République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire de signal & Communications

Thèse de Doctorat en Electronique

Présentée par

Mr RAMOU Naim

Magister en Electronique EMP Bordj El Bahri ALGER
Attaché de recherche au CSC-Chéraga ALGER

Intitulée :

Classification des Troubles de la Prononciation chez l'Enfant Algérien

Soutenue publiquement le 28 juin 2015, devant le jury composé de :

Présidente :	HAMAMI Latifa	Prof	ENP Alger
Rapporteur :	GUERTI Mhania	Prof	ENP Alger
Examinateurs :	FERGANI Belkacem	MCA	USTHB Alger
	HALIMI Mohammed	DR	CSC Chéraga
	SAYOUD Halim	Prof	USTHB Alger
	BOUSBIA-SALAH Hicham	MCA	ENP ALGER

ENP 2015

Ecole Nationale Polytechnique
10, Avenue des Frères Ouked, Hassen Badi, BP 182 16200 El-Harrach Alger Algérie
www.enp.edu.dz

DEDICACES

Je dédie cet humble travail à :

Mes Parents,

Mon épouse et mon fils Abdelrrahim,

Mes frères et mes sœurs, amis et collègues,

À qui je suis sincèrement reconnaissant pour tout ce qu'ils ont fait pour moi

A vous tous, je suis fier de vous avoir

.

REMERCIEMENTS

Plus qu'un document et un travail de recherche, une thèse est une expérience personnelle unique où les relations sociales prennent une place centrale dans sa réussite.

Je tiens tout spécialement à remercier le Bon Dieu pour tout.

J'exprime toute ma gratitude à ma directrice de thèse, le Professeur GUERTI Mhania, enseignante à l'ENP, pour son soutien et sa confiance en mon travail et en mes idées. Sa direction de thèse m'a permis, au bout de ces années, d'acquérir la confiance nécessaire pour ma future carrière de chercheur.

J'exprimer ma profonde reconnaissance à Mme HAMAMI Latifa, Professeur à l'ENP, pour l'honneur qu'elle me fait en acceptant de présider le jury de cette thèse.

Je remercie vivement les membres examinateurs pour avoir accepté de participer à mon jury de thèse, Monsieur :

HALIMI Mohammed, Dr au CSC de Chéraga ;

SAYOUD Halim, Pr à l'USTHB ;

FERGANI Belkacem, MCA à l'USTHB ;

BOUSBIA-SALAH Hicham MCA à l'ENP.

Je remercie également Mme TRABELSI Ghania., Médecin Praticienne au service ORL CHU Lamine Debaghine Alger, pour son aide lors des enregistrements des corpus.

Que ceux qui m'ont aidé de près ou de loin trouvent ici l'expression de mes profonds remerciements !

Merci !

ملخص :
الهدف من عملنا هو تكوين نظام آلي لتصنيف اضطرابات الكلام لدى الأطفال. لذلك اعتمدنا على فكرة أن هذه الاضطرابات يمكن اعتبارها كظاهرة لهجة إقليمية، حيث يمكننا استخدام أساليب التعرف الآلي على المتكلم للبرنامج lia_spkdو ALIZE. المرحلة الأولى من هذه الدراسة تتمثل في تحليل الصوت لتسجيلات المرضى قصد استخلاص الخصائص LPCC ;والتي يمكن استعمالها كمدخل للمصنِّف. المرحلة الثانية تتعلق بتطبيق المصنِّف الهجين SVM و GMM-UBM ;التسجيلات المستعملة مأخوذة من قواعد بيانات، مسجلة من طرف WMIT ، وأخرى مسجلة في إطار هذا العمل تحت إشراف طبيب متخصص في علم الأرطوفونيا بمستشفى ليمين ذباغين-الجزائر-العاصمة تتشكل من كلمات باللغة العربية الموحدة تمثل الإنسدادات الصوتية التي نريد التعامل معها sigmatisme constrictifأو occlusif. التجارب على التسجيلات أعطت نسبة 95.8 % من التعرف على الكلام المضطرب بالإضافة إلى نسبة 93,75 % من التعرف على الأصوات المضطربة

كلما ت المفاتيح : اضطرابات الكلام . تصنيف.,SVM, GMM-UBM, LPCC, ALIZE, LIA_SpkD .

Résumé : L'objectif de notre travail est l'implémentation d'un système automatique de classification des troubles articulatoires de la parole chez l'enfant algérien. Pour cela, nous nous sommes basés sur l'idée que ces troubles peuvent être considérés comme un phénomène d'accent régional. Dans ce cas nous pouvons utiliser les techniques de la Reconnaissance Automatique du Locuteur de la plateforme LIA_SpkD et ALIZE. La première étape de notre étude consiste à effectuer une analyse acoustique des corpus enregistrés par des patients afin d'extraire les caractéristiques spectrales : LPCC (**L**inear **P**redictive **C**epstral **C**oefficients) qui sont utilisées comme entrée pour le classificateur. La seconde étape concerne l'application des classificateurs hybrides SVM (**S**upport **V**ector **M**achines) et les GMM-UBM (**G**aussian **M**ixture **M**odel-**U**niversal **B**ackground **M**odel). Nos corpus sont extraits à partir de deux Bases de Données, la première enregistrée par WMIT (**W**are **M**assachusetts **I**nstitute of **T**echnology), et la seconde enregistrée sous la supervision d'un orthophoniste au CHU Lamine Debaghine d'Alger. Ces corpus sont constitués de mots en Arabe Standard représentant la pathologie que nous voulons traiter : Sigmatisme constrictif ou occlusif. Les expériences menées ont donné un TRG (**T**aux de **R**econnaissance **G**lobal) de 95.8% pour la classification de la parole normale et pathologique ainsi qu'un TRG de 93,75% pour la classification du phonème pathologique en question.

Mots clés : Troubles articulatoires de la parole, Classification, LIA_SpkD, ALIZE, LPCC, GMM-UBM, SVM.

Abstract: The objective of our work is the implementation of an automatic system of speech children articulatory disorders classification. For it we based ourselves on the idea that these disorders can be considered as a phenomenon of regional accent. In this case we can use the techniques of the Automatic Speaker Recognition of the platform LIA_SpkD and ALIZE. The first stage of our study consists in making an acoustic analysis of patients corpus to extract the spectral characteristic : LPCC (**L**inear **P**redictive **C**epstral **C**oefficients) which are used as entry for the classifier. The second stage concerns the application of the hybrid classifiers SVM (**S**upport **V**ector **M**achines) and the GMM-UBM (**G**aussian **M**ixture **M**odel-**U**niversal **B**ackground **M**odel). Our corpus are delivered by two databases, the first recorded by Ware Massachusetts Institute of Technology, the second recorded within the framework of this work under the supervision of a speech therapist CHU Lamine Debaghine-Algiers. These records constituted by words in Standard Arabic language representing the pathology that we want to treat (constrictive or occlusive Sigmatisme). The experiences led on database gave a TRG of 95.8 % recognition for the classification of the normal and pathological word as well as a TRG of 93,75 % recognition for the classification of the pathological phoneme.

