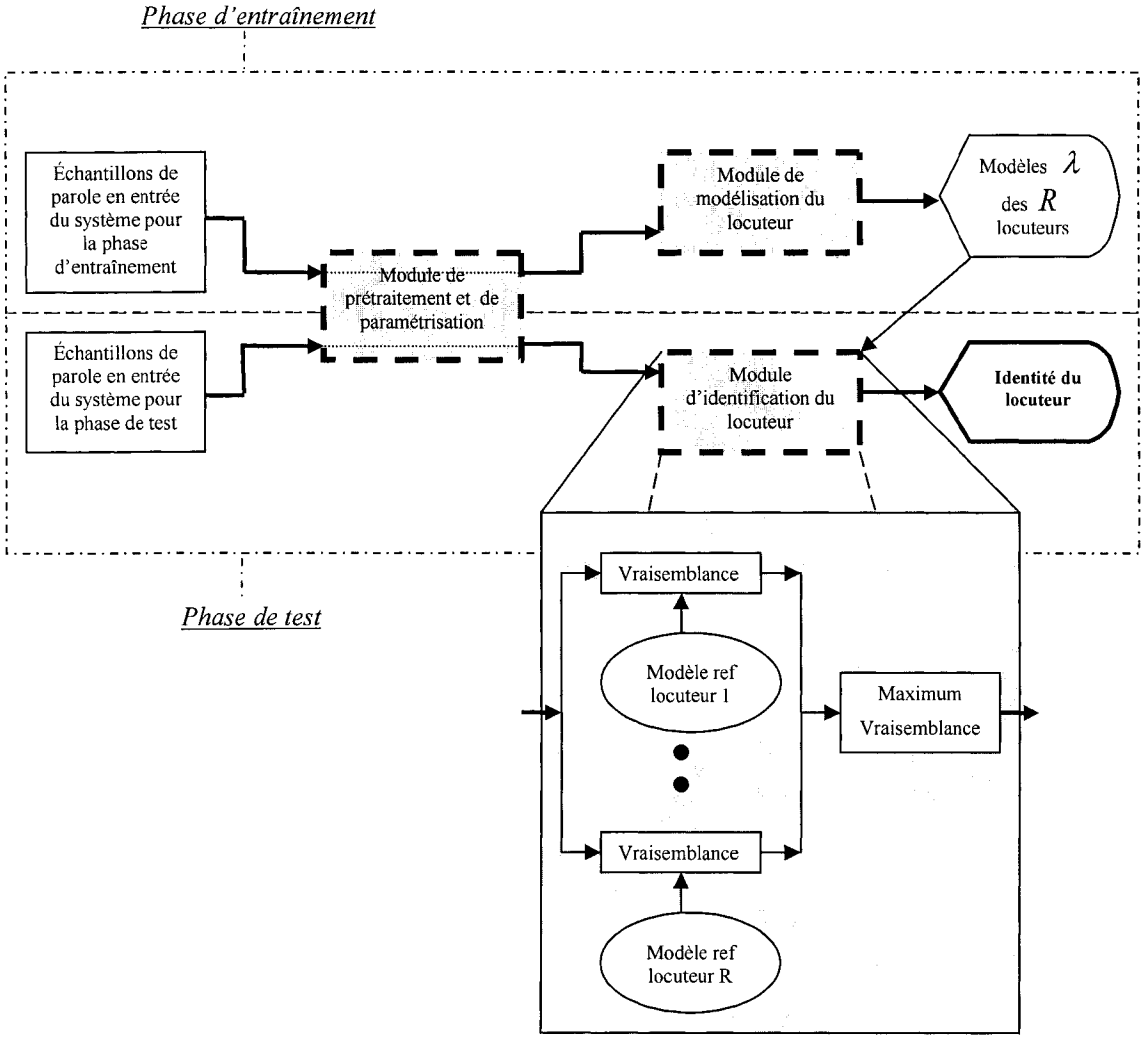


Figure 14 Structure de base des systèmes de d'identification du locuteur

Les méthodes de reconnaissance se divisent en deux domaines selon leur application : nous distinguons les méthodes indépendantes du texte [23-p294], dans lesquelles on modélise les locuteurs à partir de caractéristiques indépendantes de ce qui est prononcé (caractéristiques tirées des traits acoustiques du locuteur) et les méthodes dépendantes du texte [23-p294], [34], [35], dont la reconnaissance est basée sur les caractéristiques obtenues lors de la répétition d'une ou de plusieurs phrases spécifiques telles que des codes PIN (en anglais, Personal Identification Number), des mots de passe, ou des numéros de cartes bancaires.

Tous les systèmes de reconnaissance du locuteur doivent suivre les deux phases suivantes [36-p89] :

- La phase d'entraînement dans laquelle chaque locuteur qui sera susceptible d'être testé dans le futur, doit fournir des échantillons de parole afin que le système puisse entraîner un modèle de référence pour chacun d'entre eux. En ce qui concerne la vérification du locuteur, on doit également calculer un niveau de seuillage spécifique à chaque individu à partir de ces mêmes données.

- La phase de test où le signal de parole que l'on cherche à identifier est associé aux modèles de références collectés après l'entraînement. Ainsi, on établi la décision finale.

Par conséquent, dans un système dépendant du texte, les morceaux de parole utilisés pour l'entraînement et le test du système sont contraints à être les mêmes mots ou phrases. Dans un système indépendant du texte, il n'existe aucune contrainte.

2.3 Module de pré-traitement et de paramétrisation

Le but de ce module est tout dans un premier temps de supprimer le bias que pourrait avoir le signal après le processus d'acquisition, puis de réaliser la pré-emphase du signal, et enfin de détecter précisément les segments de parole par un détecteur d'activité vocale (VAD) pour en supprimer les silences. Puis, dans un second temps, ce module devra transformer le signal qu'on aura traité, en une sorte de représentation paramétrique que nous pourrons ensuite analyser correctement.

2.3.1 Pré-traitement

Dans le cas de notre projet, les signaux de parole que nous possédons sont dépourvus de tout offset étant donné qu'ils sont déjà plus ou moins traités lorsqu'ils sont placés dans la base de données.

Dans le chapitre précédent, nous avons pu voir que les sons voisés, chutent sensiblement de 6dB par octave (20 dB par décade) [26], [37], [65], [66]. Par conséquent, ces sons seront mal représentés par rapport aux basses fréquences. Afin de compenser le niveau faible des aigus et pour palier à une chute du rapport signal à bruit (SNR : Signal to Noise Ration), on utilise généralement un filtre passe-haut, dit de pré-accentuation [23]. Ce filtre est souvent non récursif du premier ordre et a pour fonction de transfert :

$$H(z) = 1 - az^{-1} \tag{2.1}$$

où a est une valeur proche de 1.
Dans la littérature, nous trouvons que la valeur typique est généralement $a = 0.95$ [68].

Le filtre de pré-accentuation va compenser la pente du spectre en augmentant les hautes fréquences ; il accentue ainsi les plus hauts formants qui sont importants pour la discrimination du locuteur [65].

Finalement, la dernière correction à apporter concerne les zones de silence. En effet, dans tout signal de parole, il existe toujours des régions ou il n'y a pas d'activité vocale; spécialement au début et à la fin d'une phrase. Ces régions souvent bruitées, ont très peu d'énergie, et ne contiennent aucune information sur le locuteur. C'est pourquoi il est important de les supprimer pour ne conserver que les parties essentielles pour la classification des locuteurs. L'utilisation d'un VAD nous permettra alors de réduire la taille des données tout en conservant les caractéristiques de la voix. Celui-ci discriminera les zones de silence en appliquant un seuillage sur le taux de passages par zéro pour de courts segments de parole [67], [30].

2.3.2 Extraction des points caractéristiques – Méthode des MFCC

La phase d'extraction des vecteurs de caractéristiques a pour but de créer des éléments vectoriels riches en information (sur ce qui est dit ou sur qui le dit), dépourvus de redondance, et qui présentent de fortes caractéristiques discriminatoires (entre locuteurs) et de faibles variations intra-locuteur. Cet ensemble de vecteurs doit également être aussi compacte que possible.

Le signal de parole est un signal qui varie lentement dans le temps. Il est quasi-stationnaire et peut même être considéré comme stationnaire sur un cours intervalle (entre 5 ms et 100 ms) [1], [63]. Cette propriété est très importante et va nous permettre d'appliquer des techniques d'analyse et de caractérisation propres aux signaux stationnaires. L'analyse spectrale à court terme est par conséquent la méthode de caractérisation que nous adopterons.

En analyse spectrale, il existe de nombreuses techniques permettant d'extraire les paramètres caractéristiques du signal vocal : le codage par prédiction linéaire (LPC) [38], les coefficients cepstraux de Mel (MFCC) et ceux de prédiction linéaire (LPCC) [39], [69], la fonction d'autocorrélation, etc. Les LPCC ont été très largement utilisés pour la reconnaissance, cependant, les modèles basés sur ces paramètres sont très sensibles au bruit [40]. Les MFCC sont les paramètres les plus populaires dans le domaine du traitement de la parole étant donné les bons résultats qu'ils apportent en mode indépendant du locuteur [40].

Dans la suite de cette partie, nous allons nous pencher sur la mise au point de cette dernière afin de concevoir en premier lieu un système d'identification performant et basé sur des techniques traditionnelles et efficaces.

La méthode des MFCC proposée par Davis et Mermelstein [22] pour l'extraction des paramètres caractéristiques du signal vocal, est basée sur les variations des bandes critiques de fréquence de l'oreille humaine. En fait, les coefficients cepstraux permettent de capturer les caractéristiques phonétiques que l'oreille perçoit, grâce à l'utilisation des bancs de filtres de Mel (voir chap.1, p.16). De plus, on conserve ainsi la distribution de l'énergie liée à chaque bande de fréquence. Cette technique sera utilisée pour caractériser les locuteurs tant dans la phase d'entraînement comme dans la phase de test.

Le schéma block de la figure 15 décrit grossièrement le processus de génération des coefficients MFCC :

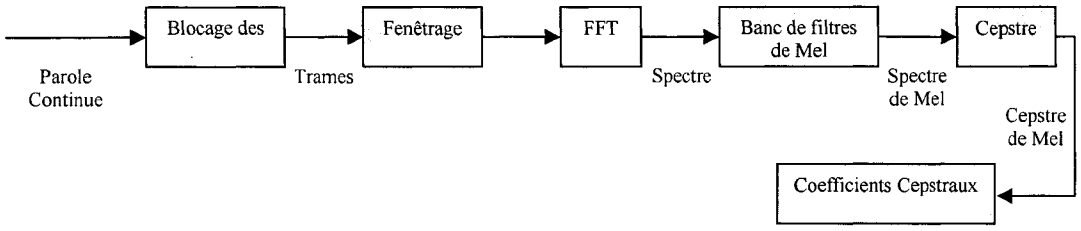


Figure 15 Principales étapes du processus de génération des coefficients MFCC

Maintenant que nous connaissons le fonctionnement général de cette procédure, nous allons à présent étudier les principaux blocs qui la constituent.

2.3.2.1 Blocage des trames

Les signaux de parole sont des signaux non stationnaires (fig. 16a). Cependant, sur un court intervalle, ils peuvent être considérés comme quasi stationnaires (fig. 16b). C'est pourquoi il est utile de diviser le signal de parole en segments successifs stationnaires.

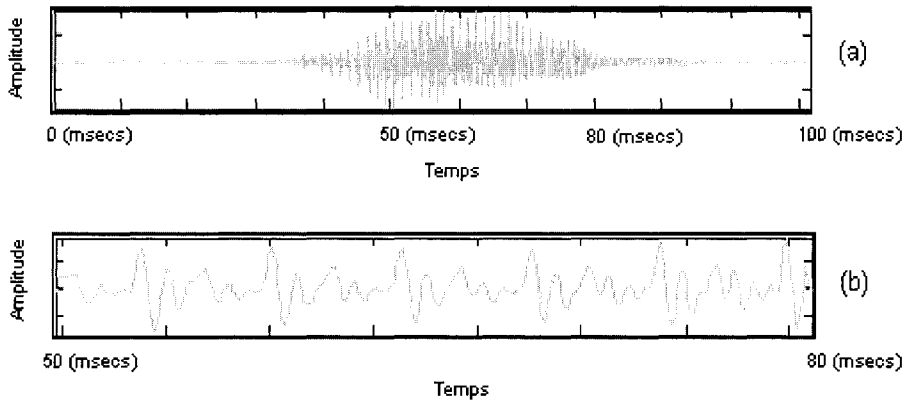


Figure 16 Signal de parole continu (a) et zoom sur 30ms de ce signal (b)

Dans cette première étape, le signal de parole continu est bloqué sous formes de trames de E échantillons. Chaque trame se chevauche à la suivante après C échantillons $(E < C)$. Par conséquent, on retrouve les E premiers échantillons dans la première trame, puis la deuxième commence C échantillons après le début de la première, ce qui correspond à un recouvrement de $(E - C)$ échantillons. Et ceci ainsi de suite jusque tout le signal soit mis sous forme de trames (fig. 20).

L'utilisation de fenêtres d'analyse est indiquée pour atténuer les discontinuités au niveau des extrémités de ces trames [41]. Ainsi, c'est au cours de l'étape de fenêtrage qu'aura lieu le découpage du signal.

On notera que les valeurs typiques de E et C pour une période d'échantillonnage de 8 kHz sont : $E = 256$ (ce qui correspond à un fenêtrage de 30 ms) ; ceci facilite la FFT ; et $C = 100$ [49].

2.3.2.2 Fenêtrage et mise en trames

L'étape suivante consiste à fenêtrer chaque segment afin de minimiser les discontinuités du signal en début et fin de trame. Le concept ici est de minimiser la distorsion spectrale en se servant de la fenêtre pour faire tendre le signal vers zéro au début et à la fin de chaque trame. Si on définit la fenêtre par $\Gamma(n)$ avec $0 \le n \le E-1$, et ou E est le nombre d'échantillons de chaque trame, le fenêtrage du signal $s(n)$ nous donne :

$$x_l(n) = s(n)\Gamma(n) \qquad \text{avec } 0 \le n \le E-1 \tag{2.2}$$

Chaque fenêtre pourra avoir une forme triangulaire, rectangulaire, gaussienne, ou autre selon nos désirs. Généralement, on utilise la fenêtre de Hamming [23-p57], [42], qui a pour équation :

$$\Gamma(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{E-1}\right) \qquad \text{avec } 0 \le n \le E-1 \tag{2.3}$$

Cette fenêtre est avantageuse par le fait que sa résolution fréquentielle est élevée et que son lobe spectral secondaire est très petit par rapport à son lobe primaire (atténuation de - 43 dB). Nous pouvons faire ses constatations à partir de la figure 18.

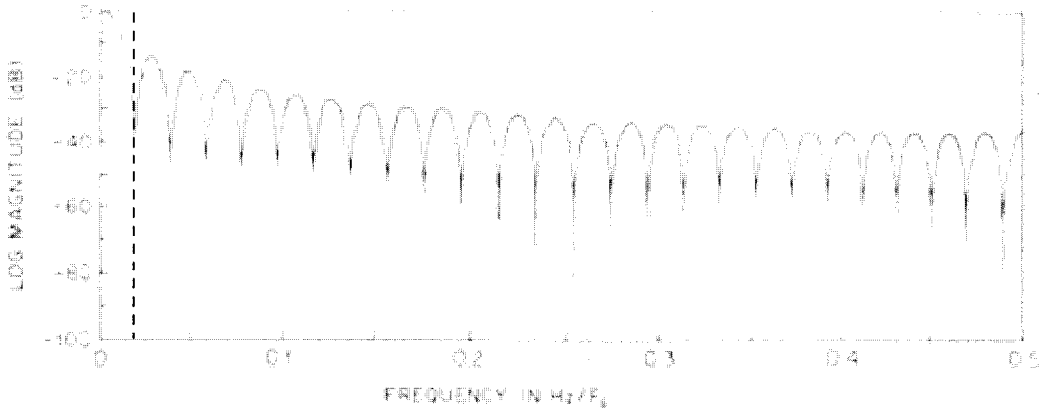


Figure 17 Fenêtre rectangulaire [64]

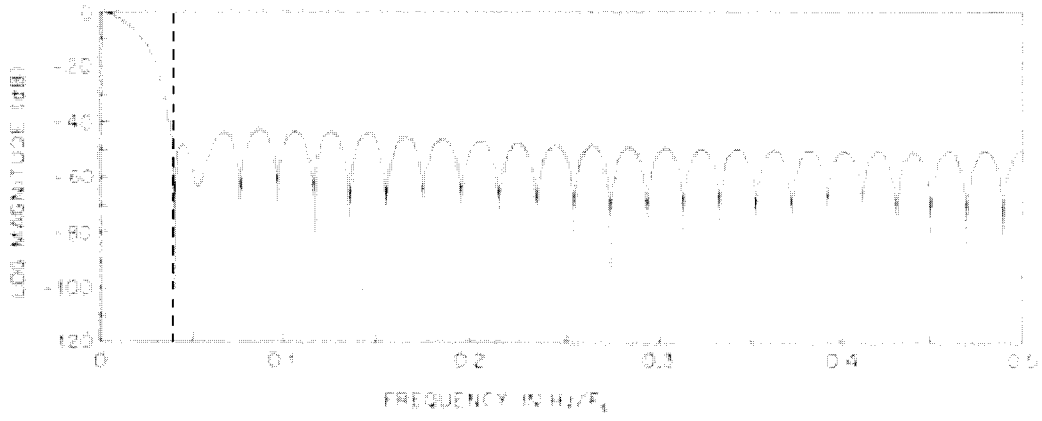


Figure 18 Fenêtre de Hamming [64]

Les figures ci-dessus (fig. 17 et fig. 18) donnent la représentation de la fenêtre de Hamming et de la fenêtre rectangulaire. Nous observons que la fenêtre de Hamming a deux fois plus de bande passante que la rectangulaire, et deux fois plus d'atténuation (pour le second lobe).

Un autre choix important à faire à présent, est celui de la taille de sa fenêtre.
Tel que le mentionne Fang [43], il existe un dilemme concernant le choix de la durée de la fenêtre : la fenêtre doit être suffisamment courte afin que les paramètres sous-jacents puissent être considérés constants sur l'intervalle observé. En revanche, elle doit être d'une longueur suffisante afin de fournir l'information adéquate pour l'extraction fiable des paramètres évalués.

Observons la figure 19 et le tableau I ci-dessous :

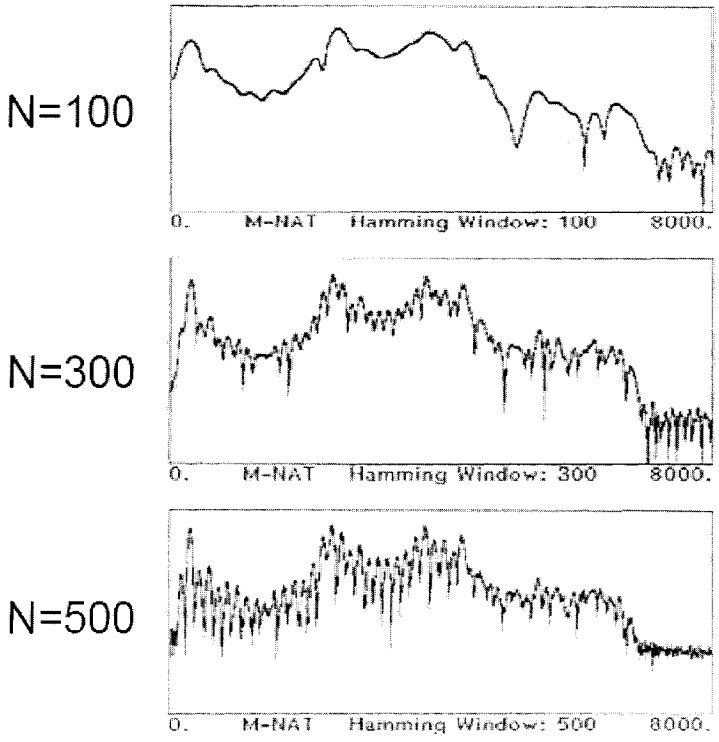


Figure 19 Spectre d'un signal de parole fenêtré par Hamming [62]

Tableau I

Bande passante de la fenêtre de Hamming en fonction de son nombre d'échantillons [62]

Taille (en échantillons)	**Bande Passante (en Hz)**
100	320
250	106.6
500	64

On constate qu'une petite fenêtre ($T_\Gamma = 100$ échantillons) ne permet pas d'obtenir les pulsations du pitch lorsqu'on observe le spectre du signal fenêtré. On remarque cependant qu'une taille de l'ordre de 250 échantillons suffit à donner de bons résultats.

Le nombre total de fenêtres N_Γ déterminera le nombres de trames $x_l(n)$ extraites du signal de parole. Cette valeur correspond par conséquent au nombre L de vecteurs de caractéristiques $\mathbf{x}_l$ qui seront extraits chez le locuteur analysé. En effet, ceci s'explique par le fait que chaque trame permettra de déterminer une série de coefficients cepstraux, et donc un vecteur paramétrique (fig. 20). L'ensemble de tous ces vecteurs sera considéré comme une séquence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_L\}$ de L vecteurs aléatoires.

En exprimant la durée[2] du signal de parole par T_s, la durée de chevauchement des trames (donc des fenêtres) par T_{ch} (avec $T_{ch} = E - C$), et enfin la durée de la fenêtre d'analyse par T_Γ (avec $T_\Gamma = E$), on peut écrire :

[2] Toutes ces durées sont exprimées en nombre d'échantillons n

$$N_\Gamma = L = \frac{T_s - T_{ch}}{T_\Gamma - T_{ch}} \tag{2.4}$$

Les valeurs de T_{ch} et de T_Γ trouvées dans la littérature sont généralement de l'ordre de 20 ms pour T_Γ, et de 10 ms pour C [62], [44], [45].

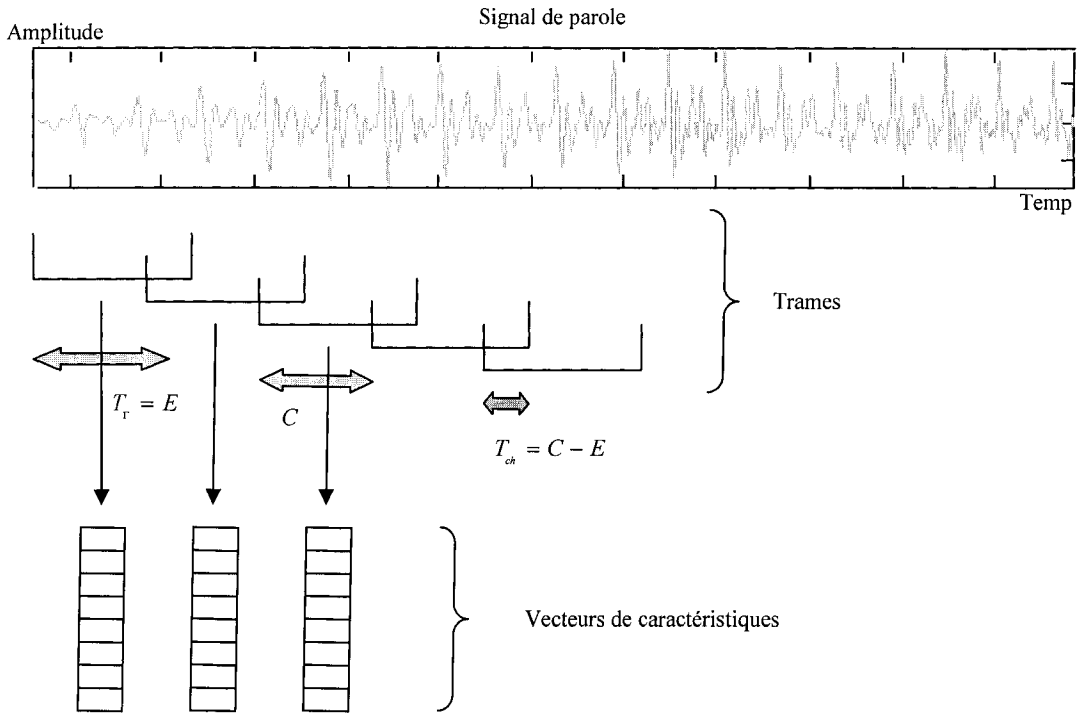


Figure 20 Fenêtrage du signal de parole pour l'obtention de vecteurs paramétriques

2.3.2.3 Transformée de Fourier Rapide (FFT)

Dans cette étape, nous sommes amenés à appliquer la transformée de Fourier rapide (TFR ou en anglais FFT : Fast Fourier Transform) sur les trames obtenues précédemment afin de les transposer dans le domaine fréquentiel. La FFT est un

algorithme rapide d'implémentation de la transformée de Fourier Discrete (TFD ou en anglais DFT : Discrete Fourier Transform) qui est définie pour un signal $\mathbf{x}_l = \{x_1, x_2, ..., x_E\}$ de E échantillons, par [107] :

$$\chi_l = \{\chi_1, \chi_2, ..., \chi_E\} \quad \text{avec} \quad \chi_n = \sum_{k=0}^{E-1} x_k e^{-j\frac{2\pi kn}{E}} \quad \text{pour} \quad 0 \le k \le E-1 \tag{2.5}$$

Le résultat obtenu à la suite de cette étape est souvent référé comme étant le spectre ou le periodogramme du signal.

2.3.2.4 Application de l'échelle de perception de Mel

Comme nous l'avons mentionné dans le premier chapitre, la perception fréquentielle de l'oreille ne suit pas une échelle linéaire. C'est pourquoi il est important de simuler ce filtrage dans notre système, pour ne pas alourdir le traitement des signaux en accumulant des données souvent inutiles. L'échelle des perceptions que nous avons choisi de schématiser est l'échelle fréquentielle de Mel [63]. Notre choix s'est tourné vers cette échelle à cause du phénomène de masquage perceptuel [46].

Afin de simuler le spectre subjectif, nous allons implémenter un banc de filtres à la suite de la FFT, chaque filtre étant attribué à chaque composante fréquentielle de Mel désirée. Le banc de filtres a une réponse fréquentielle de type passe-bande de forme triangulaire, avec un espacement et une bande passante similaire aux valeurs définies par l'échelle fréquentielle de Mel [22]. Le spectre soit disant « perçu » par l'oreille correspond par conséquent à la puissance obtenue en sortie de ces filtres.

Pour chaque trame, on calcule alors l'amplitude de son spectre (obtenu par la FFT), puis on conserve son module au carré. On passe ensuite le vecteur d'énergie à travers le banc de filtres de Mel.

2.3.2.5 Coefficients MFCC

Finalement, il ne reste plus qu'à calculer les coefficients MFCC qui reflèterons les caractéristiques spécifiques (traits acoustiques ou morphologiques) de chaque locuteur. On défini ces coefficients comme étant le résultat de la transformée en cosinus discrète (TCD ou en anglais DCT : Discrete Cosine Transform) inverse, appliquée au logarithme du vecteur d'énergie du signal obtenu en sortie du banc de filtres de Mel [47], [30], [65]. Si on note $\tilde{\chi}_{\log} = \left\{ \tilde{\chi}_{\log_1}, \tilde{\chi}_{\log_2}, ..., \tilde{\chi}_{\log_E} \right\}$ le logarithme du vecteur mentionné, N_F le nombre de filtres du banc, et D le nombre de paramètres (ou coefficients), alors on calcule les coefficients cepstraux $\tilde{C}_d$ comme défini ci-dessous [65]:

$$\tilde{C}_d = \sqrt{\frac{2}{N_F}} \sum_{j=1}^{N_F} \tilde{\chi}_{\log_j} \cos\left[d\left(j - \frac{1}{2} \right) \frac{\pi}{N_F} \right] \quad \text{avec } d = 1,2,...,D \tag{2.6}$$

On remarquera que le premier élément, $\tilde{C}_0$, est exclu de la DCT étant donné qu'il représente la valeur moyenne du signal d'entrée, valeur qui possède quelques informations spécifiques du locuteur [48].

Nous avons choisi de prendre $D = 12$ pour le nombre de coefficients cepstraux. Ce choix est fondé sur de nombreux résultats expérimentaux. En effet, la plupart des systèmes de reconnaissance basés sur les MFCC et qui font état dans la littérature scientifique, obtiennent les meilleurs aboutissements en utilisant 12 paramètres spectraux en plus du coefficient d'énergie $\tilde{C}_0$ (soit 13 coefficients au total) [1], [47].

La figure suivante schématise le processus général de production des coefficients MFCC (fig. 21):

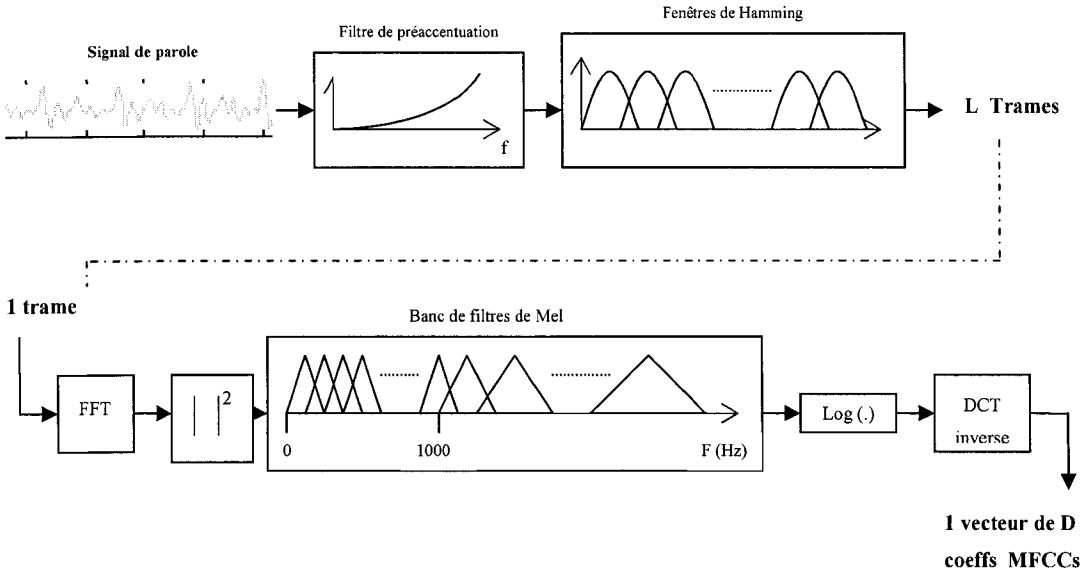


Figure 21 Processus général de production des coefficients MFCC

2.3.2.6 Dérivées temporelles des coefficients MFCC

Dans certains cas, afin de mieux paramétrer un signal de parole, on peut aussi agrandir les vecteurs de caractéristiques en incluant deux autres types de coefficients [62] :

- Les coefficients de vitesse des MFCC : (delta-cepstraux)

$$\Delta \tilde{C}_{d_t} = \sum_{\tau=-2}^{2} \Gamma[\tau](\tilde{C}_{d_{t+\tau}} - \tilde{C}_{d_{t-\tau}}) \tag{2.7}$$

- Les coefficients d'accélération des MFCC : (delta-delta-cepstraux)

$$\Delta\Delta \tilde{C}_{d_t} = \sum_{\tau=-5}^{5} \Gamma[\tau](\Delta \tilde{C}_{d_{t+\tau}} - \Delta \tilde{C}_{d_{t-\tau}}) \tag{2.8}$$

En fait, un grand nombre d'information perceptuelle nécessaire pour distinguer certains phonèmes, est contenue dans des intervalles acoustiques de courte durée qui varient rapidement. D'où l'utilité de ces coefficients [62], [67]. Cependant, cette technique n'est pas vérifiée pour tous les systèmes. Il faudra par conséquent la tester sur notre modélisation.

2.3.2.7 Motivation de l'utilisation des coefficients cepstraux MFCC

Outre la popularité de la méthode des MFCC et les excellents résultats relevés dans la littérature, il existe d'autres motivations qui nous ont poussé à adopter cette méthode de paramétrisation.

L'excitation contient de l'information prosodique ainsi que des données propres au locuteur ; cependant ces informations ne sont pas correctement modélisées dans les systèmes de reconnaissance. C'est pourquoi il est important de les filtrer afin de représenter correctement le locuteur. La déconvolution réalisée par l'opérateur logarithme a pour effet de découpler les caractéristiques du conduit vocal de celles de l'excitation glottale, et nous permet ainsi de faire la sélection des données.

Enfin, pour obtenir une représentation de bonne qualité avec la technique de modélisation que nous avons choisi (à savoir les GMM avec des matrices de covariance diagonales), il est nécessaire d'avoir des vecteurs paramétriques décorrélés [62]. La méthode des MFCC a justement cette propriété grâce à la DCT finale qui a pour effet de décorréler les éléments des vecteurs [47].

Nous venons de voir dans cette partie comment transformer un signal de parole en une séquence de vecteurs acoustiques spécifiques à chaque locuteur. Dans la partie qui suit, nous verrons comment ces vecteurs paramétriques peuvent être utilisés pour représenter un individu.

2.4 Module de modélisation et d'apprentissage

Le but du module de modélisation et d'apprentissage est de classer les objets d'intérêt, c'est-à-dire les séquences de vecteurs acoustiques extraits à partir du signal de parole, en une ou plusieurs catégories ou classes. Chaque classe est ici rattachée à un locuteur.

L'état de l'art en techniques de modélisation utilisées pour la reconnaissance du locuteur en mode indépendant du texte incluent les chaînes de Markov cachées (HMM, Hidden Markov Modeling) [41], la quantification vectorielle (VQ, Vector Quantization) [50], [51], et la modélisation par mélange de Gaussiennes, plus connue sous le nom de modélisation par GMM (Gaussian Mixture Models) [48], [52], [53]. Cette dernière est devenue depuis une dizaine d'années l'approche dominante pour les systèmes de reconnaissance du locuteur en mode indépendants du texte [54].

Nous avons choisi de représenter chaque locuteur avec la technique des GMM, qui est, selon Reynolds, une modélisation statistique robuste de l'identité d'un individu [48].

Notre choix s'est tourné vers cette méthode pour diverses raisons :

Tout d'abord, il a été montré en reconnaissance de la parole que les modèles probabilistes permettent d'aboutir à une meilleure modélisation des caractéristiques acoustiques du locuteur, ainsi qu'à une structure robuste capable de palier au bruit et à la dégradation du canal de transmission [52].

Deuxièmement, les GMM sont capable de fournir un modèle probabiliste des principaux sons contenus dans la voix d'une personne, et, contrairement aux HMM, sans imposer de contrainte entre les différentes classes de sons [52].

Enfin, la dernière motivation pour l'utilisation des GMM est basée sur des observations empiriques selon lesquelles une fonction crée par combinaison linéaire de plusieurs fonctions est capable de représenter plus fidèlement la distribution d'une grande classe d'échantillons (fig. 22) [53].

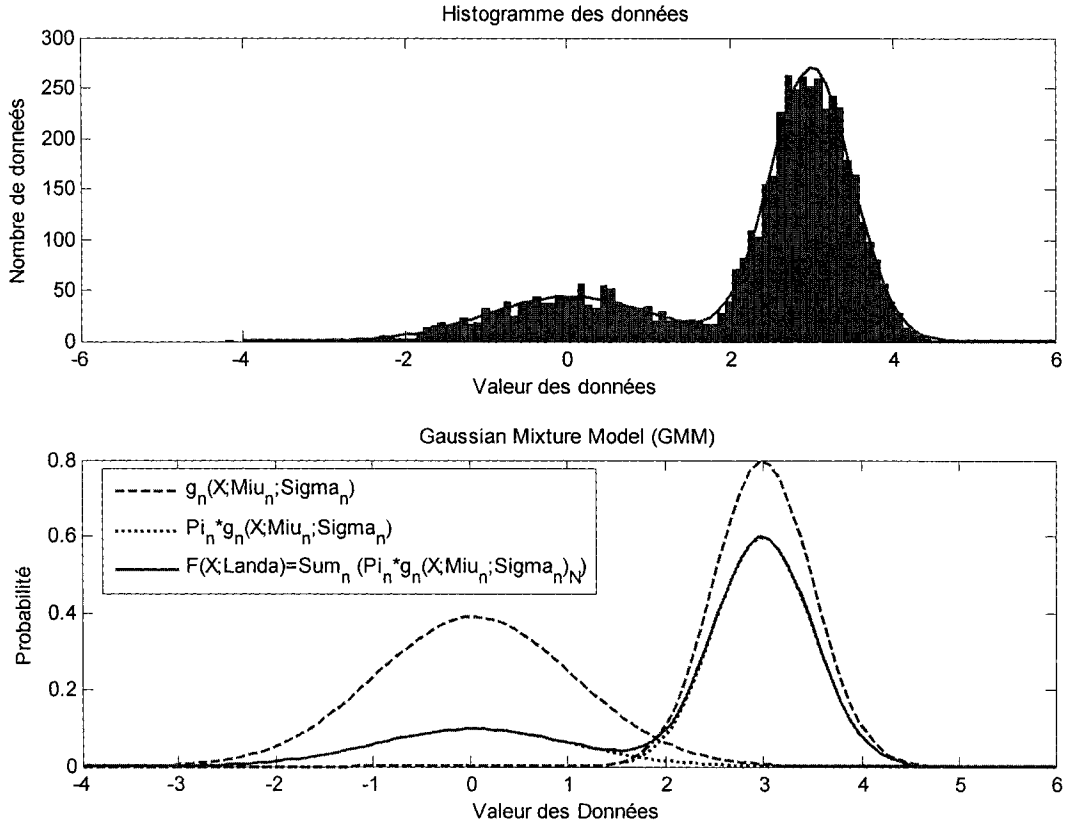


Figure 22 La fonction de modélisation de la distribution est ici créée à partir de la somme de deux fonctions Gaussiennes

Les GMM sont désormais largement utilisés et fournissent les meilleurs résultats actuels [55], [56], [62]. De plus, ils semblent également être un peu plus robustes quand les environnements d'apprentissage et de test diffèrent [57].

Nous allons donner dans ce qui suit, la définition de ce modèle, puis, nous présenterons comment sont estimés les paramètres des GMM pendant la phase d'apprentissage.

2.4.1 Modélisation du locuteur par mélange de Gaussiennes (GMM)

Les GMM sont apparus pour la première fois dans les articles de Reynolds en 1992 [48], et de Rose et Reynolds en 1995 [52], et ont démontré un taux élevé de reconnaissance pour les systèmes d'identification en mode indépendant du texte et testés sur de courts segments de parole de qualité audio identique à celle obtenue en téléphonie. De plus, selon Reynolds et Rose [52], les composantes Gaussiennes constituant les GMM modélisent parfaitement l'ensemble du contenu spectral spécifique à chaque locuteur ainsi que les principaux traits acoustiques permettant de définir l'identité d'une personne.

Le concept des GMM repose sur l'utilisation d'une série de fonctions Gaussiennes pour représenter la densité de probabilité de l'ensemble des vecteurs de caractéristiques produits par chaque locuteur (fig. 22).

2.4.2 Modèle du locuteur avec les GMM

2.4.2.1 Définition

Considérons $\underline{\mathbf{x}}_l$ [3] un vecteur aléatoire représentant un des L vecteurs de caractéristiques extraits par le module précédent. Comme on peut le voir sur la figure 23, ce vecteur est composé de D variables aléatoires telles que $\underline{\mathbf{x}}_l = \{\underline{x}_1, \underline{x}_2, ..., \underline{x}_D\}$ (de dimension $D \times 1$); chacune de ces variables correspond à un coefficient obtenu par l'analyse cepstrale effectuée dans le module précédent.

---

[3] On distinguera le caractère aléatoire d'un vecteur ou d'une variable par une lettre de notation soulignée. De plus, on distinguera les vecteurs des variables, en mettant la lettre représentative du vecteur en gras.

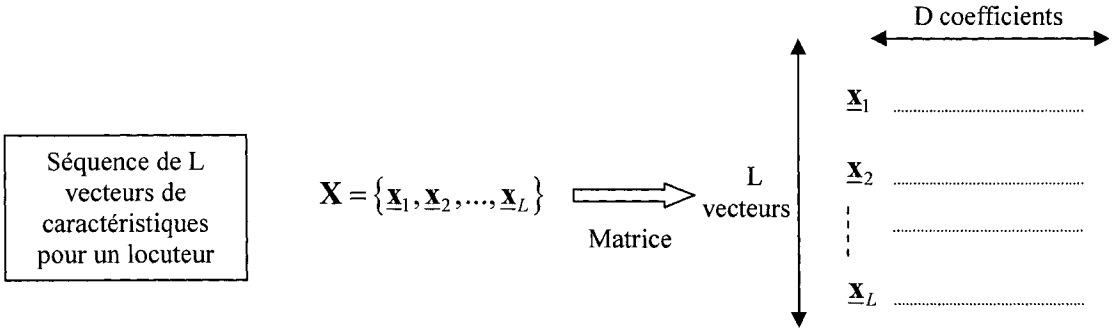


Figure 23 Composition d'une suite $\mathbf{X}$ de vecteurs de caractéristiques extraite d'un locuteur

Nous rappelons que le modèle présenté ici est similaire à celui développé par Reynolds [48], [52], [53].

La densité de probabilité du GMM λ, associé au locuteur que l'on souhaite modéliser, est définie comme étant une somme pondérée de N densités de probabilité unimodales $g_n(\underline{\mathbf{x}}_l)$ suivant une loi de probabilité Gaussienne (Fig. 24):

$$p(\underline{\mathbf{x}}_l|\lambda) = \sum_{n=1}^{N} \Pi_n p(\underline{\mathbf{x}}_l|\lambda_n) = \sum_{n=1}^{N} \Pi_n g_n(\underline{\mathbf{x}}_l) \tag{2.9}$$

De plus, notons que chacune des composantes $g_n(\underline{\mathbf{x}}_l)$ du GMM est paramétré par son vecteur de moyennes $\underline{\boldsymbol{\mu}}_n$ (de dimension $D\times 1$), et par sa matrice de covariance Σ_n (de dimension $D\times D$), tels que:

$$g_n(\underline{\mathbf{x}}_l) = \frac{1}{(2\pi)^{D/2}|\Sigma_n|^{1/2}} \times e^{-\frac{1}{2}(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n)^T (\Sigma_n)^{-1}(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n)} \tag{2.10}$$

et $$\underline{\boldsymbol{\mu}}_n = E\left[\underline{\mathbf{x}}_l\right]$$

et $$\Sigma_n = E\left[(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n)(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n)^T\right]$$

Les poids Π_n du mélange de Gaussiennes doivent de plus vérifier la contrainte suivante:

$$\sum_{n=1}^{N} \Pi_n = 1 \tag{2.11}$$

Par conséquent, on peut définir les paramètres de chaque GMM, et donc de chaque locuteur r pour la phase d'identification, par son model λ_r (fig. 25):

$$\lambda_r = \left\{\underline{\boldsymbol{\mu}}_n^r, \Sigma_n^r, \Pi_n^r\right\}, \qquad \text{avec } n = 1,...,N \tag{2.12}$$

où $\underline{\boldsymbol{\mu}}_n$, Σ_n, et Π_n , représentent la moyenne, la variance, et la pondération de la $n^{ème}$ Gaussienne du mélange.

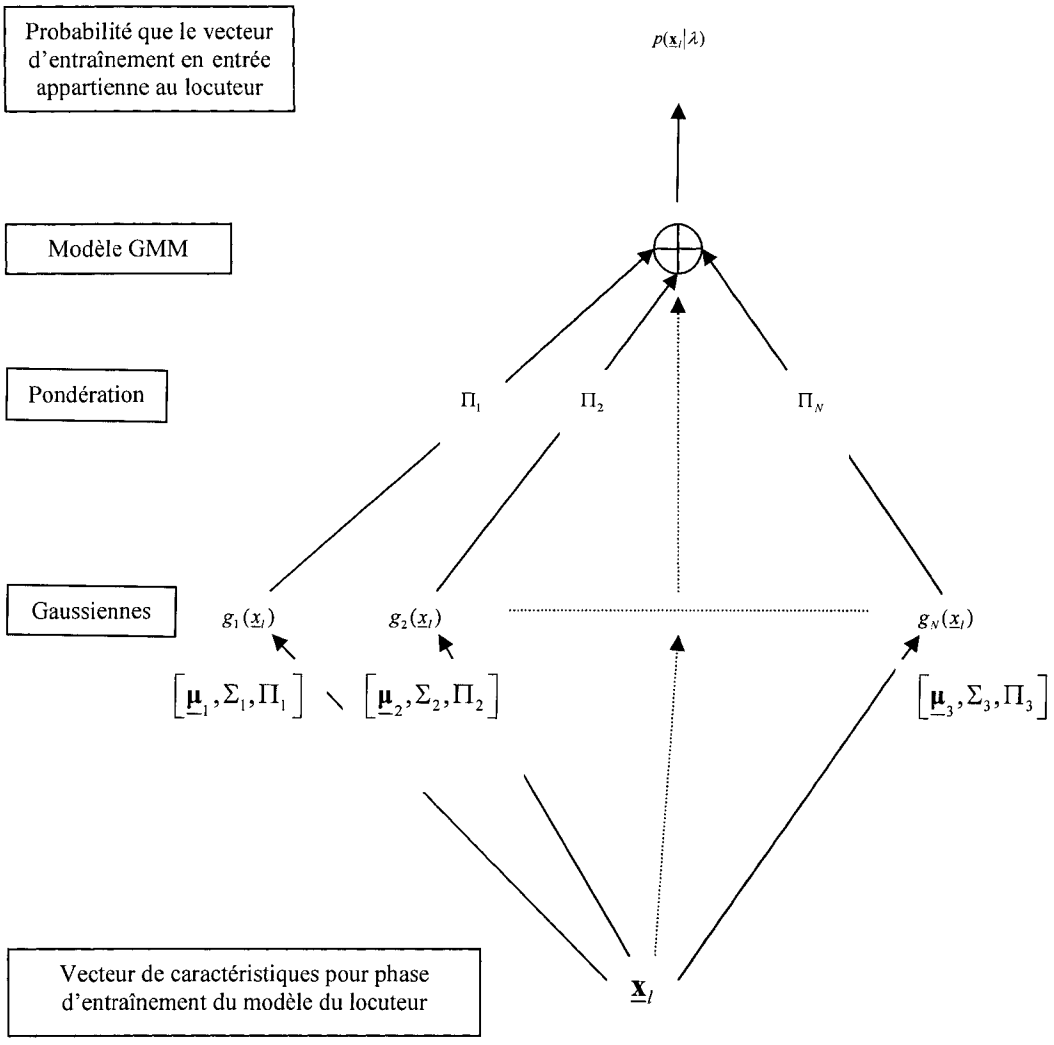


Figure 24 Schématisation de la modélisation GMM d'un locuteur

2.4.2.2 Type de modèle

Les GMM peuvent avoir différentes formes dépendamment du choix de la matrice de covariance [52]. En fait, le model peut soit avoir une matrice de covariance par composante Gaussienne comme indiqué dans l'équation (2.12) (la covariance est dite nodale), soit une matrice de covariance pour tout le modèle du locuteur (la covariance

est dite grande), soit enfin, une simple matrice de covariance partagée par tous les modèles de locuteurs (ici, la covariance est dite globale).