Key Words: Articulations disorders, Speech, Classification, LIA_SpkD, ALIZE, LPCC, GMM-UBM,SVM.

Table Des Matières

LISTE DES ABREVIATIONS VI

LISTE DES FIGURES VII

LISTE DES TABLEAUX IX

INTRODUCTION GENERALE 11

ETAT DE L'ART 13

CHAPITRE 1 : GENERALITES SUR LA PAROLE .....

1. Introduction 16

2. Fonctionnement du système phonatoire humain .....

2.1. Respiration .....

2.2. Organes Phonatoires 17

2.2.1. Larynx 18

2.2.2. Cordes vocales 19

2.2.3. Conduit vocal 20

3. Caractéristiques Acoustiques de la parole .....

3.1. Intensité Sonore 21

3.2. Fréquence fondamentale .....

3.3. Timbre vocal 22

3.4. Durée .....

3.5. Formants et Transitions Formantiques .....

4. Sons voisés et non-voisés 23

4.1. Sons voisés .....

4.2. Sons non voisés 24

5. Classification des sons du langage 25

5.1. Voyelles 26

5.2. Consonnes 27

5.3. Semi-voyelles 28

6. Description des sons de l'Arabe Standard .....

7. Troubles de la parole 29

7.1. Classification des troubles 30

7.2. Défauts de la voix détectés par l'oreille 31

8. Troubles du Sigmatisme 32

8.1. Sigmatisme des consonnes constrictives 34

8.2. Sigmatisme des consonnes occlusives .....

9. Conclusion .....

CHAPITRE 2 : ANALYSE PARAMETRIQUE DU SIGNAL VOCAL .....

1. Introduction 36

2. Production de la parole humaine .....

2.1. Production de l'onde glottique 37

2.2. Fonction résonateur du conduit vocal .....

2.3. Fonction générateur de bruit du conduit vocal 38

3. Modèle de la production de la parole par LPC .....

4. Traitement du signal vocal 39

4.1. Analyse LPCC 40

4.1.1. Échantillonnage du signal vocal 41

4.1.2. Préaccentuation 42

4.1.3. Segmentation en trames 43

4.1.4. Fenêtrage 44

4.1.5. Analyse acoustique LPCC 45

4.1.6. Détection des zones de la parole .....

4.2. Caractéristiques Dynamiques et Mesure d'Énergie 46

5. Conclusion 47

CHAPITRE 3 : TECHNIQUES DE LA RAL APPLIQUEES A LA CLASSIFICATION DES TROUBLES DE LA PAROLE..

1. Introduction 49

2. Reconnaissance Automatique du Locuteur (RAL) .....

2.1. Phase de Paramétrisation 50

2.1.1. Analyse spectrale 51

2.1.2. Analyse cepstrale .....

2.1.3. Paramètres dynamiques .....

2.2. Phase d'apprentissage des modèles 52

2.2.1. Modélisation par Mélange de Gaussiennes .....

2.2.2. Machines à Vecteurs de Support 56

2.2.3. Système hybride GMM-SVM 65

2.3. Phase de tests et mesure de ressemblance 67

3. Adaptation des techniques de la RAL à la classification des troubles de la parole 68

3.1. Modèle des troubles de la parole, Monde et Décisions 69

3.2. Tâches de classification .....

3.3. Prise de décision 70

4. conclusion 71

CHAPITRE 4 : IMPLEMENTATION DU SYSTEME DE CLASSIFICATION DES TROUBLES DE LA PAROLE (SCTP) .....

1. Introduction 73

2. Architecture de l'implémentation du SCTP .....

2.1. Paramétrisation 75

2.2. Détection d'énergie .....

2.3. Normalisation des paramètres 76

2.4. Modèle du monde 77

2.5. Modèle du locuteur 78

2.6. Tests et scores.....

3.8. Vecteurs des moyennes des GMM 79

2.7. Modèles SVM.....

3. Classification Contrôle/Pathologique 80

3.1. Description de la Base de Données (WMIT) .....

3.2. Paramétrage des corpus de WMIT 81

3.3. Modélisation des données 82

3.4. Décision et calcul des performances .....

4. Classification des phonèmes étudiés 85

4.1. Description de la Base de Données enregistrée .....

4.2. Paramétrage des corpus..... 89

4.3. Modélisation des données .....

4.4. Décision et calcul des performances .....

5. Conclusion..... 92

CONCLUSIONS ET PERSPECTIVES 94

REFERENCES BIBLIOGRAPHIQUES..... 97

Liste des Abréviations

API : **A**lphabet **P**honétique **I**nternational

EM : **E**xpectation-**M**aximization

GMM-UBM : **G**aussian **M**ixture **M**odel-**U**niversal **B**ackground **M**odel

IIR : **I**nfinite **I**mpulse **R**esponse
Filtre à **R**éponse **I**mpulsionnelle **I**nfinie

KL : **K**ullback-**L**eibler

LFC : **L**inear **F**requency **C**epstral

LFCC : **L**inear **F**requency **C**epstral **C**oefficients

LPC : **L**inear **P**redictive **C**oding

LPCC : **L**inear **P**redictive **C**epstral **C**oefficients

L_N : **L**ocuteur **N**ormal

L_p : **L**ocuteur **P**athologique

LS_N : **L**ocuteurs **N**ormaux

LS_p : **L**ocuteurs **P**athologiques

MAP : **M**aximum a **P**osteriori

MFC : **M**el **F**requency **C**epstral

MFCC : **M**el **F**requency **C**epstral **C**oefficients

ML : **M**aximum **L**ikelihood

NIST : **N**ational **I**nstitute of **S**tandards and **T**echnologies

RAL : **R**econnaissance **A**utomatique du **L**ocuteur

RAP : **R**econnaissance **A**utomatique de la **P**arole

RBF : **R**adial **B**asis **F**unction

ROC : **R**eceiver **O**perating **C**haracteristic

SVM : **S**upport **V**ector **M**achines

SCTP : **S**ystème de **C**lassification des **T**roubles de la **P**arole

TR : **T**aux de **R**econnaissance

TRG : **T**aux de **R**econnaissance **G**lobal

WMIT : **W**are **M**assachusetts **I**nstitute of **T**echnology

Liste des Figures

Figure 1.1 : Organes Phonatoires....18

Figure 1.2 : Schéma du larynx....19

Figure 1.3 : Sonagrammes des voyelles [a], [i] et [u]....23

Figure 1.4 **:** Spectre de la voyelle [a]....24

Figure 1.5 : Spectre de [ʃ]....

Figure 1.6 : Classification des sons du langage....25

Figure 1.7 : Système vocalique de l'Arabe Standard....26

Figure 1.8 : Diagramme de classement des pathologies....30

Figure 1.9 : Lieux d'articulation des phonèmes [S, ʃ, Ṣ, Θ, Z, ž]....34

Figure 2.1 : Vue schématique du système de production de la voix humaine....38

Figure 2.2 : Modèle de production de la parole....39

Figure 2.3 **:** Représentation spectrale des LPC et des fréquences des formants....40

Figure 2.4 : Etapes d'extraction des coefficients LPCC....42

Figure 2.5 : Échantillonnage du signal vocal....