On peut également distinguer deux types de matrices de covariance : elle peut être soit complète soit diagonale [52].

Pour notre projet, nous avons choisi d'utiliser des matrices de covariance nodales diagonales pour la modélisation des locuteurs. Ce choix est basé sur de premiers résultats expérimentaux qui ont démontrés de meilleurs taux de reconnaissance [52].

Aussi, comme les composantes Gaussiennes agissent simultanément pour modéliser la totalité du pdf (probability density function), les matrices complètes ne sont pas nécessaires même si les vecteurs de données ne sont pas statistiquement indépendants. De plus, la combinaison linéaire de Gaussiennes à covariances diagonales est capable de modéliser les corrélations entre les éléments de chaque vecteur [52].

2.4.3 Estimation à maximum de vraisemblance des paramètres des GMM – phase d'entraînement

Comme nous l'avons mentionné au tout début de cette partie, les systèmes de reconnaissance doivent passer par deux différentes phases lors de leur fonctionnement : la phase d'entraînement, puis la phase de test.

La phase d'entraînement consiste en fait à entraîner le model λ_r du locuteur r avec les échantillons de parole de ce dernier : à partir de la distribution des L vecteurs de caractéristiques d'entraînement, on estime les paramètres du GMM qui permettent de donner naissance au modèle qui épousera au mieux cette distribution (Fig. 25).

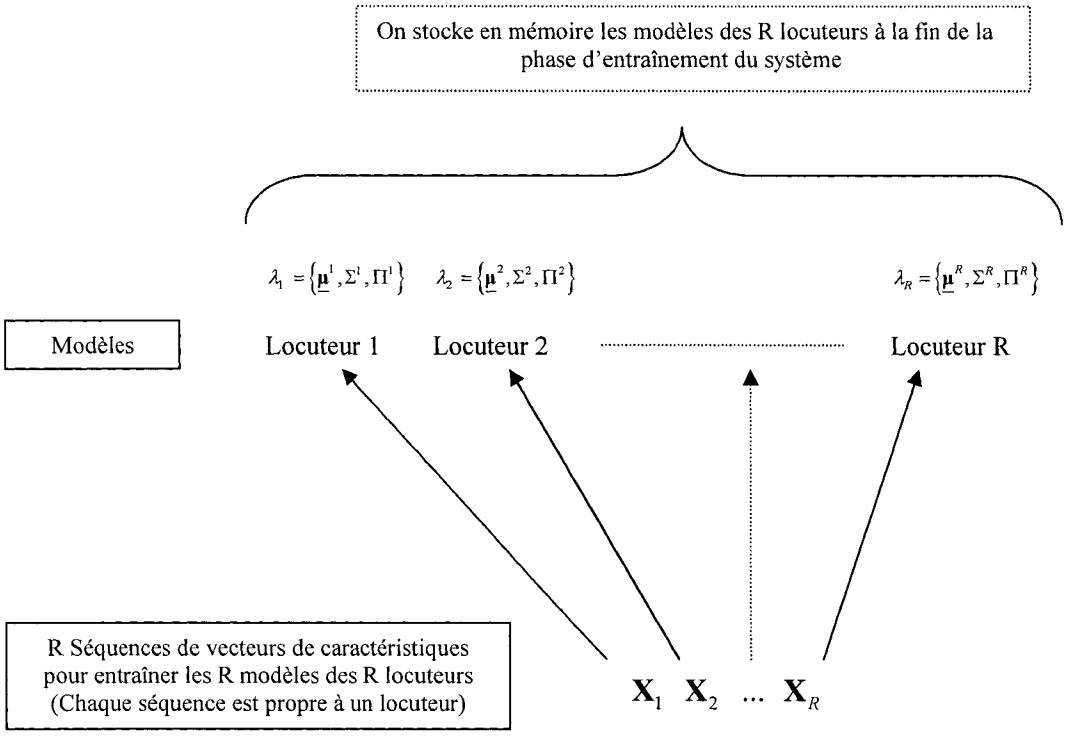


Figure 25 Phase d'entraînement du système

Il existe plusieurs techniques pour estimer les paramètres d'un GMM [58]. Nous avons choisi d'utiliser la plus populaire et la mieux établie de toutes, c'est-à-dire, l'estimation à maximum de vraisemblance (MLE, Maximum Likelihood Estimation). Cette méthode simple permet d'aboutir la plupart du temps à un bon estimateur. En fait, on peut montrer que, sous certaines conditions (qui sont quasiment toujours vérifiées en pratique), l'estimateur ML est consistent [59]; ceci explique pourquoi il est très largement utilisé.

Le but de cette estimation est d'évaluer les paramètres $\{\underline{\mu}_n, \Sigma_n, \Pi_n\}$ du modèle λ maximisant la vraisemblance du GMM, à partir des données d'entraînement.

2.4.3.1 Maximisation directe

Considérons une séquence de L vecteurs d'entraînement $\mathbf{X} = \{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, ..., \underline{\mathbf{x}}_L\}$, la vraisemblance du GMM s'écrit :

$$p(\mathbf{X}|\lambda) = \prod_{l=1}^{L} p(\underline{\mathbf{x}}_l|\lambda) \tag{2.13}$$

On notera que $\mathbf{X}$ [4] est une matrice de taille $L \times D$, chaque ligne étant un vecteur de caractéristiques $\underline{\mathbf{x}}_l$ formé de D coefficients.

Les équations (2.9) et (2.13) nous permettent d'écrire :

$$p(\mathbf{X}|\lambda) = \prod_{l=1}^{L} \sum_{n=1}^{N} \Pi_n p(\underline{\mathbf{x}}_l|\lambda_n) \tag{2.14}$$

L'estimée ML de λ est obtenue en calculant les valeurs paramétriques du modèle qui maximisent le logarithme de la séquence de L vecteurs de caractéristiques :

$$p(\mathbf{X}|\lambda) = \prod_{l=1}^{L} \sum_{n=1}^{N} \Pi_n \log[p(\underline{\mathbf{x}}_l|\lambda_n)] \tag{2.15}$$

Malheureusement, cette expression n'est pas une fonction linéaire du paramètre λ et par conséquent la maximisation directe de $p(\mathbf{X}|\lambda)$ en résolvant $\frac{dp(\mathbf{X}|\lambda)}{d\lambda} = 0$ est impossible.

---

[4] Dans le programme en langage Matlab, la matrice X correspond à la variable *data*, matrice regroupant tous les vecteurs de coefficients cepstraux pour la phase d'entraînement.

2.4.3.2 Maximisation à l'aide de l'algorithme EM

Dempster et al. proposèrent en 1977 l'algorithme itératif EM (Expectation-Maximization) qui est de nos jours très utilisé pour déterminer les paramètres du model de chaque locuteur réalisant le maximum de vraisemblance [60] [61].

L'algorithme EM possède la propriété de faire croître à chaque itération la vraisemblance $p(\mathbf{X}|\lambda)$ du GMM, ce qui permet de garantir sous certaines conditions, la convergence du résultat vers un maximum local (fig. 26).

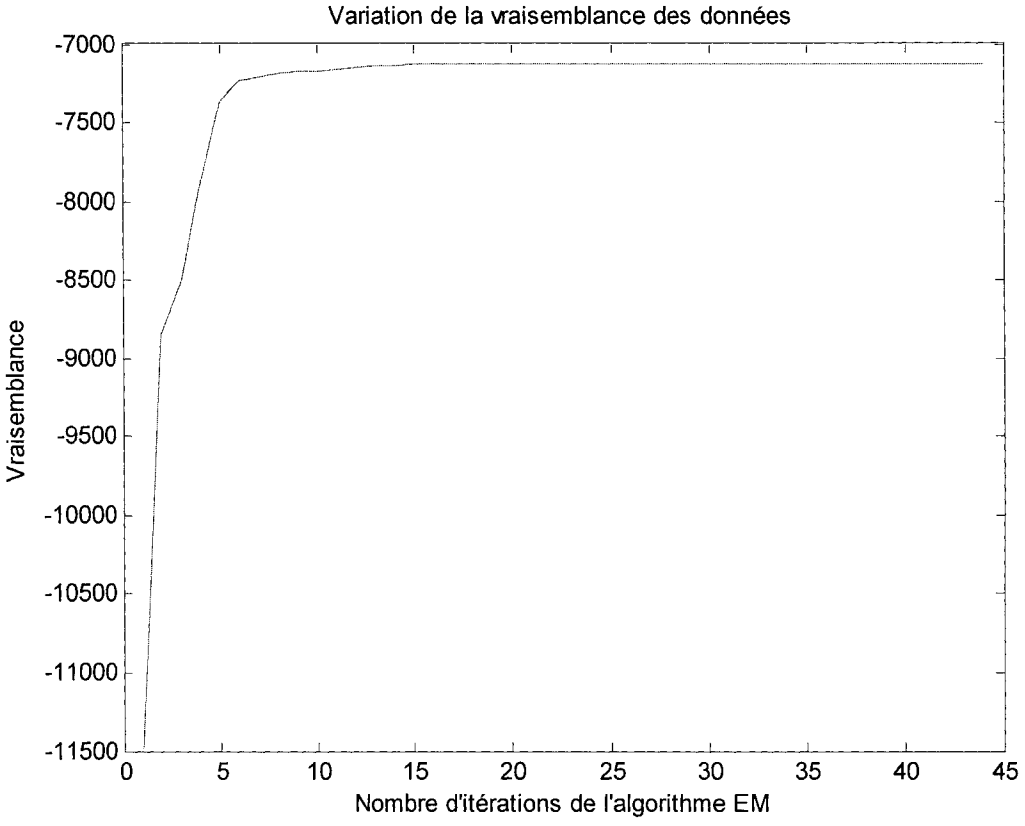


Figure 26 Variation de la vraisemblance des données

Nous observons ci-dessous que l'algorithme converge vers un maximum local en moins de 20 itérations.

Les vecteurs de caractéristiques $\underline{\mathbf{x}}_l$ sont en réalité des données incomplètes du locuteur [60]. En effet, on ne peut pas dire quelle variable aléatoire de $\underline{\mathbf{x}}_l$ fut engendrée par quelle composante. Par conséquent, les données manquantes sont les labels indiquant à quelle Gaussienne appartient chaque point du vecteur de données.

Afin de compléter ces données, nous allons attribuer un label $\underline{z}_{l;d}$ à chaque variable aléatoire $\underline{x}_d$ de $\underline{\mathbf{x}}_l$. Ainsi, on définit le nouveau vecteur paramétrique de la façon suivante:

$$\underline{\mathbf{y}}_l = (\underline{\mathbf{x}}_l; \underline{\mathbf{z}}_l) \text{ avec } \underline{y}_d = (\underline{x}_d; \underline{z}_{l;d}) \tag{2.16}$$

$$\text{et avec } \underline{z}_{l;d} = \begin{cases} 1 \text{ , si } \underline{x}_d \text{ est engendrée par la d-ième composante Gaussienne} \\ 0 \text{ , autrement} \end{cases}$$

Le vecteur $\underline{\mathbf{x}}_l$ est dit observable, tandis que le vecteur de labels $\underline{\mathbf{z}}_l$ est une donnée dite non observable.

1ère partie principale de l'algorithme : Expectative (E)
Le but ici est d'obtenir la relation qui nous donnera :

$$E_{\underline{\mathbf{z}}_l}(\ln p(\mathbf{Y}|\lambda)|\mathbf{X}, \lambda^j) = F(\lambda; \lambda^j) \tag{2.17}$$

où j est le pas de l'algorithme.

Tout d'abord, nous devons trouver l'expression $\ln p(\mathbf{Y}|\lambda)$:

$$p(\mathbf{Y}|\lambda) = \prod_{l=1}^{L} p(\underline{\mathbf{y}}_l|\lambda) \tag{2.18}$$

Nous savons que : $p(\underline{\mathbf{y}}_l|\lambda) = p(\underline{\mathbf{x}}_l;\underline{\mathbf{z}}_l|\lambda) = p(\underline{\mathbf{x}}_l|\underline{\mathbf{z}}_l,\lambda).p(\underline{\mathbf{z}}_l|\lambda) = \Pi_n p(\underline{\mathbf{x}}_l|\lambda_n)$ (2.19)

avec $\underline{z}_{l;d} = 1$ et $\underline{z}_{l;n} = 0$ pour $n \neq d$

(2.20)

Cette expression peut alors s'écrire : $p(\underline{\mathbf{y}}_l|\lambda) = \prod_{n=1}^{N}\left[\Pi_n p(\underline{\mathbf{x}}_l|\lambda_n)\right]^{\underline{z}_n}$

Alors on peut écrire l'expression (2.9) de la façon suivante

$$p(\mathbf{Y}|\lambda) = \prod_{l=1}^{L}\prod_{n=1}^{N}\left[\Pi_n p(\underline{\mathbf{x}}_l|\lambda_n)\right]^{\underline{z}_n} \tag{2.21}$$

Et finalement, on peut trouver le logarithme de cette dernière :

$$\ln p(\mathbf{Y}|\lambda) = \sum_{l=1}^{L}\sum_{n=1}^{N}\underline{z}_{l;n}\ln(\Pi_n p(\underline{\mathbf{x}}_l|\lambda_n)) = \sum_{l=1}^{L}\sum_{n=1}^{N}\underline{z}_{l;n}\ln(\Pi_n) + \sum_{l=1}^{L}\sum_{n=1}^{N}\underline{z}_{l;n}\ln(p(\underline{\mathbf{x}}_l|\lambda_n)) \tag{2.22}$$

À présent, il nous reste à calculer l'espérance de $\ln p(\mathbf{Y}|\lambda)$:

$$\begin{aligned} E\left[\ln p(\mathbf{Y}|\lambda)\right] &= E_{\underline{\mathbf{z}}_l}(\ln p(\mathbf{Y}|\lambda)|\mathbf{X},\lambda^j) \\ &= \sum_{l=1}^{L}\sum_{n=1}^{N}E\left[\underline{z}_{l;n}\right]\ln(\Pi_n^j) + \sum_{l=1}^{L}\sum_{n=1}^{N}E\left[\underline{z}_{l;n}\right]\ln(p(\underline{\mathbf{x}}_l|\lambda_n^j)) \end{aligned} \tag{2.23}$$

Puisque $E\left[z_{l;n}\right] = \sum_h z_{l;h}P(z_{l;h}) = P(z_{l;n}) = P(n|\underline{\mathbf{x}}_l)$, (2.24)

on peut dire que $E\left[z_{l;n}\right]$ correspond à la probabilité que le vecteur $\underline{\mathbf{x}}_l$ ait été généré par la n-ième composante Gaussienne, et ainsi:

$$E\left[z_{l;n}\right] = \frac{\Pi_n^j p(\underline{\mathbf{x}}_l \mid \lambda_n^j)}{\sum_{h=1}^{N} \Pi_h^j p(\underline{\mathbf{x}}_l \mid \lambda_h^j)} \tag{2.25}$$

2ème partie principale de l'algorithme : Maximisation (M)

Le but ici est de trouver la valeur des paramètres de λ qui maximisent l'expression (2.9) en tenant compte de la contrainte (2.11) portant sur les poids :

$$\lambda^{j+1} = \arg\max_{\underline{\lambda}} F(\lambda; \lambda^j) \tag{2.26}$$

Pour cela on dérive l'expression ci-dessus, ce qui nous donne :

$$F'(\lambda; \lambda^j) = \sum_{l=1}^{L}\sum_{n=1}^{N} E\left[\underline{z}_{l;n}\right] \ln(\Pi_n^j) + \sum_{l=1}^{L}\sum_{n=1}^{N} E\left[z_{l;n}\right] \ln(p(\underline{\mathbf{x}}_l \mid \lambda_n^j)) + \kappa(1 - \sum_{n=1}^{N} \Pi_n) \tag{2.27}$$

où κ est le multiplieur de Lagrange (voir annexe 2).

La contrainte (2.11) nous permet de trouver $\kappa = \sum_{l=1}^{L}\sum_{n=1}^{N} E\left[\underline{z}_{l;n}\right]$ (2.28)

Par conséquent, on peut conclure que $F'(\lambda; \lambda^j)$ est maximum si

$$\begin{cases} \dfrac{\partial F'}{\partial \Pi_n} = 0 \Rightarrow \sum_{l=1}^{L} E(\underline{z}_{l;n}) \dfrac{1}{\Pi_n} = 0 \Rightarrow \Pi_n^{j+1} = \dfrac{1}{L} \sum_{l=1}^{L} E(\underline{z}_{l;n}) & (2.29) \\ \dfrac{\partial F'}{\partial \mu_n} = 0 \Rightarrow \underline{\boldsymbol{\mu}}_n^{j+1} = \dfrac{1}{L\Pi_n^{j+1}} \sum_{l=1}^{L} E(\underline{z}_{l;n}) \underline{\mathbf{x}}_l & (2.30) \\ \dfrac{\partial F'}{\partial \Sigma_n} = 0 \Rightarrow \Sigma_n^{j+1} = \dfrac{1}{L\Pi_n^{j+1}} \sum_{l=1}^{L} E(\underline{z}_{l;n})(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n^{j+1})(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n^{j+1})^T & (2.31) \end{cases}$$

En résumé, nous pouvons exprimer l'algorithme EM en quatre étapes ; dans chaque itération de l'algorithme EM, les formules suivantes de ré-estimation sont utilisées garantissant une croissance monotone à la vraisemblance du modèle :

Initialisation:

L'initialisation est une étape importante car l'algorithme EM est très sensible aux conditions initiales. Il nous faut par conséquent adopter une procédure efficace d'initialisation des paramètres. D'après nos lectures dans la littérature, l'algorithme des K-means [52], [59] (voir annexe 3) est le plus efficace pour initialiser les centres $\underline{\boldsymbol{\mu}}_n$ du modèle[5] :

$$\lambda_n^0 = \left\{ \underline{\boldsymbol{\mu}}_n^0, \Sigma_n^0, \Pi_n^0 \right\} \tag{2.32}$$

L'initialisation des paramètres Π_n et Σ_n s'effectue conformément à [52].

Nous sommes maintenant en mesure d'estimer les paramètres d'un nouveau modèle $\hat{\lambda}$, pour lequel :

$$p(\mathbf{X}|\hat{\lambda}) \geq p(\mathbf{X}|\lambda) \tag{2.33}$$

---

[5] Nous verrons dans les chapitres suivants consacrés aux essais expérimentaux et aux résultats obtenus, qu'il existe plusieurs méthodes d'initialisation des centres, et nous verrons quelle est la plus efficace.

Expectative (E): Tout d'abord, on détermine la vraisemblance moyenne de la séquence de données en utilisant l'estimée j du ML du modèle λ : (2.25)

$$E\left[z_{l;n}\right] = \frac{\Pi_n^j p(\underline{\mathbf{x}}_l | \lambda_n^j)}{\sum_{h=1}^{N} \Pi_h^j p(\underline{\mathbf{x}}_l | \lambda_h^j)}$$

Maximisation (M): Dans cette étape, on maximise la vraisemblance du modèle pour déterminer une nouvelle estimation de λ au pas $j+1$, à savoir, λ^{j+1} [(2.29) ;(2.30) ;(2.31)] :

$$\text{Pondération: } \Pi_n^{j+1} = \frac{1}{L}\sum_{l=1}^{L} E(\underline{z}_{l;n})$$

$$\text{Moyenne: } \underline{\boldsymbol{\mu}}_n^{j+1} = \frac{1}{L\Pi_n^{j+1}}\sum_{l=1}^{L} E(\underline{z}_{l;n})\underline{\mathbf{x}}_l$$

$$\text{Covariance: } \Sigma_n^{j+1} = \frac{1}{L\Pi_n^{j+1}}\sum_{l=1}^{L} E(\underline{z}_{l;n})(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n^{j+1})(\underline{\mathbf{x}}_l - \underline{\boldsymbol{\mu}}_n^{j+1})^T$$

Convergence: Si à l'itération $j+1, \left\|\lambda^{j+1} - \lambda^j\right\| \leq \varepsilon$, alors l'algorithme converge vers un maximum local de vraisemblance, et on a par conséquent atteint la dernière itération. Sinon, le nouveau modèle devient ensuite le modèle initial pour la prochaine itération et le processus est répété (à partir de l'étape 2) jusqu'à ce que le seuil de convergence soit atteint (fig. 27).

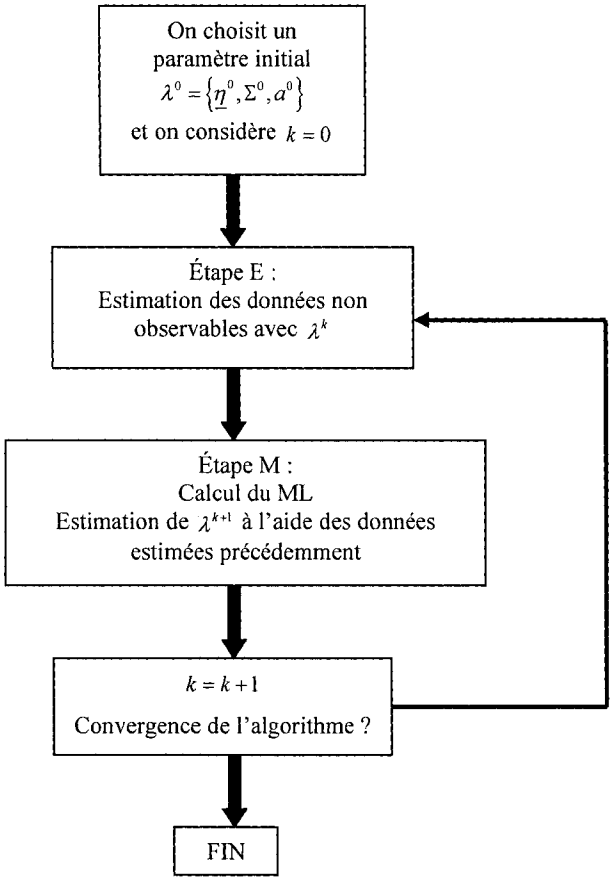


Figure 27 Schéma récapitulatif de l'algorithme EM

Une fois la phase d'entraînement terminée, nous pouvons stocker les modèles dans une base de données (fig. 25). Les paramètres ainsi obtenus permettront de modéliser fidèlement la distribution des données. À titre d'exemple, nous pouvons observer la figure 27 qui nous montre le GMM d'une séquence L d'un locuteur, ainsi que ses composantes Gaussiennes.

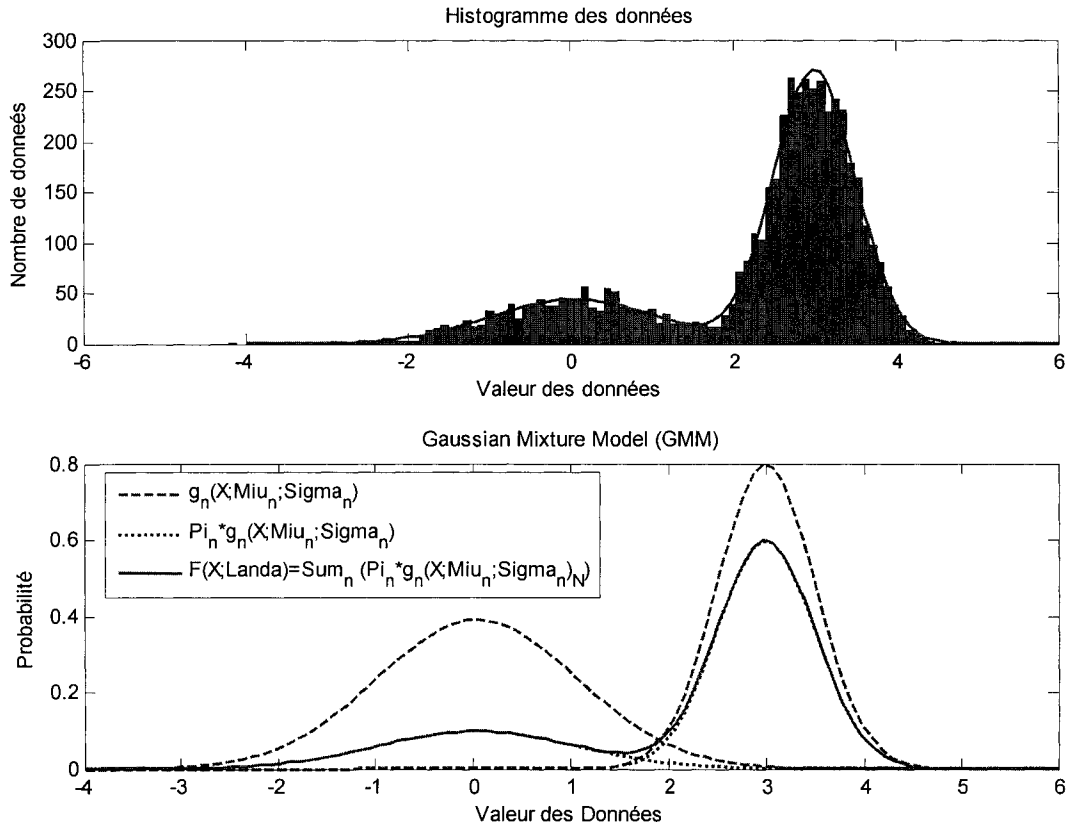


Figure 28 Modélisation GMM des données L d'un locuteur

2.5 Module de reconnaissance - phase de test

Finalement, nous abordons le dernier module constituant le système de reconnaissance. C'est ici que, au cours de la phase de test, sera prise la décision concernant l'identité du locuteur. Dans ce qui suit, nous présentons la technique utilisée dans la phase de reconnaissance.

2.5.1 Étape préliminaire

Pour la phase de test, nous considérons de nouveaux échantillons de parole. Cependant, nous ne sommes plus en mesure de dire à qui appartiennent ces échantillons. C'est le système qui devra nous fournir la réponse. Pour cela, nous devons reprendre toute la procédure de prétraitement et de paramétrisation vue précédemment. Nous obtenons ainsi de nouvelles séquences de vecteurs de caractéristiques $\mathbf{X} = \{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, ..., \underline{\mathbf{x}}_L\}$ pour chaque locuteur. Ces séquences test sont ensuite amenées à l'entrée de l'identificateur (fig. 29).

2.5.2 Identification du locuteur

Le système d'identification est purement et simplement un classificateur a maximum de vraisemblance.

Pour une base de données de R locuteurs $r = \{1, 2, ..., R\}$ représentée par les GMMs $\lambda_1, \lambda_2, ...\lambda_R$, l'objectif est de trouver le modèle du locuteur qui a la plus grande probabilité a posteriori pour la séquence $\mathbf{X} = \{\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, ..., \underline{\mathbf{x}}_L\}$ donnée en entrée.

En appliquant le minimum d'erreur pour la loi de décision de Bayes, on peut estimer :

$$\tilde{Sc} = \arg\max_{1 \le r \le R} \mathrm{P}_{priori}(\lambda_r | \mathbf{X}) = \arg\max_{1 \le r \le R} \frac{p(\mathbf{X}|\lambda_r)\mathrm{P}_{priori}(\lambda_r)}{p(\mathbf{X})} \tag{2.34}$$

En assumant l'égale probabilité a priori de tous les locuteurs :

$$\mathrm{P}_{priori}(\lambda_r) = 1/R \tag{2.35}$$

et en remarquant que les termes $P_{priori}(\lambda_r)$ et $p(\mathbf{X})$ sont constants pour tous les modèles de locuteurs, on peut alors simplifier loi de classification précédente et écrire que :

$$\tilde{Sc} = \arg\max_{1\le r\le R} p(\mathbf{X}|\lambda_r) \tag{2.36}$$

En considérant que les données sont indépendantes, et en appliquant le logarithme, on a :

$$\log[p(\mathbf{X}|\lambda_r)] = \sum_{l=1}^{L} \log p(\underline{\mathbf{x}}_l|\lambda_r) \tag{2.37}$$

Et par conséquent, le système d'identification va établir un score correspondant au logarithme du maximum de vraisemblance de la séquence de vecteurs de caractéristiques test pour chaque locuteur, à savoir :

$$\tilde{Sc}_{Log} = \arg\max_{1\le r\le R} \sum_{l=1}^{L} \log p(\underline{\mathbf{x}}_l|\lambda_r) \tag{2.38}$$

et où $p(\underline{\mathbf{x}}_l|\lambda_r)$ est donné dans l'équation (2.9).

Le GMM qui génère le plus grand score $\tilde{Sc}_{Log}$ est identifié comme étant le modèle du signal de parole testé.

Un schéma block du système d'identification du locuteur est représenté dans la figure suivante (fig. 29) :

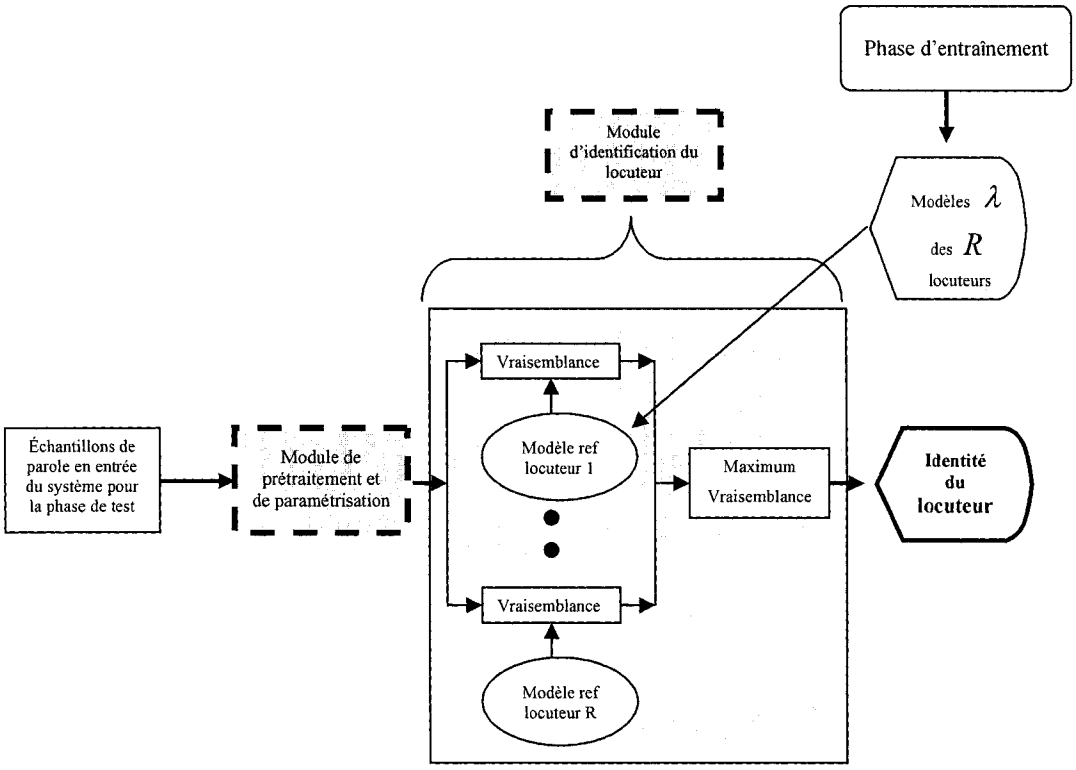


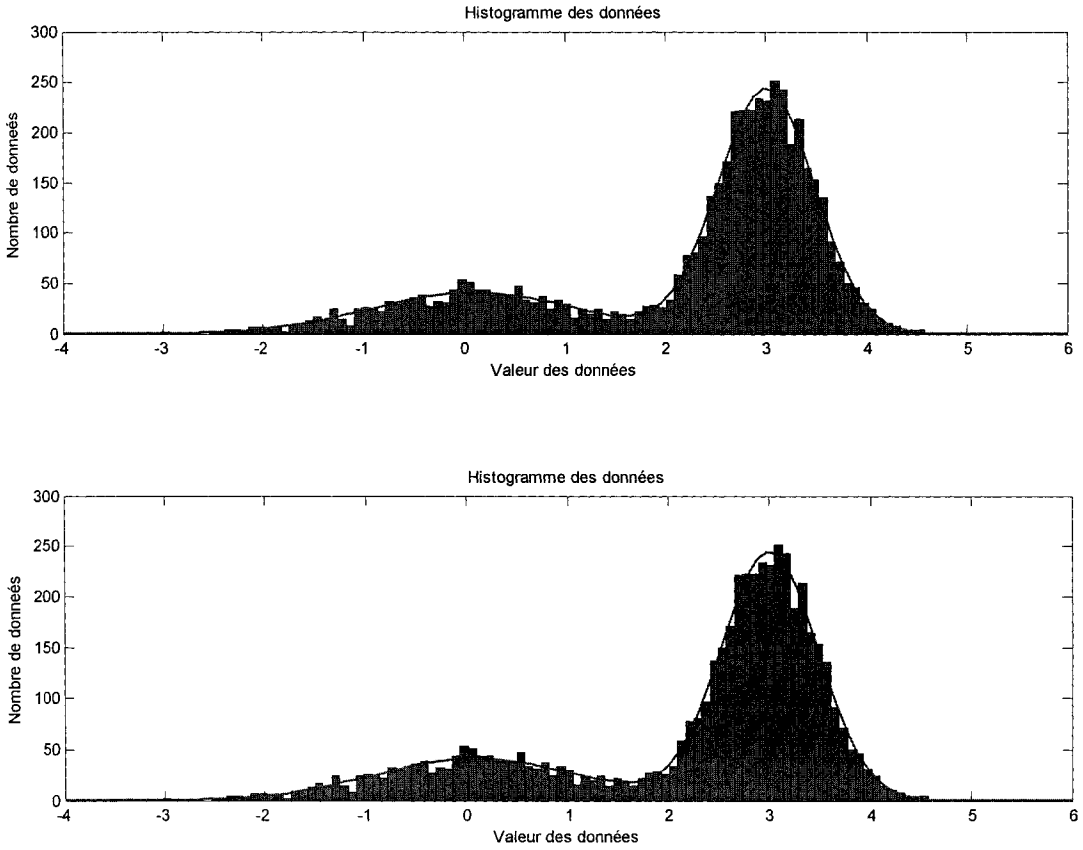
Figure 29 Schéma bloc de la phase de test du locuteur

2.6 Derniers réglages de perfectionnement pour la mise au point du système de référence pour la reconnaissance du locuteur

Il existe notamment deux facteurs critiques que l'on rencontre lors de l'entraînement du modèle GMM [52]. Tout d'abord, la sélection de l'ordre N du mélange de gaussiennes, et enfin, la méthode d'initialisation des paramètres du modèle dans l'algorithme EM. En fait, il n'existe pas de moyens théoriques qui permettent de nous aider à faire ces choix, seuls des essais expérimentaux pour des cas précis peuvent nous y aider.

2.6.1 Choix du nombre de Gaussiennes pour les GMM

Le choix du nombre de Gaussiennes N pour représenter le locuteur est très important, et est directement lié au taux de reconnaissance que nous donnera le système. Ceci est facilement explicable par le fait que plus on dispose de composantes pour modéliser la distribution des données, à savoir les coefficients cepstraux, plus précise sera la représentation (fig. 30).



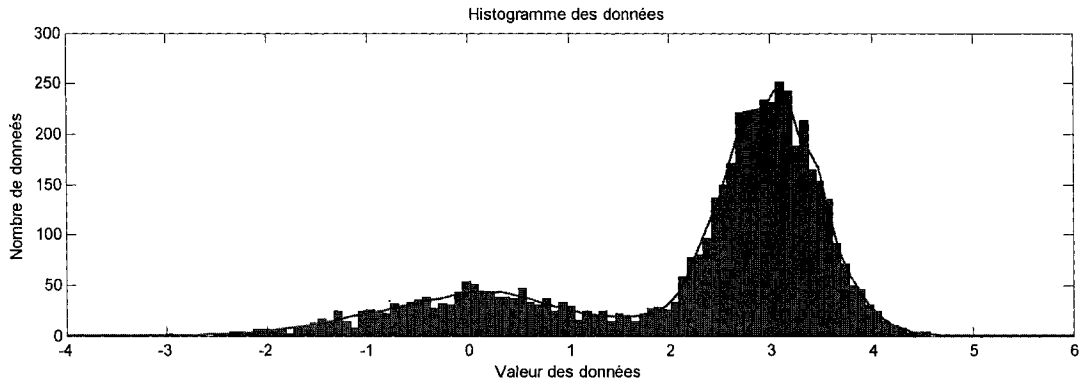


Figure 30 Modélisation GMM des vecteurs caractéristiques extraits à partir de la parole d'un locuteur (Distribution des coefficients cepstraux), avec (a) 2 Gaussienne, (b) 4 Gaussiennes, (c) 18 Gaussiennes

Donc a priori, plus on utilise de Gaussiennes et meilleurs seront nos résultats. Ceci est vrai jusqu'à un certain point. En effet, lorsqu'on augmente le nombre de Gaussiennes jusqu'à près d'une vingtaine, on augmente aussi le taux de reconnaissance. Cependant, si on continue d'augmenter le nombre de composantes, les performances du système se mettent à décroître (voir fig. 31). Ces résultats furent prouvés expérimentalement [1] par de nombreux chercheurs qui ont montré qu'à partir d'un certain point (dépendant du nombre de données d'entraînement), on ne peu plus rendre notre modèle plus précis; les effets d'un surplus d'éléments Gaussiens sera une dégradation des résultats traduite par une baisse du taux de reconnaissance et une augmentation considérable du temps d'exécution (avec des calculs inutiles).

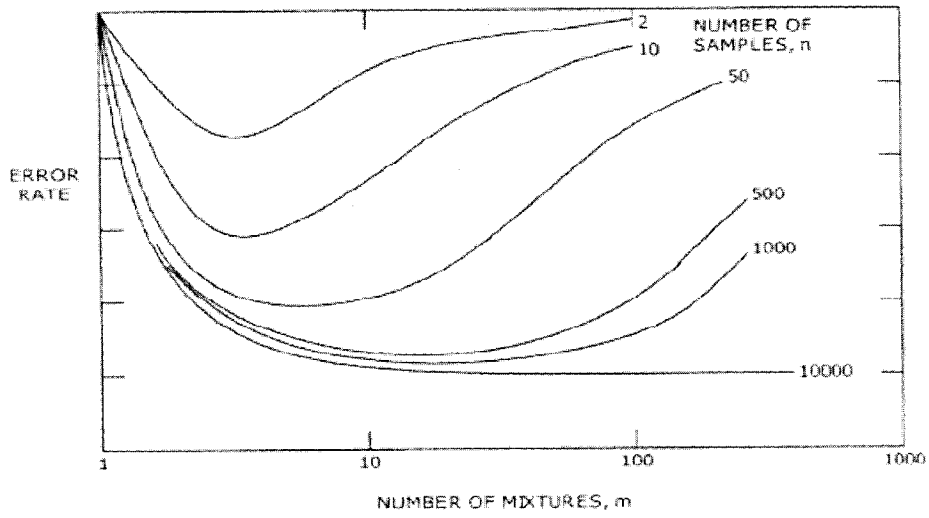


Figure 31 Entraînement et généralisation des GMM [21]

Ainsi, nous savons que 20 est un ordre de grandeur pour le nombre de Gaussiennes pour la modélisation GMM [21]. Cependant, nous ne pouvons trouver la valeur exacte du meilleur nombre qu'après avoir effectué des essais expérimentaux, car celle-ci diffère d'un système à l'autre suivant les données modélisées. En testant le système que nous venons de concevoir, on verra que dans notre cas, le meilleur taux de reconnaissance est obtenu pour 18 Gaussiennes (tableau II).

2.6.2 Initialisation des paramètres du modèle pour l'algorithme EM

L'initialisation des paramètres $\lambda_n^0 = \left\{\underline{\boldsymbol{\mu}}_n^0, \Sigma_n^0, \Pi_n^0\right\}$, peut être vue comme le point de départ de la phase d'entraînement du modèle à l'aide de l'algorithme EM.

2.6.2.1 Méthode d'initialisation des poids

Pour de simples raisons d'équiprobabilité, on initialise chaque poids avec la même valeur, à savoir [52] :

$$\Pi_n = 1/N \tag{2.39}$$

On rappelle que N correspond au nombre total de composantes Gaussiennes du mélange.

2.6.2.2 Méthode d'initialisation des centres

La méthode d'initialisation des centres que nous avons choisie, est celle des K-means. En effet, cette dernière est celle qui est la plus répandue et aussi celle qui permet d'obtenir les meilleurs résultats [59].

L'algorithme des K-means n'est autre qu'une procédure permettant de partitionner un ensemble de L vecteurs de données $\mathbf{x}_l$ en N ensembles disjoints S_j contenant L_j données de telle forme que le critère de la somme des carrés soit minimisé :

$$J = \sum_{j=1}^{N} \sum_{l \in S_j} \left| \mathbf{x}_l - \boldsymbol{\mu}_j \right|^2 \tag{2.40}$$

où $\boldsymbol{\mu}_j$ est le centre géométrique de l'ensemble S_j. Chacune des N régions d'intérêt sera modélisée par une composante Gaussienne ; par conséquent, on aura autant de centres que de Gaussiennes pour le GMM (voir fig. 32).

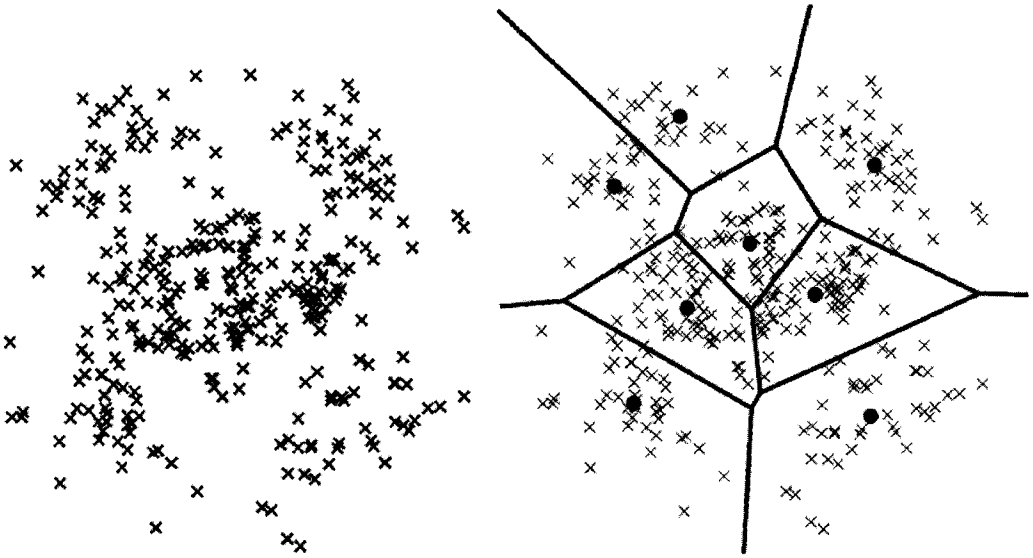


Figure 32 Division de l'espace de données en régions d'intérêt. Pour chaque région on calcule son centre géométrique [59]

En fait, la procédure adoptée est la suivante [59] :

- Premièrement, on divise aléatoirement l'espace des données en N ensembles.

- Puis, on calcule le centre de chacun de ces ensembles.

Il existe différentes manières de choisir les centres pour utiliser l'algorithme des Kmeans : on peut soit prendre aléatoirement une donnée comme étant le centre de la région, soit prendre comme centre la donnée la plus proche du vecteur moyenne des données appartenant à cette région, ou encore prendre la donnée la plus éloignée à ce même vecteur, ou enfin, prendre tout simplement le centre géométrique de ces ensembles. Expérimentalement, nous avons obtenu les meilleures performances en choisissant la deuxième hypothèse, c'est-à-dire en prenant les centres comme étant les données les plus proches des vecteurs moyenne.

- Ensuite, on assigne chaque point de donnée à la région dont le centre est le plus proche de celui-ci.

- Enfin, on ré-itère les deux dernières étapes jusqu'à ce que le critère (2.40) soit minimisé. Ceci sera traduit par une absence de changements dans la position des centres recherchés.

Pour plus de renseignements concernant cet algorithme, veuillez vous reporter à l'annexe 3.

2.6.2.3 Méthode d'initialisation des covariances

Les matrices de covariance sont calculées comme étant la covariance discrète des points associés (c'est-à-dire les plus proches) aux centres considérés.

2.6.3 Nombre d'itérations pour l'algorithme EM

Le nombre d'itérations pour l'algorithme EM représente le nombre de fois que les paramètres du modèle du locuteur vont être re-évalués afin de faire croître la vraisemblance de ce dernier. Généralement, 10 itérations suffisent pour obtenir la convergence des paramètres [42]. Ceci peut de plus être observé lors de la phase d'entraînement de notre système (fig. 33). En effet, la figure 33 nous montre que la forme générale du modèle du locuteur représenté ne change plus significativement à partir de 10 itérations.