Figure 2.6 : Segmentation en trames....44

Figure 2.7 : Représentation fréquentielle des fenêtres....45

Figure 2.8 : Détection des zones de la parole....47

Figure 3.1. Paramètres acoustiques statiques et dynamiques....51

Figure 3.2. Phase d'apprentissage d'un système de RAL....52

Figure 3.3 : EM : algorithme itératif....53

Figure 3.4 : Adaptation du modèle du monde selon les paramètres extraits du signal d'apprentissage....56

Figure 3.5 : Exemple de projection des données dans un espace plus grand....58

Figure 3.6. Données linéairement séparables....59

Figure 3.7 : Structure générale d'un système SVM-GMM....67

Figure 3.8 : Supervecteur des moyennes GMM....68

Figure 3.9 : Des vecteurs acoustiques au vecteurs SVM....69

Figure 3.10 : Distributions des scores pour un locuteur cible et des imposteurs....71

Figure 4.1 : Architecture de l'application SCTP inspirée de la plateforme ALIZE 74

Figure 4.2 : Fichier de paramétrisation des corpus. 75

Figure 4.3 : Fichier de configuration de détection d'énergie. 76

Figure 4.4 : Fichier de configuration pour la normalisation ds paramètres 77

Figure 4.5 : Génération de la liste des corpus du modèle de monde 78

Figure 4.6 : Génération des listes pour l'apprentissage des modèles de la parole .....

Figure 4.7 : Génération des fichiers pour la pahase de test.....

Figure 4.8 : Création des vecteurs des GMM.....

Figure 4.9 : Apprentissage des modèles SVM. 80

Figure 4.10 : Caractéristiques acoustiques observées de la consonne[s] dans le mot "sève" pour normal (à gauche) et pathologique (à droite) : (a) la représentation du temps, (b) l'intensité spectrale, (c) les LPC 81

Figure 4.11 : Influence de de coefficients cepstraux sur les performances de system. 83

Figure 4.12 : Distribution des scores pour le modèle du LP en utilisant les : a) GMM-UBM, b) GMM-SVM.....85

Figure 4.13 : Distribution des scores pour le modèle du LN en utilisant les : a) GMM-UBM, b) GMM-SVM.....

Figure 4.14 : Courbe ROC pour le modèle du LP en utilisant les : a) GMM-UBM ; b) GMM-SVM.....85

Figure 4.15 : Spectrogramme de [ʃaχsija] and [ΘaχΘija].....86

Figure 4.16 : Spectrogramme de [ʃamson] and [[ʃa m ʃø].....87

Figure 4.17 : Spectrogramme de [ΘalaΘa] and [salasa].....

Figure 4.18 : Segmentation semi automatique du mot [ʃaχsija] 88

Figure 4.19 : Performances de system GMM-UBM pour des LS_N et LS_P en termes de la courbe ROC.....90

Liste des Tableaux

Tableau 1.1 : Transcription des phonèmes de l'Arabe Standard en API 29

Tableau 4.1 : Performance du système de la classification Contrôle et Pathologique 83

Tableau 4.2 : Transcription phonétique des mots prononcé par un L_P.....87

Tableau 4.3 : Matrice de confusion pour des LS_N 91

Tableau 4.4 : Matrice de confusion pour des LS_P .....

Tableau 4.5 : Matrice de confusion entre les LS_N et LS_P .....

INTRODUCTION GENERALE

D'après des recherches faites aux niveaux des centres hospitaliers et des hôpitaux universitaires sur le territoire national qui possèdent des orthophonistes, un tiers des enfants en Algérie souffrent des troubles de la parole. La plupart de ces patients consultent quand la personne ou son entourage entend des changements dans le résultat vocal, uniquement sur des sensations perceptives [1]. La perception auditive est la modalité première, la plus accessible, pour évaluer la qualité vocale [2]. En effet, plusieurs auditeurs sont requis afin d'obtenir une appréciation moyenne ou consensuelle plus représentative de l'état vocal qu'un jugement isolé [3]. De ce fait, une analyse perceptive fiable s'avère consommatrice en temps et en ressources humaines. Des approches instrumentales dites : objectives, fondées sur de la mesure physique ont été proposées pour pallier les faiblesses précédemment décrites de l'évaluation perceptive [4]. Cette analyse est conçue pour qualifier les dysfonctionnements vocaux à partir des mesures acoustiques réalisées sur le patient en cours de production vocale [5-7]. Tout comme pour l'évaluation perceptive, les techniques instrumentales comportent un certain nombre de limites. Tout d'abord, la plupart des analyses sont fondées sur la production de voyelles tenues, contexte d'élocution très éloigné de la parole continue [8]. Récemment, ces restrictions ont conduit, à tester l'adaptation des techniques issues de la RAL [9] sur des locuteurs ayant des troubles de la parole [10-12]. Notre objectif est de proposer une méthode mieux adaptée au suivi de la pathologie des patients : facile et rapide à utiliser, non contraignante pour le patient et accessible pour les cliniciens. Le système conçu pour cette tâche particulière s'appuie sur l'approche de GMM-SVM [13,14]. Il est issu des outils de RAL (LIA_Spk Det et ALIZE). Nous allons, à travers ce travail contribué à la réalisation d'un **S**ystème automatique de **C**lassification des **T**roubles articulatoires de la **P**arole. Cette thèse est organisée en quatre chapitres :

- le premier expose quelques généralités sur la parole et la description de diverses pathologies vocale.
- le deuxième est consacré à l'extraction des informations utiles à la caractérisation des voix pathologiques.
- le troisième porte sur l'adaptation des techniques de la RAL à la classification des troubles de la parole.
- le dernier chapitre concerne la description du corpus des patients utilisés dans cette étude ainsi que l'implémentation du SCTP.

Finalement, nous présentons des conclusions générales et perspectives.

ETAT DE L'ART

La reconnaissance de la parole pathologique et la perception des causes de sa dégradation à travers différents indices acoustiques ont toujours été la préoccupation clinique principale des Phoniâtres. Comme dans les autres disciplines médicales, ces derniers ont été attentifs à toutes les techniques qui seraient susceptibles de leur donner des informations complémentaires, pour aider au diagnostic et évaluer les effets des traitements chirurgicaux et médicamenteux ou les progrès des rééducations

Il existe une grande variété de méthodes pour établir un bilan vocal de personnes atteintes de troubles articulatoires : interrogatoire avec le patient, examen endoscopique du larynx [15], appréciation du comportement postural du patient [16], profil psychologique et étude comportementale, questionnaire d'auto-évaluation, jugement perceptif de la qualité vocale, analyse instrumentale. La multiplication des angles d'observation s'avère nécessaire pour prendre en compte l'aspect multidimensionnel de la communication parlée, une méthode prise isolément et se révélant souvent réductrice.