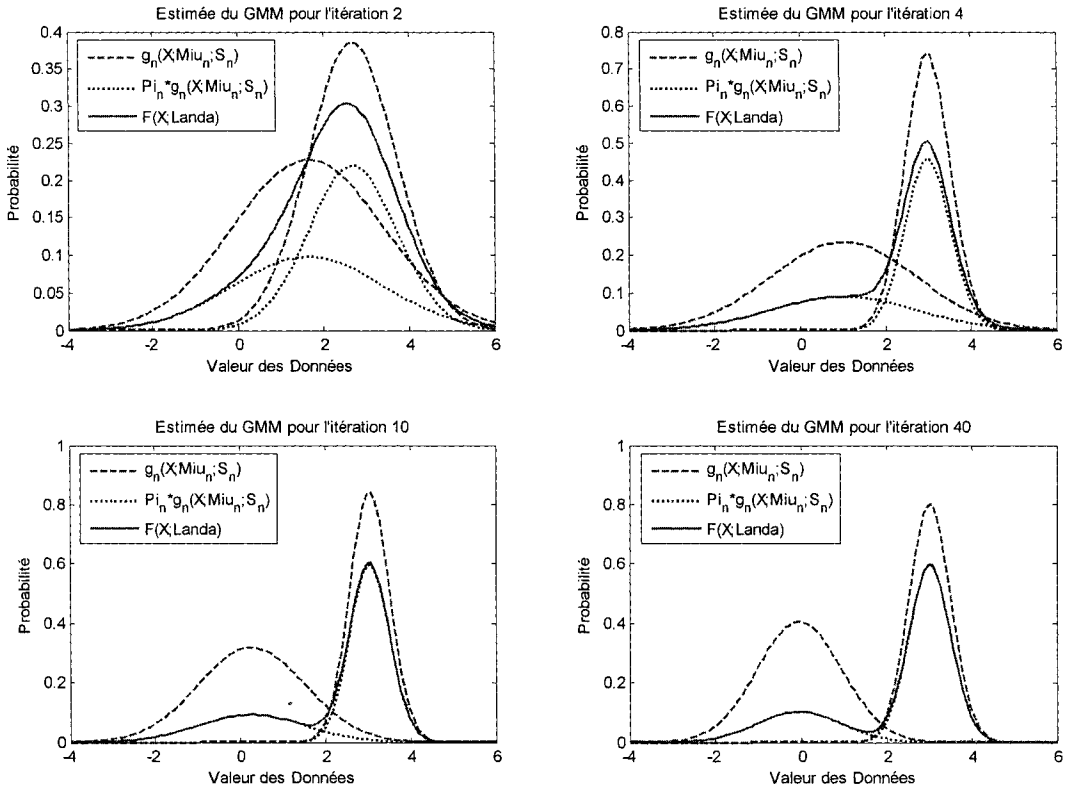


Figure 33 Estimation du GMM d'un locuteur pour 2, 4, 10, et 40 itérations

Si maintenant on s'intéresse à la valeur de la vraisemblance du GMM, on peut aboutir aux même conclusions en observant la figure 34. En effet, on remarque que la vraisemblance maximale est quasiment atteinte au bout d'une dizaine d'itérations de l'algorithme EM. Pour ces raisons on choisi de prendre une valeur légèrement supérieure, soit 20 itérations. Un plus grand nombre d'itérations ne ferait qu'augmenter le temps de calcul sans pour autant améliorer les performances.

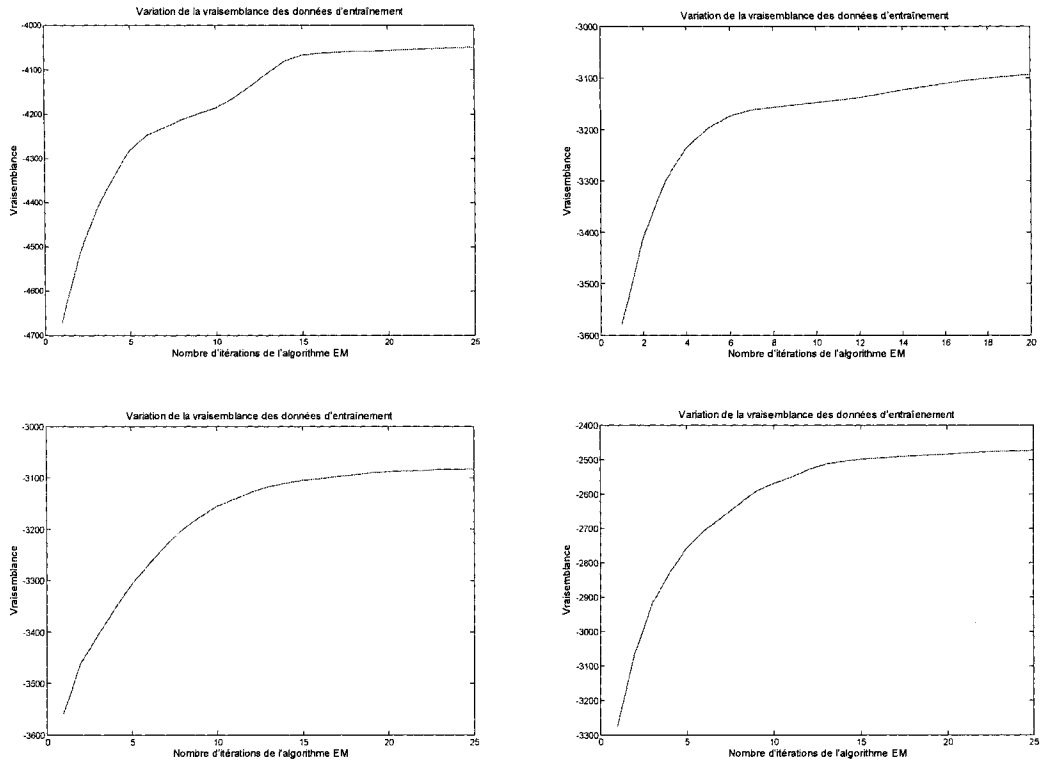


Figure 34 Variation de la vraisemblance des données pour la modélisation du locuteur 1, 2, 3, et 4

2.7 Expérimentation du système d'identification de référence

Les tests sont effectués sur des bases de données de 30 et de 50 locuteurs (possédant un SNR de 43 dB). Nous verrons dans le chapitre 4 comment se constituent ces bases de données lorsque nous aborderons notre contribution personnelle sur la reconnaissance du locuteur.

Pour l'instant, intéressons-nous uniquement aux résultats obtenus avec le système d'identification basé sur les techniques que nous venons de décrire au cours de chapitre. Voici son schéma descriptif (fig. 35):

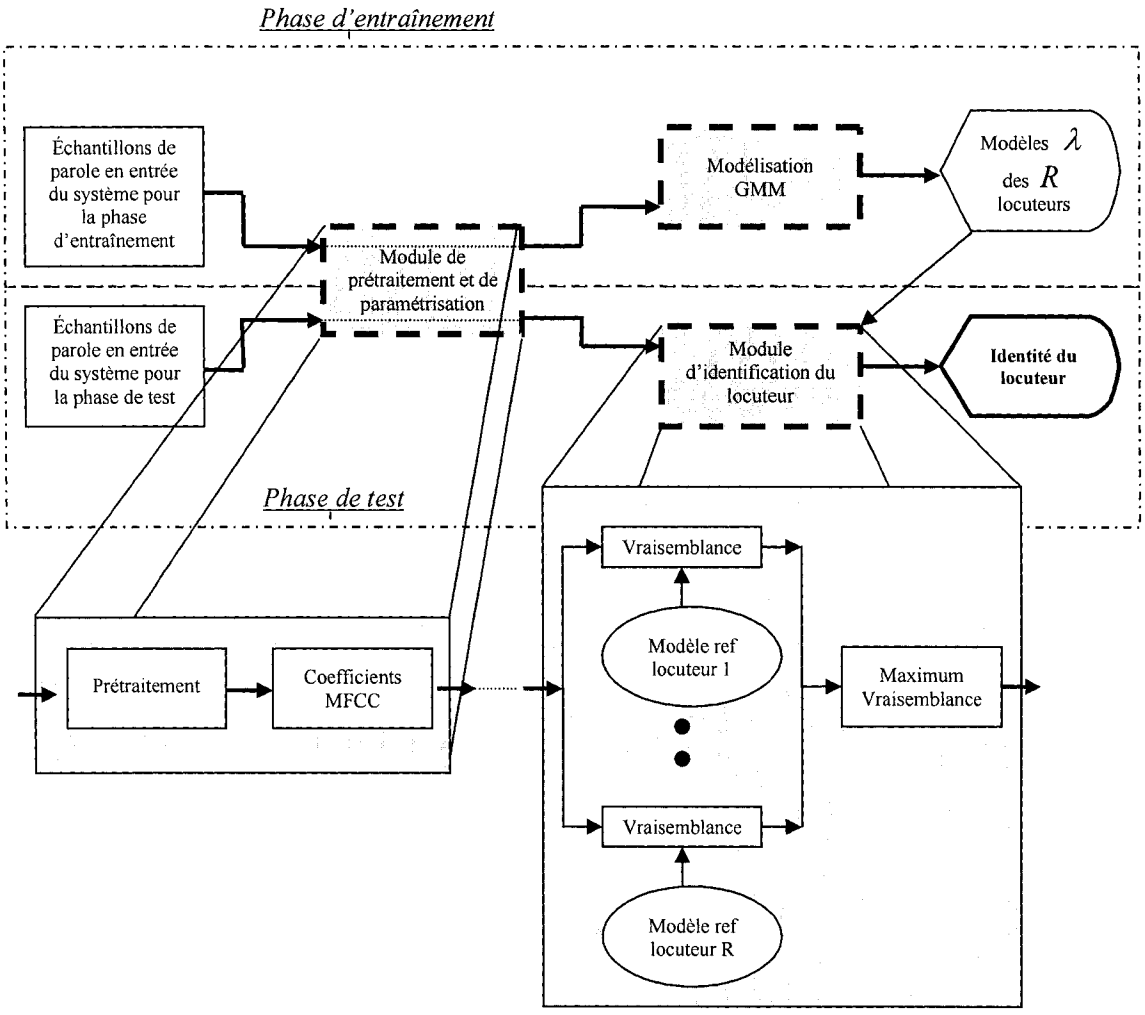


Figure 35 Schéma descriptif du système de reconnaissance de référence

Avant de vous présenter les résultats obtenus, voyons quels sont les différents paramètres expérimentaux utilisés.

2.7.1 Paramètres optimaux du système de référence

Pour trouver les paramètres optimaux, nous les avons fait varier un par un, puis par association, afin d'aboutir aux valeurs optimales indiquées ci-dessous.

2.7.1.1 Normalisation du signal de parole

Des tests préliminaires effectués tant sur le système de reconnaissance de référence comme sur les systèmes hybrides, ont démontré que le signal de parole doit être normalisé entre -1 et 1 avant tout traitement, que ça soit pour la phase d'entraînement comme pour la phase de test. Il existe une différence de 10 à 20 points sur les taux de reconnaissance selon que l'on normalise le signal d'origine ou pas. Par conséquent, le module de pré-traitement de tous nos systèmes, effectuera la normalisation des signaux avant toute autre opération.

2.7.1.2 Extraction des vecteurs de caractéristiques avec les MFCC

L'extraction des vecteurs de caractéristiques se fait à l'aide de la technique des MFCC. Nous avons choisi de prendre 13 coefficients incluant le coefficient d'énergie [47], [87].

Les différents paramètres choisis pour extraire les coefficients cepstraux sont les suivants [49], [87]:

Taille de la fenêtre :	32 ms (soit 256 échantillons avec un taux d'échantillonnage de 8 kHz).
Chevauchement :	12 ms (obtenu avec taux de trames = 1/0.020).

Nous avons également choisi de ne pas utiliser les coefficients delta cepstraux et delta-delta cepstraux puisque ces derniers sont efficaces lorsque les tests sont effectués en milieu bruité [68], et que nous effectuerons les notre en milieu quasi idéal (comme défini dans les objectifs initiaux du projet). Dans notre cas, ils risqueraient donc d'avoir des effets négatifs sur les performances du système.

2.7.1.3 Phase d'apprentissage

Chaque locuteur est modélisé par un GMM à covariance diagonale, constitué par 18 Gaussiennes et entraîné par 6 sessions d'entraînement contenant au total une moyenne de 15 secondes de parole. L'initialisation des GMM est effectuée à l'aide de l'algorithme des K-means, et les meilleurs résultats sont atteints avec la méthode d'initialisation des centres qui consiste à les prendre le plus près du vecteur moyenne.

2.7.1.4 Phase de test

Les GMM ont démontré (Reynolds et Rose, 1990; Reynolds, 1992) de très bonnes performances en matière d'identification en mode indépendant du texte à partir de courts segments de parole (pour la phase de test) obtenus en téléphonie haute qualité. Comme en réalité il n'y a pratiquement aucun contrôle sur la durée d'élocution d'une personne, nous avons concentré cette recherche sur des performances obtenues à partir de courts (<10 ms) segments de parole pour l'identification. La base de données YOHO s'est avérée tout à fait appropriée pour cette tâche.

Pour chaque locuteur, nous avons alors procédé à une seule session d'identification composée uniquement par trois combinaisons de phrases, soit approximativement une moyenne de 5 secondes de parole pour chaque individu.

2.7.2 Résultats expérimentaux trouvés pour le système de référence

Le tableau II montre que les performances augmentent avec le nombre de Gaussiennes de la modélisation. On observe que l'on atteint un maximum pour 18 Gaussiennes. Si on continue d'augmenter cette valeur, cela entraîne soit une diminution des taux de reconnaissance (cf. 20 Gaussiennes), soit de nombreuses erreurs de calculs (erreurs

signalées par Matlab) lors de la phase d'entraînement (nous avons distingué ces valeurs par une * dans les tableaux ci-dessous).

La modélisation GMM la plus appropriée compte tenu des performances et des études déjà effectuées [21], est celle qui se compose de 18 Gaussiennes. En fait, nos tests permettent juste de confirmer ce qui avait déjà été montré (fig. 31).

Tableau II

Taux de reconnaissance et IER du système de référence en fonction du nombre de Gaussiennes du modèle GMM pour une base de données de 30 locuteurs

	Taux de reconnaissance (en %)	**IER**
Nombre de Gaussiennes du GMM		
2	63.3	0.579
3	70	0.429
4	83.3	0.200
5	83.3	0.200
6	86.7	0.154
7	86.7	0.154
8	86.7	0.154
9	86.7	0.154
10	86.7	0.154
11	90	0.111
12	90	0.111
14	90	0.111
16	90	0.111
18	90	0.111
20	86.7	0.154
22*	90	0.111
24*	80	0.250
26*	90	0.111
28*	86.7	0.154
30*	76.7	0.304
32*	80	0.250

Le taux de reconnaissance est alors de 90 %, et le taux d'erreur (IER) est de 0.1111 pour une population de 30 locuteurs.

On observe les mêmes tendances pour une population de 50 locuteurs (tab. III). Pour le même nombre de Gaussiennes (18 Gaussiennes) que dans le cas de 30 locuteurs, nous trouvons les meilleures performances : on trouve également un taux de reconnaissance de 90 %, et un taux d'erreur correspondant de 0.1 pour une population de 50 locuteurs. Pour plus de Gaussiennes, on aura soit de moins bonnes performances, soit des erreurs de calculs (signalées par Matlab) lors de l'entraînement des données.

Tableau III

Taux de reconnaissance et IER du système de référence en fonction du nombre de Gaussiennes du modèle GMM pour une base de données de 50 locuteurs

	Taux de reconnaissance en (%)	**IER**
Nombre de Gaussiennes du GMM		
2	56	0.786
3	72	0.389
4	82	0.220
5	86	0.163
6	80	0.250
7	88	0.136
8	86	0.163
9	86	0.163
10	86	0.163
11	88	0.136
12	90	0.111
14	90	0.111
16	90	0.111
18	90	0.111
20*	86	0.163
22*	90	0.111
24*	84	0.190
26*	90	0.111
28*	82	0.220

2.8 Conclusions

Dans ce chapitre, nous avons vu comment concevoir un système d'identification fonctionnel, performant, et basé sur des techniques standard. Ses performances sont de 90 % pour des populations de 30 ou de 50 locuteurs. Nous allons maintenant essayer de le modifier et d'en améliorer les performances en utilisant la transformée en ondelette continue (TOC, ou en anglais CWT : Continuous Wavelet Transform). Avant de procéder à ces changements, nous présenterons dans le prochain chapitre les propriétés et les avantages de cette transformée.

CHAPITRE 3

LES ONDELETTES ET SES APPLICATIONS

3.1 Introduction

Les ondelettes sont des fonctions mathématiques qui permettent de décomposer un signal selon différentes échelles, avec des résolutions qui varient en fonction de l'échelle d'analyse. Les ondelettes ont de nombreux avantages par rapport aux méthodes traditionnelles de Fourier en ce qui concerne l'étude de certains signaux [70]. Nous le verrons en détail tout au long de ce chapitre. Les ondelettes ont été développées indépendamment dans les domaines des mathématiques, de la physique quantique, du génie électrique, ou encore de la géologie. Les échanges entre ces domaines très variés lors de la dernière décennie, ont permit d'aboutir à de nouvelles applications telles que la compression d'image, la vision par ordinateur, le radar, et même la prédiction des tremblements de terre.

Dans ce chapitre, nous allons tout d'abord donner un bref aperçu du problème qui sera étudié dans ce mémoire, afin d'expliciter en quoi les ondelettes nous seront utiles pour ce projet. Après avoir donné un court historique de ce nouvel outil mathématique, nous étudierons ses différentes caractéristiques. Nous présenterons enfin une étude détaillée de la transformée en ondelettes, et plus particulièrement de la transformée en ondelettes continue qui constituera l'outil de base de notre projet de recherche. Nous terminerons ce chapitre du rapport en passant en revue quelques unes des autres transformées qui sont souvent référées et utilisées dans la littérature.

3.1.1 Position du problème

Dans le domaine du traitement du signal, nous avons à notre disposition plusieurs types de transformées qui permettent d'extraire le contenu fréquentiel d'un signal. La transformée de Fourier (TF ou en anglais FT : Fourier Transform) analyse le signal dans son ensemble pour calculer son spectre de fréquences de ce signal.

$$F(\omega) = \int_{-\infty}^{\infty} f(t).e^{-j\omega t}.dt \tag{3.2}$$

Le gros désavantage de cette transformée est qu'elle possède uniquement une résolution fréquentielle. Elle ne permet donc pas de déterminer à quel moment les fréquences observées sont présentes dans le signal, et comment elles évoluent dans le temps. Par conséquent, cette méthode n'est valide que pour les signaux stationnaires.

Pour palier à ce problème, de nombreuses solutions ont été développées au cours des dernières décennies. Certaines d'entre elles se sont avérées plus ou moins efficaces pour représenter le signal dans le domaine temporel et fréquentiel simultanément (c'est-à-dire dans le plan temps-fréquence).

L'approche de base pour obtenir une représentation simultanée en temps et en fréquence consiste à découper le signal de données en plusieurs parties, puis à les analyser séparément. Maintenant reste à savoir effectuer ce découpage.

Supposons que l'on désire connaître exactement toutes les fréquences présentes dans le signal à un certain instant. On peut penser à extraire le segment utile à l'aide d'une impulsion de Dirac, et à le passer dans le domaine fréquentiel. En réfléchissant bien, on se rend compte qu'il y a quelque chose de faux dans cette procédure. Le problème ici est que l'extraction de la partie du signal qui nous intéresse, correspond en fait à un produit

entre le signal et la fenêtre choisie. Puisque le produit dans le domaine temporel est identique à une convolution dans le domaine fréquentiel, et puisque la TF d'une impulsion de Dirac contient toutes les fréquences possibles, les composantes fréquentielles seront dispersées sur tout l'axe des fréquences (on parle ici d'une transformée bidimensionnelle temps-fréquence et non pas d'une transformée unidimensionnelle). En fait, nous nous retrouvons dans la situation inverse de celle obtenue avec la TF standard, puisque nous avons désormais une bonne résolution temporelle mais au détriment de la résolution fréquentielle.

Le phénomène observé ci dessus est du au principe d'incertitude d'Heisenberg [71] qui dit, en matière de traitement du signal, qu'il est impossible de connaître exactement une fréquence et l'instant à laquelle elle apparaît dans un signal. Ce principe établi par Heisenberg démontre également l'importance du découpage du signal à analyser [72].

Pour palier à ce problème, une des solutions apportées fut la transformée de Fourier à fenêtre glissante (TFFG ou en anglais STFT : Short Term Fourier Transform) [100], qui peut être utilisée pour étudier les signaux non stationnaires ; elle extrait le contenu fréquentiel d'un signal à l'intérieur d'une fenêtre temporelle glissante d'une durée fixe.

$$F(\omega,b)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t).\Gamma(t-b).e^{-j\omega t}.dt \tag{3.2}$$

On considère une fenêtre $\Gamma(t)$ à support compact de largeur donnée. La TFFG consiste à multiplier le signal $f(t)$ par la fonction fenêtre $\Gamma(t)$, centrée en $t=0$, puis à calculer les coefficients de Fourier du produit $\Gamma(t)f(t)$. Ces coefficients donnent une indication sur le contenu fréquentiel du signal $f(t)$ au voisinage de $t=0$. Cette procédure est ensuite répétée avec une version translatée de la fonction fenêtre, c'est-à-dire que l'on remplace $\Gamma(t)$ par $\Gamma(t-b)$ où b est le paramètre de translation temporelle.

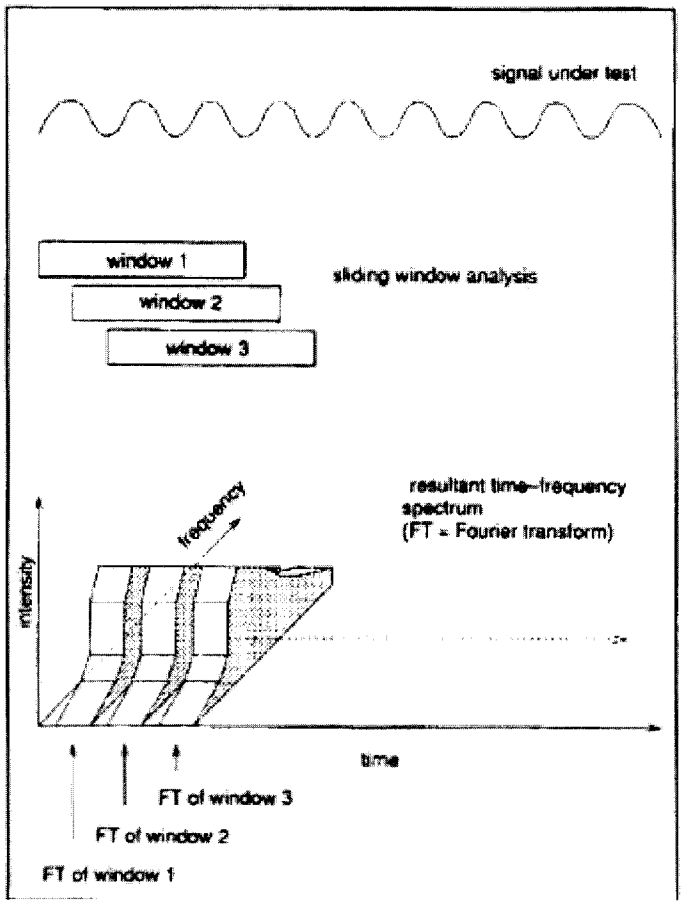


Figure 36 Analyse à fenêtre glissante dans la TFFG [71]

Nous verrons dans ce chapitre (équation 3.25), que selon le principe d'incertitude d'Heisenberg, la résolution temporelle est inversement proportionnelle à la résolution fréquentielle, ce qui entraîne que la résolution dans le plan temps-fréquence est fixe pour la TFFG. Seulement, le manque de souplesse de cette transformée fait que l'on ne peut agir sur la taille de la fenêtre, les hautes fréquences étant conséquemment mal analysées. Ceci a motivé l'introduction de la transformée en ondelettes (TO ou en anglais WT : Wavelet Transform).

L'utilisation de la TO pour l'étude de tels signaux a l'avantage d'extraire le contenu fréquentiel en utilisant une fenêtre temporelle de durée variable. La fenêtre est déplacée tout le long du signal, et, pour chaque position, on calcule le spectre de fréquence. Puis, on répète ce processus avec une fenêtre plus ou moins grande pour chaque nouveau cycle. Au final, on aboutira à une collection de représentations temps-fréquence[6] du même signal, mais toutes avec une résolution différente. Ceci correspond plus communément à une analyse multirésolution [74].

Au cœur même de cette théorie, on retrouve ainsi le concept de schéma itératif, c'est-à-dire la répétition sans fin d'une même opération à des échelles de plus en plus petites. Cela ressemble à un zoom grossissant sur un point de l'espace qui nous intéresse. La méthode des ondelettes utilise des fonctions qui se rapprochent de ce point à chaque étape pour en extraire l'information pertinente.

La TO est donc particulièrement adéquate pour l'étude de signaux non stationnaires comme la parole, qui possèdent des composantes basses fréquences de longue durée et des composantes hautes fréquences de courte durée. Cet outil s'avère donc approprié pour notre projet.

3.1.2 Introduction aux ondelettes

Les premiers travaux concernant l'analyse par ondelettes se situent autour du début des années 80, et ils ont été entrepris par Morlet, et Grossmann, [75]. En 1985, Stéphane Mallat donne un nouvel élan aux ondelettes à travers ses travaux en traitement numérique du signal [76]. En effet, ce dernier réussi à mettre en évidence des liens entre les filtres miroirs en quadrature (FMQ ou en anglais QMF : Quadrature Miror Filters),

---

[6] Nous rappelons que dans le cas des ondelettes, nous ne devrions pas employer le terme de représentation temps-fréquence, mais plutôt celui de échelle-fréquence, l'échelle étant en fait l'inverse de la fréquence; le terme de fréquence est strictement réservé à la TF.

les algorithmes pyramidaux, et les bases orthonormales d'ondelettes. Inspiré en partie par ces travaux, Y. Meyer créa les premières ondelettes à forme non triviale [77]. Contrairement à l'ondelette de Haar, les ondelettes de Meyer sont continues et intégrables. Cependant, il fallut attendre 1988 pour qu'un article d'Ingrid Daubechies [78], conclu notamment grâce aux travaux de Mallat [76], attire définitivement l'attention des ingénieurs sur les possibilités d'application de cette méthode.

Le but recherché à l'époque, était de donner une représentation des signaux permettant de faire apparaître simultanément des informations temporelles (localisation dans le temps, durée) et fréquentielles, facilitant par là l'identification des caractéristiques physiques du signal. Les ondelettes n'ont depuis lors cessé de se développer et de trouver de nouveaux champs d'application. C'est ainsi qu'est apparu un parallèle étonnant entre ces méthodes et des techniques développées à des fins totalement différentes en traitement d'images [108], mais aussi d'autres théories mathématiques poursuivant des objectifs sans aucun lien apparent (comme par exemple des problèmes d'analyse mathématique pure, ou d'autres liés au problème de la quantification de certains systèmes classiques, ou plus récemment des problèmes de statistiques) [79], [109].

De nos jours, les ondelettes sont de plus en plus utilisées dans les le domaine des nouvelles technologies. Que ce soit pour la compression d'images [76], pour le traitement du son et de l'image (téléphonie, télévision [80],...), le graphisme, la modélisation numérique ou pour la géologie, l'astronomie, le radar,... Enfin, presque partout. A titre d'exemple, la base de données d'empreintes digitales du FBI est compressée avec les ondelettes depuis le début des années 90 [81]; le format JPEG 2000 par exemple, fait également usage des ondelettes [82].

3.2 Étude de la méthode des ondelettes

L'analyse de Fourier est sans conteste l'un des outils les plus puissants mis à la disposition des mathématiciens et physiciens d'aujourd'hui et aussi l'un des plus utilisé. Néanmoins, bien que bâtie sur la base du concept physique de fréquence (spatiale ou temporelle), elle se révèle imparfaitement adaptée à la description de fonctions ou de signaux que l'on peut rencontrer couramment, tels que les sons musicaux ou vocaux par exemple.

La méthode des ondelettes permet de palier à ce problème. Dans cette partie, nous nous pencherons sur les fondements de cette nouvelle méthode.

3.2.1 Qu'appelle-t-on une ondelette ?

Nous avons tenu à vous présenter ci-dessous une synthèse de la théorie des ondelettes sans trop s'engouffrer dans les détails mathématiques. La bibliographie mentionne des articles variés de références couvrant d'avantage la théorie des ondelettes : [43], [84], [85], [74], [70], [86], [76], [88].

3.2.1.1 Présentation générale

Une ondelette est fondamentalement une petite onde (impulsion), qui commence à t=0 et s'annule à t=N. Son énergie est donc concentrée dans le temps, ce qui donne un outil d'analyse très puissant des phénomènes non stationnaires ou variant dans le temps. Nous rappelons que les transformées classiques comme la TF, utilisent des ondes infinies dans le temps telles que les sinusoïdes.

Les ondelettes $\psi_{a,b}(t)$ sont des fonctions linéairement indépendantes dans le domaine continu, et peuvent être utilisées pour reconstruire toutes les fonctions réelles $f(t)$.

$$f(t) = \sum_{a,b} p_{a,b} \psi_{a,b}(t) \tag{3.3}$$

La base formée par l'ensemble des ondelettes $\psi_{a,b}(t)$ a une particularité, celle d'être toutes construites à partir d'une unique ondelette appelée « ondelette mère » $\psi(t)$.

L'ondelette mère doit remplir deux conditions très importantes : elle doit être admissible et régulière.

3.2.1.2 Condition d'admissibilité

On a démontré [85] que les fonctions $\psi(t)$ de carré intégrable, c'est-à-dire à énergie finie, et qui satisfont à la condition d'admissibilité :

$$\int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega = K \leq \infty \qquad \text{K=constante} \tag{3.4}$$

(cette condition entraîne, en particulier, que l'ondelette est d'intégrale nulle)

peuvent être utilisées pour analyser un signal dans un premier temps, puis pour le reconstruire sans perte d'information par la suite. Nous rappelons que dans (3.4), $\Psi(\omega)$ désigne la transformée de Fourier de $\psi(t)$.

La condition d'admissibilité implique que la transformée de Fourier de $\psi(t)$ s'annule à la fréquence zéro, soit :

$$\left|\Psi(\omega)\right\|_{\omega=0} = 0 \tag{3.5}$$

Ceci signifie que l'ondelette mère doit avoir un spectre de type passe-bande. Ceci constitue une observation très importante et cette propriété sera utilisée plus loin lors de la mise au point d'un algorithme efficace pour la transformée en ondelettes.
Enfin, une valeur nulle à la fréquence zéro indique également que la valeur moyenne de l'ondelette dans le domaine temporal doit être nulle :

$$\psi(t) \text{ réelle, et } \int_{-\infty}^{\infty}\psi(t)dt = 0 \tag{3.6}$$

et par conséquent, la fonction $\psi(t)$ doit être oscillante. Ceci constitue une condition suffisante d'admissibilité, beaucoup plus simple à vérifier.

Il existe cependant d'autres conditions portant sur l'ondelette mère et qui on pour fonction de faire décroître rapidement les coefficients de la TO lorsqu'on diminue l'échelle d'analyse a .

Ces conditions sont les conditions de régularité, et elles impliquent que l'ondelette mère doit avoir simultanément une certaine fluidité et concentration dans le domaine temporel et fréquentiel.

3.2.1.3 Condition de régularité

La régularité d'une ondelette est un concept quelque peu complexe, et nous allons essayer d'éclaircir ce point en nous servant de la théorie des moments nuls.
Si l'on développe la TO en séries de Taylor pour t = 0 jusqu'à l'ordre n (prenons $b = 0$ afin de simplifier les calculs), nous obtenons [85]:

(3.7)

$$T_{W,f}(a,b=0)=\frac{1}{\sqrt{a}}\left[\sum_{p=0}^{n} f^{(p)}(0)\int_{-\infty}^{\infty}\frac{t^p}{p!}\psi_{a,b}(\frac{t}{s})dt+O(n+1)\right]$$

Ici $f^{(p)}$ correspond à la p iéme dérivée de la fonction f et O(n+1) correspond au reste de l'expansion. A présent, si l'on défini les moments de l'ondelette par M_p,

$$M_p=\int t^p\psi(t)dt \tag{3.8}$$

nous pouvons alors écrire (3.5) de la façon suivante :

$$T_{W,f}(a,b=0)=\frac{1}{\sqrt{a}}\left[\begin{array}{r} f(0)M_0a+\frac{f^{(1)}(0)}{1!}M_1a^2+\frac{f^{(2)}(0)}{2!}M_2a^3+... \\ ...+\frac{f^{(n)}(0)}{n!}M_na^{n+1}+O(a^{n+2}) \end{array}\right] \tag{3.9}$$

À partir de la condition d'admissibilité, nous avons déjà le moment d'ordre 0 $M_0=0$ de telle sorte que le premier terme de l'égalité de droite dans (3.9) est nul. Si maintenant nous parvenons à rendre nuls tous les autres moments jusqu'à M_n, nous obtiendrons par la suite des coefficients de la transformée en ondelettes $T_{W,f}(a,b)$ qui vont décroître aussi rapidement que a^{n+2} pour un signal continu f(t). Dans la littérature, ceci est appelé communément moments nuls [89] ou ordre d'approximation. Si une ondelette a N moments nuls, la TO qui lui est associée aura aussi un ordre d'approximation égal à N. Bien sûr, les moments ne doivent pas être strictement égaux à zéro, une très petite valeur reste acceptable. En fait, des recherches expérimentales on établi que le nombre de moments nuls requis dépends essentiellement du signal étudié [90].

Très grossièrement, la régularité d'une fonction traduit la rapidité avec laquelle elle varie. Plus une fonction est régulière, plus ses coefficients d'ondelettes décroissent vite quand l'échelle décroît, et vice versa.

Pour résumer, la condition d'admissibilité nous donne l'onde, la régularité et les moments nuls nous donnent la vitesse de décroissance ou la pente, et quand ces deux conditions sont réunies, elles nous permettent d'obtenir une ondelette. Pour plus d'explications sur la régularité d'une fonction, on peut consulter [74] et [70].

3.2.1.4 Compression et dilatation d'une ondelette

Le paramètre d'échelle a dans l'analyse par ondelettes, possède la même fonction que celui utilisé en cartographie: les grandes échelles correspondent à une vue globale du signal non détaillée, et les petites échelles correspondent à une vue détaillée. De même, en terme de fréquence, les basses fréquences (hautes échelles) correspondent à une information globale du signal (qui souvent sont présentes sur tout le signal), tandis que les hautes fréquences (petites échelles) correspondent à une information détaillée d'une caractéristique cachée dans le signal (qui souvent ne dure qu'un court instant).
L'opération mathématique d'échelonnage, permet de dilater ou de compresser un signal. Les grandes échelles correspondent aux signaux dilatés et les petites échelles correspondent aux signaux compressés.

Une ondelette typique est compressée a fois et translatée b fois et elle a pour expression :

$$\psi_{a,b}(t) = \psi(2^a t - b) \tag{3.10}$$

Ainsi, on voit bien que l'ondelette $\psi(2^a t - b)$ est obtenue à partir de $\psi(t)$ (l'ondelette mère) par dilatation de 2^a (en fréquence) et par une translation de b (dans le temps).

Nous voyons dans ce qui suit comment on peut générer les ondelettes à partir de cette base.

3.2.2 Exemples classiques d'ondelettes continues 1D

Voici quelques exemples d'ondelettes 1D obtenues avec le logiciel Matlab:

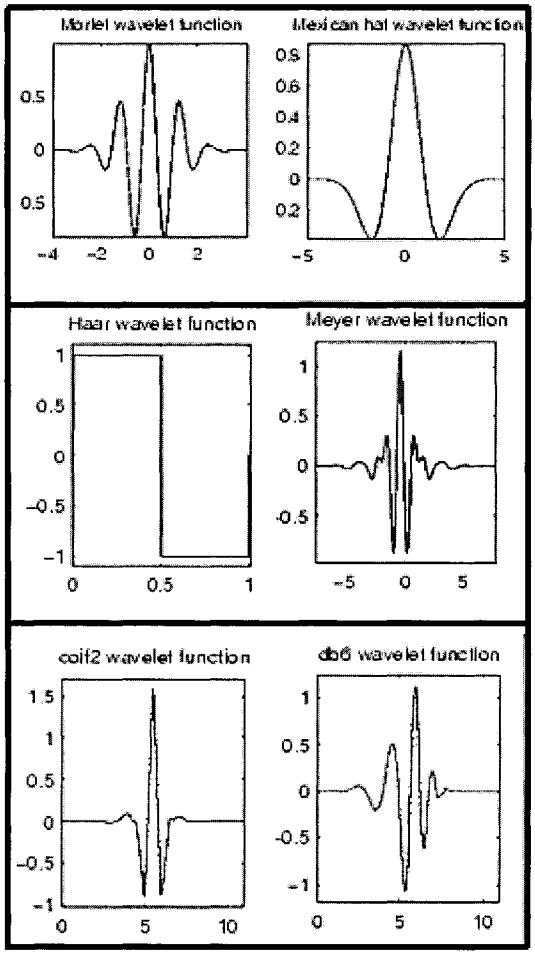


Figure 37 Représentation d'ondelettes continues usuelles en 1D [Matlab]

La figure 37 montre la forme de différentes ondelettes, ainsi que leurs caractéristiques propres. Par exemple, nous pouvons observer que les ondelettes de Morlet et du chapeau mexicain sont symétriques. L'ondelette coif2 (Coifman-2) laisse apparaître des contours anguleux et rigides, tandis que les ondelettes db6 (Daubechies-6) et sym6 possèdent des contours arrondis. L'ondelette de Haar quant à elle, est la seule fonction non continue avec 3 points de discontinuité, à savoir (0, 0.5, 1).

Nous observons également sur ces figures, que toutes les fonctions convergent rapidement vers zéro.

3.2.2.1 Présentation de l'ondelette de Morlet

Certaines applications ont de meilleurs résultats avec une ondelette mère qu'avec une autre. Par exemple pour la détection de discontinuités, l'ondelette appelée « chapeau mexicain » donne les meilleurs résultats. L'ondelette de Morlet est aussi bien utilisée pour l'analyse de la parole que pour l'analyse de signaux biomédicaux [36], [73]. D'après ces articles de B.Corona, l'ondelette de Morlet permet d'obtenir une bonne localisation temps-fréquence ; elle est donc la plus appropriée pour faire ressortir les points caractéristiques contenus dans le signal de parole. Par conséquent, nous avons choisi cette ondelette pour nos recherches sur la reconnaissance du locuteur.

Voyons dans ce qui suit comment se caractérise une telle ondelette.

L'ondelette de Morlet peut être considérée comme le précurseur de toutes ses congénères. En effet, bien que la première ondelette mise en œuvre fût l'ondelette de Haar, il n'en reste pas moins que ce fut grâce aux travaux de Morlet que les fondements mathématiques de la théorie des ondelettes permirent une meilleure formulation des méthodes d'analyse temps-fréquence ou temps-échelle.

L'ondelette de Morlet n'est en fait qu'une gaussienne modulée :

$$\psi(t) = e^{j\omega_0 t} e^{\frac{-t^2}{2\sigma_0^2}} \tag{3.11}$$

Sa transformée de Fourier est :

$$\Psi(\omega) = e^{\frac{[(\omega-\omega_0)\sigma_0]^2}{2}} \tag{3.12}$$

Le terme additionnel ξ permet de corriger la fonction $\psi(t)$ afin que celle-ci soit toujours admissible. Ce terme peut être négligeable si on choisit $\omega_0 > 5{,}5$.

La TO de Morlet est complexe, il est donc possible d'interpréter séparément sa phase et son module. Dans notre cas, nous nous intéresserons uniquement à sa partie réelle.

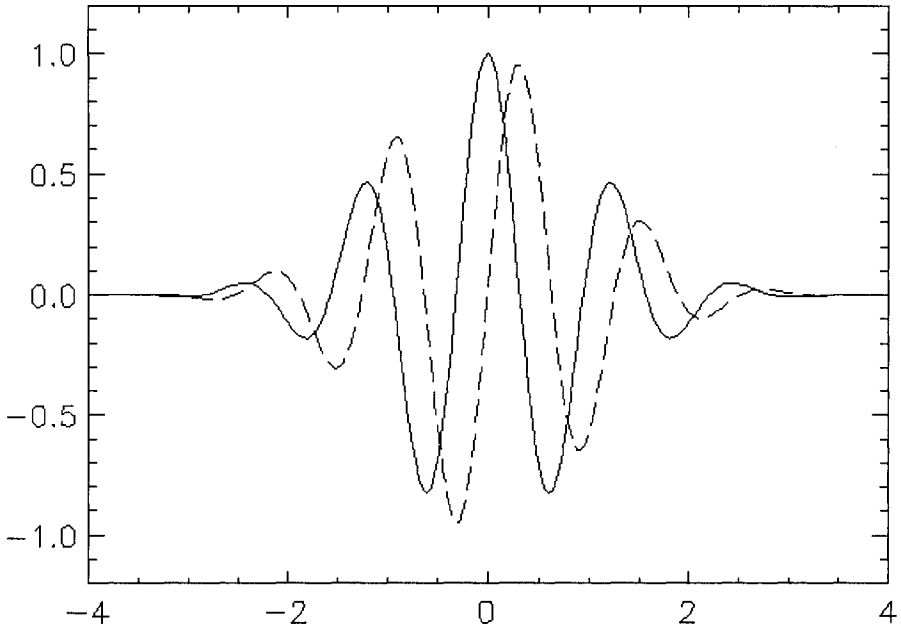


Figure 38 Représentation de la partie réelle (tracé continu) et imaginaire (tracé en pointillés) de l'ondelette de Morlet

Analytiquement, on doit soustraire une fonction non oscillante afin de satisfaire la condition d'admissibilité de l'ondelette. Pour une fonction sinus de fréquence unitaire dans une enveloppe de largeur $2\sigma_0 = \frac{z_0}{\pi}$, nous avons : (en remplaçant dans (3.11))

$$\psi(t, z_0) = [\cos(2\pi t) + j\sin(2\pi t)].e^{(\frac{-2t^2\pi^2}{z_0^2})} - e^{(-\frac{z_0^2}{2} - \frac{-2t^2\pi^2}{z_0^2})} \tag{3.13}$$

Le choix de z_0 reflète un compromis entre localisation temporelle et localisation fréquentielle: la valeur $z_0 = 5$ est recommandée en pratique, mais elle peut très bien être modifiée.

Clairement, la TO aura une partie réelle et une partie imaginaire. Tout comme dans le cas de la TF, les deux parties devront être connues afin de calculer la transformée inverse et reconstruire le signal d'origine.

3.3 Étude de la transformée en ondelettes continue

La vaste majorité des méthodes de traitement de la parole basées sur la TO rapportées dans la littérature fait usage de bancs de filtres discrets, d'arbres dyadiques et de paquets d'ondelettes. En revanche, il existe à ce jour très peu de textes de référence sur les décompositions continues en ondelettes pour le traitement de la parole. Le but ici sera de faire le point sur la transformée en ondelettes continue (TOC ou en anglais CWT : Continuous Wavelet Transform) afin de pouvoir l'appliquer par la suite dans notre projet.

3.3.1 La transformée en ondelettes continue

La transformée en ondelettes continue ou TOC, est en théorie infiniment redondante. Elle se caractérise par un paramètre de dilatation qui rétrécit ou élargit l'ondelette analysante.

Une famille d'ondelettes définies sur R $\{\psi_{a,b}(t)\}$ associée à un paramètre d'échelle a et à un paramètre de translation b peut être exprimée par :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \tag{3.14}$$

La transformée en ondelettes continue $TOC_{\psi,f}(a,b)$ d'une fonction $f(t)$ est définie de la façon suivante :

$$TOC_{\psi,f}(a,b) = \int_{-\infty}^{\infty} \psi_{a,b}(t) f(t) dt = \langle \psi_{a,b}(t), f(t) \rangle \tag{3.15}$$

La TOC d'une fonction $f(t)$ peut aussi être introduite comme suit :

$$TOC_{\psi,f}(a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} \psi(\frac{t-b}{a}) f(t) dt \tag{3.16}$$

3.3.2 Transformée de Fourier d'une ondelette ; analyse temps-fréquence

Soit $\Psi(\omega)$, la TF de $\psi(t)$, alors pour $a \geq 0$ la TF de $\frac{1}{\sqrt{a}}\psi(\frac{t}{a})$ sera :

$$\sqrt{a}\Psi(a\omega) \tag{3.17}$$

La TF de l'ondelette $\psi_{a,b}(t)$ est donc donnée par :

$$\Psi_{a,b}(\omega) = \sqrt{a}\Psi(a\omega)e^{-j\omega} \tag{3.18}$$

Par conséquent la fonction en ondelettes $\psi_{ab}(t)$ se contracte en temps et se dilate en fréquence pour une grande valeur de a, ou alors, elle se dilate en temps et se contracte en fréquence pour une petite valeur de a. La transformée en ondelettes permet donc une résolution temps-fréquence flexible.

3.3.3 Avantages de la TOC

L'opérateur d'échelle a agit directement sur les résolutions temporelle Δt et fréquentielle $\Delta\omega$ de l'ondelette ψ [93]. En effet on démontre facilement en agissant sur a, que ces deux notions sont antagonistes. Ceci s'explique du fait que les pavés élémentaires de la représentation temps-fréquence de la TO s'élargissent et se contractent selon les variations de l'opérateur d'échelle a (fig. 39). L'inconvénient majeur de la TFFG, est qu'elle ne peut donner de résolutions temporelles et fréquentielles variables. Par conséquent les pavés élémentaires de la représentation temps-fréquence de la TFFG sont identiques (fig. 38) [37].