On peut regrouper ces méthodes dans trois catégories :

- l'évaluation perceptive : consiste à faire juger la qualité vocale de patients par des experts (phoniâtres, orthophonistes, …). Le principe est de faire lire au patient un texte normalisé dont l'énoncé enregistré est ensuite soumis en aveugle à divers juges expérimentés qui attribuent une note sur une échelle par catégorisation directe ou à travers des échelles analogiques visuelles interprétées. Une telle démarche montre une amélioration des performances du juge en réduisant la variabilité inter juges et en renforçant la concordance avec les mesures instrumentales ;
- l'évaluation instrumentale des mesures instrumentales sont effectuées sur les patients à l'aide du dispositif EVA [8] qui permet d'obtenir des mesures acoustiques primaires (F_0, intensité en dB), des mesures de stabilité laryngée (jitter, shimmer, coefficient de Lyapounov), des estimations de performance pneumo-phonatoire (étendue vocale, temps maximal de phonation) et des grandeurs aérodynamiques qui explorent de façon directe et sélective certains mécanismes comme la fuite glottique (par mesure de débit d'air oral) ou la tension de la source (par estimation de la pression sous-glottique).

- les techniques de la RAL adaptées à la reconnaissance de la parole pathologique : Au départ, ces techniques sont conçues pour vérifier ou identifier automatiquement un locuteur, dans une certaine mesure, à partir d'une de ses productions vocales. Ces techniques sont fondées sur l'hypothèse que les troubles articulatoires de la parole peuvent être considérés de façon similaire à un accent régional, et que les techniques utilisées en reconnaissance automatique du locuteur étaient capables d'appréhender un tel phénomène. Nous manquons actuellement de recul du fait de la nouveauté de cette démarche [6]. Toutefois, nous pouvons penser que, l'intérêt résiderait alors en la non nécessité de réunir un jury d'experts et de multiplier les écoutes, configuration difficile à obtenir en pratique. L'autre avantage est l'aspect déterministe de la méthode, écartant ainsi toute inconstance que l'on peut observer chez un auditeur.

CHAPITRE 1 :

GENERALITES SUR LA PAROLE

1. Introduction

Lorsqu'on parle de pathologies vocales, une référence intuitive nous fait penser à un médecin, c'est un réflexe très logique. Toutefois le domaine de la pathologie de la parole fait intervenir d'autres disciplines telle que l'électronique, la mécanique, la physique, etc. Le traitement des pathologies langagières se situe essentiellement à détecter la zone à traiter, mesurer l'ampleur de la maladie.

Dans ce chapitre avons exposé en premier l'appareil phonatoire humain. Cela nous a permis de mettre l'accent sur les éléments essentiels qui peuvent entraîner une pathologie de la parole. Nous avons décris les sons de l'Arabe Standard et les phénomènes spécifiques à cette langue Nous avons ensuite donné une brève description des diverses pathologies de la parole les plus fréquentes. A la fin, nous avons consacré une attention particulière à un cas pathologique, à savoir le sigmatisme des consonnes constrictives et occlusives.

2. Fonctionnement du système phonatoire humain

Les structures de la phonation assurent la vie par la respiration et l'alimentation. La phonation est donc une fonction superposée. Elle est faite par trois systèmes qu'il faut connaître pour faire distinction entre voix normale et pathologique respiratoire, phonatoire et le système de résonance.

2.1. Respiration

La respiration assure le renouvellement de l'air dans les poumons, qui permet l'hématose (échanges gazeux entre alvéoles et sang) et module l'ampliation thoracique, leur rôle dans la phonation est de fournir la pression d'air pour production du son [17]. Cette fonction se base sur plusieurs systèmes anatomiques. Les deux phases respiratoires sont :

- l'inspiration : c'est un terme utilisé en physiopathologie ou thérapeutique, phénomène actif, augmente l'ampliation thoracique, faisant entrer l'air frais dans les poumons ;
- l'expiration : phénomène passif, déprime l'ampliation thoracique par simple relâchement de l'action musculaire, expulsant l'air vicié.

Seule l'inspiration représente un temps respiratoire actif, commandé de façon réflexe avec une périodicité automatique dont la fréquence augmente spontanément avec

l'effort, l'émotion ou la peur. Elle module aussi le travail cardiaque. Enfin, l'air étant évacué sous pression, l'inspiration fournit l'énergie nécessaire à la phonation. Sa facilité est conditionnée par l'état de liberté ou d'encombrement des voies respiratoires inférieures et supérieures qui sont constituées respectivement par :

- l'ensemble de l'arbre trachéo-bronchite. Celui-ci est formé d'un squelette tubulaire fait d'anneaux cartilagineux empilés, unis entre eux par des membranes fibro élastiques, plus étroites que les anneaux, lui conférant une certaine souplesse et extensibilité, lui-même revêtu intérieurement d'une muqueuse ayant des propriétés sécrétoires associées aux poumons. Elles forment le soufflet respiratoire ou générateur d'énergie pour la phonation.
- le larynx, le pharynx, les fosses nasales ou accessoirement la cavité buccale, elles constituent le modulateur phonatoire et ses résonateurs.

La technique respiratoire en phonation est différente de celle au repos :

- la respiration au repos : flux d'air libre, inspiration nécessite contraction musculaire. Cependant, l'expiration est passive (forces élastiques).
- la respiration à la phonation : L'air doit être sous pression au niveau sous glottique, cela se fait par forces expiratoires et par la résistance créée par les cordes vocales. Inspiration et expiration prennent respectivement 10% et 90% du temps respiratoire. La phonation dépend de la quantité des mots ou syllabes par phase expiratoire et le contrôle de la coordination entre respiration et phonation.

La contraction des muscles continue lors de l'expiration pour contrôler la pression alvéolaire nécessaire à la phonation.

2.2. Organes Phonatoires

Les sons de la parole se produisent lors de la phase de l'expiration grâce à un flux d'air contrôlé, en provenance des poumons et passant par la trachée-artère. Il va rencontrer sur son passage plusieurs obstacles potentiels qui vont le modifier de manière plus ou moins importante. Tous ces éléments représentent le système phonatoire humain (figure 1.1)

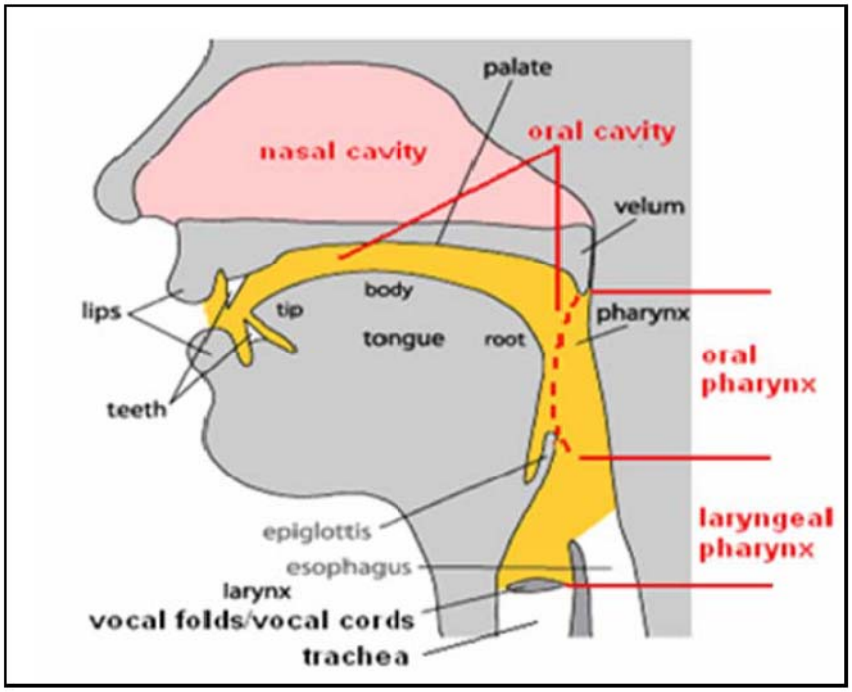

Figure 1.1 : Organes Phonatoires

2.2.1. Larynx

Le larynx est un organe creux dans lequel la voix est produite, situé dans la partie antérieure du cou, au-dessus de la trachée, au-dessous de la partie moyenne du pharynx, en arrière de la bouche et en avant de la partie inférieure du pharynx [18]. Le larynx est une structure cartilagineuse, se compose de quatre cartilages différents; dont le cartilage thyroïde prend le nom de pomme d'Adam et l'épiglotte cartilage en forme de lame, pouvant fermer par un mouvement de bascule en arrière l'entrée du larynx afin d'empêcher le bol alimentaire d'entrer dans le larynx et la trachée artère (figure 1.2). Les dimensions du larynx sont variables avec l'âge et le sexe ; petit chez l'enfant, plus grand chez l'homme que chez la femme. Le larynx peut se déplacer vers le bas ou vers le haut ; s'élève lors de l'émission des sons aigus, il s'abaisse à l'émission des sons graves, de ce fait la longueur de la cavité pharyngienne peut se trouver modifiée [19]. Au niveau physiologique, le larynx a une triple fonction :

- respiratoire marquée par le passage de l'air à travers le larynx ;
- de protection des voies aériennes inférieures par la fermeture de la glotte ;
- phonatoire : l'émission du son.

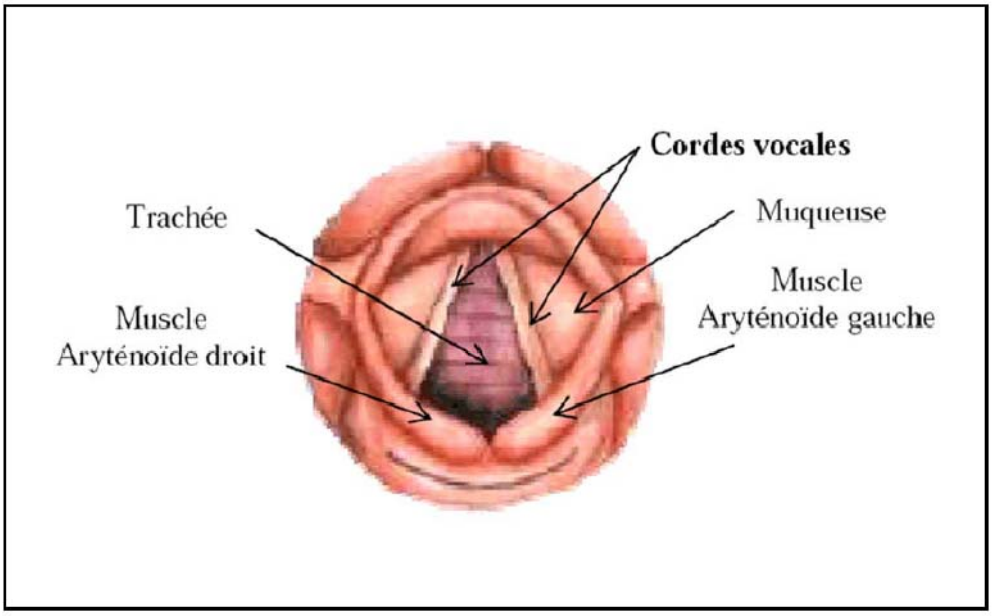

Figure 1.2 : Schéma du larynx

2.2.2. Cordes vocales

Ce sont des organes vibratoires constitués de tissu musculaire et de tissu conjonctif résistant; en effet, les fonctions sphinctériennes du larynx (respiratoire et phonatoire) dépendent de l'aspect et de l'état de ces deux éléments. Les cordes vocales situées en avant du larynx jusqu'à la base du cartilage aryténoïde. Leurs dimensions sont variables surtout en fonction du sexe puisqu'elles sont estimées à 22 mm chez l'homme et 18 à 20 mm chez la femme. Lors de l'émission vocale ; les cordes vocales vont d'abord se rapprochez en position de fermeture grâce aux cartilages aryténoïdes, la pression de la conne d'air expiratoire se heurte à un obstacle (fermeture des cordes). Elle va augmenter et contraindre les bords libres des cordes à s'écarter l'égerment et se positionner sous la forme d'un "V" appelée la glotte, pour laquelle laissant passer une petite quantité d'air, aussitôt libérée, les bords libres vont à nouveau se rapprocher à la fois :

- sous l'action de la diminution de la pression sous glottique ;
- par effet de Bernoulli (effet de rétro aspiration de la muqueuse cordale) ;
- grâce à l'élasticité propre des cordes vocales.

La glotte ne possède pas qu'un rôle de phonation, elle joue également un rôle de protection des voies aériennes supérieures grâce à la fermeture des cordes vocales lors de la déglutition [20].

2.2.3. Conduit vocal

C'est un ensemble de cavités situées entre la glotte et les lèvres reliées entre elles. Il mesure en moyenne entre 17 et 18 cm. Les cavités constituent les résonateurs qui doivent trier les harmoniques du son de base, leurs fréquences propres dépendront de leur volume, de leurs orifices et de leur couplage, on peut distinguer [21] :

- la cavité pharyngale : conduit musculeux membraneux situé en bouche l'œsophage d'une part et les fosses nasales d'autre part ; la paroi du pharynx est constituée de muscles constricteurs. En effet d'une modification du diamètre du pharynx, la racine de la langue peut reculer ou avancer et donc agir sur le volume de cette première cavité supra glottique ;
- la cavité nasale : deux cavités uniformes séparées par une cloison verticale médiane, recouvertes de muqueuses, relient les narines au pharynx. L'air passe par le nez lorsque le voile du palais est rabaissé (passage oro-nasal ouvert) ;
- la cavité buccale : sépare les fosses nasales par une cloison appelée le palais. Dans cette cavité se situent des articulateurs fixes et d'autres mobiles ;
- la cavité labiale : une cavité que l'on crée lorsqu'on projette en avant les lèvres (progression labiale), les lèvres jointes constituent un organe vibratoire accessoire intervenant dans la formation des consonnes ;
- la langue : est une structure frontière, appartenant à la fois à la cavité buccale pour sa partie dite mobile et au glossopharynx pour sa partie dite fixe ; elle a de l'importance pour la phonation.