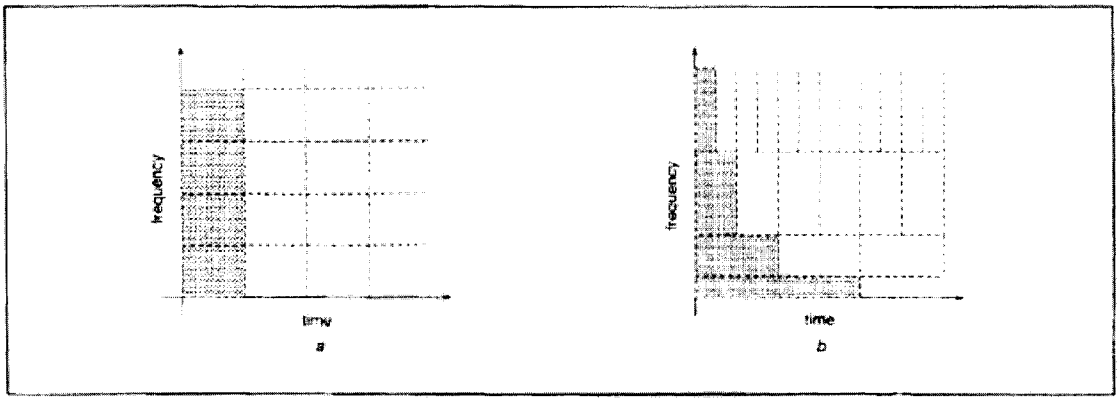


Figure 39 Plan temps-fréquence (a) TFFG (b) TO [33]

Le principe d'incertitude d'Heisenberg met en évidence que la localisation temps-fréquence idéale ($\Delta t = 0, \Delta\omega = 0$) est impossible et que le produit $\Delta t \Delta\omega$ est borné inférieurement par [71]:

$$\Delta t \Delta\omega \geq \frac{1}{4\pi} \tag{3.19}$$

Ceci explique que si nous augmentons la résolution temporelle (c'est-à-dire que si nous diminuons Δt), alors nous nous retrouvons à diminuer la résolution fréquentielle (c'est-à-dire à augmenter $\Delta\omega$) et vice-versa.

3.3.4 La transformée en ondelettes continue inverse

Soit la résolution de l'identité pour les ondelettes appliquées aux signaux $f(t)$ et $g(t)$ définie de la façon suivante [94] :

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} TOC_{\psi f}(a,b)TOC_{\psi g}(a,b)\frac{dadb}{a^2} = \left(\int_{0}^{\infty}\frac{|\Psi(\omega)|^2}{\omega}d\omega\right)\langle f(t), g(t)\rangle \tag{3.20}$$

La transformée en ondelettes est inversible si et seulement si la résolution de l'identité est vérifiée. On peut alors reconstruire $f(t)$ à partir de sa transformée en ondelettes de la façon suivante :

$$f(t)=\left(\int_0^{\infty}\frac{|\Psi(\omega)|^2}{\omega}d\omega\right)^{-1}\int_{-\infty}^{\infty}\int_0^{\infty}TOC_{\psi f}(a,b)\psi_{ab}(t)\frac{dadb}{a^2} \tag{3.21}$$

On remplace g par f dans l'équation (3.20) et on obtient

$$\int_{-\infty}^{\infty}|f(t)|^2\,dt=\left(\int_0^{\infty}\frac{|\Psi(\omega)|^2}{\omega}d\omega\right)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|TOC_{\psi f}(a,b)|^2\frac{dadb}{a^2} \tag{3.22}$$

Ceci permet d'établir une relation entre l'énergie d'une fonction f(t) et l'énergie de sa transformée en ondelettes.

Soient a_1 et a_2 deux échelles différentes, nous pouvons établir la relation suivante entre l'énergie de la fonction d'ondelette à l'échelle a_1 et celle à l'échelle a_2 [55] :

$$\int_{-\infty}^{\infty}\frac{1}{|a_1|}|\psi_{ab}(t)|^2\,dt=\int_{-\infty}^{\infty}\frac{1}{|a_2|}|\psi_{a_1b}(t)|^2\,dt \tag{3.23}$$

3.3.5 Qu'y a-t-il de continu dans la TOC ?

N'importe quel processus de traitement de signal appliqué à des données tirées de l'environnement qui nous entoure, et que l'on traite sur un ordinateur, doit être effectué sur un signal discrétisé. Ceci nous porte donc à nous interroger sur ce qu'on appelle

alors « continu » dans notre TOC. Pourquoi parle-t-on de transformée continue si le signal analysé est en réalité discret?

En fait, ce qui différencie la TOC de la transformée en ondelettes discrète (TOD ou en anglais DWT : Discrete Wavelet Transform) (que l'on traitera dans la suite de cette partie), c'est le jeu d'échelles et de positions temporelles à laquelle ces deux transformées opèrent. À la différence de la TOD, la TOC peut opérer à n'importe quelle échelle, à partir de celle du signal d'origine jusqu'à celle que l'on voudra en fonction de nos besoins (pour une analyse très détaillée). La TOC est aussi « continue » en terme de décalage temporel : pendant le calcul des coefficients d'ondelette, l'ondette analysante est décalée doucement tout au long de l'axe du temps du signal analysé.

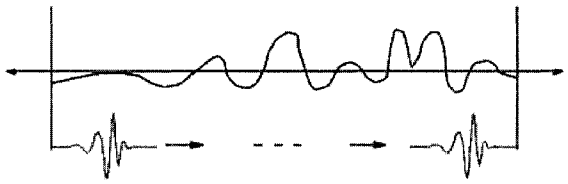


Figure 40 Décalage de l'ondelette analysante [Matlab]

3.4 Passage en revue des autres transformées en ondelettes

Calculer les coefficients d'ondelettes pour toutes les échelles représente un travail considérable, et cela génère un nombre important de valeurs. D'autant plus qu'une grande partie de ces valeurs sont redondantes et par conséquent inutiles. Il serait donc intéressant d'un point de vue de la vitesse de traitement comme de la taille de stockage, de réduire ces valeurs obtenues, en ne sélectionnant que les échelles et les positions dont nous avons besoins. C'est l'avantage des autres TO; elles permettent plus ou moins d'obtenir de bons résultats selon l'application donnée [71], [74], [91], [93], [95]. Nous allons ici voir un aperçu des plus pertinentes en traitement de la parole.

Au total, il existe 4 différents types de TO et comme il n'existe toujours pas d'appellations standard pour ces dernières, leur terminologie dans la littérature peut souvent porter à confusion. Nous les avons récapitulées ci dessous, et nommées de la façon qui nous a semblé la plus appropriée.

Pour un signal d'entrée continu, les paramètres de temps et d'échelle peuvent être continus, et on parle de transformée en ondelettes continue (TOC ou CWT). C'est ce dont nous avons parlé jusqu'à présent. Nous avons vu que cette transformée était très assimilable à la TF.

3.4.1 Décomposition discrète en série d'ondelettes

Les paramètres de temps et d'échelle peuvent aussi être discrets [78], [97], [98], [94], et on parle alors de décomposition discrète en série d'ondelettes (DDSO ou en anglais DPWT : Discrete parameter Wavelet Transform).

La formule de la DDSO est la suivante [78] :

$$DDSO_{\psi,f}(m,n) = a_0^{-\frac{m}{2}} \int_{-\infty}^{\infty} \psi_{m,n}(a_0^{-m}t - nb_0) f(t) dt \tag{3.24}$$

où les paramètres a et b sont discrétisés de la façon suivante :

$$a = a_0^m \text{ et } b = nb_0a_0^m \tag{3.25}$$

et où a_0 et b_0 sont des intervalles d'échantillonnage et m et n des entiers.

La fonction $f(t)$ et l'ondelette $\psi(a_0^{-m}t)$ sont encore continues ici.

Cette décomposition est très semblable à celle des séries de Fourier ou seule la fréquence est un paramètre discret. Pour de meilleures performances de calcul, on choisi habituellement $a_0 = 2$ et $b_0 = 1$, ce qui résulte en une dilatation binaire de 2^{-m} et en une translation dyadique de $2^m n$ [100].

La TO peut également être définie pour les signaux à temps discret [78], [101], [88], et on parle ici de transformée en ondelettes à temps discret (TOTD) et de transformée en ondelettes discrète (TOD).

3.4.2 Transformée en ondelettes à temps discret

La transformée en ondelettes à temps discret (TOTD ou en anglais DTWT : Discrete Time Wavelet Transform) qui a pour formule [78]:

$$TOTD_{\psi,f}(m,n) = a_0^{-\frac{m}{2}} \sum_k \psi_{m,n}(a_0^{-m}k - nb_0) f(k) \tag{3.26}$$

n'est en fait qu'une discrétisation de la DDSO (3.24), avec $t = kT$ et l'intervalle d'échantillonnage $T = 1$. Cette transformée est assimilable à la décomposition en séries discrètes de Fourier. On remarque que pour $a_0 = 2$, il y a seulement une sortie tous les 2^m échantillons lorsque $2^{-m}k$ est un entier (on dit alors que les paramètres de temps et d'échelle sont dyadiques). Dans cette transformée, on parle d'ondelette discrète.

Dans notre projet de mémoire, nous avons opté pour les ondelettes continues car elles sont plus robustes au bruit. Pour un développement plus long sur les ondelettes discrètes, reportez-vous aux ouvrages de Kumar et Foufoula Georgiou [103] pour une rapide revue, ou, pour des développements plus spécifiques, reportez-vous aux articles de Mallat [76], [97], Daubechies [70], [78], Meyer[77], et Chui [100].

3.4.3 Transformée en ondelettes discrète

La transformée en ondelettes discrète (TOD ou DWT) qui se positionne dans un contexte totalement discret [97], se défini de la manière suivante :

$$TOD_{\psi,f}(m,n) = 2^{-\frac{m}{2}} \sum_{k} \psi_{m,n}(2^{-m}k - n) f(k) \tag{3.27}$$

Ici, l'ondelette discrète $\psi(k)$ peut être, mais pas nécessairement, une version échantillonnée de sa version continue (ceci signifie qu'il est possible que $\psi(k)$ puisse ne pas avoir d'équivalent continu dans le temps). Lorsque $\psi(k)$ est une discrétisation de $\psi(t)$, la TOD est identique à la TOTD, avec la $\psi(t)$ de (3.24). Dans ce cas, il existe un parallèle entre la TOD et la transformée de Fourier discrète.

Pour l'implémentation de cette transformée, Stéphane Mallat proposa dans les années 80 un algorithme extrêmement rapide de décomposition et de reconstruction en utilisant une méthode classique de traitement du signal : l'implémentation d'un banc de filtres à deux canaux et à reconstruction parfaite [97], [76]. Cette technique permet d'obtenir la transformée en ondelettes rapide (TOR ou en anglais FWT : Fast Wavelet Transform).

3.5 Conclusion

Les ondelettes sont un outil très puissant en traitement du signal et la diversité de leurs applications rend cette technique d'autant plus intéressante.

Ce chapitre nous a permis de comprendre dans un premier temps comment se définissait une ondelette et quelles étaient leurs différentes propriétés ; dans un second temps, nous avons étudié en détail une des applications des ondelettes, à savoir la transformée en ondelettes continue (TOC), qui sera par la suite notre outil de base concernant nos recherches en reconnaissance du locuteur; enfin, nous avons survolé quelques techniques

additionnelles de traitement du signal telle que la transformée en ondelettes discrète (TOD), et ce, dans le but de pouvoir comparer leurs performances (complexité d'implémentation vs taux de reconnaissance) avec celles de la TOC. La revue de la littérature nous a permis de faire le point sur le fonctionnement de ces nouvelles techniques, et de distinguer les avantages de la TO par rapport à d'autres transformées, notamment sa grande flexibilité pour la représentation temps-fréquence. La TO est particulièrement appropriée pour les signaux qui possèdent des composantes hautes fréquences de courte durée ou des composantes basses fréquences de longue durée. Ces propriétés tout à fait intéressantes pour des signaux non stationnaires, nous ont incités à choisir la TOC pour tenter d'améliorer les performances en matière de traitement de la parole.

CHAPITRE 4

MODIFICATION DU SYSTÈME DE RECONNAISSANCE DE RÉFÉRENCE À L'AIDE DE LA TOC : UTILISATION D'UNE GRILLE DE SÉLECTION DES COEFFICIENTS DE LA TOC

4.1 Introduction

Les systèmes de reconnaissance hybrides que nous allons vous présenter ici, restent très similaires au système de référence (sans TOC) décrit à la fin du second chapitre. Ce sont des systèmes d'identification basés sur la modélisation par mélange de Gaussiennes où les vecteurs paramétriques sont extraits à partir de la transformée en ondelettes continue du signal de parole. Ces systèmes seront évalués sur une base de données publique que nous avons choisie, à savoir YOHO (Higgins et al., 1991; Campbell, 1992). Les échantillons de parole qui y sont stockés ont été enregistrés dans un milieu non bruité à travers un canal de transmission de qualité audio comparable à celle du téléphone (SNR de 43 dB).

4.2 Présentation des bases de données utilisées dans la phase de tests

Un facteur qui représente une difficulté majeure dans la tache d'identification, est la taille de la population de locuteurs. En effet, plus le nombre de sujets étudiés augmente, et plus il y a de risques d'erreur lors de la phase d'identification du locuteur.

La similitude entre locuteurs dans une population représente également un facteur de difficulté. Cette constatation est triviale, étant donné qu'un ensemble de locuteurs avec des caractéristiques vocales très différentes (par exemple, une population constituée à 50% par des femmes) produit généralement un meilleur taux de reconnaissance qu'un ensemble de locuteurs plus homogène (par exemple, 100% d'hommes) [52]. C'est

pourquoi, nous avons procédé à des expérimentations sur nos systèmes en faisant varier la taille de la population essentiellement constituée d'hommes.

4.2.1 Base de données YOHO

Toute base de données possède différentes caractéristiques, soit dans le domaine fonctionnel (par exemple, la dépendance du texte, le nombre d'individus, ...), soit dans le domaine qualitatif (par exemple, la parole bruitée ou pas).

Notre choix s'est tourné vers la base YOHO car celle-ci est utilisée pour vérifier les performances des systèmes en mode dépendant ou indépendant du texte dans un environnement non bruité (SNR de 43 dB) [48], [65]. Pour un même locuteur, elle offre différentes allocutions formées par la prononciation consécutive et aléatoire de divers nombres entiers.

La base de données YOHO est un regroupement de données de 138 locuteurs, 108 hommes et 30 femmes, enregistrées sur une période de 3 mois dans un environnement de travail réel. Cette base possède déjà un scénario de reconnaissance (phase d'entraînement / phase d'identification) bien défini [48]. En effet, chaque locuteur possède quatre séances d'apprentissage où il est amené à prononcer une série de phrases de 24 combinaisons de trois nombres entiers (par exemple, « 12 – 34 – 41 »). Il existe 10 séances de vérification par locuteur constituées de 4 phrases. Dans le tableau IV, nous avons décrit les principales caractéristiques des bases de données souvent employées en reconnaissance vocale :

Tableau IV

Caractéristiques des principales bases de données utilisées en traitement de la parole [48] (PSTN = Public Switched Telephone Network)

Base de données	Population	Phrases par locuteur	Canal	Environnement acoustique	Combiné
TIMIT	630	10 phrases lues	Propre	Son en cabine	Microphone à bande large
NTIMIT	630	10 phrases lues	PSTN local et longue distance	Son en cabine	Bouton en carbone fixe
Switchboard	500	1 – 25 conversations	PSTN longue distance	Maison et bureau	variable
YOHO	138	4 phrases d'entraînement 10 phrases de test à combinaison fermée	Propre	Bureau	Microphone téléphonique haute qualité

Les échantillons de parole contenus dans cette base de données sont obtenus à partir de nombres allant de 21 à 97, prononcés dans un environnement de bureau, et enregistrés sous un taux d'échantillonnage de 8kHz avec une bande passante légèrement supérieure à celle utilisée en téléphonie, soit 3.8kHz [48]. Il n'y a aucune dégradation dans la transmission. Pour plus de renseignements concernant le conditionnement des signaux, leur acquisition et leur sauvegarde, veuillez consulter l'annexe 4. La base YOHO a pour avantage d'offrir des données de qualité audio identiques à celles obtenues en téléphonie, et ce, sans bruit additionnel; elle permet de démontrer les performances des systèmes indépendants du texte, mais également de ceux dépendants du texte en opérant de simples modifications dans la sélection des données d'entraînement et d'identification. Ainsi, nous pouvons vérifier les performances de notre système pour différents modes si nous le désirons.

4.2.2 Bases de données dérivées de YOHO

Pour les expériences que nous avons réalisées, nous ne nous sommes pas servis de la base YOHO dans son intégralité. En réalité, nous voulions tester les performances de nos systèmes sur de plus petites populations afin de ne pas perdre de longues semaines de calculs inutilement si jamais les résultats obtenus sur peu de locuteurs s'avéraient déjà décevants. C'est pourquoi nous avons crée plusieurs bases à partir de YOHO, avec un nombre de locuteurs variant de 10 à 50.

Les sous bases de données de YOHO que nous avons considéré pour mener à bien la phase expérimentale sont des regroupements de locuteurs mixtes choisis et répartis comme indiqué dans l'annexe 5. En voici leur composition (tab. V):

Tableau V

Composition des bases de données dérivées de la base YOHO

Base de donnée composée d'une population de	**Répartition Hommes / Femmes**
10 locuteurs	9 Hommes et 1 Femme
20 locuteurs	19 Hommes et 1 Femme
30 locuteurs	28 Hommes et 2 Femmes
50 locuteurs	43 Hommes et 7 Femmes

4.3 Modification du système de reconnaissance de référence avec la TOC

Dans le premier chapitre, nous avons vu que l'évolution temporelle d'un signal ne fournit pas directement des traits acoustiques du signal de parole d'un individu. Pour cela, nous devons passer dans le domaine fréquentiel pour faire apparaître ces caractéristiques. C'est en partie ce que réalise la FFT dans le processus d'extraction des coefficients MFCC (voir fig. 20). Plus la variabilité inter-locuteurs exprimée dans le signal de parole de base sera naturellement importante, et plus on pourra extraire de coefficients cepstraux pertinents pour la caractérisation du locuteur, et meilleure sera la modélisation de chaque individu.

Comme nous venons de le montrer dans le chapitre précédent, la TOC permet d'obtenir une représentation simultanée en temps et en échelle du signal de parole. Le résultat de cette transformée est une représentation du même signal à différentes échelles d'analyse (donc à différentes fréquences) et à tout instant.

L'hypothèse que nous avons posée est celle selon laquelle les différentes échelles de la TOC peuvent servir à mettre en évidence la variabilité inter-locuteurs d'une population, ceci en faisant ressortir d'autres caractéristiques qui apparaissent à certaines échelles et qui ne s'expriment pas de la même manière dans le signal de parole d'origine.

En appliquant la TOC sur le signal de parole, il pourrait être possible de dégager des informations pouvant être utiles à la discrimination du locuteur.

Le module d'extraction des paramètres va donc fonctionner comme s'il opérait sur un signal de parole normal. La seule différence repose sur le fait que ce nouveau signal spécifique au locuteur est obtenu à partir de la TOC du signal d'origine.

On pourrait se demander pourquoi on ne se sert pas directement des coefficients de la TOC pour former des vecteurs de caractéristiques et laisser ainsi de côté l'utilisation des MFCC. En fait, des études supplémentaires et effectuées en parallèle de celles qui sont présentées dans ce mémoire, ont montré que lorsque la TOC est utilisée seule, le système présente des temps de traitement trop importants pour rendre cette technique intéressante (environ 10 fois supérieurs à ceux que l'on obtient en combinant la TOC avec les MFCC). L'analyse cepstrale a un rôle important car elle permet de réduire les coefficients de la TOC à 13 coefficients cepstraux par trame [87]. De plus, nous avons également aboutit à de trop faibles taux de reconnaissance pour continuer à exploiter cette idée (lorsque nous implémentons la TOC sans MFCC, nous obtenons 72 points de moins sur le taux de reconnaissance par rapport au système le plus performant que nous vous présenterons et qui lui utilise la TOC avec les MFCC). C'est pour ces raisons que nos systèmes hybrides seront basés sur la combinaison TOC - MFCC pour le module de paramétrisation.

4.3.1 Mise en place de la TOC pour essayer d'améliorer les performances du système

Un signal vocal se compose de différentes fréquences, chacune d'entre elles étant engendrée par une partie spécifique de la cavité vocale du locuteur. Par conséquent, toutes ces fréquences sont susceptibles de contribuer à la caractérisation d'un individu. Dans les techniques standard de reconnaissance, les scientifiques utilisent la bande de base du signal pour extraire les vecteurs de caractéristiques. Maintenant, en appliquant la TOC au signal avant de procéder à la phase de paramétrisation, on obtient une représentation du signal à différentes échelles (correspondant à différentes fréquences), ce qui permet de faire ressortir d'autres caractéristiques qui ne s'expriment pas dans l'analyse spectrale du signal.

4.3.1.1 Empreinte graphique du locuteur

La représentation obtenue peut être vue comme une empreinte graphique de la voix d'un individu : la TOC d'un signal bi-dimensionnel peut être interprétée comme une image [104]. Elle pourrait même être utilisée en reconnaissance de la parole en travaillant avec des techniques d'extraction développées pour le traitement de l'image [104]. C'est ce que propose Brian Damiano et al., dans l'article « Recognizing a Voice from its Model », 2000 [104] : la TOC est capable de révéler des caractéristiques du système vocal humain qui peuvent être utilisées pour modéliser la voix d'un locuteur. De plus, ils émettent l'hypothèse que cela pourrait palier aux problèmes d'identification survenant dans le cas d'une large population. Pour mieux comprendre ce type de représentation, nous vous présentons ci-dessous (fig. 41 et fig. 42) la TOC d'un signal de parole extrait de la base YOHO ; la phrase « zéro » est prononcée par une femme ; l'analyse est faite à l'aide de l'ondelette de Morlet, pour les échelles 1 à 128 (fig. 41), puis pour les échelles 1 à 64 (fig. 42).

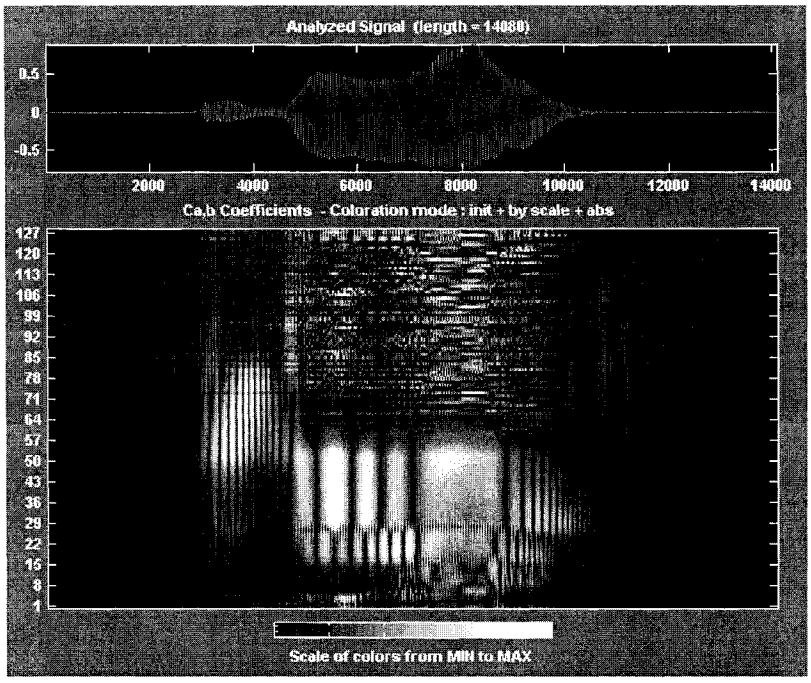


Figure 41 Analyse par l'ondelette de Morlet sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 128

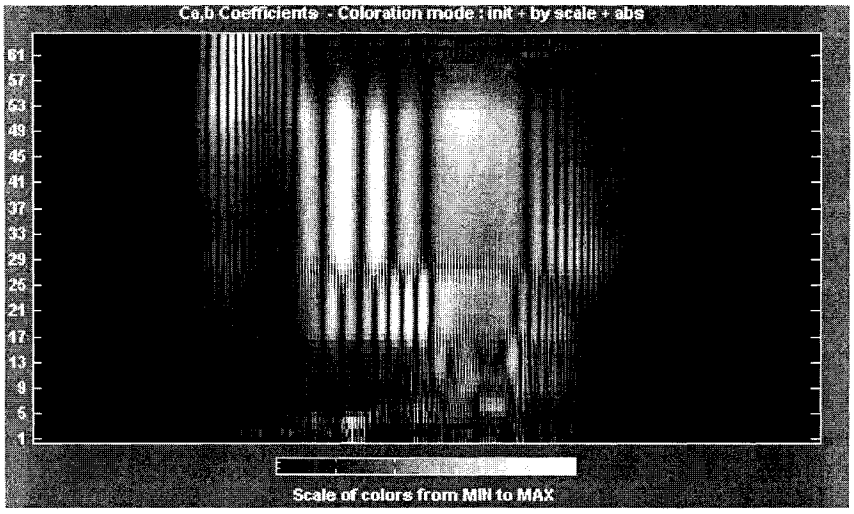


Figure 42 Analyse par l'ondelette de Morlet sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 64

Ces figures sont très assimilables à une empreinte vocale, et sont comparables à un spectrogramme (utilisé en reconnaissance de la parole) (fig. 43). Nous pourrons nous servir des points de cette image pour caractériser un locuteur.

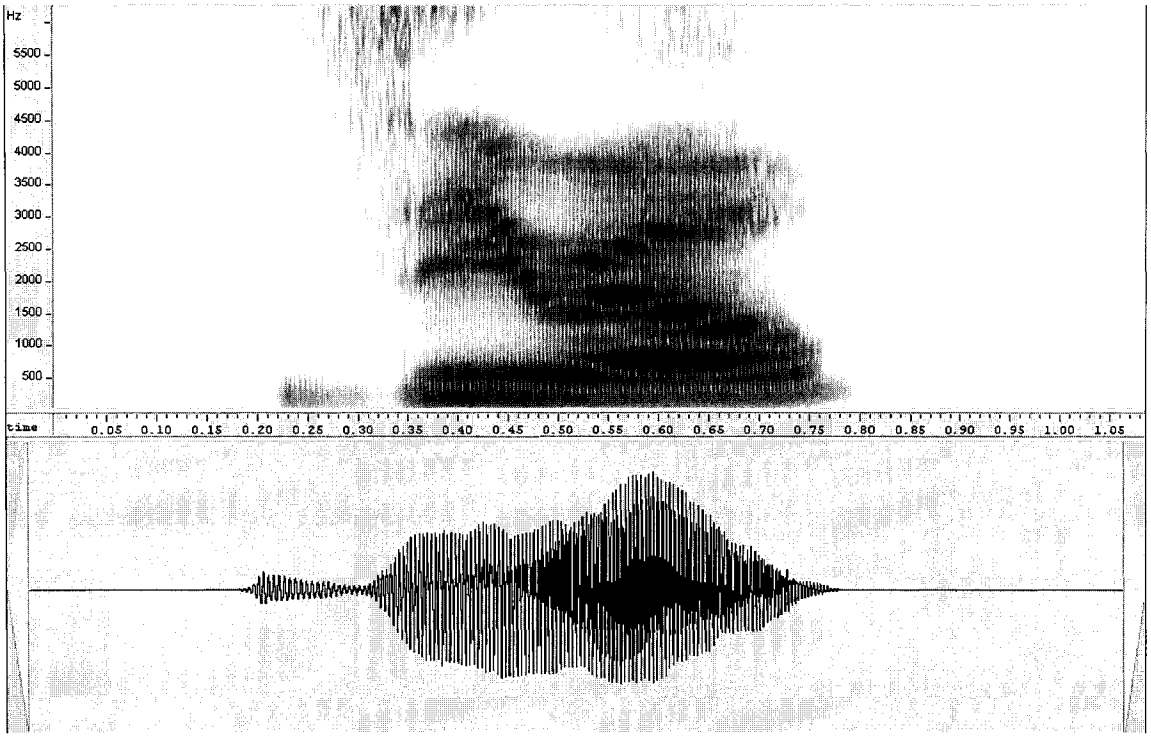


Figure 43 Spectrogramme du signal de parole « zéro » (YOHO) calculé sur WaveSurfer

Les recherches réalisées sur ces empreintes vocales nous ont amenés à faire certaines constatations qui méritent tout de même d'être citées bien que ce ne soit pas le sujet de notre projet. Nous les évoquerons donc brièvement dans le paragraphe qui suit.

4.3.1.2 Proposition d'utilisation de la TOC pour faire de la reconnaissance de mots isolés

En fait, nous savons que le spectrogramme d'un signal de parole peut être utilisé pour déterminer les formantes (fig. 44), mais aussi pour reconnaître des phonèmes ou bien des mots isolés [29], [105]. Pour les formantes, ceci est réalisé en localisant la position fréquentielle des zones de plus forte énergie (zones foncées) du spectrogramme de parole (fig. 44). En ce qui concerne la reconnaissance de phonèmes ou de mots, il suffit d'étudier la variation et la position de ces formantes.

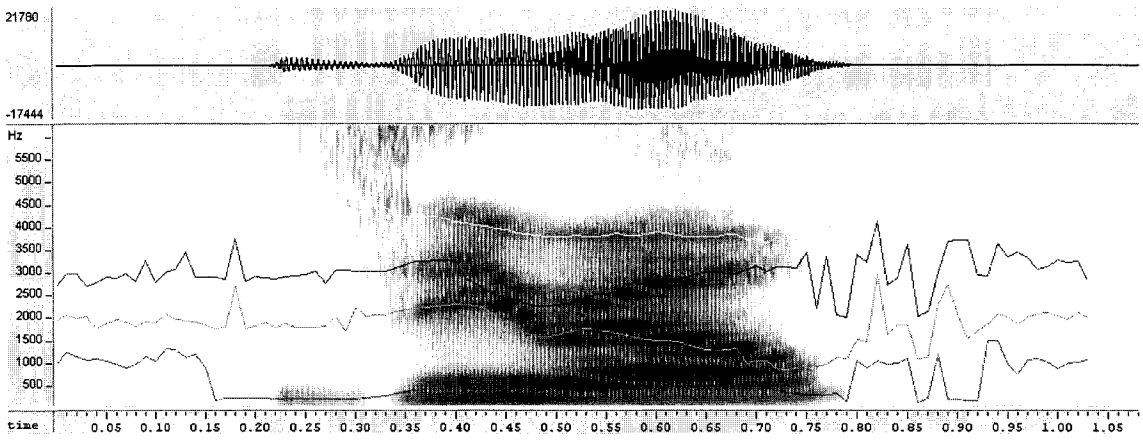


Figure 44 Spectrogramme du signal de parole « zéro » (YOHO) calculé sur WaveSurfer, avec détermination des quatre premières formantes

Nous avons observé des possibilités similaires avec la TOC. En fait, l'image 2D que nous obtenons est très semblable à celle donnée par le spectrogramme. Nous ne pouvons pas déterminer les formantes aussi facilement, par contre, nous pouvons reconnaître des phonèmes ou bien des segments de parole d'une manière bien plus simple : en effet, pour caractériser un phonème à l'aide d'un spectrogramme, nous devions analyser les trois premières formantes; avec l'image obtenue avec la TOC, il suffirait d'analyser une seule « courbe ». En fait, la TOC d'un signal de parole donne une empreinte vocale qui

se répète indéfiniment lorsque les échelles augmentent [104] (voir fig. 45). C'est un phénomène de redondance.

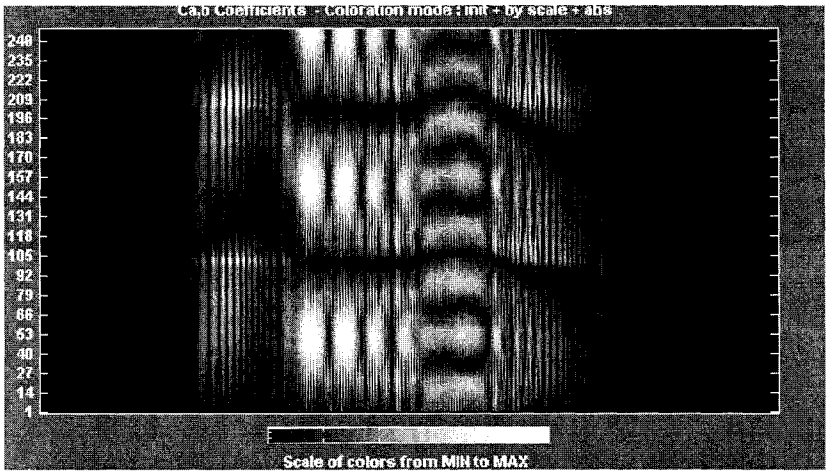


Figure 45 Analyse par l'ondelette de Haar sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 256

Entre chacune de ces répétitions, il existe une sorte de courbe noire séparatrice, une zone de coefficients de très faible énergie qui, graphiquement, se distingue fortement du reste de l'espace de la transformée. Cette courbe est parfaitement définie lorsque l'on utilise l'ondelette de Haar (fig. 41 et fig. 46). Elle varie au cours du temps selon les phonèmes prononcés, la vitesse d'élocution, et l'intonation. Le comportement est similaire à celui d'une formante. Il suffirait par conséquent de l'analyser pour caractériser quel est le phonème prononcé et donc faire de la reconnaissance de mots isolés.

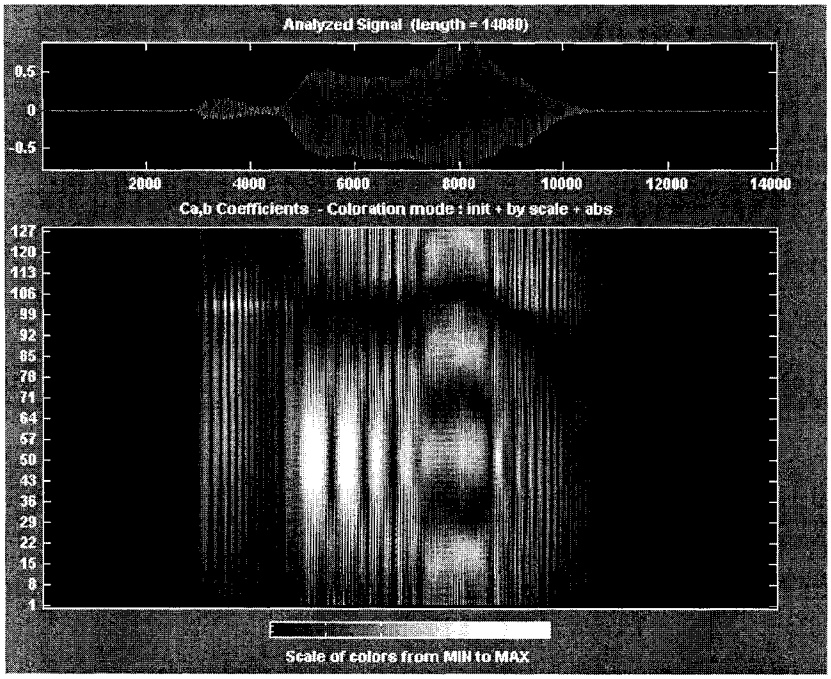


Figure 46 Analyse par l'ondelette de Haar sur le signal de parole « zéro » (YOHO) effectuée sur Matlab avec la toolbox « wavelets » pour les échelles 1 à 128

Cette technique ne permet pas de faire de la reconnaissance du locuteur étant donné que cette courbe de faible énergie dépend fortement de ce qui est dit. C'est la raison pour laquelle nous avons laissé ceci de côté. Nous avons tout de même tenu à le mentionner.

4.3.1.3 Méthode proposée pour exploiter l'empreinte du locuteur

À présent, il existe plusieurs possibilités pour traiter cette empreinte et extraire des vecteurs caractéristiques en utilisant uniquement des techniques de traitement du signal. Nous en avons retenu une qui nous semble triviale. Nous allons donc commencer par l'étudier et la tester, puis nous y apporterons les modifications nécessaires pour améliorer les performances du système.

Le choix de la méthode n'est pas basé sur des procédés existants; nous nous sommes quelque peu inspiré des bases enseignées en imagerie et plus précisément dans le domaine de la vision industrielle ou de la reconnaissance de formes [113].

Voici le raisonnement que nous avons suivi pour exploiter la TOC du signal de parole en procédant à de simples modifications dans le système de reconnaissance traditionnel:

On considère tout d'abord l'empreinte d'un locuteur obtenue en appliquant la TOC sur ses échantillons de parole. Dans le domaine du traitement de l'image, pour caractériser cette emprunte, on pourrait chercher à déterminer ses contours, c'est-à-dire les points ou s'expriment les plus fortes variations d'énergie dans un faible voisinage. Par analogie, il suffirait de considérer les points de la TOC du signal où s'expriment les plus fortes variations d'énergie. Cependant, il serait trop fastidieux de déterminer ces points sans employer les techniques d'imagerie. D'autant plus que de nombreux autres points contenus à l'intérieur de l'empreinte pourraient également être intéressants pour la caractérisation. C'est pourquoi, il nous a paru évident que l'utilisation d'une grille de points, tel un masque que l'on appliquerait sur l'image de la TOC, pourrait s'avérer intéressante. De plus, cette méthode serait très simple à implémenter et à utiliser. Cette grille pourrait avoir différentes formes, la plus simple serait une grille uniforme tel un quadrillage, et on pourrait ensuite la contracter en temps et en échelle au cours de la phase expérimentale. Ainsi, on pourrait déterminer par essai-erreur quel serait le meilleur maillage possible. C'est la méthode que nous avons choisie d'étudier. Elle est tout à fait innovante et paraît assez pertinente.

Les points ainsi retenus après ce filtrage seront ensuite recombinés pour former un nouveau signal, signal à partir duquel seront ensuite extrait les vecteurs paramétriques servant à caractériser le locuteur et à modéliser ce dernier.

4.3.2 Présentation des systèmes de reconnaissance hybrides proposés

Les systèmes de reconnaissance que nous proposons, sont en grande partie identiques au système d'identification de référence (sans TOC) que nous avons détaillé à la fin du chapitre 2, excepté le fait que ces derniers font usage de la TOC dans le module de prétraitement et de paramétrisation. Pour ces raisons, nous qualifierons nos systèmes de systèmes hybrides combinant ainsi la TOC et les MFCC pour l'extraction des vecteurs de caractéristiques.

4.3.2.1 Procédure

Quelle que soit la phase dans laquelle on se trouve (phase d'entraînement ou phase de test) le signal de parole est tout d'abord prétraité conformément à ce qui a été dit dans le deuxième chapitre. Puis on lui applique la transformée en ondelette continue. Le résultat de cette opération aboutit à une matrice 2D contenant les coefficients de la TOC.

Nous procédons par la suite à la phase de sélection des coefficients de la TOC. (nous détaillerons cette phase pour chaque système que nous testerons). Les points retenus après ce filtrage sont à nouveaux représentés sous forme d'une matrice 2D. Nous devons alors passer par une phase de recombinaison de ces coefficients pour aboutir à la formation d'un nouveau signal 1D. Ce nouveau signal a l'avantage de présenter des caractéristiques du locuteur différentes de celles contenues dans le signal vocal de base (et donc peut-être plus diversifiées). L'hypothèse que nous proposons est que le signal ainsi créé pourrait exprimer plus de variabilité intra-locuteur. Finalement, il ne reste plus qu'à procéder à l'extraction des coefficients cepstraux à partir du nouveau signal. Le schéma ci-dessous (fig. 47) résume les étapes que nous venons de citer :

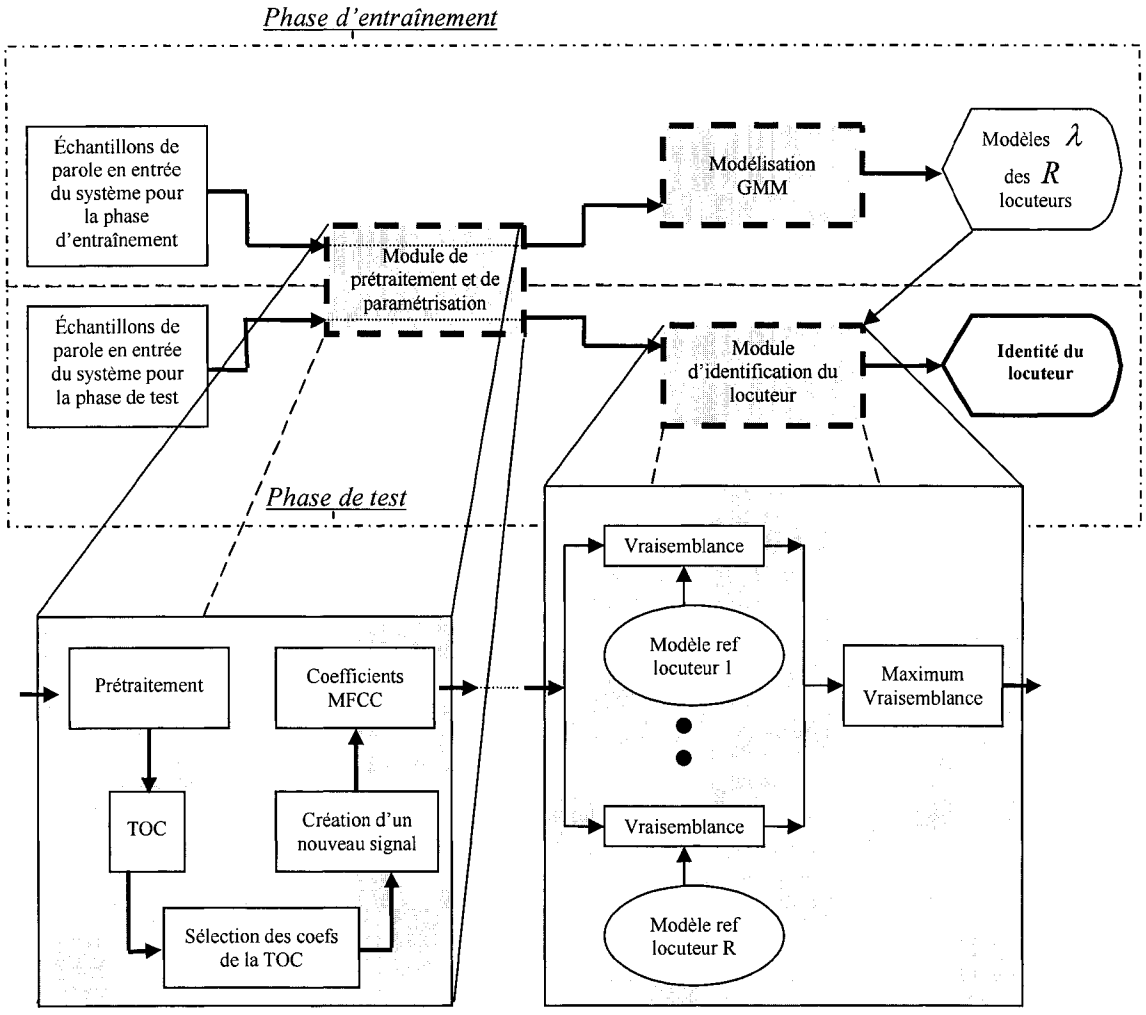


Figure 47 Principales étapes constituant un système de reconnaissance hybride

Dans ce chapitre, nous allons présenter une première catégorie de systèmes hybrides. Au fur et à mesure de nos modifications en vue de leur amélioration, nous les nommerons différemment pour mieux les distinguer.

4.3.2.2 Systèmes de reconnaissance hybrides utilisant une grille pour la sélection des coefficients de la TOC : systèmes hybrides G

Les systèmes d'identification hybrides décrits dans ce chapitre font usage d'une grille pour la sélection des coefficients de la TOC. Pour des facilités de compréhension, nous désignerons par G tout système hybride utilisant une grille comme outil de sélection.

La grille que nous utilisons pour sélectionner les points caractéristiques de la TOC du signal de parole est un simple maillage, plus ou moins resserré suivant les pas (pour l'axe des échelles et l'axe du temps) que nous considérons.

Nous appellerons Δa le pas existant entre les nœuds de l'axe des échelles, et Δb celui de l'axe du temps (fig. 48). Nous ferons ensuite varier ces paramètres dans la phase expérimentale.

Dans les tableaux récapitulatifs des tests effectués, nous avons fait varier ces deux pas à partir de la plus petite valeur des paramètres a et b, à savoir $a = 2$ (première échelle de la TOC) et $b = 1$ (premier échantillon temporel), jusqu'à leur valeur maximale, à savoir $a = 61$ (dernière échelle de la TOC) et $b = T$ (dernier échantillon temporel pour un signal constitué de T échantillons).

On notera qu'on ne pourra pas quantifier les grilles selon le nombre de points qu'elles considèrent étant donné que les échantillons de parole n'ont pas tous la même taille puisqu'ils dépendent de l'élocution de chaque locuteur. Les grilles devrant couvrir l'intégralité du signal, on ne pourra pas par conséquent disposer de grilles de tailles fixes. On peut juste définir leur maillage.

Enfin, on notera que le choix des pas Δa et Δb n'est pas basé sur une technique précise. Ne sachant pas comment choisir ces paramètres, nous pouvons seulement effectuer

quelques essais pour dégager une tendance générale et voir si cette méthode pour choisir les coefficients de la TOC peut s'avérer intéressante et mériter une étude plus approfondie par la suite.