3. Caractéristiques Acoustiques de la parole

La parole, très souvent considérée comme activité propre de l'homme et rarement étudiée comme fonction biologique, elle est un moyen de communication avec les autres. Elle met en jeu des organes de phonation et une véritable gymnastique des muscles du larynx, du pharynx, de la langue et des parois de la cavité buccale d'une façon générale. La voix représente le support acoustique de la parole, c'est un ensemble des sons produits par le larynx, lorsque l'air expiré fait vibrer les cordes vocales. Tous les sons simples peuvent être décrits par les caractéristiques suivantes [22, 23].

3.1. Intensité Sonore

L'intensité d'un son, appelée aussi volume, permet de distinguer un son fort d'un son faible. Elle correspond à l'amplitude de l'onde qui est donnée par l'écart maximal de la grandeur qui caractérise l'onde de compression. Cette grandeur correspond à la pression d'air. L'amplitude sera donc donnée par l'écart entre la pression la plus forte et la plus faible exercée par l'onde acoustique. Lorsque l'amplitude de l'onde est grande, l'intensité est grande et donc le son est plus fort. L'intensité du son se mesure en décibels (dB). On distingue différentes façons de mesurer l'amplitude d'un son :

- la puissance acoustique : La puissance acoustique est associée à une notion physique. Il s'agit de l'énergie transportée par l'onde sonore par unité de temps et de surface. Elle s'exprime en Watt par mètre carré ($W.m^2$) ;
- l'addition de sons : l'échelle des décibels est une échelle dite logarithmique, ce qui signifie qu'un doublement de la pression sonore implique une augmentation de l'indice d'environ 3 : avec 3 dB de plus, l'intensité est en fait doublée [24].

3.2. Fréquence fondamentale

La fréquence de la voix, au cours d'une conversation varie selon les personnes, elle est dépend essentiellement de la dimension et de la tension des cordes vocales, ainsi que des dimensions des résonateurs. Elle peut être volontairement modifiée dans certaines limites, par l'intermédiaire des muscles respiratoires, en faisant varier la pression de l'air. L'association de ces éléments détermine la fréquence de vibration des cordes vocales, appelée fréquence fondamentale ou pitch "F_0". Elle est variable selon l'âge et le sexe. Alors que la fréquence fondamentale de la voix parlée est :

- 100 à 150 Hz pour une voix masculine ;
- 200 à 300 Hz pour une voix féminine ;
- 300 à 450 Hz pour une voix d'enfant.

La mesure de la fréquence fondamentale "F_0" s'effectue soit à partir d'un signal microphonique (fréquencemètre, Glottal Fréquency Analyser ou GFA) soit à partir d'un signal électro laryngographique ou électro glottographique. Cependant le traitement numérique des signaux offre certaines méthodes d'estimation du pitch que l'on peut classer en trois catégories :

- méthodes temporelles ;
- méthodes spectrales ;
- méthodes mixtes (temporelles et spectrales).

Toutes ces méthodes peuvent se ramener à une détection de voisement associée à une mesure de périodicité [25].

3.3. Timbre vocal

C'est la couleur du son vocal à partir de laquelle, nous pouvons identifier une personne à la simple écoute de sa voix (par exemple, lors d'une conversation téléphonique, etc.). Le timbre vocal dépend de trois critères essentiels : l'accolement des cordes vocales, leur épaisseur et enfin les caractéristiques anatomiques des différentes cavités de résonance de l'appareil phonatoire. Par ailleurs, selon que les ouvertures glottiques se font plus ou moins rapidement, le spectre vocal est plus riche en aigus et inversement. Les cavités de résonances contribuent également à la couleur de la voix, car en modifiant leurs volumes, nous obtenons telle ou telle voyelle. Les éléments physiques du timbre comprennent :

- la répartition des fréquences dans le spectre sonore ;
- les relations entre les parties du spectre, harmoniques ou non ;
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence) ;
- l'évolution dynamique globale du son ;
- l'évolution dynamique de chacun des éléments les uns par rapport aux autres.

3.4. Durée du son

La durée d'un son représente le laps de temps des silences et des phonèmes. Il est difficile de les extraire car en un mot prononcé d'une façon naturelle, sans aucun traitement, donne un mélange de phonèmes chevauchés entre eux avec un silence d'intensité non nulle (le bruit). Une durée erronée peut produire une parole chaotique et parfois difficilement intelligible, pouvant provoquer un changement de sens du mot ou de la phrase. C'est le cas en langue arabe, où ce paramètre est pertinent. Ainsi, les deux mots [ğamal] (chameau) et [ğamāl] (beauté) présentent deux sens différents même s'ils ne diffèrent que par la durée temporelle de la dernière voyelle.

3.5. Formants et Transitions Formantiques

L'air est amplifié et subit différentes transformations dues aux degrés d'ouverture et de fermeture au niveau de chaque cavité, à la position de la langue, des lèvres, etc. Ces cavités possèdent des fréquences de résonance qui renforcent certaines régions du spectre des sources excitatrices. En phonétique acoustique, les fréquences renforcées aux régions des fréquences de résonance correspondant aux cavités vocales sont désignées par le terme de "formants". Ainsi, ces derniers sont les paramètres acoustiques qui permettent d'étudier et d'expliquer les phénomènes physiologiques que subit le son laryngé lors de son passage à travers les différentes cavités vocales. Les valeurs des formants varient selon le volume de la cavité et la surface de l'ouverture du résonateur. De façon générale, un formant a une valeur de fréquence inversement proportionnelle au volume de la cavité. Plus le volume de cette dernière est grand, plus cette fréquence est basse, et vice versa [26] (figure1.3).

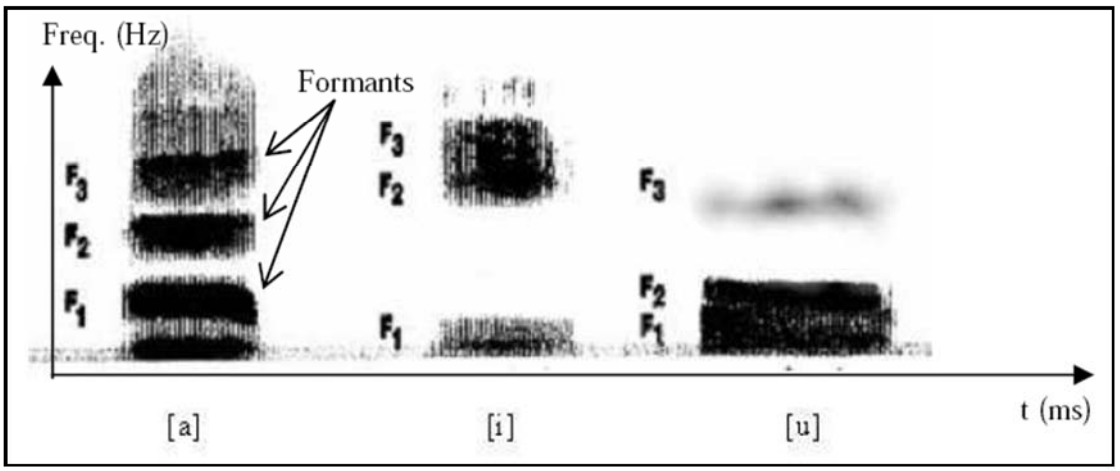

Figure 1.3 : Sonagrammes des voyelles [a], [i] et [u]

4. Sons voisés et non-voisés

Les différents sons de la parole sont classés en deux catégories principales selon que les cordes vocales vibrent ou ne vibrent pas.

4.1. Sons voisés

Pendant l'articulation de certains sons, la glotte s'ouvre brusquement libérant ainsi la pression accumulée en amont sous forme d'impulsions périodiques. Ces impulsions mettent les cordes vocales en vibration quasi-périodique. Le spectre d'un son voisé présente des raies correspondantes à l'harmonique du fondamental (structure de

pitch) c'est le cas des voyelles, l'enveloppe de ces raies présente des maximums appelés les formants, les trois ou quatre premiers formants sont essentiels pour caractériser le spectre vocal (figure 1.4).

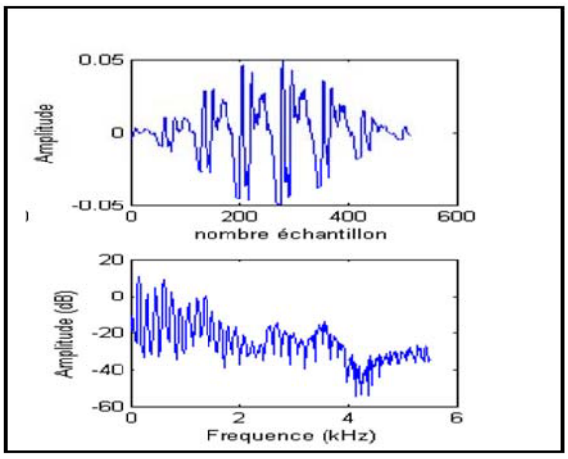

Figure 1.4 : Spectre de la voyelle [a]

4.2. Sons non voisés

Si les cordes vocales sont écartées, une turbulence quasi aléatoire d'air est produite dans le conduit vocal par diminution de sa section, ou bien le conduit est momentanément fermé complètement pour augmenter la pression et rouvert instantanément produisant une transitoire décroissante. Les sons ainsi produits sont appelés sons non voisés. Le son non voisé ne présente pas une structure périodique, il peut être considéré comme un bruit blanc, ainsi son spectre ne présente pas une structure de pitch (figure 1.5) [27,28].

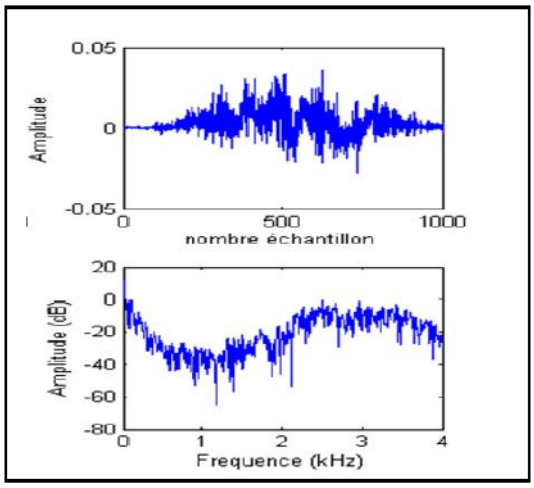

Figure 1.5 : Spectre de [ʃ]

5. Classification des sons du langage

D'un point de vue linguistique, la production des sons ou d'un mot réside dans la génération en série de tous les phonèmes constituant ce mot. Ces phonèmes forment les unités phonétiques qui sont classées en voyelles, consonnes et semi-voyelles, etc. [29, 30]. Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode et lieu d'articulation. Dans la cavité buccale, le point d'articulation est l'endroit où se trouve un obstacle au passage de l'air. D'une manière générale, nous pouvons dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air.

Nous distinguons généralement trois classes principales : les voyelles, les semi-voyelles et les consonnes (Figure. 1.6).