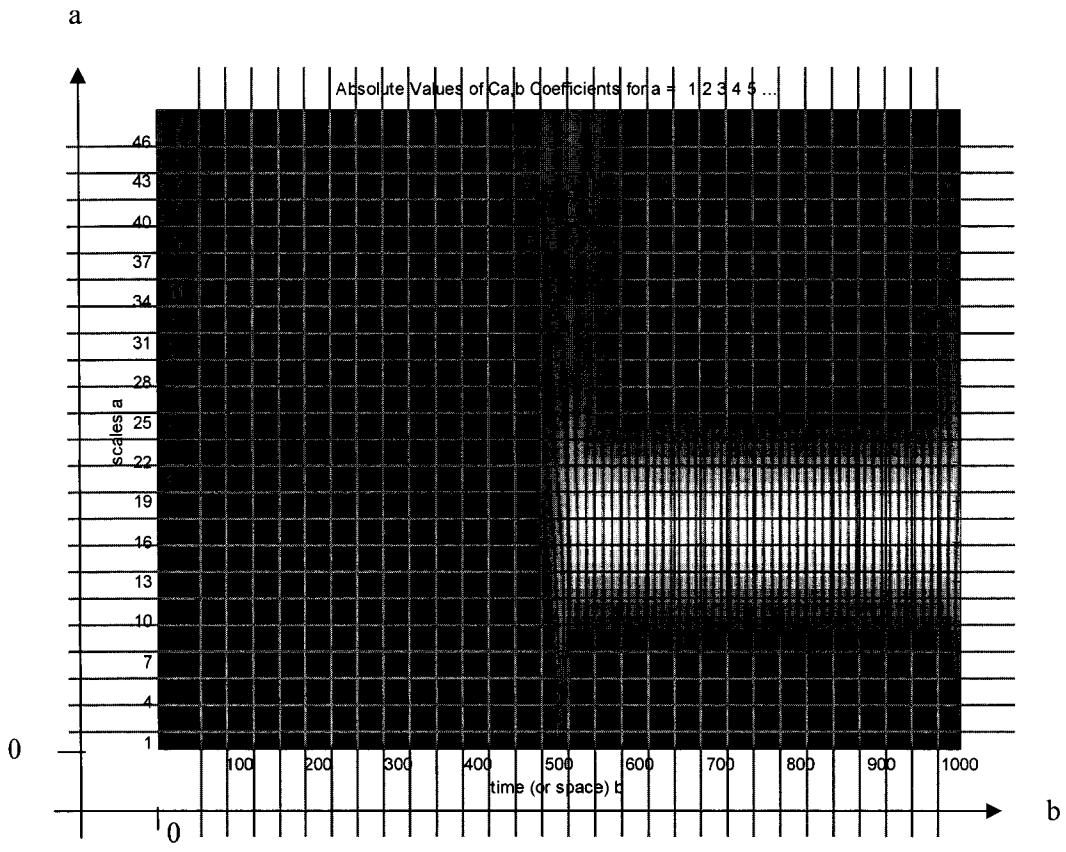


Figure 48 Sélection des coefficients de la TOC à l'aide d'une grille

4.3.2.3 Recombinaison des coefficients de la TOC en un nouveau signal 1D

Nous avons proposé deux possibilités de recombinaison des coefficients de la TOC pour obtenir un nouveau signal 1D spécifique au locuteur. À partir de la matrice de coefficients d'après sélection, nous pouvons :

- soit faire une recombinaison verticale en prenant les coefficients les uns après les autres suivant l'axe du temps, et ce, pour chaque échelle considérée; ainsi, on va créer un vecteur 1D formé par tous ces coefficients. On appellera ceci la recombinaison 1 (fig. 49). Si on utilise ce type de recombinaison après une sélection faite à l'aide d'une grille, on désignera le système par G1.

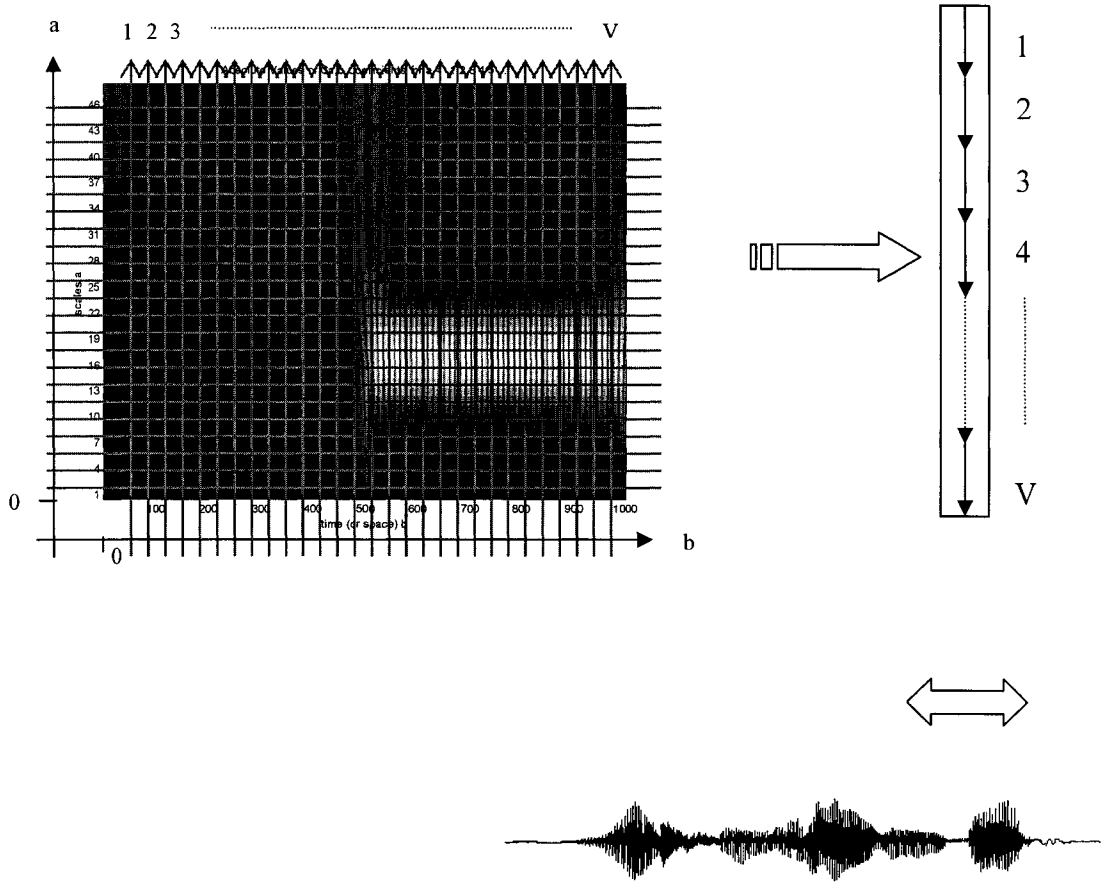


Figure 49 Recombinaison verticale des points extraits de la grille (recombinaison 1)

- soit faire une recombinaison horizontale en prenant les coefficients les uns après les autres suivant l'axe des échelles, et ce, pour chaque instant temporel du signal d'origine; ceci sera la recombinaison 2 (fig. 50). Si on utilise ce type de recombinaison après une sélection faite à l'aide d'une grille, on désignera le système par G2.

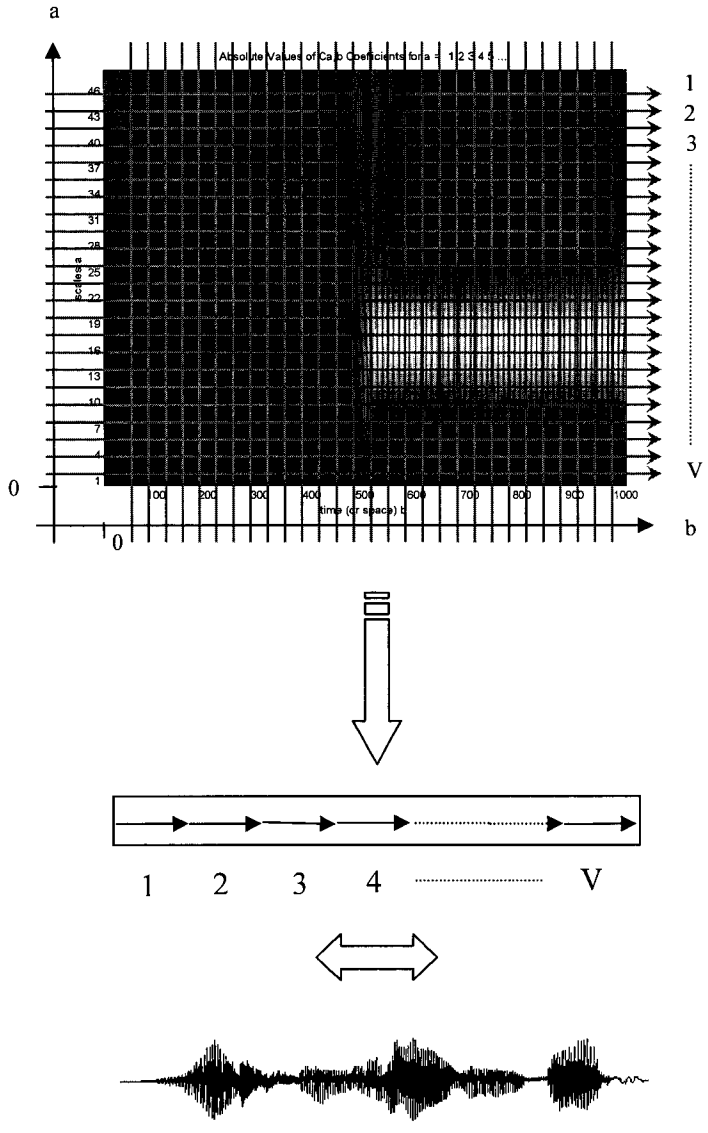


Figure 50 Recombinaison horizontale des points extraits de la grille (recombinaison 2)

On peut dire qu'un signal recombiné de cette manière est équivalent à la concaténation de la TOC du signal de parole de départ, calculée pour V échelles différentes. On a par conséquent V fois le même signal d'origine, mais présenté pour diverses échelles d'analyse.

Dans la partie qui suit, nous testerons les performances des systèmes hybrides G pour différents maillages, et pour les deux techniques de recombinaison.

4.4 Phase expérimentale

Les expériences menées ci-dessous permettent de mettre en évidence les performances des systèmes en milieu non bruité dans le cadre d'une population de locuteurs croissante.

Nous avons calculé le taux de reconnaissance ainsi que le taux d'erreur d'identification IER (en anglais, Identification Error Rate) pour chaque population étudiée. Ces mesures se calculent de la façon suivante [42]:

$$Taux\ de\ reconnaissance = \frac{Nombre\ de\ locuteurs\ identifiés}{Taille\ de\ la\ population} x100 \tag{4.1}$$

$$IER = \frac{Nombre\ d'erreurs\ d'identification}{Nombre\ de\ locuteurs\ identifiés} \tag{4.2}$$

Un système peut être qualifié de bon en matière d'identification si l'IER est relativement faible (le plus proche de zéro possible). Dans le cas contraire, plus l'IER est proche de 1, moins le système est performant. Bien entendu, les performances d'un système de reconnaissance dépendent fortement de la taille de la population considérée. Lorsque la

population augmente, les capacités d'identification décroissent naturellement, ce qui fait croître le taux d'erreur IER [42].

Avant de procéder au test des systèmes hybrides que nous avons conçu, nous avons dans un premier temps procédé à l'évaluation des performances du système de référence (sans TOC) en calculant son taux de reconnaissance et son IER (voir chap.2). Ainsi, nous pourrons comparer les résultats obtenus au cours de cette phase expérimentale avec ceux obtenus dans les mêmes conditions avec le système de référence.

4.4.1 Paramètres optimaux additionnels pour les systèmes hybrides

Seuls quelques paramètres additionnels doivent être tenus en compte dans le cas des systèmes hybrides. Ce sont ceux concernant le nouveau module de paramétrisation faisant usage de la TOC. En effet, nous devons déterminer quelle est la meilleure ondelette analysante et quelles sont les échelles d'analyse à considérer.

4.4.1.1 Ondelette analysante

Conformément aux résultats d'études donnés dans la littérature (voir p. 86), nous utiliserons l'ondelette analysante réelle de Morlet. Cette ondelette est la plus adéquate pour étudier les signaux de parole [87]. De plus, au cours d'une phase d'essais préliminaire, nous avons trouvé que les ondelettes de Haar et de Daubechie donnent également de bons résultats mais légèrement inférieurs à ceux obtenus avec Morlet. Les autres ondelettes testées, à savoir Gauss, Chapeau Mexicain, Gauss complexe et Morlet complexe, ne permettent pas d'aboutir de bonnes performances.

4.4.1.2 Échelle d'analyse

Les échelles sur lesquelles nous avons choisi de calculer la TOC des signaux de parole, sont celles comprises entre les échelles 2 (fréq. 3400 Hz) et 61 (fréq. 100 Hz). En fait, ce choix vient du fait que YOHO est une base de données de haute qualité audio téléphonique, où les segments de parole ont été filtrés par un filtre passe-bas de fréquence maximale 3800 Hz. Il est donc inutile de calculer la TOC des signaux sur des échelles correspondant à des fréquences autres que celles appartenant à cet intervalle.

4.4.2 Essais expérimentaux pour les systèmes hybrides G

Testons à présent les performances des systèmes hybrides G qui utilisent une grille pour sélectionner les coefficients de la TOC. Les différentes méthodes de recombinaison sont successivement évaluées.

4.4.2.1 Tests des systèmes d'identification hybrides G1

Nous testons ici les performances des systèmes hybrides qui utilisent une grille pour sélectionner les coefficients de la TOC, et dont la recombinaison de ces points en un nouveau signal 1D se fait de façon verticale. Les tests sont menés avec des maillages différents. On rappelle que les choix du maillage à l'aide des paramètres Δa et Δb a été fait par essais-erreur, et nous avons reporté ci-dessous quelques grilles intéressantes du point de vue de leurs résultats.

Les tableaux VI et VII présentent une partie des résultats obtenus pour des populations de 10 et 30 locuteurs :

Tableau VI

Performances (en %) des systèmes hybrides G pour différentes tailles de grilles et pour une recombinaison verticale (1) des coefficients de la TOC

	Taux de reconnaissance (en %)					
Base de données (population)	**Système de référence[7] (sans TOC)**	**Systèmes hybrides G1 (avec TOC)**				
		$\Delta a = 4$ $\Delta b = 80$	$\Delta a = 8$ $\Delta b = 80$	$\Delta a = 4$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 1$
10	100	10	10	100	90	50
30	90	10	3.3	73.3	66.7	23.3

Les résultats des tests effectués sont globalement faibles en matière d'identification excepté par exemple pour la grille $[\Delta a = 4;\ \Delta b = 8]$ (grille dont les nœuds ont un pas de 4 selon l'axe des échelles quand a varie de 2 à 61, et un pas de 8 échantillons selon l'axe du temps quand b varie de 1 à T), ou encore la grille $[\Delta a = 8;\ \Delta b = 8]$. Cependant, pour une large population, les taux de reconnaissance demeurent assez éloignés de ceux atteints sans utiliser la TOC : 73.3 % et 66.7 % respectivement pour une population de 30 locuteurs, contre 90 % sans faire usage de la TOC.

Nous avons vu que plus le maillage se resserre (donc plus les pas sont petits), et meilleurs sont les résultats. Ceci est valable jusqu'à un certain point. Plus on considère d'échelles et meilleures sont les performances. Cependant, en prenant trop de points sur l'axe du temps, nous faisons chuter le taux de reconnaissance. Ceci s'explique par le fait que la TOC est une transformée très redondante, donc en prenant trop de points on aura

[7] Le système de référence auquel on se réfère est celui qui est décrit à la fin du chapitre 2 et qui ne fait pas usage de la TOC

plus de risques de prendre en compte de l'information inutile et donc de faire une mauvaise modélisation. De plus, les temps de calculs pour les phases d'entraînement et de test du système deviennent très lourds. Quelques exemples sont reportés dans le tableau VII ci-dessous; nous les avons mesurés à l'aide des commandes *tic* / *toc* de Matlab (qui ont été placé en début et fin de programme) :

Tableau VII

Temps de traitement (en secondes) des systèmes d'identifications hybrides G pour les phases d'entraînement et de test, en fonction de la taille de la grille et pour une recombinaison verticale (1) des coefficients de la TOC

	Temps de traitement (en secondes)		
	Système de référence (sans TOC)	**Systèmes hybrides G1 (avec TOC)**	
Base de données (population)		$\Delta a = 4$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 1$
10	Entraînement : 13 Test : 1	Entraînement : 574 Test : 15	Entraînement : 6830 Test : 309
30	Entraînement : 44 Test : 4.5	Entraînement : 1720 Test : 55	Entraînement : 78 876.9 Test : 2188.8

Testons à présent les performances des systèmes hybrides G pour une recombinaison horizontale des coefficients de la TOC.

4.4.2.2 Tests des systèmes d'identification hybrides G2

Nous testons ici les performances des systèmes hybrides qui utilisent une grille pour sélectionner les coefficients de la TOC, et dont la recombinaison de ces points en un nouveau signal 1D se fait de façon horizontale.

Les tableaux VIII et IX présentent quelques résultats obtenus pour des populations de 10 et 30 locuteurs :

Tableau VIII

Performances (en %) des systèmes hybrides G pour différentes tailles de grilles et pour une recombinaison horizontale (2) des coefficients de la TOC

	Taux de reconnaissance (en %)						
	Système de réf. (sans TOC)	**Systèmes hybrides G2 (avec TOC)**					
Base de données		$\Delta a = 4$ $\Delta b = 80$	$\Delta a = 8$ $\Delta b = 80$	$\Delta a = 4$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 1$	$\Delta a = 34$ $\Delta b = 1$
10	100	10	10	80	80	90	100
30	90	6.7	3.3	50	50	73.3	86.7

Nous observons également ici que de meilleurs résultats sont obtenus pour des maillages plus resserrés.

En comparant les résultats des grilles avec une recombinaison des coefficients de la TOC verticale avec ceux obtenus ici pour une recombinaison horizontale, on observe qu'une recombinaison verticale donne de meilleurs résultats lorsqu'on ne considère pas tous les coefficients suivant l'axe du temps. Lorsque tous les coefficients temporels sont

pris en compte, la recombinaison horizontale est nettement meilleure (un gain de 50 points en reconnaissance sur une population de 30 locuteurs). La recombinaison de type horizontale donne de plus une ouverture vers de nouvelles possibilités bien plus avantageuses : on observe que pour tout type de population, plus on prend de coefficients suivant l'axe du temps (même en considérant tous les coefficients temporels), et meilleurs sont les résultats. De plus, les deux dernières colonnes du tableau VIII nous montrent que pour une telle recombinaison, on peut obtenir de meilleures performances en diminuant le nombre d'échelles. On aura ainsi moins de coefficients à traiter et on améliore aussi les temps de calculs du système. On atteint ainsi un taux de 86.7 % pour 30 locuteurs juste en considérant 2 échelles différentes (échelles 2 et 36). Ces résultats sont fortement encourageants étant donné qu'ils sont très proches de ceux donnés par le système d'identification de référence (90 %).

Ce type de recombinaison nous semble la mieux adaptée étant donnée qu'il permet de palier le problème de la TOC : en considérant moins d'échelles et tout en considérant tous les coefficients de celles sélectionnées, on réduit les risques de sélectionner des coefficients redondants. On améliore ainsi la modélisation, donc les taux de reconnaissance, ainsi que les temps de traitement (comme indiqué au tableau IX):

Tableau IX

Temps de traitement (en secondes) des systèmes d'identifications hybrides G pour les phases d'entraînement et de test, en fonction de la taille de la grille et pour une recombinaison horizontale (2) des coefficients de la TOC

	Temps de traitement (en secondes)		
	Système de référence	**Systèmes hybrides G2 (avec TOC)**	
Base de données (population)	**(sans TOC)**	$\Delta a = 4$ $\Delta b = 8$	$\Delta a = 8$ $\Delta b = 1$
10	Entraînement : 13 Test : 1	Entraînement : 92 Test : 8	Entraînement : 342 Test : 10
30	Entraînement : 44 Test : 4.5	Entraînement : 278 Test : 30	Entraînement : 1160 Test : 53

4.5 Conclusions

Les tests effectués sur les systèmes hybrides de type G nous permettent d'apporter les conclusions suivantes :

Les essais portant sur l'utilisation de la transformée en ondelettes continue n'ont pas permis pour l'instant d'aboutir à des performances supérieures à celles des systèmes actuels décrits dans la littérature et ne faisant pas usage de cette transformée. Cependant, elle permet dans certains cas de s'en approcher.

Les coefficients de la TOC peuvent effectivement être utiles pour faire de la reconnaissance du locuteur étant donné qu'on aboutit à des taux de reconnaissance de

l'ordre des 80%. Seulement, tout dépend comment nous nous en servons par la suite pour caractériser le locuteur. En effet, nous avons pu observer que des recombinaisons verticales des coefficients de la TOC en un vecteur 1D, sont dans certains cas plus performantes que des recombinaisons horizontales. Cependant, une recombinaison horizontale permet d'obtenir des performances supérieures si l'on considère tous les coefficients suivant l'axe temporel pour les échelles choisies. Peu d'échelles suffisent pour aboutir à de meilleurs résultats. On aura donc des maillages constitués par moins de points, ce qui a pour effet de réduire considérablement les temps de calculs.

Enfin, nous avons vu que le type grille utilisée joue également un rôle important dans les performances du système. En effet, il semble exister un type de maillage plus adéquat qu'un autre, mais la justification d'une telle sélection est très improbable. Il est impossible de tester tous les types de maillages imaginables pour déterminer le plus performant. Cependant, étant donné les résultats que nous avons obtenus, nous pouvons penser qu'il existerait un type de maillage utilisant peu de points et dont la sélection des pas serait simple à déterminer, et qui permettrait de donner de bons taux de reconnaissance avec un faible temps de traitement. Par exemple en utilisant une grille de points aléatoires. Ce point pourrait d'ailleurs faire l'objet d'études futures.

Finalement, on remarquera que les grilles (des systèmes G2) $[\Delta a = 8;\ \Delta b = 1]$, et $[\Delta a = 34;\ \Delta b = 1]$ qui sont les plus performantes sont en fait des combinaisons de lignes de coefficients pour des échelles prédéfinies plutôt que des grilles proprement dites. En effet, ces dernières prennent en compte tous les échantillons de l'axe temporel (puisque le pas de b est égal à 1), et considèrent respectivement 8, et 2 échelles distinctes. Ceci revient par conséquent à sélectionner non plus des points sur une grille, mais plusieurs lignes de coefficients sur l'axe des échelles.

Dans la suite du mémoire, nous considérerons donc uniquement la recombinaison de type horizontale.

Ces résultats nous amènent alors à formuler de nouvelles hypothèses :

- La combinaison de plusieurs lignes dans leur intégralité, sélectionnées à des échelles différentes, peut aboutir à de bons résultats. Ceci entraînerait de plus une baisse des temps de calculs si on prend tout au plus 10 échelles (voir tableau X) puisque l'on traite moins de données.

- La sélection d'une seule et unique ligne de coefficients pourrait même s'avérer suffisante. Ceci serait extrêmement intéressant étant donné la simplicité d'implémentation que cela requiert et les gains en temps de calcul que cela pourrait engendrer.

Nous allons par conséquent nous pencher sur ce nouveau type de sélection des coefficients de la TOC dans le chapitre qui suit. Nous essayerons ainsi d'améliorer les performances du système hybride que nous avons conçu et nous continuerons de vérifier si l'utilisation de la TOC permet ou pas d'aboutir à de meilleurs résultats que ceux obtenus avec le système de référence (sans TOC).

CHAPITRE 5

MODIFICATION ET AMÉLIORATION DU SYSTÈME DE RECONNAISSANCE HYBRIDE PROPOSÉ : UTILISATION DE LIGNES POUR SÉLECTIONNER LES COEFFICIENTS DE LA TOC

5.1 Introduction

Dans ce chapitre, nous allons modifier le système hybride que nous venons de concevoir en utilisant désormais des lignes pour sélectionner les coefficients de la TOC. Nous pourrons ainsi vérifier une des hypothèses posée précédemment, et selon laquelle un tel mode de sélection par combinaison de plusieurs lignes dans leur intégralité, sélectionnées à des échelles différentes, pourrait donner de bons résultats.

5.2 Présentation du système de reconnaissance hybride amélioré

Tout d'abord, nous allons donc nous intéresser à l'expérimentation de ces « nouvelles grilles de filtrage » que nous nommerons dès maintenant « combinaisons de lignes ». Nous testerons ici les performances des systèmes hybrides lorsque l'on prend tous les coefficients de la TOC pour une gamme d'échelles prédéfinie.

5.2.1 Systèmes de reconnaissance hybrides utilisant une combinaison de lignes pour la sélection des coefficients de la TOC : systèmes hybrides C

Pour cela, il nous suffit de modifier le module de sélection des coefficients de la TOC (cf. fig. 47) en prenant tous les points sur l'axe du temps pour seulement quelques échelles, et en conservant désormais la recombinaison 2 de type horizontale. Nous nommerons par C les systèmes basés sur une telle combinaison de lignes.

Une fois de plus, il n'existe pas de méthode qui puisse nous aider à choisir la combinaison de lignes parfaite. On peut juste émettre l'hypothèse qu'une bonne combinaison serait celle qui prend en compte les lignes où se manifeste le plus d'activité vocale, ou autrement dit, les lignes qui passent par le plus grand nombre de maxima (lorsqu'on considère l'énergie de la TOC du signal). Mais on peut également supposer que les zones de faible activité possèdent également un important pouvoir discriminant. Ceci n'étant pas vérifié, il nous faut par conséquent tester différentes possibilités (fig. 51).

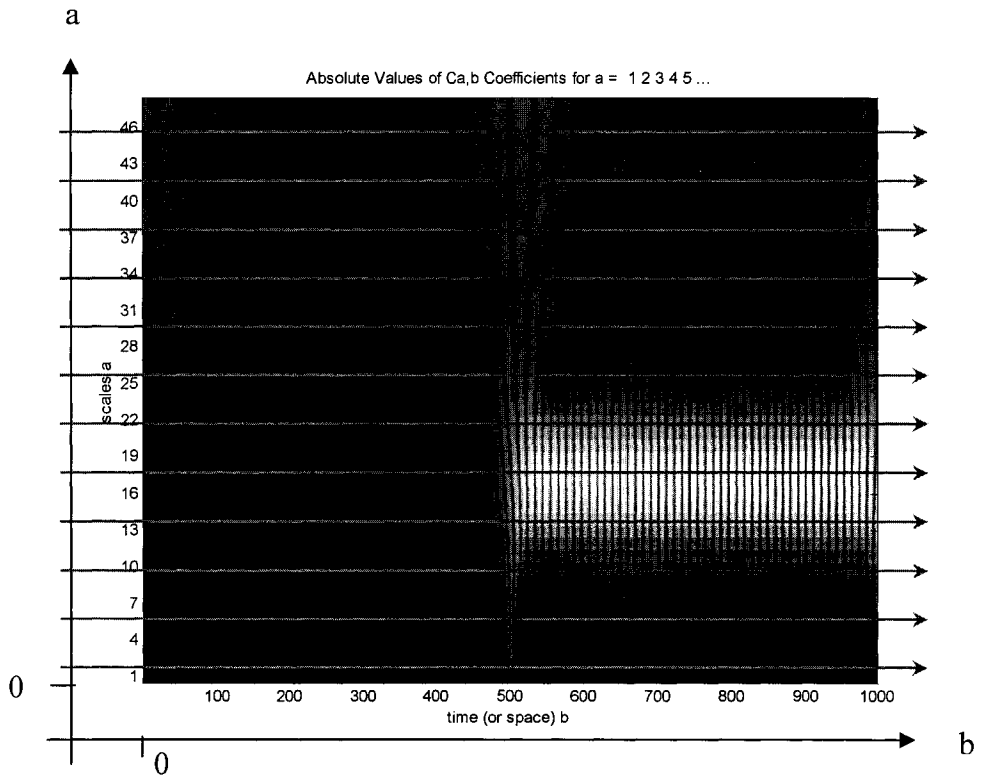


Figure 51 Sélection des coefficients de la TOC à l'aide d'une combinaison de plusieurs lignes de coefficients

5.2.2 Tests du système d'identification hybride C

Dans cette phase expérimentale, nous avons testé plusieurs types de combinaisons de lignes. Tout d'abord, nous avons procédé en prenant des lignes de manière consécutive (pour un pas $\Delta a = 1$).

Nous avons voulu tester d'abord certaines bandes à caractère particulier :

- La bande de lignes pour les échelles 2 à 21 qui correspond aux fréquences comprises entre 3400 Hz et 300 Hz. C'est la bande des fréquences filtrées en téléphonie. C'est ici que se trouvent les principales caractéristiques audibles de la voix.

- La bande de lignes pour les échelles 12 à 34 qui correspond grossièrement à la zone où l'on observe la majorité de l'empreinte vocale du locuteur sur l'image donnée par la TOC du signal de parole.

- La bande de lignes pour les échelles 2 à 64 qui correspond à la zone des fréquences que nous avions choisies pour appliquer la TOC.

Et nous avons ensuite réalisé les mêmes tests en prenant les bandes sans signification particulière :

- La bande de lignes pour les échelles 20 à 31

- La bande de lignes pour les échelles 25 à 36

Les tableaux X et XI présentent quelques résultats obtenus pour des populations de 10 et 30 locuteurs :

Tableau X

Performances des systèmes hybrides C pour différentes bandes de lignes consécutives à caractère particulier

	Taux de reconnaissance (en %)			
	Système de référence (sans TOC)	**Systèmes hybrides C (avec TOC)**		
		[Bandes de lignes consécutives à caractère particulier]		
Base de données (population)		$a \in [2;61]$ $\Delta a = 1$ $\Delta b = 1$	$a \in [2;21]$ $\Delta a = 1$ $\Delta b = 1$	$a \in [12;34]$ $\Delta a = 1$ $\Delta b = 1$
10	100	90	80	90
30	90	80	70	73.3

Tableau XI

Performances des systèmes hybrides C pour différentes bandes de lignes consécutives sans signification particulière

	Taux de reconnaissance (en %)		
	Système de référence (sans TOC)	**Systèmes hybrides C (avec TOC)**	
		[Bandes de lignes consécutives sans signification particulière]	
Base de données (population)		$a \in [20;31]$ $\Delta a = 1$ $\Delta b = 1$	$a \in [25;36]$ $\Delta a = 1$ $\Delta b = 1$
10	100	100	100
30	90	86.7	90

Les résultats obtenus pour des populations de 10 et de 30 locuteurs sont très satisfaisants. En effet, pour la base de 10 locuteurs, le système hybride C atteint les 100 % de reconnaissance pour deux bandes différentes, égalant ainsi les performances du système de référence. Pour la base de 30 locuteurs, on atteint également un taux de 90 % pour une de ces bandes.

Ce qui ressort de ces résultats, c'est que les zones que nous pensions intéressantes en matière de discrimination (bandes de lignes pour les échelles [2 à 21] et [12 à 34]) ne permettent pas d'atteindre de bons taux de reconnaissance. La sélection de toutes les échelles de la TOC améliore un peu les résultats mais au dépend d'une forte augmentation des temps de traitement. En revanche, nous remarquons qu'on peut égaler les performances du système de référence en choisissant une bande qui apriori ne présente aucune signification particulière (bande d'échelles [25 à 36] par exemple).
On pourrait alors penser que d'autres types de combinaisons de lignes (non forcément consécutives) permettraient également une bonne identification. C'est ce que nous avons vérifié et retranscrit dans les tableaux ci-dessous (tab. XII et tab. XIII) :

Tableau XII

Performances des systèmes hybrides C pour différentes combinaisons aléatoires de plusieurs lignes de coefficients de la TOC

	Taux de reconnaissance (en %)				
	Système de référence (sans TOC)	**Systèmes hybrides C (avec TOC)**			
		[Combinaison aléatoire de plusieurs lignes]			
Base de données		$a \in [2;61]$ $\Delta a = 10$ $\Delta b = 1$	$a \in [2;22]$ $\Delta a = 2$ $\Delta b = 1$	$a = [11;21;31;43]$ $\Delta b = 1$	$a = \begin{bmatrix} 7;13; \\ 21;\ 27;30 \end{bmatrix}$ $\Delta b = 1$
10	100	80	90	90	100
30	90	80	83.3	86.7	90

Nous observons à partir des résultats de ce dernier tableau, que des combinaisons aléatoires de plusieurs lignes permettent aussi d'égaler les performances obtenues sans l'utilisation de la TOC (Combinaison des lignes pour les échelles 7;13;21;27; et 30).

Tableau XIII

Performances des systèmes hybrides C pour différentes combinaisons aléatoires de peu de lignes de coefficients de la TOC

	Taux de reconnaissance (en %)				
	Système de référence	**Systèmes hybrides C (avec TOC)**			
		[Combinaison aléatoire de peu de lignes]			
Base de données (population)	**(sans TOC)**	$a=[48;51]$ $\Delta b=1$	$a=[21;43]$ $\Delta b=1$	$a=[22]$ $\Delta b=1$	$a=[2]$ $\Delta b=1$
10	100	90	100	100	90
30	90	83.3	80	90	76.7

Le résultat le plus prometteur est celui obtenu pour les combinaisons avec peu de lignes. En effet, en ne prenant en compte qu'une seule échelle de la TOC, on montre qu'il est possible d'égaler les performances du système de référence (par exemple, pour la ligne à l'échelle 22).

De plus, nos observations semblent indiquer qu'il n'existe pas de corrélation entre le nombre d'échelles considérées (et par conséquent le nombre de lignes, ou encore le nombre de points) et la valeur du taux de reconnaissance obtenue.

5.2.3 Conclusions sur le système hybride C

La sélection d'une seule ligne peut nous permettre d'obtenir de bons résultats et ceci en améliorant considérablement les temps de calculs par rapport à la sélection de plusieurs lignes simultanées (cf. tab. XIV). Il existe en effet moins de donnée à traiter aussi bien pendant la phase d'apprentissage que pendant la phase de test du système.

Tableau XIV

Temps de traitement des systèmes d'identifications hybrides C pour les phases d'entraînement et de test, pour une combinaison formée par une seule ligne de coefficients de la TOC

	Temps de traitement (en secondes)	
Base de données (population)	**Système de référence (sans TOC)**	**Systèmes hybrides C (avec TOC)** $a=[22]$ $\Delta b=1$
10	Entraînement : 13 Test : 1	Entraînement : 17 Test : 1.5
30	Entraînement : 44 Test : 4.5	Entraînement : 55 Test : 6.5

Nous avons des temps de calculs très proches de ceux du système de reconnaissance de référence (sans TOC).

Nous avons vu lors des tests, que l'obtention de bons résultats dépend directement du choix de l'échelle (cf. tab. XIII, ligne à l'échelle 22 : 90 % pour 30 locuteurs; ligne à l'échelle 2 : 76.7 % pour 30 locuteurs). Certaines de ces options permettent d'égaler les

performances du système traditionnel. Mais il faudrait maintenant procéder par essai-erreur pour vérifier si certaines d'entre elles ne permettent pas de les surpasser puisque nous ne savons pas, pour l'instant, comment les choisir.

5.3 Présentation d'un nouveau système hybride amélioré

Après avoir débuté la phase expérimentale en sélectionnant les coefficients de la TOC à l'aide d'une grille, et, en allant petit à petit d'amélioration en amélioration, nous avons aboutit à un système basé sur une sélection rectiligne à échelle fixe, offrant de bonnes perspectives en matière de reconnaissance. Nous allons le vérifier ci-dessous.

5.3.1 Systèmes de reconnaissance hybrides utilisant une seule ligne pour la sélection des coefficients de la TOC : systèmes hybrides L

Le nouveau système hybride amélioré est par conséquent strictement identique au précédent, excepté pour le module de sélection des coefficients de la TOC (voir fig. 47) dont la tache est ici réalisée en prenant tous les points d'une seule échelle d'analyse. Ces systèmes basés sur la sélection d'une ligne seront nommés systèmes L.

Le seul problème qui se pose à nos yeux concerne le choix de l'échelle de la TOC. Ce problème n'est pas nouveau, et beaucoup de scientifiques se sont déjà posé la question dans leur propre domaine de recherche. Il n'existe pas de méthode permettant de déterminer la meilleure échelle d'analyse pour le signal de parole. Nous devrons par conséquent trouver une technique permettant de justifier le choix de l'échelle (fig. 52).

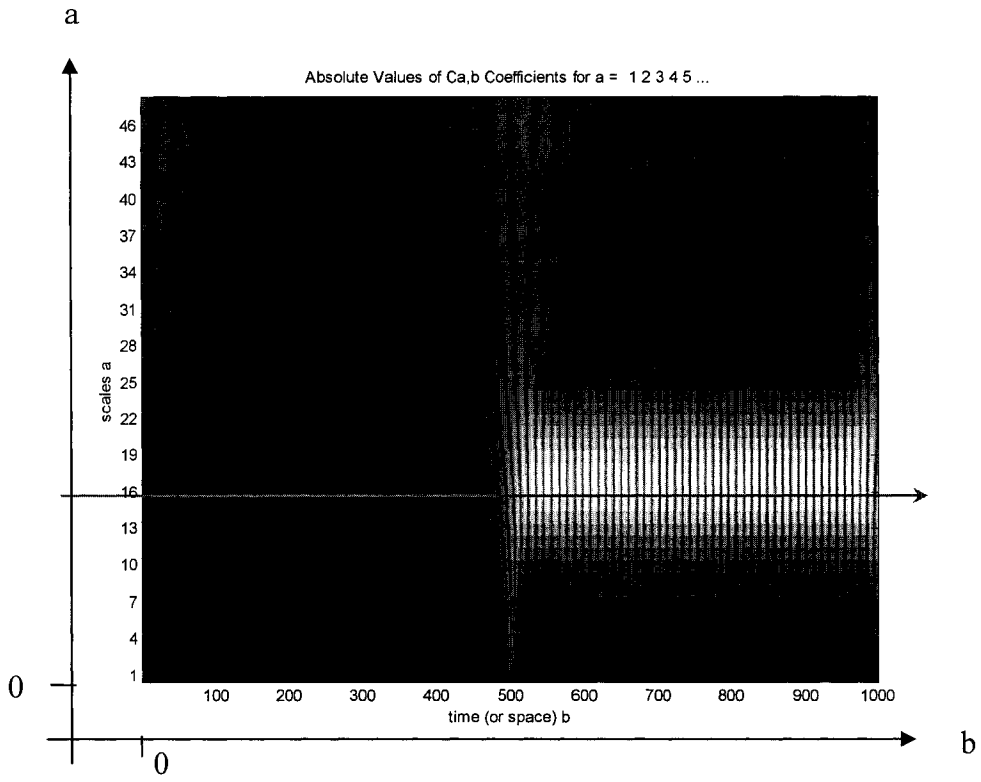


Figure 52 Sélection des coefficients de la TOC à l'aide d'une seule ligne de coefficients

Pour commencer, nous testerons les performances du système L pour toutes les lignes (correspondant à des échelles à valeur entière) de la TOC du signal de parole. Nous chercherons par la suite à justifier le choix de l'échelle retenue.

Dans les essais ci-dessous, nous ferons donc varier l'échelle d'analyse de 2 à 61 (ce sont les bornes sur lesquelles on avait décider de calculer la TOC).

5.3.2 Tests du système d'identification L

Nous avons décidé de regrouper dans les tableaux XV et XVI ci-dessous une partie des résultats obtenus dans le but de montrer certains points intéressants. La totalité des résultats de l'expérience est regroupée dans l'annexe 6. Les tests ont été effectués pour des populations de 10, 20, 30, 40, et 50 locuteurs.

Tableau XV

Performances des systèmes hybrides L pour différentes lignes de coefficients de la TOC

	Taux de reconnaissance du système (en %)						
Base de données (population)	**Système de référence (sans TOC)**	**Système hybride L (avec TOC)**					
		Ligne de coefficients pour l'échelle					
		a=3	**a=16**	**a=17**	**a=21**	**a=33**	**a=42**
10	100	100	80	100	100	100	100
20	90	85	70	90	90	95	90
30	90	86.7	53.3	90	93.3	93.3	90
40	85	90	37.5	90	92.5	85	90
50	90	92	40	86	88	84	88

Les zones ombragées correspondent à des résultats du système hybride L (avec TOC) supérieurs à ceux du système de référence (sans TOC).

Tableau XVI

IER des systèmes hybrides L pour différentes lignes de coefficients de la TOC

	IER (Identification Error Rate) du système						
	Système de référence (sans TOC)	Système hybride L (avec TOC)					
		Ligne de coefficients pour l'échelle					
Base de données		a=3	a=16	a=17	a=21	a=33	a=42
10	0	0	0.25	0	0	0	0
20	0.111	0.176	0.429	0.111	0.111	0.053	0.111
30	0.111	0.154	0.875	0.111	0.071	0.071	0.111
40	0.176	0.111	1.667	0.111	0.081	0.176	0.111
50	0.136	0.087	1.500	0.16	0.136	0.190	0.136

Les résultats obtenus en choisissant une seule échelle de la TOC nous permettent enfin d'aboutir à d'excellentes performances en matière de reconnaissance avec la TOC (voir par exemple les cases ombragées des tableaux).

Pour donner une idée des résultats globaux, nous avons tracé dans le graphique ci-dessous (fig. 53) l'ensemble des taux de reconnaissance obtenus en fonction des échelles des lignes, et selon la taille de la population. (ces courbes sont obtenues à partir des résultats reportés dans l'annexe 6) :

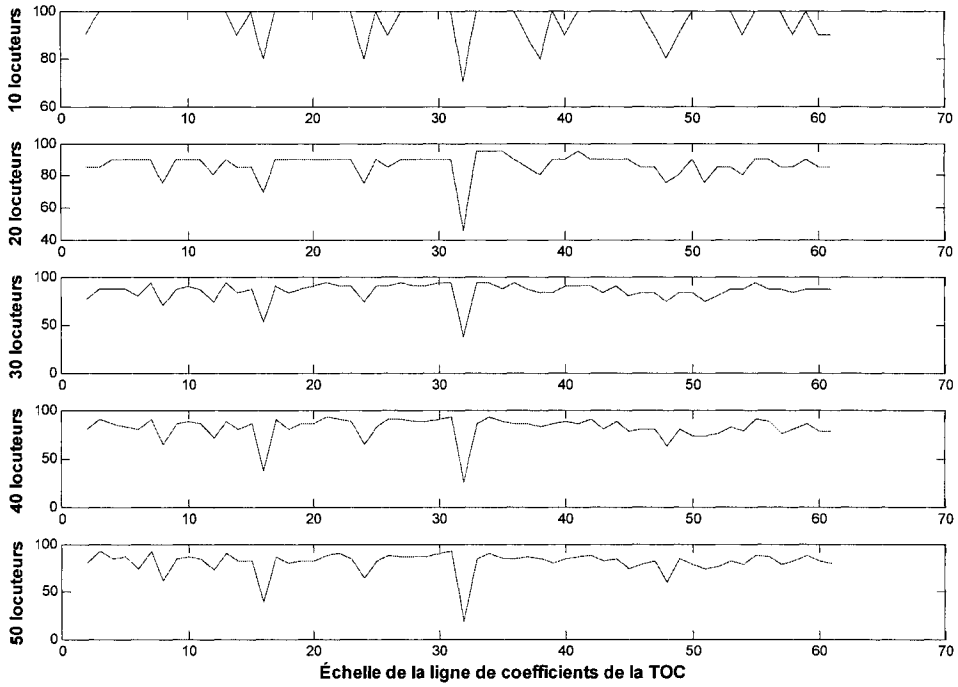


Figure 53 Performances des systèmes hybrides L en fonction des lignes de coefficients de la TOC et suivant la taille de la population

Le système d'identification hybride L permet d'améliorer les performances du système traditionnel en augmentant dans le meilleur des cas le taux de reconnaissance de 3.3 points pour une population de 30 locuteurs, soit un taux de 93.3%, et de 2 points pour celle de 50 locuteurs, équivalent à un taux de 92 %.

De plus d'autres tests viennent confirmer les résultats trouvés plus haut pour les systèmes de type G (voir tab. XVII) : en effet, nous avions vu dans les cas des grilles, que plus le pas b était petit, donc plus on avait de coefficients sur l'axe du temps, et meilleurs étaient les résultats.

Prenons par exemple la ligne de coefficients à l'échelle 17 pour une population de 30 locuteurs, et faisons varier son pas Δb avec des valeurs entières.

Tableau XVII

Performances des systèmes hybrides L pour une même ligne de coefficients de la TOC, mais avec un pas temporel variable (population de 30 locuteurs)

	Taux de reconnaissance du système (en %) pour 30 locuteurs					
Ligne pour l'échelle	**Système de référence (sans TOC)**	**Système hybride L (avec TOC)**				
		Pas des coefficients sur l'axe temporel				
		$\Delta b = 1$	$\Delta b = 2$	$\Delta b = 4$	$\Delta b = 6$	$\Delta b = 8$
17	90	90	70	50	33.3	3.3

Nous observons que plus il y a de points pour représenter le signal de parole analysé par TOC pour une échelle d'analyse donnée, et plus le taux de reconnaissance du système augmente. Par conséquent nous devons sélectionner les lignes dans leur intégralité afin d'être performant.

Si l'on considère maintenant la répartition de ces taux sous forme d'un histogramme (fig. 54), nous pouvons observer que le choix de la ligne n'est pas critique puisqu'il existe une très faible proportion de d'échelles donnant des taux de reconnaissance peu élevés quelle que soit la population considérée.

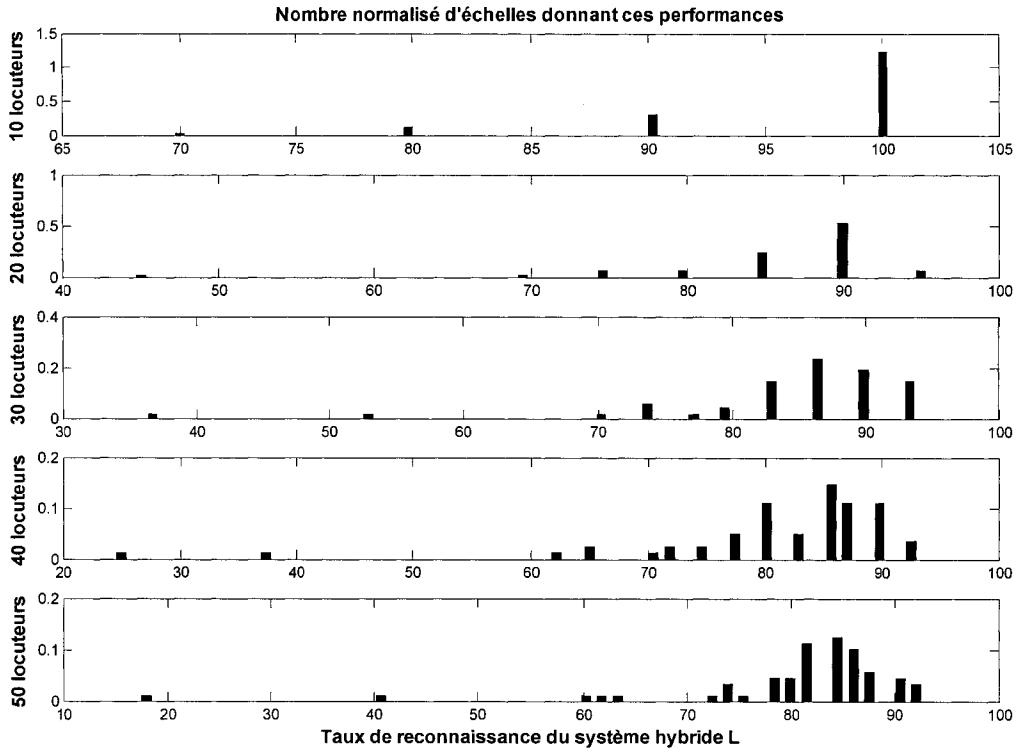


Figure 54 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 1 de 2 à 61, et pour une population de 10, 20, 30, 40, et 50 locuteurs

5.3.3 Conclusions sur le système hybride L

Les diverses conclusions que nous pouvons tirer de ces tests sont les suivantes :

- Il existe plusieurs lignes qui permettent d'aboutir à un taux de reconnaissance supérieur à celui obtenu par le système de référence (sans TOC).

- Pour obtenir un taux de reconnaissance supérieur, ces lignes doivent être considérées dans leur intégralité.

- En observant les résultats reportés à l'annexe 6, on peut définir la zone des coefficients de la TOC calculés pour la bande d'échelles allant de 25 à 36, comme une zone de lignes à fort taux de reconnaissance étant donné que la majorité de ces lignes donnent des résultats supérieurs ou égaux à 90 % (pour 30 locuteurs) et que le taux d'erreur (IER) minimum est obtenu dans cette zone.

- Il n'existe pas de relation apparente entre les échelles des lignes qui donnent de bons résultats et celles qui en donnent des mauvais. Une « bonne » ligne et une « mauvaise » ligne peuvent être côte à côte.

- L'échelle qui permet d'obtenir le meilleur taux de reconnaissance pour une population donnée, est propre à cette même population. Celle-ci diffère selon les individus considérés.

Les séries de test et de modifications précédentes nous ont permit d'aboutir à un système de reconnaissance hybride amélioré et performant vis-à-vis de son homologue basé sur des techniques standard.

Le système hybride L, dont la sélection des coefficients de la TOC s'effectue en prenant une ligne complète de coefficients pour une échelle pré-définie, est le système le plus performant que nous proposons, tant en matière d'identification comme en temps de calcul. Le seul inconvénient reste comment choisir les échelles. Nous nous pencherons sur ce sujet plus loin dans ce chapitre lorsque nous mettrons en place une méthode automatique de reconnaissance du locuteur (ASR, Automatic Speaker Recognition).

Nous venons donc de voir que la transformée en ondelettes continue permet d'améliorer la reconnaissance du locuteur. Maintenant, est-t-il possible d'améliorer d'avantage les performances du système L en diminuant le pas a pour la recherche de la meilleure ligne de coefficients? En effet, jusqu'à présent, nous avons uniquement considéré les

échelles à valeur entières de la TOC, et nous avons procédé avec un pas constant et unitaire. Cependant, il existe une infinité d'échelles possibles étant donné que la transformée est continue. Dans la partie qui suit, nous allons procéder de la même manière pour vérifier s'il existe des lignes de coefficients à des échelles non entières qui délivrent un meilleur taux d'identification.

5.4 Amélioration des performances du système hybride L par un raffinement d'échelles

Cette phase de tests consiste en fait en un « raffinement » des échelles de la TOC, dans le but de vérifier s'il est encore possible d'obtenir un meilleur taux de reconnaissance.

5.4.1 Sélection des coefficients de la TOC : premier raffinement des échelles

Nous considérerons maintenant les échelles à un chiffre significatif après la virgule. Bien entendu, il existe bien trop de choix à tester ici si on souhaite procéder à nouveau par essais-erreur. C'est pourquoi nous devons focaliser nos recherches dans une zone bien distincte.

Nous avions défini précédemment une zone de lignes à fort taux de reconnaissance. Cette zone commence à l'échelle 25 et se termine à l'échelle 36. Nous pouvons émettre l'hypothèse que s'il existe des lignes à meilleur rendement, celles-ci se trouvent dans cette zone très performante. Les résultats précédents nous ont montré que ces lignes ne se trouvent pas nécessairement dans un voisinage proche ou immédiat; en effet nous avons vu que deux lignes à performances opposées peuvent être voisines. Cependant c'est tout de même dans cette zone où l'on a le plus de probabilités de les détecter. Par conséquent, il y a une centaine de nouvelles lignes à exploiter si l'on prend un pas de a égal à 0.1.

5.4.2 Performances

Nous avons regroupé ci-dessous dans le tableau XVIII quelques résultats intéressants. La totalité des résultats de l'expérience est regroupée dans l'annexe 7. Les tests ont été réalisés sur des populations de 30 et de 50 locuteurs.

Tableau XVIII

Performances des systèmes hybrides L améliorés, pour différentes lignes (à 1 chiffre significatif après la virgule) de coefficients de la TOC

	Taux de reconnaissance du système (en %)								
Base de données	**Syst. de réf.**	**Système hybride L (avec TOC)**							
		Ligne de coefficients pour l'échelle							
		a=25.1	**a=25.2**	**a=31**	**a=32.7**	**a=32.5**	**a=34.9**	**a=35.4**	**a=36**
30	90	90	93.3	93.3	93.3	93.3	96.7	96.7	93.3
50	90	92	90	92	94	92	88	88	84

De plus, avec la totalité des résultats reportés à l'annexe 7, nous avons tracé le graphique ci-dessous (fig. 55) montrant l'ensemble des taux de reconnaissance obtenus en fonction des échelles des lignes, et selon la taille de la population:

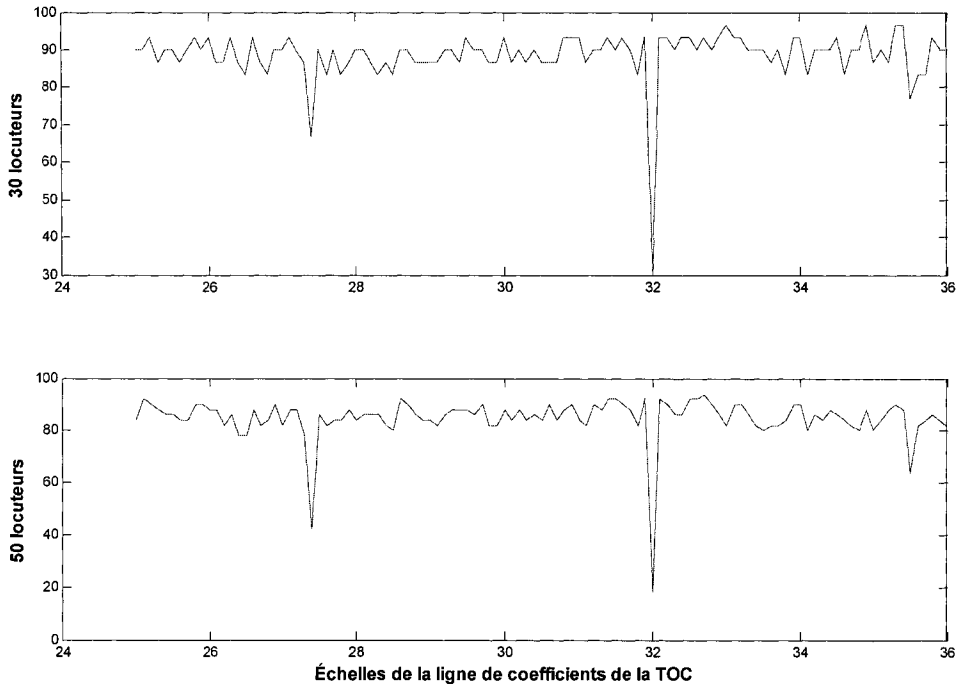


Figure 55 Performances des systèmes hybrides L améliorés, en fonction des lignes (à 1 chiffre significatif après la virgule) de coefficients de la TOC et suivant la taille de la population

L'ensemble des résultats nous permet d'aboutir à des performances encore meilleures (par exemple, cases du tableau XVIII ombragées) en matière de reconnaissance en utilisant la TOC. Ceci vérifie les hypothèses que nous avions posées.

Le système d'identification hybride L basé sur une sélection de lignes à un chiffre significatif après la virgule, permet de faire croître d'avantage les performances du système traditionnel dans le cadre d'une population formée par 30 ou de 50 locuteurs. Le taux de reconnaissance est amélioré de 3.3 points et de 2 points respectivement, atteignant ainsi les 96.7 % et 94 % de reconnaissance.

De plus, les essais montrent qu'en diminuant le pas de l'échelle de la TOC, on peut trouver davantage de lignes à fort taux d'identification. On observe effectivement qu'il existe une grande quantité de lignes fournissant de bonnes performances (fig. 56). Ceci est très encourageant puisque cela nous permet de voir qu'il n'existe pas une seule configuration précise dans laquelle doit fonctionner notre système.

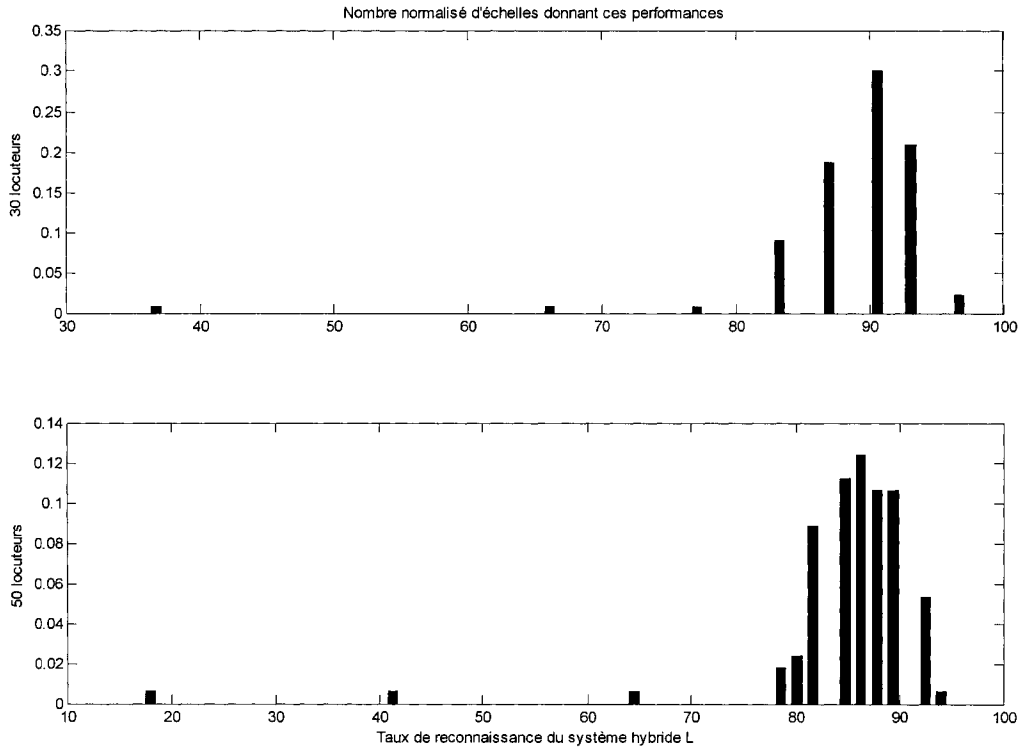


Figure 56 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.1 de 25 à 36, et pour une population de 30, et 50 locuteurs

5.4.3 Conclusions sur le premier raffinement des échelles

Les diverses conclusions que nous pouvons tirer de ces tests sont les suivantes :

- Même en diminuant le pas Δa de l'échelle de la TOC, et en observant les voisinages des meilleures lignes que nous avions trouvées pour des échelles à valeur entière, on remarque que les lignes de coefficients permettant d'obtenir les meilleurs taux de reconnaissance ne sont pas nécessairement voisines et que de « bonnes » et de « mauvaises » lignes peuvent se trouver côte à côte. Elles dépendent essentiellement du signal de parole et de ses caractéristiques.

- Plus on « raffine » les échelles de la TOC, et plus on trouve de « bonnes » lignes à faible taux d'erreur d'identification.

- Il peu exister une ligne de coefficients dont l'échelle n'est pas entière, et qui permet d'atteindre un taux de reconnaissance supérieur à celui obtenu pour une échelle à valeur entière. Cette ligne à de très fortes probabilités de se trouver dans une zone d'échelles à valeur entière où se trouvent la plupart des lignes qui donnent les plus grandes performances.

Ceci nous amène à émettre l'hypothèse qu'il existe presque toujours une ligne de coefficients, se trouvant à une échelle inconnue, et qui permet d'aboutir au meilleur taux de reconnaissance. On pourrait penser que plus on diminue le pas de la TOC, et plus on peut trouver de lignes performantes, et plus grande est la probabilité de trouver une ligne à taux maximal. Si cette hypothèse était vérifiée, on trouverait donc de meilleurs résultats avec un « raffinement » supplémentaire (c'est-à-dire une diminution supplémentaire du pas de l'échelle d'analyse).

Essayons donc de voir s'il est encore possible d'améliorer les résultats précédents.

5.4.4 Nouvelle sélection des coefficients de la TOC : second raffinement des échelles

Raffinons à nouveau nos recherches sur les meilleures lignes de coefficients en diminuant le pas de l'échelle d'analyse à 0.01. Nous considérerons donc des échelles à deux chiffres significatifs après la virgule. Étant donné que l'étendu du choix des échelles s'agrandi à chaque fois que nous réduisons le pas de la TOC, nous focaliserons une fois de plus nos recherches dans des zones de lignes à fort taux de reconnaissance.

Nous avions défini précédemment une zone de lignes à fort taux de reconnaissance pour les bases de données de 30 et de 50 locuteurs: échelles 30.6 à 35.4. C'est une région qui permettait d'obtenir en majorité des taux de 93.3% (pour 30 locuteurs) et de 92 % (pour 50 locuteurs). Pour chaque population, nous avons trouvé des résultats maximums dans des sous-régions de cette dernière, à savoir entre l'échelle 34.5 et l'échelle 35.5 pour 30 locuteurs et entre l'échelle 32.1 et l'échelle 32.8 pour 50 locuteurs. Nous adopterons ces deux derniers sous-ensembles.

Nous pouvons à nouveau penser que s'il existe des lignes à meilleur rendement, celles-ci se trouvent dans cette zone très voisine à celles-ci. Par conséquent, nous exploiterons ces nouvelles lignes avec un pas de raffinement de 0.01.

5.4.5 Performances

Seule une partie des résultats est regroupée ci-dessous dans les tableaux XIX et XX. La totalité des résultats de l'expérience est donnée dans les annexes 8 et 9. Les tests ont été accomplis pour 30 et 50 locuteurs.

Tableau XIX

Performances des systèmes hybrides L améliorés, pour différentes lignes (à 2 chiffres significatifs après la virgule) de coefficients de la TOC dans la zone d'analyse définie pour 30 locuteurs

	Taux de reconnaissance du système (en %)							
	Système de référence	**Système hybride L (avec TOC)**						
Base de données		**Ligne de coefficients pour l'échelle**						
		33.47	**33.48**	**34.81**	**35.18**	**35.29**	**35.31**	**35.32**
30	90	96.7	96.7	96.7	96.7	96.7	90	86.7
50	90	96	96	88	86	90	82	84

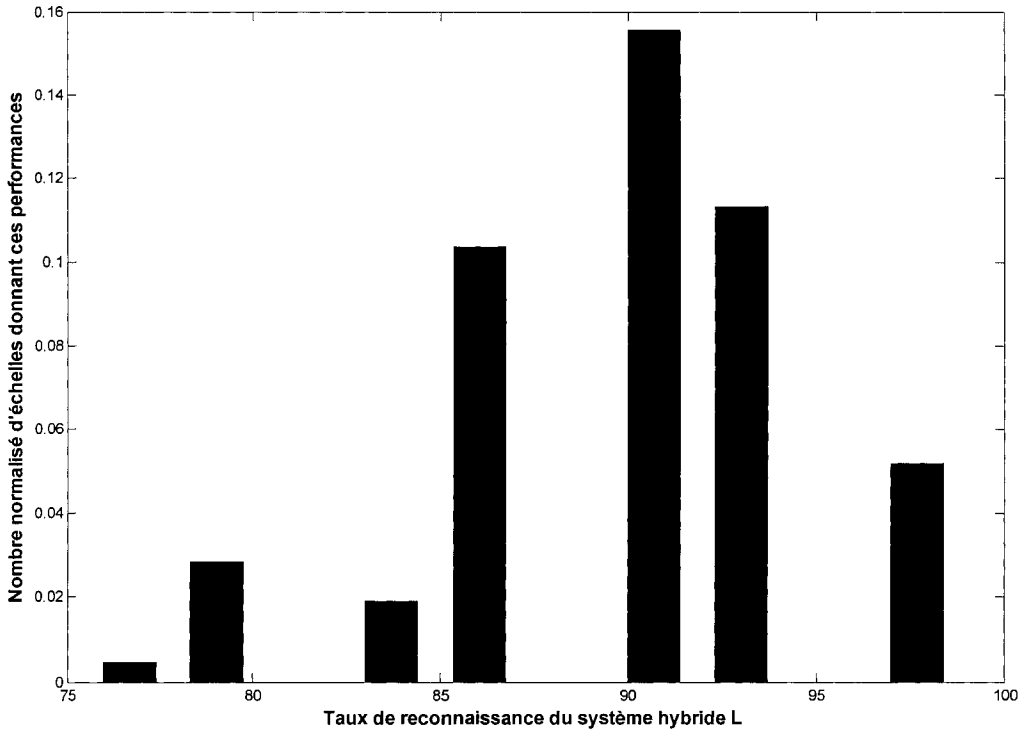


Figure 57 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.01 de 34.5 à 35.5, et pour une population de 30 locuteurs

Tableau XX

Performances des systèmes hybrides L améliorés, pour différentes lignes (à 2 chiffres significatifs après la virgule) de coefficients de la TOC dans la zone d'analyse définie pour 50 locuteurs

	Taux de reconnaissance du système (en %)							
	Système de référence	**Système hybride L (avec TOC)**						
Base de données (population)		**Ligne de coefficients pour l'échelle**						
		32.12	**32.48**	**32.52**	**32.57**	**32.58**	**32.61**	**32.63**
50	90	90	88	90	90	92	92	92

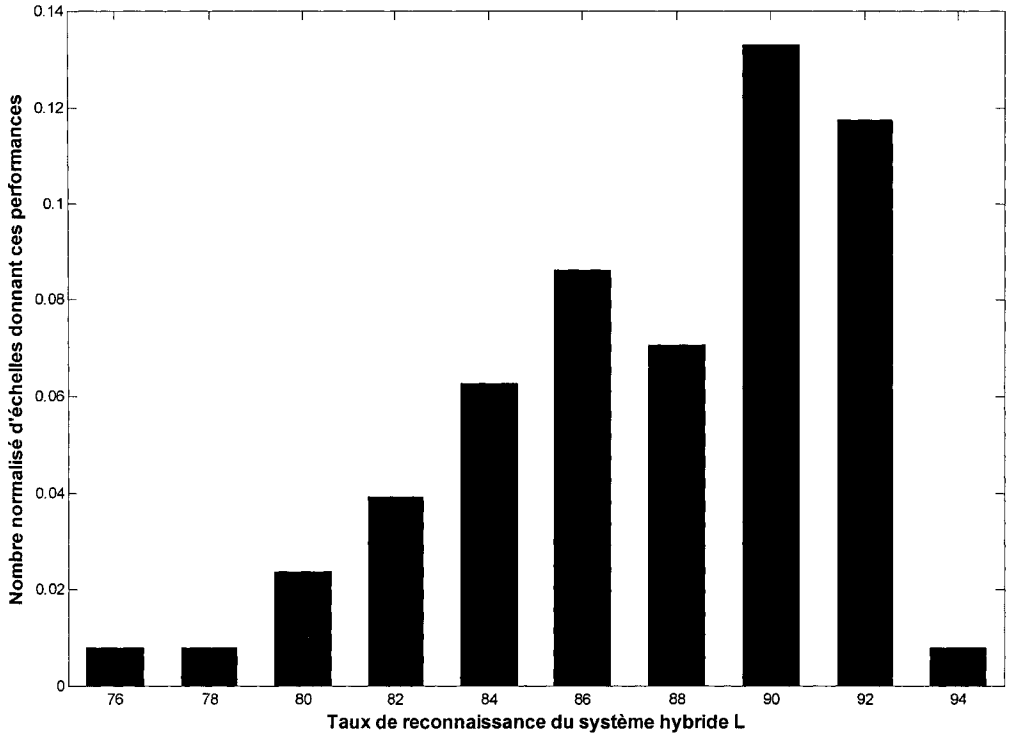


Figure 58 Répartition des taux de reconnaissance du système hybride L selon la ligne de coefficients de la TOC pour des échelles variant par pas de 0.01 de 32.1 à 32.8, et pour une population de 50 locuteurs

Nous observons que l'on ne peut plus, à priori, améliorer le taux de reconnaissance dans le cas d'une population de 30 locuteurs. Cependant, pour une population de 50 locuteurs, nous réussissons à augmenter le taux de reconnaissance à 96 %, soit 2 points de plus par rapport à la dernière amélioration. Ces résultats sont très bons et très encourageants; ils permettent de vérifier l'hypothèse selon laquelle on peut encore améliorer le taux d'identification du système en raffinant les échelles des lignes, c'est-à-dire, en diminuant le pas de la TOC. Cependant, ils nous montrent aussi que pour trouver de meilleurs résultats que des précédents, on ne les trouvera pas forcément dans un voisinage proche à des lignes très performantes ou bien dans une zone regroupant de nombreuses lignes à forts taux de reconnaissance. En effet, dans le cas de 50 locuteurs, on ne trouve pas le nouveau taux maximal dans la zone étudiée, mais dans ne zone qui ne donnait pas de performances très élevées lors des recherches précédentes (voir annexe 7).

5.4.6 Conclusions sur le second raffinement des échelles

Nous venons de montrer qu'il est très difficile de déterminer une méthode qui nous permettrait de trouver l'échelle de la TOC qui donnerait les meilleures performances. La difficulté se trouve dans le fait que les échelles donnant de bonnes performances ne sont pas nécessairement voisines. Nous avons pu voir cependant qu'en « raffinant » les recherches en diminuant progressivement le pas de l'échelle de la TOC, nous trouvons de plus en plus de lignes à fortes performances (performances supérieures à celles du système de référence). En résumé, si on considère un certain nombre d'échelles consécutives, plus on choisit un pas petit, et plus on trouvera de lignes de coefficients de la TOC performantes.

5.5 Vérification des tests

Selon John G. Harris, il est courant de rapporter deux scores de reconnaissance lorsqu'on utilise des classificateurs paramétriques [110]: on mentionne le taux de

reconnaissance sur les données de test, mais également le taux de reconnaissance sur les données d'entraînement. Ce deuxième score est calculé pour confirmer le bon fonctionnement du système testé. John G. Harris explique que le taux de reconnaissance pour les données d'entraînement doit être supérieur à celui obtenu avec des données inconnues pour la phase de test puisque les paramètres du modèle GMM sont estimés à partir de ces données. Voici les résultats de nos vérifications (tab. XXI) :

Tableau XXI

Vérification du système L avec les données d'entraînement

	Taux de reconnaissance du système (en %)			
	Système hybride L (avec TOC)			
	Base de 30 locuteurs		**Base de 50 locuteurs**	
	Ligne de coefficients pour l'échelle			
	35.4	**35.5**	**33.48**	**35.31**
Sur les données d'entraînement	100 %	100 %	100 %	100 %
Sur les données de test	96.7 %	76.7 %	96 %	82 %

Les résultats obtenus lorsqu'on test deux lignes de coefficients dont les performances sont très différentes, sont maximums. On obtient des taux de 100 %, ce qui confirme la fiabilité de notre système hybride. De plus, John G. Harris ajoute que la différence entre les deux scores est un indicateur de généralisation de notre classificateur ; étant donné la différence minime qui existe, on peut conclure que le module de modélisation par GMM est bien généralisé.

5.6 Conclusions

Les essais expérimentaux que nous venons d'effectuer nous ont permit d'aboutir à un système hybride, créé à partir d'un système de reconnaissance de référence, et faisant usage de la TOC, et qui permet de surpasser les performances de ce dernier. Les résultats que nous obtenons sont très satisfaisants étant donné que l'on augmente le taux de reconnaissance de 3.3 points pour une population de 30 locuteurs (93.3 %), et de 2 points pour une population de 50 locuteurs (92 %). De plus, il est encore possible d'améliorer ce taux en « raffinant » les échelles d'analyse de la TOC. En effet, en choisissant une bande de lignes de la TOC à forts taux d'identification, et en faisant fonctionner notre système sur toutes les échelles avec un pas de 0.1 ou de 0.01, nous avons obtenus un taux de 96.7 % pour une population de 30 locuteurs, et 96 % pour 50 locuteurs, soit des gains de 6.7 et 6 points par rapport aux performances du système de référence.

Ce système hybride performant est fondé sur une technique de sélection des coefficients de la TOC innovante et simple : on choisit l'échelle qui correspond à la ligne qui permet d'obtenir les meilleurs résultats. Cette ligne de coefficients est ensuite vue par le système comme un nouveau signal spécifique au locuteur, et on extrait alors les coefficients MFCC pour la modélisation GMM.

Il est bon à présent de nous intéresser à un point très important afin de mettre en place une méthode de reconnaissance automatique (ASR):

Tout d'abord, une question nous vient naturellement à l'esprit : que possèdent de plus ces nouveaux signaux que le signal de parole initial, et qui fait en sorte de donner un meilleur taux de reconnaissance? Dans le même style, nous aimerions aussi savoir ce qu'ont de particulier ces lignes de coefficients performantes par rapport à d'autres lignes

moins efficaces. Ce point crucial se résume en fait à comment choisir l'échelle d'analyse de la TOC.

Pour répondre à cette question, nous allons étudier ces lignes de coefficients dans le dernier chapitre, et analyser ce qu'elles ont de si particulier.

CHAPITRE 6

CHOIX DE L'ÉCHELLE D'ANALYSE DE LA TOC POUR LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR

6.1 Introduction

Nous venons de voir que certaines lignes de coefficients de la TOC permettent d'obtenir de bien meilleurs taux de reconnaissances que d'autres. Nous allons donc étudier ces lignes dans ce chapitre afin de déterminer ce qu'elles ont de particulier, et comment s'exprime leur pouvoir discriminatoire. Ceci permettra de justifier le choix de l'échelle pour laquelle nous devons calculer la TOC du signal de parole, et on pourra alors mettre au point une technique de reconnaissance automatique du locuteur.

6.2 Étude des lignes de coefficients de la TOC

Cette étude sera menée dans le cadre d'une population de 30 locuteurs afin d'écourter les temps de calculs.

6.2.1 Analyses statistiques du premier ordre

Afin de déceler certaines particularités statistiques de ces lignes à fort taux de reconnaissance, étudions leur moyenne, médiane, déviation standard, et la répartition des valeurs des coefficients dans le cadre de populations de 30 locuteurs.

Pour cela, nous avons considéré plusieurs lignes dont les performances sont très diverses. Puis nous avons déterminé les statistiques moyennes de ces vecteurs de

données pour une population entière afin de les comparer les uns aux autres et voir ainsi quelles sont leurs particularités (tab. XXII).

Tableau XXII

Statistiques moyennes pour diverses lignes de coefficients pour une population de 30 locuteurs

	Statistiques moyennes des lignes de coefficients pour une population de 30 locuteurs						
	Système de référence (90 %)	**Système hybride L (avec TOC et avec MFCC)**					
		Ligne à 96.7 %	**Ligne à 93.3 %**	**Ligne à 73.3 %**	**Ligne à 56.7 %**	**Ligne à 53.3 %**	
		a=34.81	**a=31**	**a=12**	**a=93**	**a=16**	**a=96**
Moyenne	$1.83.e^{-4}$	$8.63.e^{-7}$	$8.23.e^{-8}$	$-1.45.e^{-6}$	$3.78.e^{-7}$	$1.26.e^{-6}$	$2.70.e^{-7}$
Médiane	$7.41.e^{-4}$	$8.96.e^{-5}$	$1.27.e^{-4}$	$2.54.e^{-4}$	$-3.53.e^{-6}$	$-1.17.e^{-5}$	$4.69.e^{-6}$
Dév. Stand.	$1.28.e^{-1}$	$7.59.e^{-2}$	$7.14.e^{-2}$	$2.53.e^{-1}$	$2.58.e^{-2}$	$3.15.e^{-1}$	$2.88.e^{-2}$
Max	1	0.521	0.409	2.287	0.448	2.071	0.475
Min	-1	-0.503	-0.409	-2.255	-0.394	-2.15	-0.453

Les lignes de coefficients obtenues aux échelles 16 et 96 qui toutes deux permettent de trouver un taux de reconnaissance de 53.3 %, montrent qu'il n'y a pas une valeur statistique en particulier qui pourrait être à l'origine d'une bonne ligne ou pas. Il pourrait cependant exister une combinaison précise de statistiques, cependant ceci serait trop fastidieux à vérifier.

L'analyse graphique des histogrammes simple et cumulatif des signaux ne permet pas non plus de déceler le moindre indice. Globalement, ils sont ont la même allure que ceux du signal de parole traité par le système traditionnel, comme représenté ci-dessous (fig. 59) :

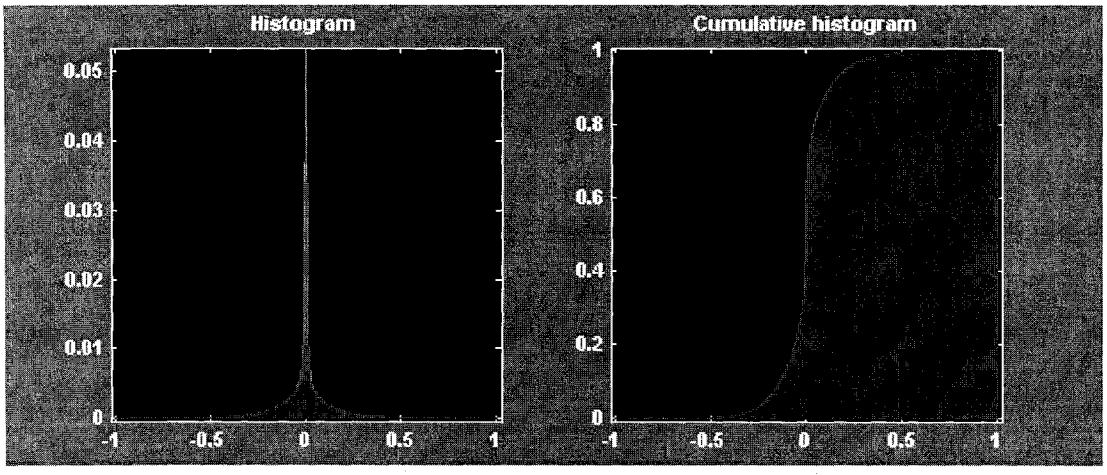


Figure 59 Histogrammes simple et cumulatif des lignes de coefficients de la TOC

Voyons maintenant ce que pourrait nous apporter une analyse énergétique.

6.2.2 Analyses énergétiques

Nous avons procédé de la même manière que précédemment pour les analyses énergétiques.

Tableau XXIII

Énergie moyenne pour diverses lignes de coefficients pour une population de 30 locuteurs

	Énergie moyenne des lignes de coefficients pour une population de 30 locuteurs						
	Système de référence (90 %)	**Système hybride L (avec TOC)**					
		Ligne à 96.7 %	**Ligne à 93.3 %**	**Ligne à 73.3 %**	**Ligne à 56.7 %**	**Ligne à 53.3 %**	
		a=34.81	**a=31**	**a=12**	**a=93**	**a=16**	**a=96**
Énergie	1932.3	677.4	599.6	7564.2	78.4	11685	97.8

On pourrait effectivement croire que l'énergie du signal serait responsable des bonnes performances du système. Cependant, l'étude des deux lignes de coefficients aux échelles 16 et 96 montre que le niveau d'énergie est indépendant des performances du système; tout du moins, lorsqu'on isole ce paramètre.

Nous avons également vérifié s'il existait une quelconque corrélation entre le signal de parole d'origine et les lignes de coefficients de la TOC, en calculant l'énergie de la corrélation des deux signaux. Nous en avons conclu qu'il n'existait aucune relation entre ces deux signaux.

Vérifions à présent si une autre marque d'activité vocale, le taux de passage par zéro, serait à l'origine de bons résultats.

6.2.3 Analyse du taux de passage par zéro

En analysant le taux de passage par zéro, nous avons voulu vérifier si les signaux pris en compte étaient sujets à une forte activité vocale ou non.

Le calcul de la moyenne du taux de passage par zéro pour les locuteurs d'une même population a montré que non. À titre d'exemple, nous avons regroupé quelques résultats dans le tableau ci-dessous (tab. XXIV):

Tableau XXIV

Taux de passage par zéro moyen pour diverses lignes de coefficients pour une population de 30 locuteurs

	Taux de passage par zéro moyen des lignes de coefficients pour une population de 30 locuteurs				
	Système de référence (90 %)	**Système hybride L (avec TOC)**			
		Ligne à 96.7 %	**Ligne à 76.7 %**	**Ligne à 53.3 %**	
		a=35.4	**a=35.5**	**a=16**	**a=96**
Taux de passage par zéro	11841	10620	10542	10635	9843

Nous pouvons observer qu'il n'existe pas de lien apparent entre le taux de reconnaissance et le taux de passage par zéro.

6.2.4 Conclusions sur l'étude des lignes de coefficients de la TOC

L'étude des lignes de coefficients de la TOC n'a pas été fructifiante et ne nous a pas permis de conclure sur le choix des échelles de la transformée. Penchons-nous maintenant une autre étude que nous avons réalisée.

6.3 Étude des coefficients MFCC extraits à partir des lignes de la TOC

Pour obtenir un système de reconnaissance capable de donner de bons résultats, il faut que les locuteurs concernés aient des modèles GMM aussi différentiables les uns des autres que possible. D'où l'importance d'une grande variabilité inter locuteurs (comme nous l'avons déjà évoqué au début du chapitre) dans les signaux de parole que nous considérons et à partir desquels on extrait les vecteurs de caractéristiques

Au cours de nos expérimentations, nous avons pu constater que la TOC du signal de parole d'origine, permet de mieux différentier les locuteurs et donc de mettre en évidence cette variabilité pour aboutir à un meilleur taux d'identification.

Dans l'étude précédente, nous n'avons pu déceler ce qui donne aux lignes de coefficients de la TOC un plus grand pouvoir discriminatoire. C'est pourquoi nous avons été contraints de prendre le problème à l'envers pour déterminer ce qui est à l'origine de ces performances.

Pour avoir de bons résultats, nous venons de parler de l'importance de la différentiation entre les modèles de locuteurs. Par conséquent, nous devons générer des modèles les plus fidèles et précis possibles pour éviter toute erreur d'identification. Le module de modélisation modélise chaque individu à partir des vecteurs de caractéristiques extraits à l'aide de la méthode des MFCC. Nous devons donc étudier le comportement des vecteurs paramétriques extraits à partir des lignes de coefficients de la TOC en fonction

des performances du système pour établir une corrélation, et mettre au point une méthode de sélection automatique des échelles.

Cette étude sera menée dans le cadre d'une population de 30 locuteurs afin d'écourter les temps de calculs.

6.3.1 Analyses statistiques du premier ordre

En suivant la même démarche que pour l'étude des lignes de coefficients de la TOC, nous étudierons l'histogramme, la moyenne, la médiane, et la déviation standard, des coefficients dans le cadre de populations de 30 locuteurs.
Pour cela, nous avons considéré ici deux lignes dont les performances sont opposées. Puis nous avons déterminé les statistiques moyennes de ces vecteurs de données pour une population entière afin de les comparer et voir ainsi quelles sont leurs particularités.

De plus, ces lignes sont choisies pour des échelles très proches l'une de l'autre afin d'obtenir un niveau d'énergie le plus proche possible, et mettre ainsi de côté ce paramètre pour le moment.

Tableau XXV

Statistiques moyennes pour les coefficients MFCC extraits à partir de diverses lignes pour une population de 30 locuteurs

	Statistiques moyennes des MFCC des lignes de coefficients pour une population de 30 locuteurs		
	Système de référence (90 %)	**Système hybride L (avec TOC)**	
		Ligne à 96.7 %	**Ligne à 76.7 %**
		a=35.4	**a=35.5**
Moyenne	-1.136	-1.082	-1.131
Médiane	-0.1776	0.346	0.441
Dév. Stand.	3.818	5.271	6.122
Max	3.333	3.407	4.164
Min	-20.71	-27.24	-31.09

Les conclusions que l'on peut apporter sont les suivantes :

Les lignes de coefficients obtenues aux échelles 35.4 et 35.5 qui permettent de trouver respectivement un taux de reconnaissance de 96.7 % et 76.7 %, montrent qu'il n'y a toujours pas de valeur statistique en particulier qui pourrait être à l'origine d'une bonne ligne ou pas.

Nous avons également procédé à de nouvelles analyses du taux de passage par zéro, mais aucune expérimentation ne fut concluante. Effectuons à présent une analyse graphique de la répartition des coefficients MFCC.

6.3.2 Analyses graphiques des histogrammes

L'analyse graphique des histogrammes simple et cumulatif des coefficients cepstraux calculés à partir d'une ligne de coefficients de la TOC, permet de déceler indice primordial. En effet, en comparant les histogrammes d'une même personne pour deux lignes bien distinctes et ce, pour chaque locuteur, on remarque globalement qu'ils sont très différents en certains points :

Pour la ligne à l'échelle 35.4 qui donne un de taux de reconnaissance de 96.7 %:

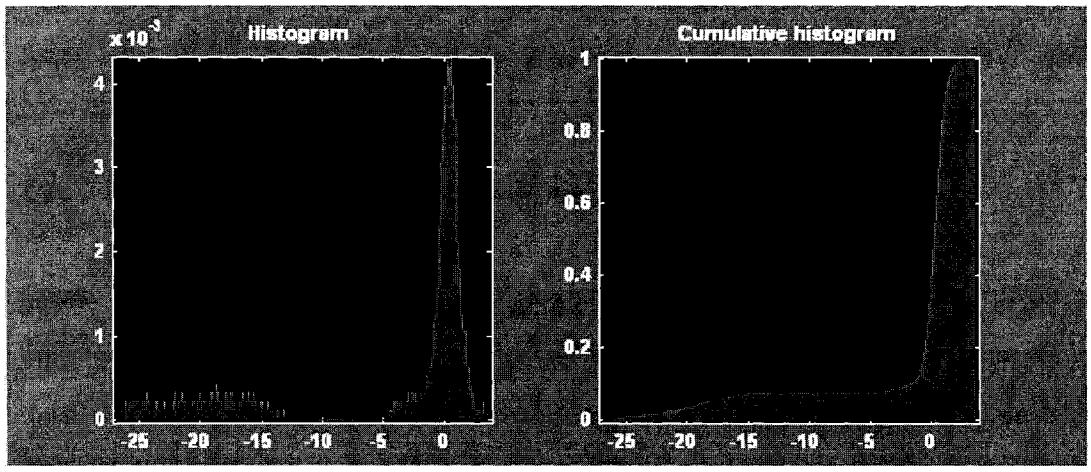


Figure 60 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 35.4

Pour la ligne à l'échelle 35.5 qui donne un taux de reconnaissance de 76.7 %:

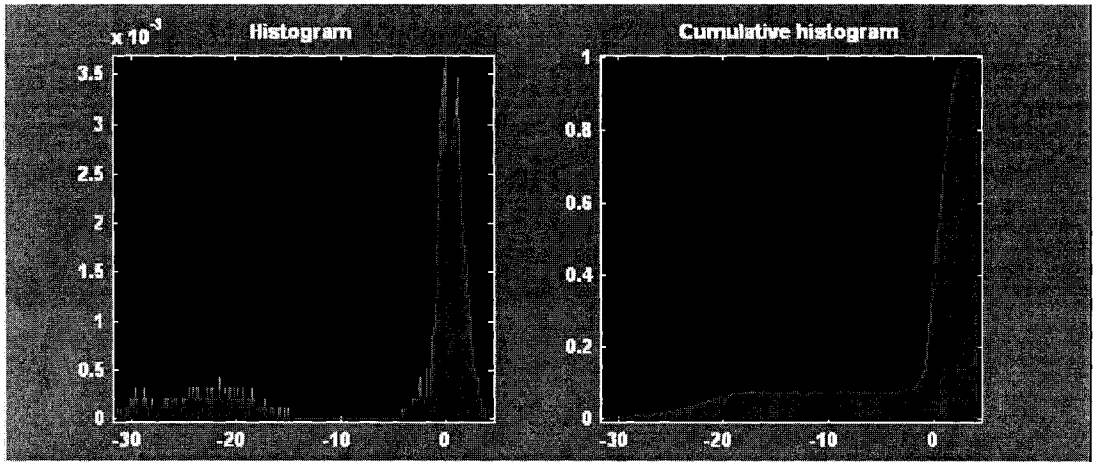


Figure 61 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 35.5

En effet, on observe que les deux répartitions ont approximativement le même niveau d'énergie, mais par contre la distribution des données présente une forme moins compacte, moins régulière et moins lisse dans le cas de la ligne pour l'échelle 35.5 (dont les performances sont faibles). Comparons maintenant ces résultats à ceux donnés par le système de référence.

Pour le signal de parole de base qui donne un taux de reconnaissance de 90 %:

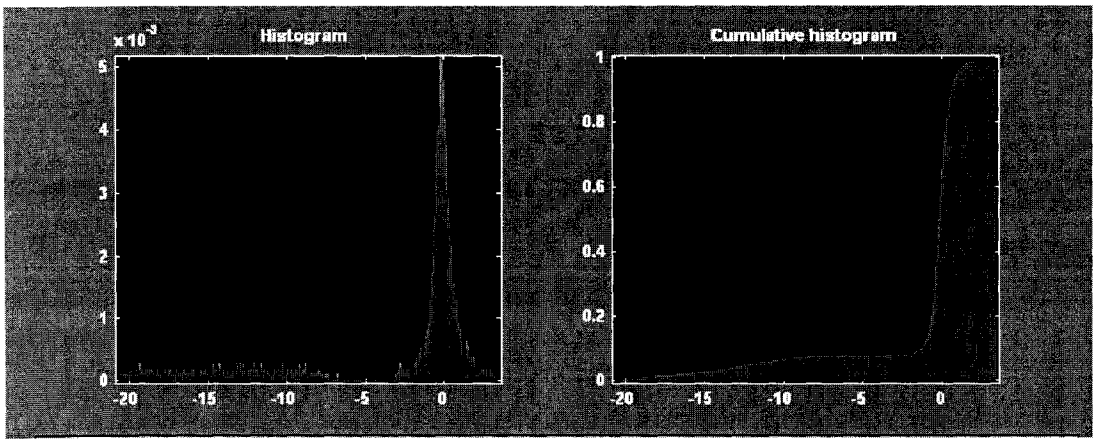


Figure 62 Histogrammes de la répartition des coefficients MFCC extraits à partir du signal de parole d'origine (celui du système de référence)

On observe que dans le cas du système de reconnaissance de référence, nous avons des histogrammes très proches de ceux obtenus pour la ligne d'échelle 35.4. Cependant, on relève ici un niveau d'énergie nettement supérieur. Donc le niveau d'énergie semble jouer un rôle important, tout comme la régularité de la distribution. Vérifions ces hypothèses avec des observations supplémentaires :

Pour la ligne à l'échelle 16 qui donne un taux de reconnaissance de 53.3 %:

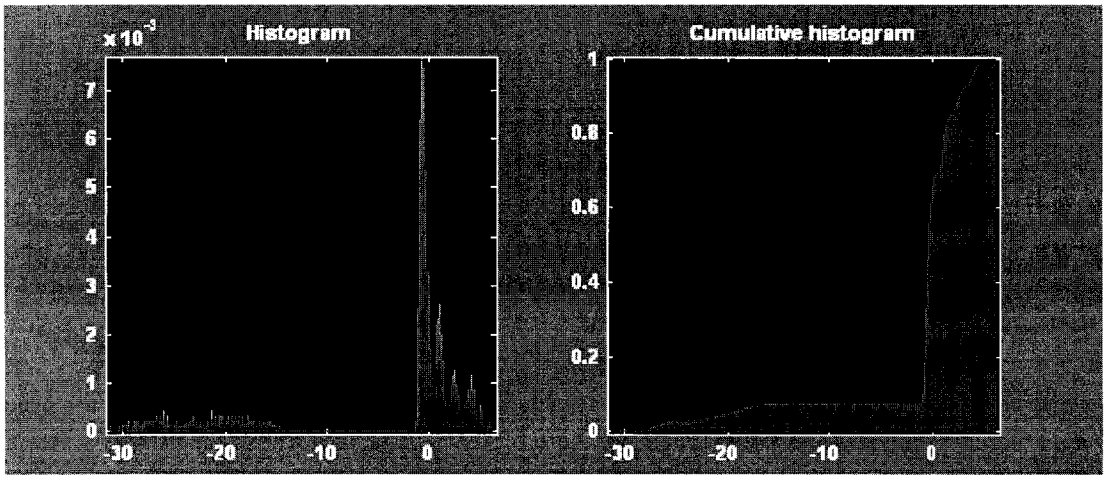


Figure 63 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 16

Pour la ligne à l'échelle 96 qui donne un taux de reconnaissance de 53.3 %:

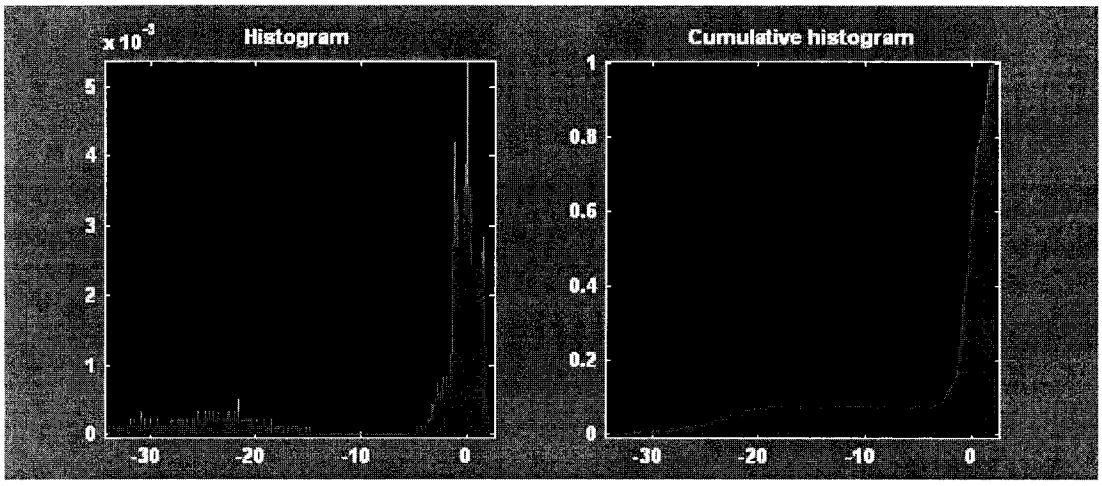
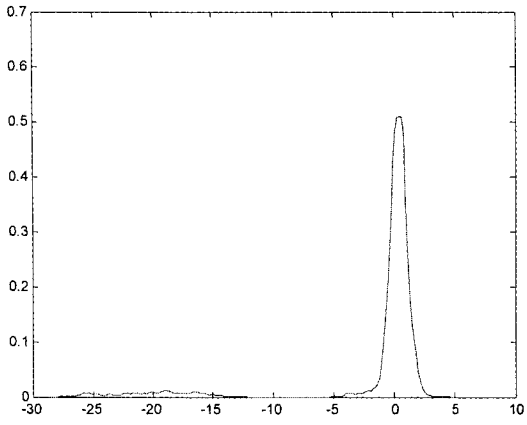


Figure 64 Histogrammes de la répartition des coefficients MFCC extraits à partir de la ligne de coefficients de la TOC d'échelle 96

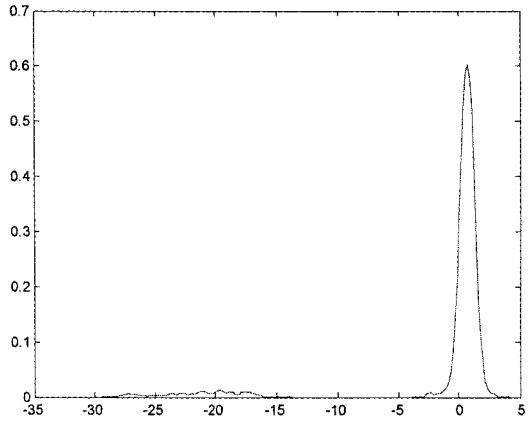
Les histogrammes cumulatifs montrent que lorsque l'énergie est trop importante, elle dénature les performances du système. Les histogrammes simples montrent quant à eux que le niveau d'énergie n'est pas le seul paramètre à intervenir. L'entropie des données pourrait également influencer l'allure de ces répartitions : trop d'irrégularités pourraient détériorer les performances du système.

La méthode de l'histogramme pour estimer la densité des données, est une méthode simple et intuitive, mais ce n'est pas la meilleure : la densité générée est très irrégulière et ne donne pas de représentation continue.

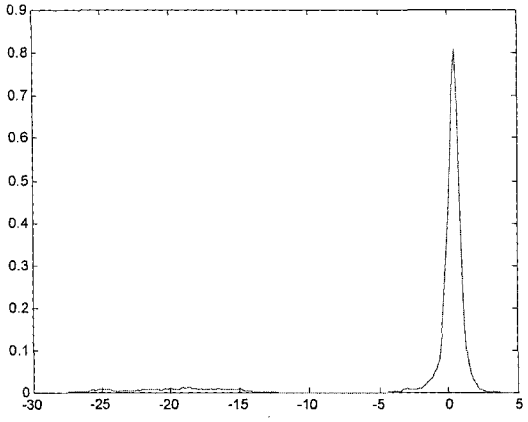
Une approche plus générale qui génère des estimations de densité plus lisses est appelée l'estimateur de densité de Kernel (ou en anglais, KDE pour Kernel Density Estimator) (voir annexe 10). En représentant cette densité, il sera plus facile de confirmer les observations précédentes :



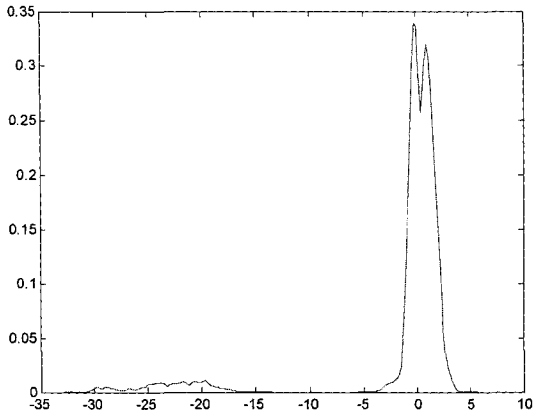
Pour la ligne à l'échelle 35.4 qui donne 96.7 %



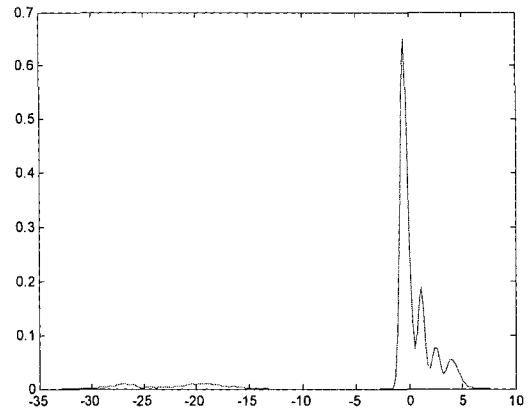
Pour la ligne à l'échelle 31 qui donne 93.3 %



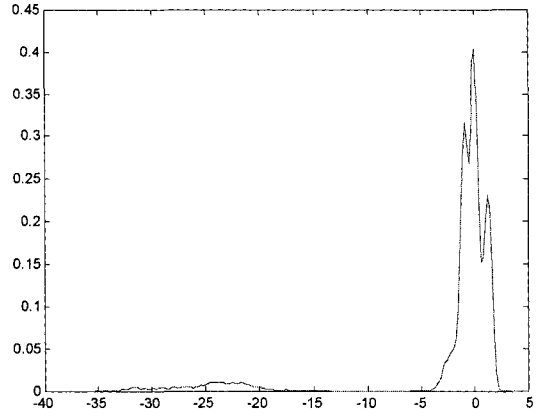
Pour la ligne à l'échelle 30 qui donne 93.3 %



Pour la ligne à l'échelle 48 qui donne 73.3 %



Pour la ligne à l'échelle 16 qui donne 53.3 %



Pour la ligne à l'échelle 96 qui donne 53.3 %

Figure 65 Représentations des densités de Kernel des coefficients MFCC extraits à partir de différentes lignes de coefficients de la TOC

6.3.3 Hypothèse

Lors de la phase d'apprentissage du modèle du locuteur, on estime les paramètres de chaque GMM à partir des données d'entraînement. Pour chaque locuteur, lors de l'estimation, l'algorithme calcule à chaque itération le maximum de vraisemblance entre

le modèle et les données. Le vraisemblance du modèle croît logarithmiquement jusqu'à atteindre sa valeur maximale, lorsque l'erreur entre les deux dernières valeurs estimées sera inférieure à 0.1 %. A ce moment là, on sera en présence du modèle GMM le plus représentatif des données d'entraînement. Ce sera donc le modèle qui épousera au mieux la répartition des vecteurs de caractéristiques du locuteur, c'est-à-dire, les vecteurs de coefficients cepstraux MFCC.

L'hypothèse que nous posons à partir de ces analyses graphiques est la suivante :
Lorsque la distribution des coefficients cepstraux de chaque locuteur est la plus simple possible (donc la plus lisse, régulière, et compacte), alors elle est plus facilement modélisable, et elle permettra donc de commettre moins d'erreurs d'identifications étant donné que le GMM paramétré sera plus précis. On aboutira donc au meilleur taux de reconnaissance.

Il existe donc plusieurs paramètres théoriques qui interviennent dans la mise en place d'une bonne distribution des coefficients cepstraux : l'entropie qui permettra de vérifier si la distribution est régulière ou bien très morcelée; ou encore l'énergie qui peut indiquer si la distribution est compacte ou pas.

À présent, nous nous intéresserons à chacun de ces paramètres afin d'étudier leur comportement en fonction des variations du taux de reconnaissance du système.

6.3.3.1 Analyse de l'énergie

L'étude de ce paramètre permettra de mieux caractériser la distribution des coefficients cepstraux.

Dans les analyses énergétiques que nous avons réalisées ici, nous avons calculé pour chaque ligne de coefficients de la TOC, la somme des énergies de chaque locuteur de la

population. Ceci a pour but de nous montrer le rôle global de l'énergie pour l'ensemble des locuteurs, sur les performances du système de reconnaissance.

Nous avons rapporté les résultats dans un graphique (fig. 66) afin de visualiser simultanément le taux de reconnaissance et la somme des énergies en fonction de la ligne de coefficients choisie. Cela nous permet de voir directement les effets de ce paramètre sur les performances du système.

Sur ces graphiques, nous observons des similitudes concernant les lignes qui donnent de mauvais résultats. En effet, pour les échelles 8, 16, 32, ou 64 par exemple, on remarque que la somme énergies est très élevée par rapport à la moyenne.

On observe également que les bonnes performances ($\geq 90\%$) sont obtenues pour des niveaux d'énergie pour la population inférieurs à 8.10^6. Nous voyons clairement que pour des valeurs supérieures, le taux de reconnaissance chute sous les 80 %.

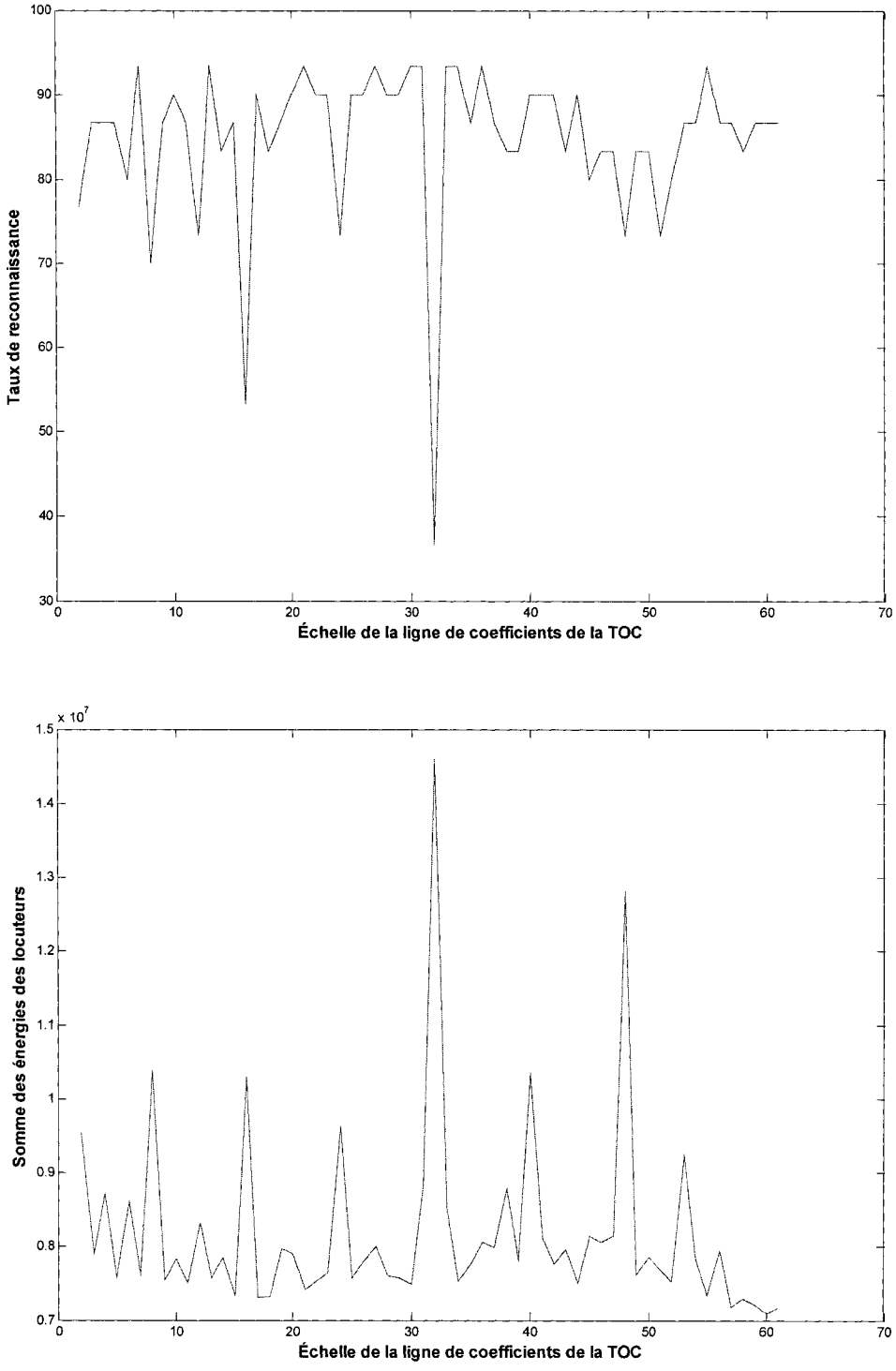


Figure 66 Effets de l'énergie des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybrid L

En conclusion, l'analyse énergétique ci-dessus nous a permit de caractériser d'avantage les distributions des coefficients cepstraux des locuteurs d'une même population, et d'appuyer les analyses graphiques précédentes ainsi que l'hypothèse que nous en avions tirés. Une distribution possédant trop d'énergie fait chuter les performances du système. Cependant, ce paramètre n'est pas le seul à intervenir. Intéressons-nous à présent à l'entropie de ces lignes.

6.3.3.2 Analyse de l'entropie

L'entropie est un outil fondamental lorsque l'on souhaite mesurer le contenu en information d'une distribution de probabilité donnée [111], [115]. La mesure de l'entropie peut de plus être adaptée au traitement du signal pour quantifier l'information contenue dans les signaux [112].

Williams et al. ont étendu ces mesures théoriques vues en probabilités, au domaine des signaux représentés dans le plan temps-fréquence tout simplement en traitant les distributions temps-fréquence (TFDs) comme des pdf. Nous avons alors eu l'idée de faire la même chose en amenant ceci dans le plan temps-quéfrences (coefficients cepstraux).
L'entropie va nous être utile pour quantifier notre « distribution » des coefficients cepstraux en matière de régularité. En effet, comme le dit Hemant Misra dans [114], lorsque l'on analyse l'entropie d'une distribution (pmf ou pdf), cet outil peut servir à mesurer les irrégularités de cette distribution. Étant donné que notre distribution ne répond pas aux critères d'un pdf, nous devons la convertir en la normalisant. Nous devons pour cela diviser chaque composant par la somme de tous les composants. Ceci assure que l'aire de la répartition normalisée des coefficients cepstraux est égale à 1.

Pour chaque trame de coefficients MFCC, on calcule l'entropie de Boltzmann-Shannon conformément à [114] de la manière suivante :

$$H_S = -\sum_{i=1}^{N} x_i . \log_2 x_i \tag{4.3}$$

Cette mesure d'entropie fut la première proposée en tant qu'outil pour l'information dans les systèmes de communication [115].

D'après H. Misra et al. [114], une distribution très irrégulière, présentant de nombreux pics, aura une entropie faible, tandis qu'une distribution plate, régulière, ou lisse, aura une forte entropie. L'entropie nous donne alors un indicateur de confiance sur notre classificateur. Donc plus forte sera l'entropie, et meilleure sera la modélisation des locuteurs, et donc meilleur sera le taux de reconnaissance du système.

Nous avons rapporté les résultats dans un graphique (fig. 67) afin de visualiser simultanément le taux de reconnaissance et la somme des entropies en fonction de la ligne de coefficients choisie.

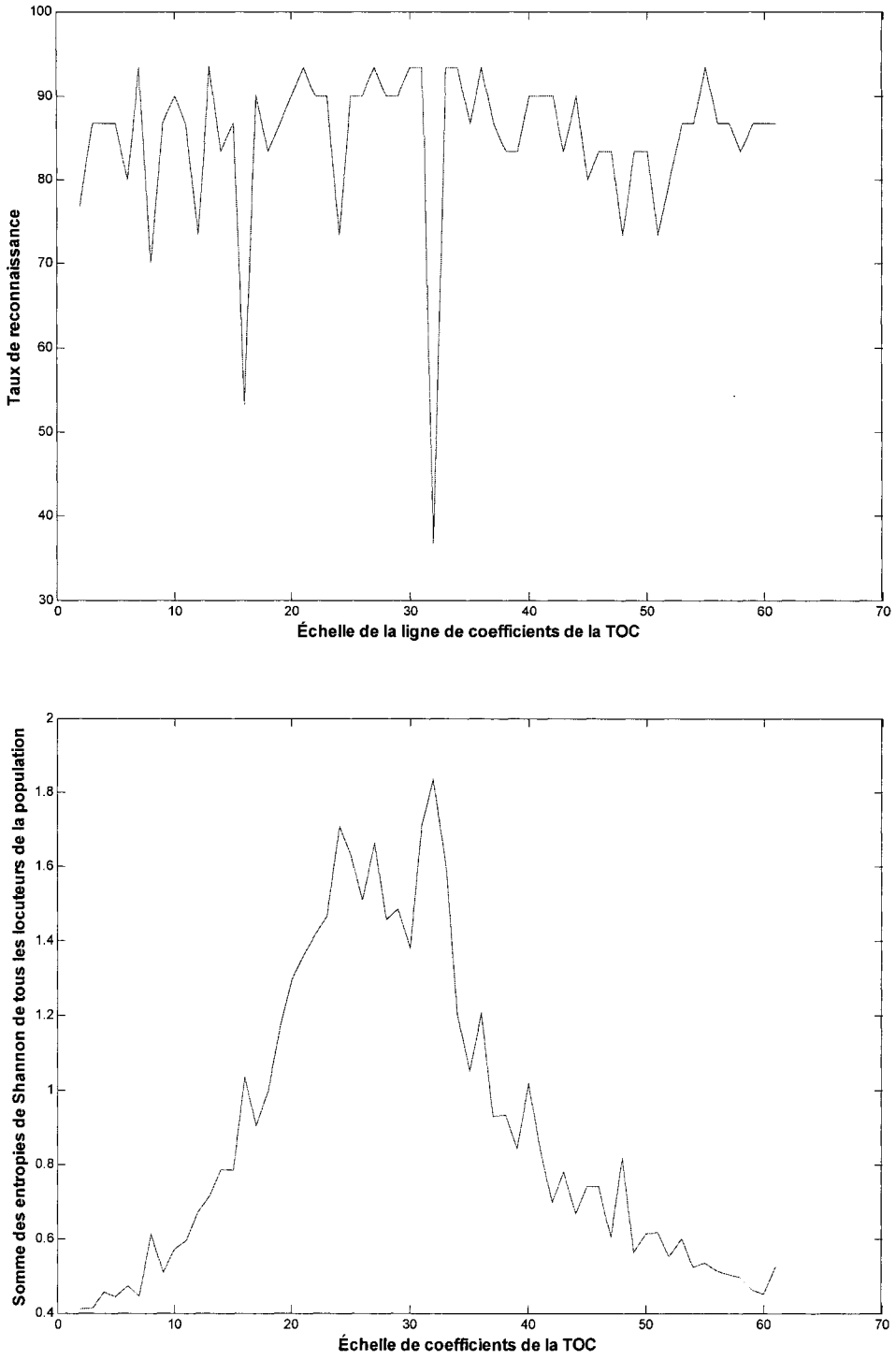


Figure 67 Effets de l'entropie de Shannon des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybride L

La figure 67 donnant les variations de l'entropie de Shannon montre deux choses importantes. La première est celle que le paramètre d'entropie joue bien un rôle dans les performances du système d'identification. En effet, on observe par exemple pour les lignes de coefficients pour les échelles 31, 33, ou 34 (à 93.3 %) que ces dernières possèdent une forte entropie; ceci correspond donc à une bonne régularité de la fonction de densité associée à leur histogramme. La deuxième chose est que ce paramètre n'intervient pas seul. En effet, si on considère la ligne de coefficients pour l'échelle 32 (30 %), on observe que l'entropie est également élevée mais avec des performances totalement opposées. Si on prend en compte simultanément l'entropie et l'énergie de la distribution des coefficients cepstraux, on voit que pour cette échelle la ligne présente un niveau d'énergie trop élevé. Quant aux lignes performantes citées plus haut, elles possèdent un niveau d'énergie faible. Ces deux paramètres semblent donc jouer simultanément un rôle important dans les performances de reconnaissance.

Enfin, on pourra remarquer que la zone que nous avions définie (p.135) comme une zone dont les échelles permettent d'aboutir à de forts taux de reconnaissance (lignes d'échelles 25 à 36), est la région sur la figure 67 où l'entropie est la plus élevée (entropie supérieure à 1).

Nous venons donc de voir que plus la somme des entropies des coefficients cepstraux de toute la population est élevée, et plus on aura une distribution lisse et régulière et plus on a de probabilités d'avoir de bons résultats. En effet, nous parlons de probabilité car l'entropie n'est pas la seule à intervenir. Il faut que les meilleures conditions soient réunies simultanément pour toutes les grandeurs physiques responsables de l'allure des distributions (voir fig. 65, fig. 66, et fig. 67).

6.3.4 Conclusions sur l'étude des coefficients MFCC extraits à partir des lignes de la TOC

L'étude des coefficients MFCC extraits à partir des lignes de coefficients de la TOC nous a permit de déterminer théoriquement les facteurs qui interviennent dans les performances du système de reconnaissance.

Tout d'abord, l'analyse des histogrammes a démontré graphiquement que de bonnes performances sont obtenues lorsque la répartition est lisse, régulière, compacte (présentant la forme d'une Gaussienne). Ceci s'explique par le fait que la modélisation dont nous nous servons est une modélisation par mélange de Gaussiennes. Donc plus la distribution des données aura une forme simple à représenter, une forme proche de celle d'une Gaussienne, et plus précise sera la modélisation puisque le modèle pourra d'autant mieux épouser la forme des données. L'erreur de modélisation sera plus petite et donc il sera plus difficile de confondre deux locuteurs lors de la phase de test du système.

Par la suite, nous avons procédé à des vérifications théoriques en utilisant des outils de mesure tels que l'énergie et l'entropie afin de caractériser ces distributions. Les résultats ont montré que les deux grandeurs jouent un rôle dans les performances du système. Plus l'entropie est faible ou bien plus l'énergie est grande et moins bons sont les résultats de reconnaissance. En conclusion, on pourra remarquer que les deux agissent simultanément pour donner de bonnes performances. En effet, une entropie forte ne suffit pas si l'énergie est également élevée. Les meilleures conditions semblent concorder à être un faible niveau d'énergie et un fort niveau d'entropie. Ceci est observable sur la figure 68 ci-dessous (pour des échelles variant de 15 à 40 par pas unitaire) :

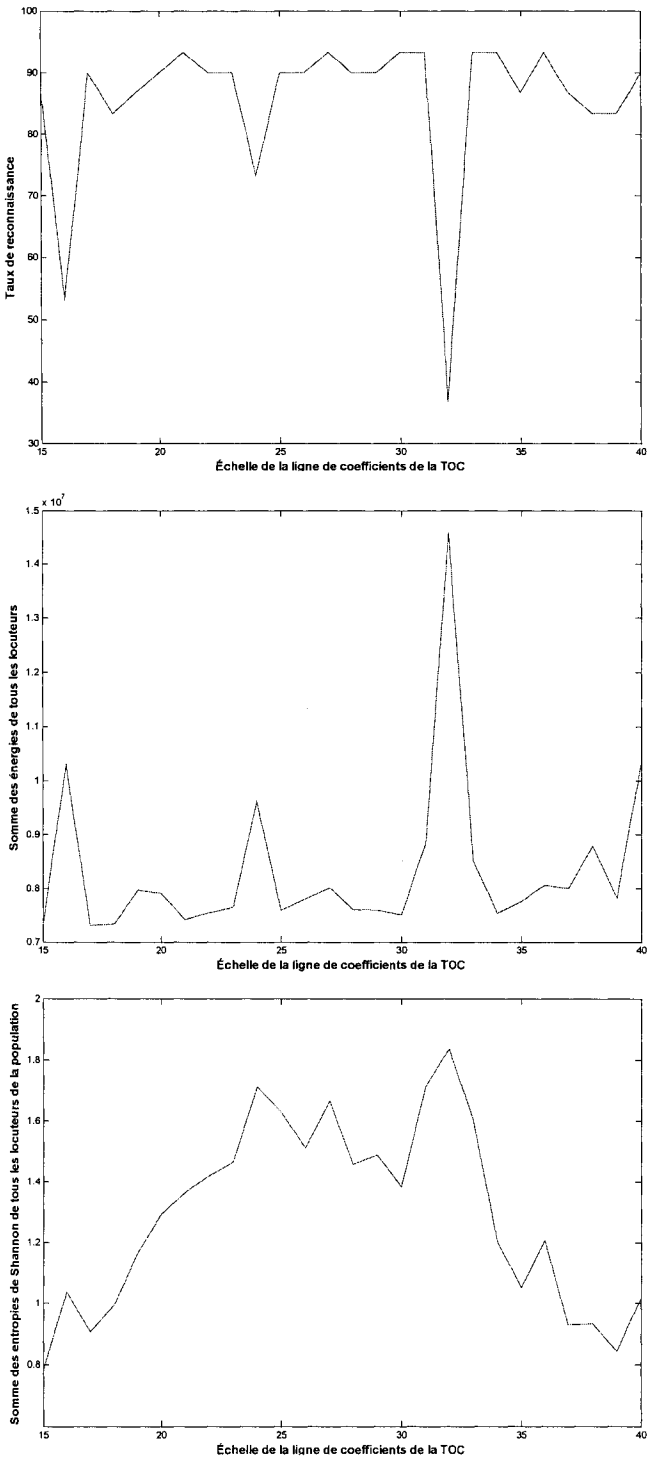


Figure 68 Pour les échelles 15 à 40, variation du taux de reconnaissance, et la somme des énergies et des entropies de la population en fonction de l'échelle de la ligne de coefficients de la TOC

Cependant, l'entropie et l'énergie ne doivent pas être antagonistes; elles doivent juste présenter un « bon dosage », un niveau moyen d'énergie pour obtenir une distribution compacte, et une forte entropie pour que la répartition des données soit lisse et régulière. Ce dosage théorique est plutôt difficile à déterminer. Ceci pourrait faire l'objet d'études futures plus approfondies sur le sujet.

Pour l'instant, nous nous proposerons une méthode plus empirique pour choisir la meilleure ligne de la TOC afin de mettre au point un système de reconnaissance automatique du locuteur (ASR) basé sur l'utilisation de cette transformée.

6.4 Proposition d'une technique pour la reconnaissance automatique du locuteur utilisant la TOC

Nous considérons une population de 30 locuteurs. Nous venons de voir au cours des différents tests, qu'il est possible d'aboutir à des taux maximums de 93.3 % de reconnaissance en tenant compte uniquement des échelles entières de la TOC. Nous considérerons donc ici que toutes les échelles sont entières afin de faciliter les calculs.

6.4.1 Méthodologie

Pour déterminer quelle est la meilleure ligne de coefficients de la TOC, nous nous servirons uniquement des données d'entraînement.

Lors de la phase de vérification du bon fonctionnement de notre système (paragraphe 4.3.4.6), nous avons vu que quelle que soit l'échelle sélectionnée, le taux d'identification est égal à 100 %. Ce « cas particulier » de la phase de test du système va nous permettre de déterminer une bonne échelle pour la ligne de coefficients à retenir. En effet, voici l'idée développée : puisque les modèles GMM ont été établi avec ces données, lorsqu'on les réutilisera à nouveau dans la phase de test pour calculer le maximum de

vraisemblance entre chaque set de données relatif à un locuteur, et chaque GMM, on devrait trouver théoriquement la valeur 0 pour le bon modèle. En pratique, on trouvera en fait la valeur de l'erreur de modélisation[8]. On observe de plus que la valeur des autres maximums de vraisemblance pour les autres locuteurs est très faible par rapport à celui obtenu avec le bon locuteur. Parmi les plus grande de toutes ces valeurs, plus grand sera l'écart avec celle du locuteur à identifier et moins il y a de risque d'erreur au niveau de la reconnaissance. Cela signifie qu'une bonne échelle de la TOC sera celle dont les données extraites auront le plus grand écart de vraisemblance avec le second locuteur le plus proche. Lorsque les données de test seront différentes de celles prises pour la phase de modélisation, on observera que les écarts vont se réduire considérablement, et même parfois s'inverser (c'est-à-dire que le locuteur à identifier ne sera pas reconnu). Cependant, on aura considéré l'échelle qui donne la plus grande marge d'erreur lors de l'identification; celle qui offre une meilleure variabilité inter locuteurs. Par conséquent, bien que n'ayant pas un taux de reconnaissance maximal, on aura le meilleur taux auquel nous pouvions aboutir.

Le lecteur pourrait être amené à se dire : si la meilleure échelle obtenue reste la meilleure pour cette même population, et ce, quelles que soient les données testées, alors pourquoi ne pas inclure une phase de test préliminaire immédiatement à la suite de l'entraînement du système pour trouver quelle est la meilleure échelle à prendre; on testerais alors toutes les échelles, et on garderait celle qui donnerait les meilleures performances. Et bien la raison est la suivante : la meilleure échelle de la TOC permettra d'obtenir les meilleures performances en matière de reconnaissance quelles que soient les données de test, cependant, elle peut, selon la nature de ces mêmes données, être égalée par d'autres échelles. Elle n'est pas unique en son genre. Cette technique ci

---

[8] On pourrait se demander pourquoi ne pas sélectionner la meilleure ligne de coefficients en se basant tout simplement sur l'échelle qui permet d'aboutir à un modèle présentant la plus petite erreur de modélisation. Ceci ne tient cependant pas compte d'une chose : la variabilité inter locuteur. En effet, la modélisation peut très bien avoir peu d'erreur, mais si il existe un autre locuteur possédant des caractéristiques trop proches (dans le cas particulier du test avec les données qui ont servi a modéliser), il risque d'être reconnu à la place de la bonne personne (lorsque l'on aura des données de test différentes).

permettrait donc de trouver toutes les meilleures échelles pour un seul cas. Il faudrait procéder à plusieurs tests consécutifs avec des données différentes pour procéder à l'éliminations de toutes les possibilités. L'avantage de la méthode que nous proposons, est qu'elle trouvera directement une seule possibilité. Notre méthode va en fait considérer toutes les échelles comme les meilleures puisqu'elles donnent toutes un taux de 100%. L'idée que nous avons développée consiste alors à trouver ce que l'une d'entre elles possède d'unique : elle a le plus important des écarts entre les maximums de vraisemblance pour le locuteur à identifier et le locuteur le plus proche de ce dernier; ceci est donc caractéristique d'une plus grande variabilité inter locuteurs.

6.4.2 Essais expérimentaux

Pour expérimenter cette méthode, nous avons donc calculé le taux de reconnaissance du système en reprenant les données d'entraînement dans la phase de test. Nous avons ensuite déterminé pour chaque échelle de la TOC, la somme des écarts pour chaque locuteur de la population testé. Voici les résultats obtenus (fig. 69):

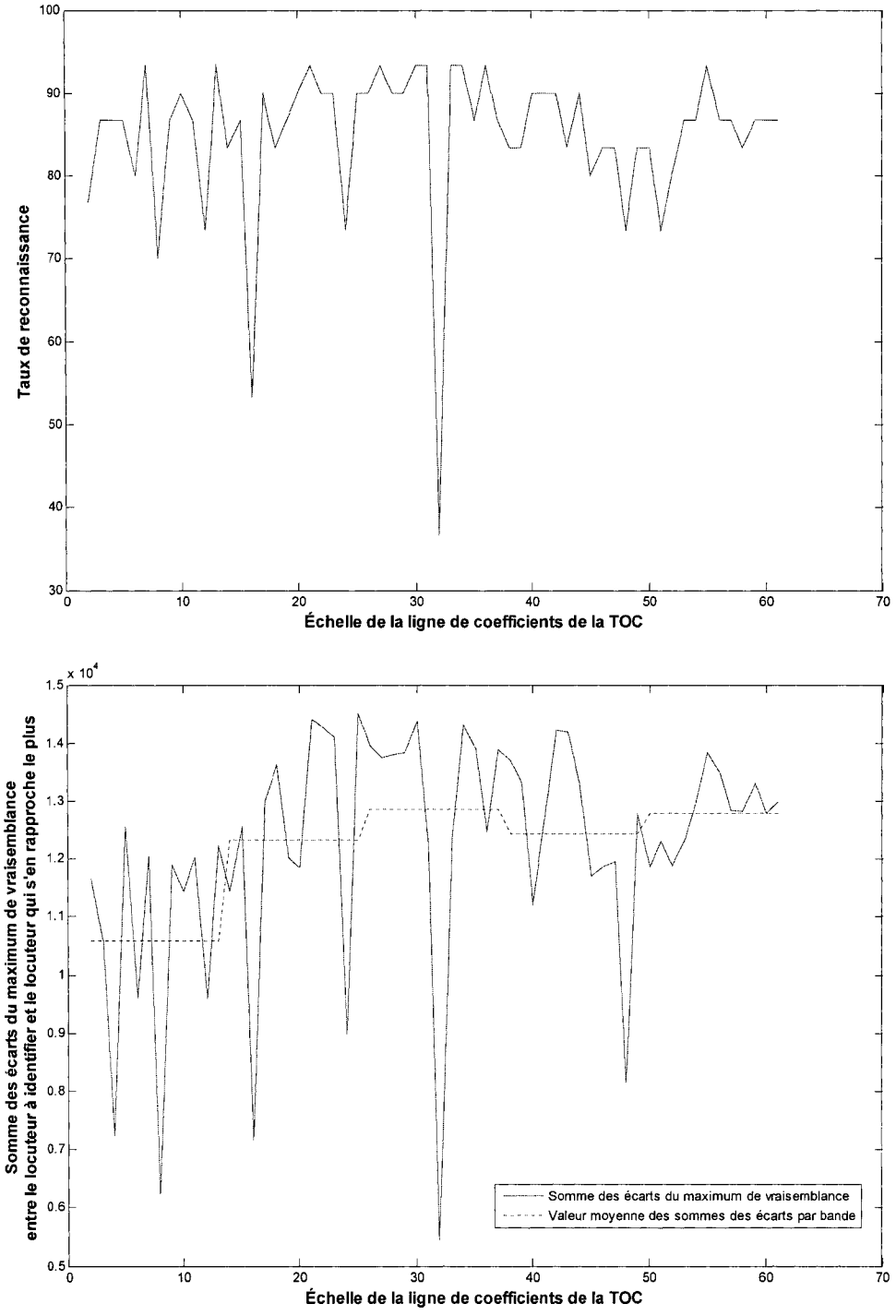


Figure 69 Effets des écarts du maximum de vraisemblance des locuteurs d'une population en fonction de l'échelle de la ligne de coefficients de la TOC, sur les performances du système hybride L

Nous divisons ensuite l'intervalle des échelles en 5 sous-ensembles égaux de 12 échelles chacun. Pour chaque sous ensemble, on calcule la somme de tous les écarts afin de déterminer la bande d'échelles pour laquelle les écarts sont les plus importants (voir fig. 69). Ceci permettra de diminuer les temps de calculs du système. Pour la bande offrant les plus fortes valeurs, on sélectionne l'échelle qui permet d'obtenir le plus grand écart entre le locuteur à identifier et le locuteur qui s'en rapproche le plus.

Cette méthode permet de détecter la ligne pour l'échelle 30 comme étant la ligne de coefficients de la TOC permettant d'aboutir aux meilleures performances du système. Ses performances sont de 93.3 % sur 30 locuteurs.
Notons que la ligne détectée en seconde position est celle pour l'échelle 34 qui permet d'aboutir également aux même performances.

6.4.3 Conclusions sur la méthode pour la reconnaissance automatique du locuteur

Cette méthode empirique que nous proposons pour la sélection de la meilleure échelle afin de choisir la ligne de coefficients de la TOC a été vérifiée par les tests expérimentaux que nous venons de vous montrer. Celle-ci n'est certainement pas la meilleure, mais elle est en tout cas très simple à implémenter. Elle a le désavantage de devoir tester toutes les échelles entières de 2 à 61, et donc de demander beaucoup de temps de calcul pour la phase d'entraînement. Cependant, cette tache ne s'effectue qu'une seule fois, donc ce désavantage n'est pas très contraignant. Nous pouvons également proposer une méthode basée sur le même principe pour choisir la meilleure échelle à valeur non entière. En effet, une fois la meilleure valeur entière déterminée, on peut faire une étude semblable autour de celle-ci, en divisant un court intervalle centré dessus.

6.5 Conclusions

Le choix de l'échelle d'analyse de la TOC reste le point crucial de notre système.

La méthode de sélection des coefficients de la TOC est une étape importante puisqu'elle conditionne les performances de notre système en matière d'identification, mais aussi en temps de traitement. Nous avons vu qu'en choisissant adéquatement l'échelle de la transformée en ondelettes continue, nous pouvions dans certains cas améliorer les performances du système de reconnaissance de référence. Nous sommes alors en présence d'un nouveau signal contenant toujours les caractéristiques du locuteur, mais cette fois-ci exprimées autrement, c'est-à-dire, exprimées à un niveau d'échelle permettant une meilleure mise en évidence de la variabilité inter-locuteurs.

Les améliorations des performances que présente notre système sont dues à une meilleure modélisation et à une meilleure différentiation des locuteurs de la population. En effet, l'utilisation de la TOC a pour but de représenter les caractéristiques vocales d'un locuteur à plusieurs échelles d'analyse; certaines de ces échelles possèdent la particularité de mieux faire ressortir des caractéristiques qui s'exprimer moins dans le signal de parole de base. Par conséquent, en considérant une échelle pour laquelle les locuteurs seront plus différentiables, on aura un plus petit taux d'erreur d'identification. Une meilleure modélisation des individus aura donc lieu si les coefficients cepstraux exprimant les caractéristiques de chacun sont mieux modélisables. Des analyses d'histogramme nous ont montré les différences dans la distribution de ces coefficients : pour une distribution régulière, lisse, compacte, ayant une allure proche de celle d'une Gaussienne, ont aboutira à un fort taux de reconnaissance; en revanche, une distribution irrégulière, très morcelée ou étalée, donnera un faible taux de reconnaissance. Ceci est explicable par le fait que l'on utilise une modélisation par mélange de Gaussiennes pour modéliser la distribution de ces données. Donc plus elle est « proche » d'une Gaussienne, plus il sera facile à la représenter fidèlement et précisément, et donc moins

il y aura d'erreur de modélisation et d'identification par la suite. Des analyses de l'énergie et de l'entropie de ces répartitions sont venues confirmer nos observations graphiques et confirmer ce qui influençait les performances du système. En effet, avec des coefficients de la TOC présentant une forte entropie et une faible énergie, on pourra aboutir aux meilleures performances du système. Le seul détail que nous ne connaissons pas et que nous n'avons pas réussi à déterminer, est l'explication « théorique » existant entre les coefficients de la TOC retenus et la distribution de ses coefficients cepstraux : quels sont les paramètres physiques qui font ainsi varier cette distribution? Beaucoup d'analyses ont été effectuées mais aucune n'a permit d'aboutir à un résultat concret. Ceci pourrait faire l'objet de futures études plus approfondies. On pourra alors améliorer notre système en mettant au point une méthode de reconnaissance automatique du locuteur qui sélectionnerait la meilleure échelle de la TOC à l'aide de ces nouvelles découvertes.

CONCLUSION

La mise au point de systèmes d'identification est en plein essor, dû en l'occurrence à la nécessité croissante du besoin de sécurité de chacun (dans les domaines privé, professionnel ou public).

Les scientifiques montrent de plus en plus d'intérêt à la reconnaissance du locuteur étant donné que la voix est le seul outil biométrique qu'ils ont à leur disposition via des liaisons de téléphoniques. C'est pourquoi beaucoup d'efforts sont entrepris pour rendre les systèmes de reconnaissance les plus robustes et fiables que possible, notamment ces dernières décennies avec l'apparition des ondelettes.

Nos travaux sur l'identification du locuteur ont montré que la transformée en ondelettes continue (très peu utilisée en traitement de la parole) permet d'obtenir des résultats intéressants dans ce domaine. Si elle est utilisée en conjonction avec les MFCC pour le paramétrage initial du signal, elle améliore alors les performances en termes de taux de reconnaissance, sans augmenter considérablement les temps de calcul.

À travers ce mémoire, nous avons testé différentes utilisations de la TOC pour la reconnaissance du locuteur. La méthode de sélection des coefficients de la TOC est une étape importante puisqu'elle conditionne les performances de notre système en matière d'identification, mais aussi en temps de traitement. Les différents systèmes hybrides que nous avons conçu et testé ont permit de révéler que la transformée en ondelettes continue pouvait dans certains cas améliorer les performances du système de reconnaissance de référence. Pour cela, nous devons appliquer la TOC au signal de parole d'entraînement, puis, considérer une seule échelle de cette transformée. Nous sommes alors en présence d'un nouveau signal contenant toujours les caractéristiques du

locuteur, mais cette fois-ci exprimées autrement, c'est-à-dire, exprimées à un niveau d'échelle permettant une meilleure mise en évidence de la variabilité inter-locuteurs. À partir de ce signal on extrait les vecteurs paramétriques pour la modélisation du locuteur (avec la même méthode que pour le système de référence). Nous avons vu qu'en sélectionnant tous les coefficients de la TOC pour cette même échelle d'analyse, on aura un signal permettant d'aboutir à de bonnes performances, et à un taux de reconnaissance supérieur à celui du système de référence (conçu comme celui de Reynolds [52]). Pour des populations de 30 ou de 50 locuteurs (de YOHO), dans un environnement non bruité (SNR=43 dB), les taux obtenus sont respectivement de 96.7 % et de 96 %, soit des gains respectifs de 6.7 points et 6 points. Nous devons de plus ajouter que ces performances sont réalisées en mode indépendant du texte sur des populations essentiellement constituées par des hommes.

Le choix de l'échelle d'analyse de la TOC est un point crucial. Dans la littérature, très peu de chercheurs expliquent comment la choisir. Cependant, les divers histogrammes du chapitre 4 ont permis de montrer que ce choix n'est pas critique comme on pourrait le penser puisqu'il existe en réalité très peu d'échelles aboutissant à de mauvaises performances. Dans nos travaux, nous avons apporté quelques éléments de réponse à ce problème. Nous avons montré qu'une bonne échelle est celle dont le signal correspondant (lorsqu'on prend tous les coefficients temporels de la TOC pour cette même échelle) possède une distribution de ses coefficient cepstraux lisse, régulière (forte entropie), et compacte (niveau d'énergie moyen) avec une forme qui s'apparenterait à une Gaussienne.

Nous avons également montré que l'utilisation d'une seule ligne de coefficients de la TOC pour une échelle fixe, permet de palier aux inconvénients de la TOC. En effet, la TOC est une transformée fortement redondante. Il est très difficile d'en extraire les données pertinentes en choisissant les bons coefficients. En choisissant les coefficients en les prenant par lignes (dans leur intégralité), on passe ainsi outre les désavantages de

la TOC puisqu'on élimine substantiellement la redondance de celle-ci, et on retrouve des temps de calcul comparables à ceux du système de référence. De plus, l'utilisation de la TOC permet au système d'utiliser moins de parole pour les phases d'entraînement et de reconnaissance.

Nous avons ainsi mis au point un système de reconnaissance du locuteur performant et qui fait usage de la TOC. Cependant, plusieurs points mériteraient d'être approfondis, le plus important étant la caractérisation des meilleures échelles d'analyse de la TOC. Par ailleurs, il serait également intéressant d'étudier de nouvelles formes de sélection des coefficients de la TOC, comme la sélection par grilles de points aléatoires; ceci permettrait entre autres d'éviter le choix de l'échelle d'analyse. Enfin, le dernier point concerne la robuste de notre système. Nous avons conçu un système permettant d'étudier la viabilité de la TOC pour la reconnaissance du locuteur. C'est pourquoi nous ne nous sommes pas intéressés à sa robustesse lors de sa conception et nous avons effectué les tests en milieu quasi idéal. Il pourrait être intéressant dans des recherches futures de rendre notre système robuste et d'étudier son comportement pour voir si la TOC permet également d'améliorer les performances en milieu bruité.

ANNEXE 1

REPRÉSENTATION DES ÉCHELLES DE MEL ET DE BARK PAR UN BANC DE FILTRES

Voici la représentation des échelles de Mel et de Bark par un banc de filtres :

Tableau XXVI

Valeurs du banc de filtres pour la modélisation de l'échelle des fréquences de Mel et de Bark [99]

Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

ANNEXE 2

L'OPTIMISATION DE LAGRANGE

Supposons que l'on souhaite maximiser la fonction $f(x)$ en tenant compte d'une contrainte telle que $g(x) = 0$.

Pour trouver le maximum de la fonction f, nous devons tout d'abord commencer par créer la fonction Lagrangienne :

$$L(x, \lambda) = f(x) + \lambda . g(x) \tag{B.1}$$

où λ est appelé le multiplieur de Lagrange

Maintenant, calculons la dérivée de l'expression précédente, et résolvons là lorsqu'elle est égale à zéro :

$$\frac{\partial L(x, \lambda)}{\partial x} = \frac{\partial f(x)}{\partial x} + \lambda . \frac{\partial g(x)}{\partial x} = 0 \tag{B.2}$$

Finalement, il ne reste plus qu'à résoudre l'équation pour la valeur de λ et celle de x qui maximisent $f(x)$.

ANNEXE 3

L'ALGORITHME DES K-MEANS

L'algorithme des K-means est un des plus simples algorithmes d'apprentissage qui permet de résoudre les problèmes de partitionnement [116]. La procédure est basée sur une simple manière de répartir n échantillons de données à travers K sous-ensembles. L'idée principale consiste alors à définir K centres, chacun appartenant à un sous-ensemble. Ces centres doivent être situés à des endroits stratégiques puisque différentes localisations entraîneront des modélisations différentes, et par conséquent, des performances également différentes. Il existe diverses manières d'initialiser les centres. L'étape suivante consiste à prendre chaque point des données et à l'associer avec le centre le plus proche. Ceci représente la première itération de l'algorithme des K-means. Dans l'itération suivante, on recalcule les K nouveaux centres comme étant les barycentres des ensembles résultants de l'étape précédente. Après avoir obtenu ces K nouveaux centres, on procède à nouveau à une répartition des données. Ces étapes se répètent à la suite les unes des autres jusqu'à ce que les centres ne changent plus de position.

Cet algorithme à en fait pour but de minimiser la somme du carré des distances entre les échantillons de données $\mathbf{x}_l$ et les centres des sous-ensembles $\boldsymbol{\mu}_j$:

$$J = \sum_{j=1}^{N} \sum_{l \in S_j} \left| \mathbf{x}_l - \boldsymbol{\mu}_j \right|^2 \qquad \text{(C.1)}$$

Pour résumer, cet algorithme est composé comme suit:

Étape 1 : Initialiser les K centres dans l'espace représenté par les données que l'on veut répartir.

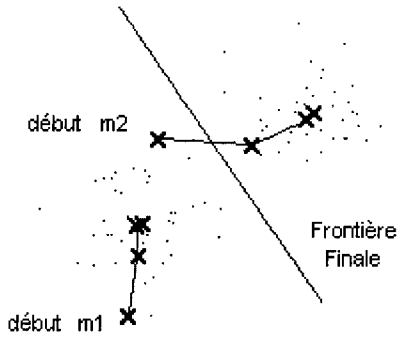
Étape2 : Assigner chaque échantillon de données au sous-ensemble dont le centre est le plus proche.

Étape 3 : Lorsque toutes les données ont été réparties, recalculez les positions des K centres.

Étape 4 : Répétez les étapes 2 et 3 jusqu, à ce que les centres ne changent plus de position.

Il est par ailleurs prouvé que cet algorithme converge quelles que soient les données []. Cependant, on ne trouvera pas nécessairement la meilleure répartition correspondant à la fonction que l'on souhaite minimiser. Enfin, cet algorithme a le désavantage d'être très sensible à l'initialisation des centres, d'où le besoin de bien effectuer cette tache [].

Voici un exemple montrant comment les points m1 et m2 se déplacent pour devenir les centres de deux sous-ensembles :



ANNEXE 4

BASE DE DONNÉES DE YOHO

Following is an excerpt from:

Campbell, J. P., Jr. "Features and Measures for Speaker Recognition." Ph.D. Dissertation, Oklahoma State University, 1992.

After this, we also include portions of an appendix ("Database Description") from:

Higgins, A., J. Porter and L. Bahler. YOHO Speaker Authentication Final Report. ITT Defense Communications Division, 1989.

We would like to express our deepest gratitude to Joseph Campbell for his assistance in preparing these materials for publication with the corpus.

YOHO DATABASE

The YOHO database was collected by ITT under a U.S. Government contract administered by the author.

The signal conditioning and acquisition was designed by the author using a 4-times oversampling method to provide bandwidth and linear phase up to 3.8 kHz. First, the analog signal is low pass filtered at approximately 5 kHz by a mild 4th order elliptic analog antialiasing filter that has negligible effect on the signal below 4 kHz. This analog antialiasing filter sufficiently limits the bandwidth to 16 kHz to prevent aliasing when it is then oversampled at 32 kHz with 12 bits of precision. Next, the 32-kHz sampled signal is passed through a 255-tap, finite duration impulse response (FIR) digital bandpass

filter. This digital filter limits the bandwidth of the signal so that it can be decimated by 4:1 to arrive at the final desired sampling frequency of 8 kHz. Using iterative inverse- and forward- Fourier transforms, the author designed a frequency-sampling symmetric-FIR filter-design routine to determine the 255 coefficients that best approximate a secure voice terminal's input characteristics in a least mean-square magnitude-response error sense. The resulting response models the STU-III secure voice terminal's input characteristics very closely and is given in Table II-1.

TABLE II-1

FREQUENCY RESPONSE OF YOHO DECIMATION FILTER

Frequency (Hz)	Response (dB)
0	-25
< 50	-21
100	-7
150	-2
200	-0.2
200 - 3600	-0.2 to +0.3 peak ripple
3600	-0.2
3800	-3
4000	-25
4400	-42
> 5000	-50
16,000	-57

The key to oversampling is that the analog antialiasing filter need not have steep skirts in the vicinity of the half sampling frequency, as in the Nyquist sampling methods, whereas the symmetric digital FIR filter has linear phase and can have arbitrarily flat magnitude

response. The advantage of the oversampling method is that the magnitude and phase distortions near the half sampling frequency are far less than is common in traditional Nyquist sampling methods. For example, Digital Sound Corporation's -3 dB analog bandwidth for 8 kHz sampling is only 3.6 kHz, as opposed to the 3.8 kHz, -3 dB bandwidth achieved by this method. This additional 200 Hz of bandwidth is vital for listeners to be able to distinguish between sounds concentrated in high frequencies (e.g., the affricate sounds differentiating "chew" and "jew").

The YOHO database is the only large scale, scientifically controlled and collected, high-quality speech database for speaker authentication testing at high confidence levels. Table II-2 describes the YOHO database (Higgins 1990).

TABLE II-2

THE YOHO DATABASE

* "Combination lock" phrases (e.g., 36-24-36)
* 138 subjects: 108 males, 30 females
* Collected over 3 month period in a real-world office environment
* 4 enrollment sessions per subject with 24 phrases per session
* 10 test sessions per subject with 4 phrases per session
* Total of 1932 validated sessions
* 8 kHz sampling with 3.8 kHz analog bandwidth
* 1.2 gigabytes of data (when uncompressed)

In a text-dependent speaker verification scenario, phrases are prompted and the claimant is requested to say them. The syntax used in the YOHO database is "combination lock" phrases. For example, the prompt might read: "Say: thirty-six, twenty-four, thirty-six." Where the claimant is to speak the phrase as three doublets.

REFERENCES

Campbell, J. P., Jr. "Features and Measures for Speaker Recognition." Ph.D. Dissertation, Oklahoma State University, 1992.

Higgins, A., J. Porter and L. Bahler. YOHO Speaker Authentication Final Report. ITT Defense Communications Division, 1989.

Higgins, A. "YOHO Speaker Verification." Baltimore: 1990.

Higgins, A., L. Bahler, and J. Porter. "Speaker Verification Using Randomized Phrase Prompting." Digital Signal Processing 1, no. 2 (1991): 89 - 106.

DATABASE DESCRIPTION

1. Introduction

The YOHO Speaker Verification Database was collected while testing a prototype speaker verification system by ITT Defense Communications Division under contract with the U.S. Departement of Defense. The database is the largest supervised speaker verification database known to the authors. The number of trials and the number of test subjects were determined to allow testing at the 80% confidence level to determine whether the system met specified performance requirements. The required error rates were 1% false rejection and 0.1% false acceptance.

The testing and database collection were conducted at ITTDCD's head- quarters in Nutley, New Jersey. The system consisted of a Sun-3 workstation with added processor boards for real-time processing, a 19-inch monitor for prompting, and a telephone handset with a high-quality microphone. When the system was used in an enrollment or verification session, a sampled waveform file was created for each phrase-length utterance. The collection of these waveform files is contained on the database tape.

2. Equipment Setup

The system was set up in the corner of a large room (approximately 25 x 25 feet) which was mostly empty. Low level noise could be heard from the adjoining office space, from people walking through the room, and from occasional paging on the public address system. Noise from the fan in the Sun workstation could also be heard. Subjects could stand or sit in a chair while using the system. Prompts were displayed on the console using "gallant" point-size 19 font. Upon completion of each phrase, the system automatically prompted the next phrase, with a delay of about one second. A "beep" was produced as each phrase was prompted to attract the user's attention. These beeps can be heard at the beginning of many of the waveform files.

On completion of each test session, a message "ACCEPTED" or "REJECTED" was displayed on the console. No other feedback or motivation was provided.

2.1. Data Acquisition

A handset containing an omnidirectional non-noise-canceling electret microphone is connected to an external audio amplifier. The microphone signal first passes through a 6-pole passive elliptic filter with a cutoff frequency of 4.3 kHz. It is then amplified and applied as input to an analog to digital converter (ADC). The amplified output is also returned to the handset to supply sidetone.

A 4-times oversampling scheme is implemented, which works as follows. The ADC operates at a sampling rate of 32 kHz, producing 12-bit samples. The analog lowpass filter prevents aliasing in the sampled signal without attenuating frequencies below 4 kHz. The sampled signal is passed through a 255-tap FIR bandpass filter. The upper band edge of this filter is approximately 4 kHz, or one fourth of the original Nyquist frequency. Therefore, the filtered signal can be represented without aliasing at one fourth of the original sampling rate. This is accomplished by downsampling, with a 4 to 1 ratio, to the final sampling rate of 8 kHz.

The advantage of this method is that the frequency response of the frontend is entirely controlled by the FIR decimation filter, assuming the net frequency response of the analog components is constant below 4 kHz. This allows the frequency response of the frontend to be more precisely controlled than would be possible using a conventional analog anti-aliasing filter and ADC conversions at an 8 kHz rate. The frequency response of the decimation filter is shown in Table I.

Frequency (Hz)	Response (dB)
0	-25
<50	-21
100	-7
150	-2
200	-0.2
200-3600	-0.2 to +0.3 peak ripple
3600	-0.2
3800	-3
4000	-25
4400	-42
>5000	-50
16000	-57

Table I: Frequency Response of Decimation Filter

3. Subjects

A total of 189 subjects began the testing program. Three subjects dropped out before the enrollment sessions were complete, leaving 186 sub- jects who completed all the enrollment sessions. Of these subjects, 156 were male, and 30 were female. Members of several departments including engineering and program management, and support staff such as secretaries and draftsmen, were asked to participate in the test. Subjects spanned a wide range of ages, job descriptions, and educational backgrounds. Most subjects were from the New York area, although there were many exceptions, including some non-native English speakers. [NB: Only 138 speakers have been included in the CD-ROM version published by the Linguistic Data Consortium.]

Subjects were introduced to the system by watching a 5-minute video tape which demonstrated the intended usage of the system. This tape, which was delivered to the Government, documents the user's view of the system. A test monitor was present during all enrollment and test sessions. His primary responsibilities were to maintain a continuous flow of subjects using the system, and to perform daily tape backups of the sampled waveform files. The test monitor also provided further instruction or assistance if needed in enrollment sessions, but did not interfere in test sessions except to take note of sessions in which the subject claimed to be someone else (which was allowed), or in which the prompts were not read correctly. A total of 57 such sessions were reported. These sessions are not present in the database.

4. Speech Material

The speech material consists of "combination-lock" phrases. An example prompt is: "35 - 72 - 41", pronounced "thirty five, seventy two, forty one". Each phrase consists of three number doublets. The doublets are chosen from a list which includes all the doublets from 21 to 99 with the following exceptions: (1) no exact decades (30, 40,

etc.), (2) no double digits (22, 33, etc.), and (3) no numbers ending in "8" (28, 38, etc.). Pausing between the doublets is optional, but not encouraged. An enrollment session consists of 24 such phrases. A verification trial or session con- sists of 4 such phrases.

5. Sessions

Subjects were asked to participate in 14 sessions over a 3-month time interval. The first 4 sessions were enrollment sessions, which required about 3 minutes each, and the following 10 sessions were test sessions, which took about 20 seconds each.

Each subject in the test completed the test sessions at his or her own rate. The nominal separation between sessions was 3 days. However, this varied to suit individuals' schedules. Table II shows the earliest, median, and latest dates of subjects' first, fifth, and tenth sessions.

Session	Earliest	Median	Latest
First	3/07/89	3/27/89	5/08/89
Fifth	3/17/89	4/21/89	5/26/89
Tenth	3/29/89	5/03/89	5/26/89

Table III: Test Session Date

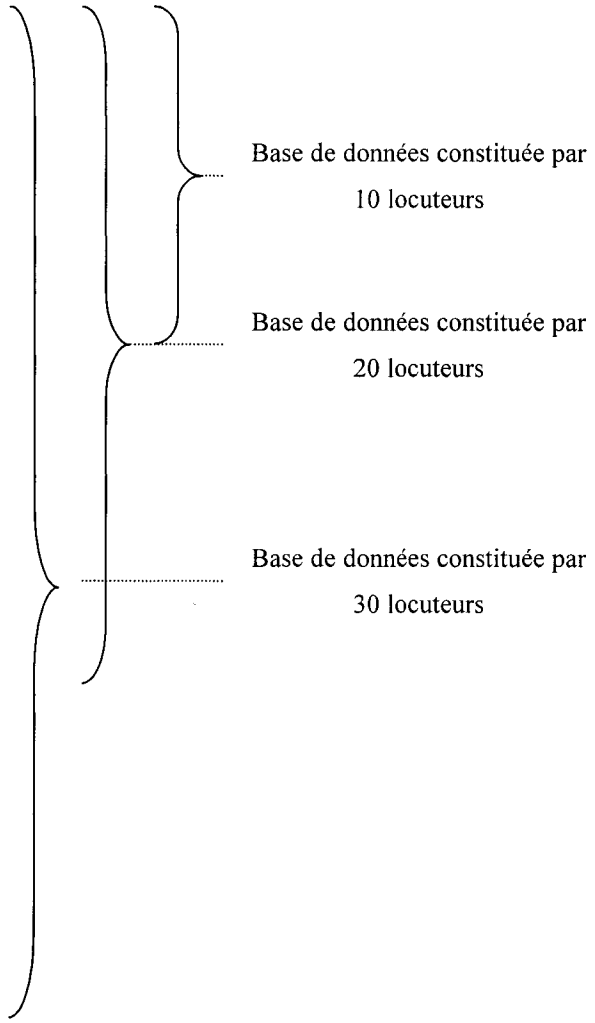
ANNEXE 5

SOUS BASES DE DONNÉES DE YOHO

Les bases de données que nous considérons pour ce projet sont des regroupements d'échantillons de parole de 10, 20, 30, et 50 locuteurs.

Ces ensembles de locuteurs sont obtenus à partir de la base de données publique YOHO. Pour effectuer la sélection des locuteurs, nous avons tout simplement considéré les 50 premiers locuteurs enregistrés dans YOHO pour créer une base de 50 locuteurs. Ainsi, nous avons effectué les répartitions suivantes :

Locuteur	Sexe	Origine
1	M	new york
2	M	south
3	M	new york
4	M	new jersey
5	F	new jersey
6	M	new york
7	M	new jersey
8	M	new jersey
9	M	new jersey
10	M	midwest
11	M	new york
12	M	new york
13	M	new york
14	M	new york
15	M	midwest
16	M	new jersey
17	M	new york
18	M	new york
19	M	new york
20	M	midwest
21	M	new york
22	F	new jersey
23	M	unknown
24	M	new york
25	M	New York
26	M	new york
27	M	ny-brooklyn
28	M	new york
29	M	new york
30	M	new york



31	M	N.Y.,LongIsland
32	M	new york
33	F	new york
34	M	pittsburgh
35	F	new york
36	M	midwest
37	M	northeast
38	M	hong kong
39	F	new york
40	M	new jersey
41	M	new jersey
42	M	south
43	M	upstate NY
44	F	unknown
45	M	estonia
46	M	new york
47	M	new york
48	M	new york
49	F	new jersey
50	M	new york

Tableau XXVII Base de données de 50 locuteurs formée par des sous ensembles de 10, 20, et 30 locuteurs

Les répartitions au sein de ces bases de données en fonction du sexe du locuteur sont les suivantes :

Base de données de 10 locuteurs : 9 Hommes et 1 Femme

Base de données de 20 locuteurs : 19 Hommes et 1 Femme

Base de données de 30 locuteurs : 28 Hommes et 2 Femmes

Base de données de 50 locuteurs : 43 Hommes et 7 Femmes

ANNEXE 6

RÉSULTATS COMPLETS DES TESTS DU SYSTÈME DE RECONNAISSANCE HYBRIDE L

Voici les résultats complets obtenus pendant la phase de test du système de reconnaissance hybride L. Les performances sont calculées pour chaque échelle entière de la TOC, de l'échelle 2 à l'échelle 61, et dans le cas de populations de 10, 20, 30, 40, et 50 locuteurs, pour un milieu dépourvu de bruit.

	Taux de reconnaissance du système hybride L (en %)				
Ligne à l'échelle	**Base de données (population)**				
	10	**20**	**30**	**40**	**50**
2	90	85	76.7	80	80
3	100	85	86.7	90	92
4	100	90	86.7	85	84
5	100	90	86.7	82.5	86
6	100	90	80	80	74
7	100	90	93.3	90	92
8	100	75	70	65	62
9	100	90	86.7	85	84
10	100	90	90	87.5	86
11	100	90	86.7	85	84
12	100	80	73.3	70	72
13	100	90	93.3	87.5	90
14	90	85	83.3	80	82
15	100	85	86.7	85	82
16	80	70	53.3	37.5	40
17	100	90	90	90	86
18	100	90	83.3	80	80
19	100	90	86.7	85	82
20	100	90	90	85	82
21	100	90	93.3	92.5	88
22	100	90	90	90	90
23	100	90	90	87.5	84
24	80	75	73.3	65	64
25	100	90	90	82.5	82
26	90	85	90	90	88
27	100	90	93.3	90	86
28	100	90	90	87.5	86

29	100	90	90	87.5	86
30	100	90	93.3	90	90
31	100	90	93.3	92.5	92
32	70	45	36.7	25	18
33	100	95	93.3	85	84
34	100	95	93.3	92.5	90
35	100	95	86.7	87.5	84
36	100	90	93.3	85	84
37	90	85	86.7	85	86
38	80	80	83.3	82.5	84
39	100	90	83.3	85	80
40	90	90	90	87.5	84
41	100	95	90	85	86
42	100	90	90	90	88
43	100	90	83.3	80	82
44	100	90	90	87.5	84
45	100	90	80	77.5	74
46	100	85	83.3	80	78
47	90	85	83.3	80	82
48	80	75	73.3	62.5	60
49	90	80	83.3	80	84
50	100	90	83.3	72.5	78
51	100	75	73.3	72.5	74
52	100	85	80	75	76
53	100	85	86.7	82.5	82
54	90	80	86.7	77.5	78
55	100	90	93.3	90	88
56	100	90	86.7	87.5	86
57	100	85	86.7	75	78
58	90	85	83.3	80	82
59	100	90	86.7	85	88
60	90	85	86.7	77.5	82
61	90	85	86.7	77.5	80

Tableau XXVIII Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à valeur entière, et pour des populations de 10, 20, 30, 40, et 50 locuteurs

Les cases ombragées correspondent à des résultats du système hybride L (avec TOC) supérieurs à ceux du système de référence (sans TOC).

ANNEXE 7

RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (1ère partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L

Voici les résultats complets obtenus pendant les tests de raffinement du système de reconnaissance hybride L. Les performances sont calculées pour chaque échelle de la TOC, par pas de 0.1, de l'échelle 25 à l'échelle 36, et dans le cas de populations de 30, et de 50 locuteurs, pour un milieu dépourvu de bruit.

	Taux de reconnaissance du système hybride L (en %)	
Ligne à l'échelle	**Base de données (population)**	
	30	**50**
25	90	82
25.1	90	92
25.2	93.3	90
25.3	86.7	88
25.4	90	86
25.5	90	86
25.6	86.7	84
25.7	90	84
25.8	93.3	90
25.9	90	90
26	90	88
26.1	86.7	88
26.2	86.7	82
26.3	93.3	86
26.4	86.7	78
26.5	83.3	78
26.6	93.3	88
26.7	86.7	82
26.8	83.3	84
26.9	90	90
27	93.3	86
27.1	93.3	88
27.2	90	88
27.3	86.7	78

27.4	66.7	42
27.5	90	86
27.6	83.3	82
27.7	90	84
27.8	83.3	84
27.9	86.7	88
28	90	86
28.1	90	86
28.2	86.7	86
28.3	83.3	86
28.4	86.7	82
28.5	83.3	80
28.6	90	92
28.7	90	90
28.8	86.7	86
28.9	86.7	84
29	90	86
29.1	86.7	82
29.2	90	86
29.3	90	88
29.4	86.7	88
29.5	93.3	88
29.6	90	86
29.7	90	90
29.8	86.7	82
29.9	86.7	82
30	93.3	90
30.1	86.7	84
30.2	90	88
30.3	86.7	84
30.4	90	86
30.5	86.7	84
30.6	86.7	90
30.7	86.7	84
30.8	93.3	88
30.9	93.3	90
31	93.3	92
31.1	86.7	82
31.2	90	90
31.3	90	88
31.4	93.3	92
31.5	90	92

31.6	93.3	90
31.7	90	88
31.8	83.3	82
31.9	93.3	92
32	36.7	18
32.1	93.3	92
32.2	93.3	90
32.3	90	86
32.4	93.3	86
32.5	93.3	92
32.6	90	92
32.7	93.3	94
32.8	90	90
32.9	93.3	86
33	93.3	84
33.1	93.3	90
33.2	93.3	90
33.3	90	86
33.4	90	82
33.5	90	80
33.6	86.7	82
33.7	90	82
33.8	83.3	84
33.9	93.3	90
34	93.3	90
34.1	83.3	80
34.2	90	86
34.3	90	84
34.4	90	88
34.5	93.3	86
34.6	83.3	84
34.7	90	82
34.8	90	80
34.9	96.7	88
35	86.7	84
35.1	90	84
35.2	86.7	88
35.3	96.7	90
35.4	96.7	88
35.5	76.7	64
35.6	83.3	82
35.7	83.3	84

35.8	93.3	86
35.9	90	84
36	93.3	84

Tableau XXIX Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à un chiffre significatif après la virgule, et pour des populations de 30 et 50 locuteurs

Les cases ombragées correspondent à des résultats du système hybride L (avec TOC) supérieurs à ceux du système de référence (sans TOC).

ANNEXE 8

RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (2ème partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L

Voici les résultats complets obtenus pendant les seconds tests de raffinement du système de reconnaissance hybride L. Les performances sont calculées pour chaque échelle de la TOC, par pas de 0.01, de l'échelle 34.5 à l'échelle 35.5, et dans le cas de populations de 30 locuteurs, pour un milieu dépourvu de bruit.

	Taux de reconnaissance du système hybride L (en %)
Ligne pour l'échelle	**Base de données de 30 locuteurs (population)**
34.5	93.3
34.51	83.3
34.52	90
34.53	86.7
34.54	90
34.55	90
34.56	90
34.57	96.7
34.58	93.3
34.59	86.7
34.6	83.3
34.61	90
34.62	90
34.63	90
34.64	90
34.65	90
34.66	93.3
34.67	93.3
34.68	93.3
34.69	93.3
34.7	90
34.71	80
34.72	90
34.73	86.7
34.74	90
34.75	93.3

34.76	86.7
34.77	90
34.78	93.3
34.79	86.7
34.8	90
34.81	96.7
34.82	86.7
34.83	93.3
34.84	86.7
34.85	90
34.86	80
34.87	80
34.88	80
34.89	90
34.9	96.7
34.91	86.7
34.92	83.3
34.93	86.7
34.94	86.7
34.95	83.3
34.96	80
34.97	93.3
34.98	86.7
34.99	90
35	86.7
35.01	90
35.02	86.7
35.03	86.7
35.04	93.3
35.05	90
35.06	90
35.07	90
35.08	86.7
35.09	90
35.1	90
35.11	93.3
35.12	86.7
35.13	86.7
35.14	90
35.15	93.3
35.16	93.3
35.17	93.3

35.18	96.7
35.19	90
35.2	86.7
35.21	96.7
35.22	86.7
35.23	93.3
35.24	96.7
35.25	86.7
35.26	93.3
35.27	93.3
35.28	96.7
35.29	96.7
35.3	96.7
35.31	90
35.32	86.7
35.33	80
35.34	96.7
35.35	90
35.36	90
35.37	90
35.38	93.3
35.39	93.3
35.4	96.7
35.41	93.3
35.42	90
35.43	86.7
35.44	93.3
35.45	90
35.46	93.3
35.47	90
35.48	93.3
35.49	90
35.5	76.7

Tableau XXX Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à deux chiffres significatifs après la virgule, et pour une population de 30 locuteurs

Les cases ombragées correspondent à des résultats du système hybride L (avec TOC) supérieurs à ceux du système de référence (sans TOC).

ANNEXE 9

RÉSULTATS COMPLETS DES TESTS DE RAFFINEMENT (3ème partie) DU SYSTÈME DE RECONNAISSANCE HYBRIDE L

Voici les résultats complets obtenus pendant les seconds tests de raffinement du système de reconnaissance hybride L. Les performances sont calculées pour chaque échelle de la TOC, par pas de 0.01, de l'échelle 32.1 à l'échelle 32.8, et dans le cas de populations de 50 locuteurs, pour un milieu dépourvu de bruit.

	Taux de reconnaissance du système hybride L (en %)
Ligne pour l'échelle	**Base de données de 50 locuteurs (population)**
32.1	92
32.11	82
32.12	90
32.13	86
32.14	88
32.15	80
32.16	92
32.17	92
32.18	88
32.19	84
32.2	90
32.21	84
32.22	90
32.23	86
32.24	82
32.25	80
32.26	92
32.27	84
32.28	92
32.29	86
32.3	86
32.31	88
32.32	90
32.33	92
32.34	88
32.35	84

32.36	80
32.37	82
32.38	86
32.39	78
32.4	86
32.41	90
32.42	82
32.43	90
32.44	90
32.45	90
32.46	88
32.47	88
32.48	88
32.49	90
32.5	92
32.51	86
32.52	90
32.53	92
32.54	92
32.55	92
32.56	88
32.57	90
32.58	92
32.59	90
32.6	92
32.61	92
32.62	90
32.63	92
32.64	90
32.65	90
32.66	90
32.67	86
32.68	84
32.69	84
32.7	94
32.71	92
32.72	86
32.73	84
32.74	88
32.75	76
32.76	86
32.77	84

32.78	86
32.79	82
32.8	90

Tableau XXXI Performances du système hybride L en fonction de l'échelle de la ligne de la TOC, pour des lignes dont les échelles sont à deux chiffres significatifs après la virgule, et pour une population de 50 locuteurs

Les cases ombragées correspondent à des résultats du système hybride L (avec TOC) supérieurs à ceux du système de référence (sans TOC).

ANNEXE 10

L'ESTIMATEUR DE DENSITÉ DE KERNEL

L'idée de cet estimateur est de centrer une fonction de densité locale, appelée Kernel, pour chaque point des données. La somme de ces estimées de densité locales donne l'estimée de la fonction de densité de l'ensemble des données.

Il existe deux paramètres de modélisation : la fonction de Kernel, et sa largeur (que l'on appelle également paramètre de lissage).

Théoriquement, on écrit :

$$\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right) \tag{J.1}$$

ou K est la fonction de Kernel,
et h le paramètre de lissage.

Généralement, on utilise une fonction de Kernel Gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \tag{J.2}$$

BIBLIOGRAPHIE

[1] Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing : a guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall PTR.

[2] Martin A., Przybocki M., 1997 speaker recognition evaluation NIST workshop, Maritime Institute Linthicum, Maryland. 25-26 June 1997.

[3] J. Trémolière, Electronique Applications N°62, « La synthèse de la parole », octobre 1988.

[4] www.r.battault.free.fr/probatoire

[5] Campbell, J. P. J. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE, 85*(9), 1437-1462.

[6] Van Den Heuvel H., Rietveld T., Cranen B., Methodological aspects of segment and speaker-related variability. A study of segmental durations in Dutch. Journal of Phonetics, n° 22, pp 389-406. 1994.

[7] Atal, B. S. (1972). *Text-independent speaker recognition.* Paper presented at the Program of the 83rd meeting of the Acoustical Society of America. Abstracts only, 18-27 April 1972, Buffalo, NY, USA.

[8] Matsui T., Furui S., Text-independent speaker recognition using vocal tract and pitch information. In Proceedings ICSLP 90, pp 137-140, 1990.

[9] Calliope (nom collectif représentant les 36 auteurs de cet ouvrage), La parole et son traitement automatique, Editions Masson, 1989.

[10] R Carré, J.F. Dégremont, M. Gross, J.M. Pierrel et G.Sabah, C.N.R.S plus, « Langage humain et machine », p217-p230, 1991.

[11] Boucheteau R., Probatoire CNAM, « Analyse et synthèse de la parole », avril 1989.

[12] Zue, course notes at MIT: 6.345 Automatic Speech recognition, 2005

[13] http://www.speech.kth.se/wavesurfer

[14] Thierry Dutoit, Introduction au traitement automatique de la parole, notes de cours, faculté polytechnique de Mons, 2000.

[15] Jean Hennebert, Traitement de la parole, SE 2005, Université de Fribourg, Suisse.

[16] Allen "Harvey fletcher" The ASA edition of speech and hearing communication 1995

[17] Fletcher "auditory patterns" , 1940,12, pp 47-65

[18] Carter Paul, Structured Variation in British English Liquids : the role of resonance, 2002

[19] Stevens, Volkman, " the relation of pitch to frequency", Journal of psychology, 1940,53,pp 329.

[20] www.en.wikipedia.org/wiki/Mel_scale

[21] Psaromiligkos Y., course ECSE 509, Probability and Random Signals II, McGill University, Autumn 2004.

[22] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 28*(4), 357-366.

[23] Furui, S. (2001). *Digital speech processing, synthesis, and recognition* (2nd ed.). New York: Marcel Dekker.

[24] Hollien, H. (1990). Phonetician as expert witness. Ethics and responsibilities. *Symposium on the Use of the Language Scientist as Expert in the Legal Setting, Apr 23 1988 Annals of the New York Academy of Sciences, 606*, 33.

[25] Kunzel, H. J. (1994). *Current approaches to forensic speaker recognition.* Paper presented at the Proceedings of Workshop on Automatic Speaker Recognition, Identification and Verification, 5-7 April 1994, Martigny, Switzerland.

[26] Speech Communications : Human & Machine by Douglas O'Shaughnessy, IEEE Press, Hardcover 2nd edition, 1999; ISBN: 0780334493.

[27] Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE, 64*(4), 460-475.

[28] Doddington, G. R. (1985). Speaker recognition-identifying people by their voices. *Proceedings of the IEEE, 73*(11), 1651-1664.

[29] O'Shaughnessy, D. (1986). Speaker recognition. *ASSP Magazine, IEEE [see also IEEE Signal Processing Magazine], 3*(4), 4-17.

[30] Furui, S. (1994). *An overview of speaker recognition technology.* Paper presented at the Proceedings of Workshop on Automatic Speaker Recognition, Identification and Verification, 5-7 April 1994, Martigny, Switzerland.

[31] Assessment of speaker verification systems, In EAGLES Spoken Language Systems, Eagles Document EAG-SLWG-Handbook Phase 2. February 1995.

[32] Naik, J. M. (1990). Speaker verification: A tutorial. *IEEE Communications Magazine, 28*(1), 42-48.

[33] Rosenberg, A. E. (1976). Automatic speaker verification: a review. *Proceedings of the IEEE, 64*(4), 475-487.

[34] Bimbot F., Paoloni A., Chollet G., Assessment Methodology for Speaker Identification and Verification Systems. Technical report – Task 2500 – Report 19, SAM-A ESPRIT Project 6819. 1993.

[35] Bimbot, F., & Mathan, L. (1994). *Second-order statistical measures for text-independent speaker identification.* Paper presented at the Proceedings of Workshop on Automatic Speaker Recognition, Identification and Verification, 5-7 April 1994, Martigny, Switzerland.

[36] B. Corona, J. Torry, M. Hind, R. Vincent, *«Effects of respiration on heart sounds using time-frequency analysis»*, IEEE Trans., pp. 457-460, 2001.

[37] Badri, N. (2002). *Utilisation de la transformee de Fourier et de la transformee en ondelettes pour la reconnaissance du locuteur (French text).* Unpublished MIng, ECOLE DE TECHNOLOGIE SUPERIEURE, Montreal.

[38] Tierney J., "A study of LPC analysis of speech in additive noise". IEEE Trans 1980;ASSP-28:389–397.

[39] Premakanthan, P., & Mikhael, W. B. (2001). *Speaker verification/recognition and the importance of selective feature extraction: Review.* Paper presented at the 2001 Midwest Symposium on Circuits and Systems (MWSCAS 2001), Aug 14-17 2001, Dayton, OH.

[40] Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing, 2*(4), 639-643.

[41] Rabiner, L. R. (1989). A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. *Proceedings of the Ieee, 77*(2), 257-286.

[42] B. Wildermoth and K.K. Paliwal, ``GMM based speaker recognition on readily available databases'', Proc. Microelectronic Engineering Research Conference, Brisbane, Australia, Nov. 2003

[43] Y. Fang, A GMM speaker Recognition System Using HTK V1.5, (1997).

[44] Shridhar, M., Mohankrishnan, N., & Baraniecki, M. (1981). *Text-independent speaker recognition using orthogonal linear prediction.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '81.

[45] Akay, M., & Mello, C. (1997). *Wavelets for biomedical signal processing.* Paper presented at the Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. `Magnificent Milestones and Emerging Opportunities in Medical Engineering', 30 Oct.-2 Nov. 1997, Chicago, IL, USA.

[46] Flanagan, J. L. (1972). *Speech analysis; synthesis and perception* (2nd ed.). Berlin, New York,: Springer-Verlag.

[47] Padsta, Radova, Comparison of several speaker verification procedures based on GMM.

[48] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication, 17*(1-2), 91-108.

[49] course ECE 8463, Fundamentals of speech recognition, College of Engineering, Mississipi State University, 2005.

[50] Soong, F. K., & Rosenberg, A. E. (1986). *ON THE USE OF INSTANTANEOUS AND TRANSITIONAL SPECTRAL INFORMATION IN SPEAKER RECOGNITION.* Paper presented at the ICASSP 86 - Proceedings, IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing., Tokyo, Jpn.

[51] Matsui, T., & Furui, S. (1992). Text-independent speaker recognition using vocal tract and pitch information. *Transactions of the Institute of Electronics, Information and Communication Engineers A, J75-A*(4), 703-709.

[52] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72-83.

[53] Reynolds, D. A. (1992). *A Gaussian mixture modeling approach to text-independent speaker identification.* Unpublished 9303140, Georgia Institute of Technology, United States -- Georgia.

[54] Schmidt M., Golden J., Gish H., GMM sample statistic log-likelihood for Text independent speaker recognition. In Proc. Eurospeech 97, Rhodes, Greece. September 1997.

[55] Lamel, L., & Gauvain, J.-L. (1997). *Speaker recognition with the Switchboard Corpus.* Paper presented at the Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. Part 2 (of 5), Apr 21-24 1997, Munich, Ger.

[56] Schmidt, M. S. (1997). *Identifying speakers with support vector networks.* Paper presented at the Proceedings of 28th Symposium on the Interface of Computing Science and Statistics (Graph-Image-Vision), 8-12 July 1996, Sydney, NSW, Australia.

[57] Van Vuuren, S. (1996). *Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch.* Paper presented at the Proceedings of the 1996 International Conference on Spoken Language Processing, ICSLP. Part 3 (of 4), Oct 3-6 1996, Philadelphia, PA, USA.

[58] McLachlan, G. J., & Basford, K. E. (1988). *Mixture models : inference and applications to clustering.* New York, N.Y.: M. Dekker.

[59] L. Schwardt, course TW414 / PR 813 Clustering, University of Stellenbosch, March 2003.

[60] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society series B*, 39:1-38, 1977.

[61] Moon, T. K. (1996). The expectation-maximization algorithm. *Signal Processing Magazine, IEEE, 13*(6), 47-60.

[62] Rose R. C., course ECSE 570, Automatic Speech Recognition, McGill University, Winter 2005.

[63] Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition.* Englewood Cliffs, N.J.: PTR Prentice Hall.

[64] Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals.* Englewood Cliffs, N.J.: Prentice-Hall.

[65] D. Chow, W. H. Abdulla, Robust speaker identification based on perceptual log area ratio and Gaussian mixture models, IEEE Transactions.

[66] Linde, Y., Buzo, A., & Gray, R. (1980). An Algorithm for Vector Quantizer Design. *Communications, IEEE Transactions on [legacy, pre - 1988], 28*(1), 84-95.

[67] Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, 34*(1), 52-59.

[68] Wildermoth B. R., Paliwal K., GMM based speaker recognition on readily available databases.

[69] Soong, F., Rosenberg, A., Rabiner, L., & Juang, B. (1985). *A vector quantization approach to speaker recognition.* Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.

[70] Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

[71] Bentley, P. M., & McDonnell, J. T. E. (1994). Wavelet transforms: an introduction. *Electronics & Communication Engineering Journal, 6*(4), 175-186.

[72] Hubbard, ondes et ondelettes, la saga d un outil mathématique, Belin, 2000.

[73] B. Corona, J. Torry, «Time-frequency representation of systolic murmurs using wavelets», IEEE Trans., pp. 601-604, 1998.

[74] Burrus, C. S., Gopinath, R. A., & Guo, H. (1998). *Introduction to wavelets and wavelet transforms : a primer*. Upper Saddle River, N.J.: Prentice Hall.

[75] Grossmann, A., and Morlet, J. : "Décomposition of Hardy functions into suqare integrable wavelets of constant shape", SIAM J. Math. Anal., 1984, 15, pp.723-736.

[76] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 11*(7), 674-693.

[77] Meyer, Y., & Salinger, D. H. (1992). *Wavelets and operators*. Cambridge England ; New York: Cambridge University Press.

[78] Daubechies, I., *«Orthogonal bases of compactly supported wavelets»*, Communications on pure and applied mathematics, vol 41, pp 909-996, november 1988.

[79] Special issue of IEEE Proceedings, 84, 4, April 1996.

[80] Strang, G. (1994). Wavelets. *American Scientist, 82*(3), 250-255.

[81] Bradley, J. N., & Brislawn, C. M. (1994, 1994//). *The wavelet/scalar quantization compression standard for digital fingerprint images.* Paper presented at the Proceedings of IEEE International Symposium on Circuits and Systems - ISCAS '94, 30 May-2 June 1994, London, UK.

[82] Adams, M. D., & Ward, R. (2001). *Wavelet transforms in the JPEG-2000 standard.* Paper presented at the 2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 26-28 Aug. 2001, Victoria, BC, Canada.

[83] Kaiser, G. (1994). *A friendly guide to wavelets*. Boston: Birkh©Þuser.

[84] Weiss, L. G. (1994). Wavelets and wideband correlation processing. *Signal Processing Magazine, IEEE, 11*(1), 13-32.

[85] Sheng, Y. WAVELET TRANSFORM. In: The transforms and applications handbook. Ed. by A. D. Poularikas. P. 747-827. Boca Raton, Fl (USA): CRC Press, 1996. The Electrical Engineering Handbook Series.

[86] Hubbard, B. B. (1998). *The world according to wavelets : the story of a mathematical technique in the making* (2nd ed.). Wellesley, Mass: A.K. Peters.

[87] Tan B. T., Fu M. Spray A., The use of wavelet transforms in phoneme recognition, IEEE Transactions.

[88] Vetterli, M., & Herley, C. (1992). Wavelets and filter banks: theory and design. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on], 40*(9), 2207-2232.

[89] Kœrner, T. W. (1988). *Fourier analysis*. Cambridge Cambridgeshire ; New York: Cambridge University Press.

[90] Calderbank, R., Daubechies, I., Sweldens, W., & Yeo, B.-L. (1998). Wavelet transforms that map integers to integers. *Applied and Computational Harmonic Analysis, 5*(3), 332-369.

[91] T. Edwards, *«Discrete Wavelet Transforms : Theory and implementation»*,Rapport de recherche, Stanford University, Palo Alto, juin 1992.

[92] Sarkar, T. K., & Su, C. (1998). A tutorial on wavelets from an electrical engineering perspective .2. The continuous case. *IEEE Antennas and Propagation Magazine, 40*(6), 36-49.

[93] Truchelet, F. , *«Ondelettes pour le signal numérique»*, Hermes, 1998.

[94] Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on, 36*(5), 961-1005.

[95] Szekely, G., Eide, A., Lindblad, T., Lindsey, C., Minerskjold, M., & Sekhniaidze, G. (1996). Wavelets and signal processing. *Proceedings of the SPIE - The International Society for Optical Engineering Applications and Science of Artificial Neural Networks II, 9-12 April 1996, 2760*, 625-632.

[96] Daubechies, I., « Orthonormal Bases of Compactly Supported Wavelets", Comm. In Pure and Applied Math., Vol.41. No.7. pp.909-996, 1988.

[97] Markowitz, J. A. (1998). Speaker recognition. *Information Security Technical Report, 3*(1), 14-20.

[98] Meyer. Y., "Orthonormal wavelets", pp.21-37, 1989

[99] E. Zwicker (1961) "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)", *J. Acoust. Soc. Am. 33*: 248. Also in E. Zwicker und R. Feldtkeller (1967) *Das Ohr als Nachrichtenempfänger* 2. Aufl., Stuttgart: Hirzel, p. 74.

[100] Chui, C. K. (1992). *An introduction to wavelets*. Boston: Academic Press.

[101] Rioul, O. (1993). A discrete-time multiresolution theory. *IEEE Transactions on Signal Processing, 41*(8), 2591-2606.

[102] B. Torrésani, *«Analyse continue par ondelettes»*, CNRS éditions, Septembre 1995.

[103] Foufoula-Georgiou, E., & Kumar, P. (1994). *Wavelets in geophysics*. San Diego: Academic Press.

[104] Damiano, B., Kercel, S. W., Tucker, R. W., Jr., & Brown-VanHoozer, S. A. (2000). *Recognizing a voice from its model.* Paper presented at the Systems, Man, and Cybernetics, 2000 IEEE International Conference on.

[105] J. Lindh, Handling the "voiceprint" issue, Proceedings, FONETIK 2004, Dept. of Linguistics, Stockholm University.

[106] Hachette Multimedia, 2004.

[107] Gargour C., Ramachandran V., Traitement numérique des signaux, Montréal : École de technologie supérieure , 2001.

[108] Goswami, J. C. (1999). *Fundamentals of wavelets : theory, algorithms, and applications* (J. Wiley and Sons ed.). New York, N.Y.

[109] Mertins, A. (1999). *Signal analysis : wavelets, filter banks, time-frequency transforms and applications* (J. Wiley ed.). Chichester, England.

[110] Harris, J. G. (Spring 2004). *EEL 6586: Automatic Speech Processing, University of Florida*, from http://www.cnel.ufl.edu/hybrid/harris.html

[111] Aviyente, S., & Williams, W. J. (2005). Minimum entropy time-frequency distributions. *Signal Processing Letters, IEEE, 12*(1), 37-40.

[112] Williams, W. J., Brown, M. L., & Hero, A. O., III. (1991). Uncertainty, information and time-frequency distributions. *Proceedings of the SPIE - The International Society for Optical Engineering Advanced Signal Processing Algorithms, Architectures and Implementations II, 24-26 July 1991, 1566*, 144-156.

[113] Digital signal and image processing, Bose, Tamal, Hoboken, N.J. : J. Wiley, c2004

[114] Misra, H., Ikbal, S., Bourlard, H., & Hermansky, H. (2004). *Spectral entropy based feature for robust ASR.* Paper presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on.

[115] Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

[116] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297.