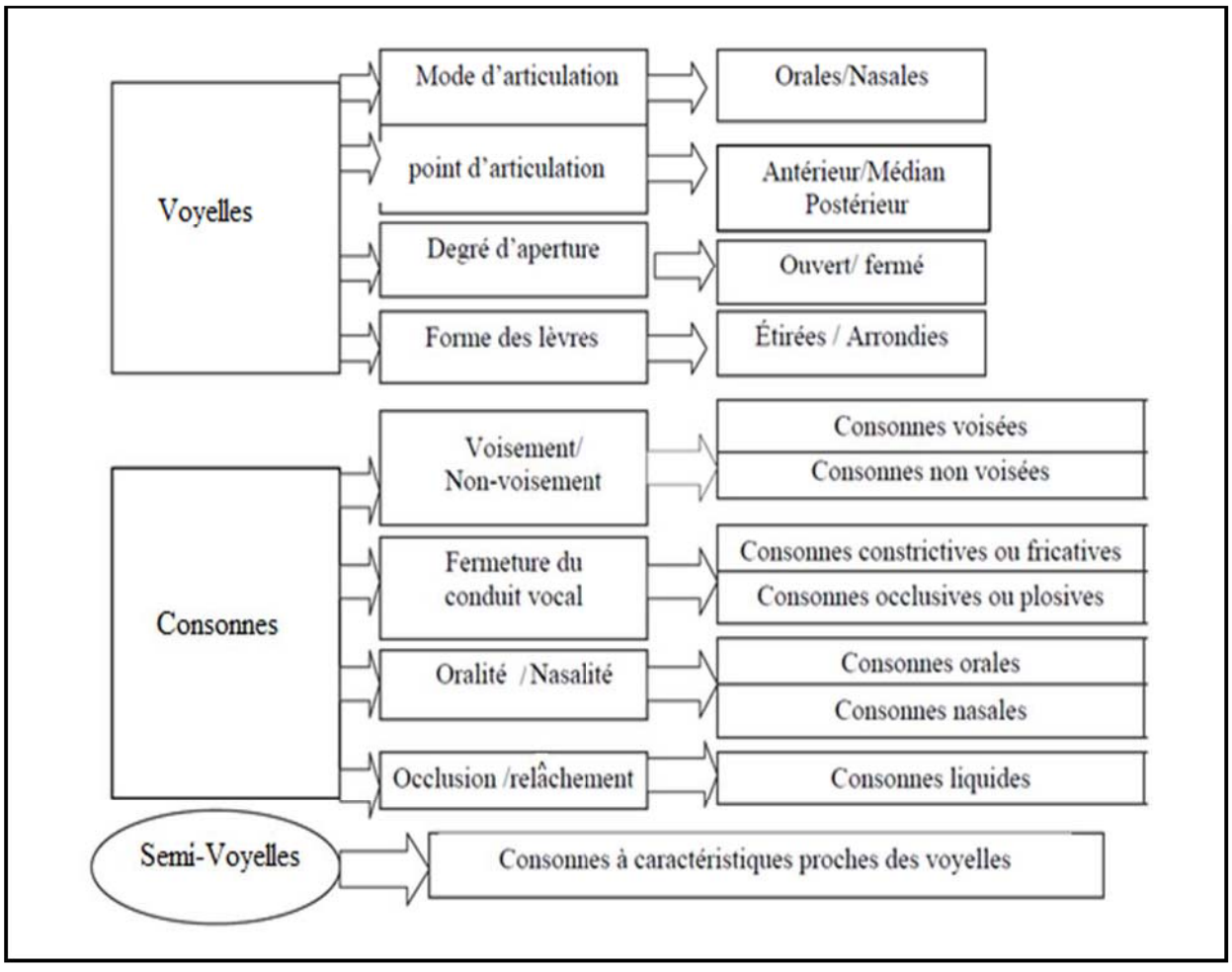

Figure 1.6 : Classification des sons du langage

5.1. Voyelles

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Quand ce dernier est suffisamment ouvert pour que l'air expiré par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la cavité buccale se réduit alors à une modification du timbre vocalique. Une voyelle se caractérise par un passage libre de l'air dans le conduit vocal et par la vibration des cordes vocales. Elles se différencient principalement les unes des autres par leur lieu d'articulation (position de la langue), leur degré d'ouverture (espace compris entre la pointe de la langue et le palais), et leur nasalisation. Nous distinguons ainsi, selon la localisation de la masse de la langue, les antérieures, les moyennes, et les voyelles postérieures, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles fermées et ouvertes. Les voyelles orales sont dues à une élévation du palais qui détermine la fermeture des fosses nasales ainsi qu'à l'écoulement de l'air expiratoire à travers la cavité buccale. Par contre les voyelles nasales sont caractérisées par l'écoulement d'une partie de l'air à travers la cavité nasale. L'AS ne possède pas de voyelles nasales. Elles tracent alors un triangle dont les extrémités sont occupées par les voyelles [i, u, a]. Ce triangle représente également les positions de la langue dans la cavité buccale selon deux axes : antérieur à postérieur (avant et arrière) et fermé à ouvert (Figure 1.7).

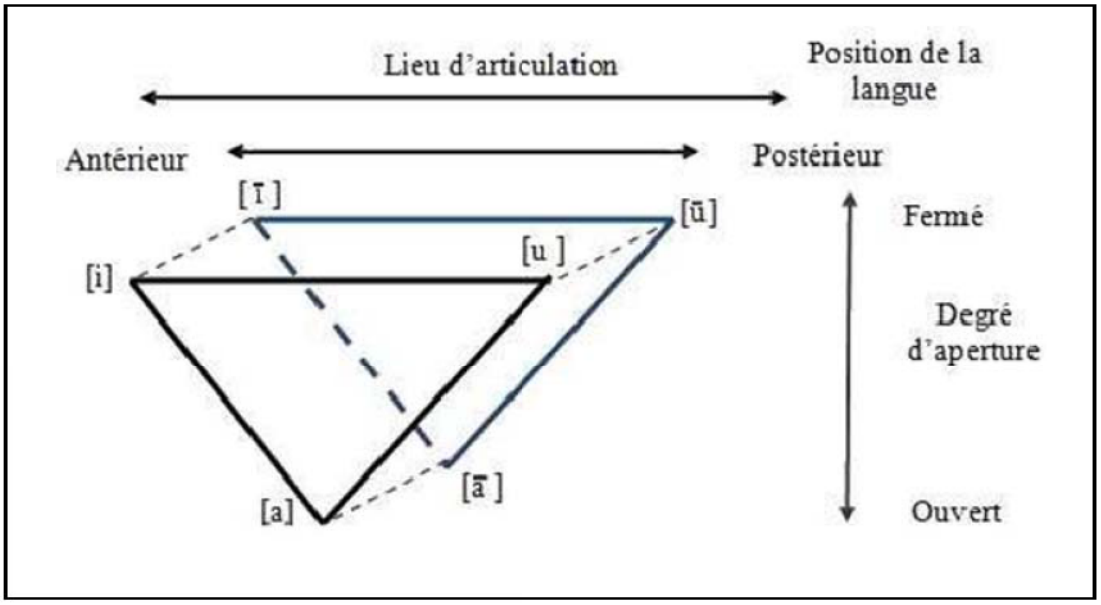

Figure 1.7 : Système vocalique de l'Arabe Standard [30].

5.2. Consonnes

Les consonnes se caractérisent par une fermeture partielle du conduit vocal ou constriction (constrictives ou fricatives) ou totale du conduit vocal (occlusion) : occlusives ou plosives. Nous classons principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- intervention ou mise en vibrations des cordes vocales : articulation sonore ;
- fermeture momentanée du passage de l'air suivie d'une ouverture brusque (explosion): articulation occlusive ;
- rétrécissement du passage de l'air qui produit un bruit de friction ou de frôlement : articulation fricative ;
- position abaissée du voile du palais: articulation nasale ;
- contact de la langue au milieu du canal buccal; l'air sort des deux côtés ;
- une série d'occlusions brèves ; séparées de la luette: articulation vibrante.

La distinction du mode d'articulation conduit à deux classes : les fricatives ou constrictives et les occlusives ou plosives. Les consonnes fricatives appelées également spirantes sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais, les dents ou les lèvres. Les fricatives non voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et d'autres turbulentes : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète. Les consonnes occlusives ou plosives sont reconnues grâce au silence provenant de la fermeture totale du conduit vocal ou occlusion. Cette dernière comporte trois phases :

- l'implosion ou fermeture ;
- l'occlusion proprement dite tenue de la fermeture ;
- l'explosion ou détente.

Les consonnes liquides combinent une occlusion et une ouverture simultanée du conduit vocal. Elles sont caractérisées par un degré de sonorité proche de celui des voyelles. Enfin, les consonnes nasales font intervenir la cavité nasale par abaissement

du voile du palais. Elles sont produites par l'écoulement de l'air phonatoire dans le conduit nasal.

5.3. Semi-voyelles

Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, elles sont assez difficiles à classer.

6. Description des sons de l'Arabe Standard

L'Arabe est la langue du Coran, des médias, de la science, de l'enseignement, de la littérature, etc. Elle est structurée d'une manière différente relativement aux autres langues, les consonnes ou [huru:f] et les voyelles ou [haraka:t]. L'Arabe n'a pas une écriture strictement phonétique, car à une même graphie correspondent plusieurs images phoniques selon le contexte. Son écriture, comme celle des autres langues sémitiques est consonantique. Les sons de l'Arabe se composent de :

- vingt huit consonnes qui peuvent prendre des formes légèrement différentes selon qu'elles sont situées en position isolée ou initiale, médiane ou finale dans le mot ;
- trois voyelles brèves kasra, damma et fatha [i, u, a]. Ces voyelles ne sont pas notées. Mais pour faciliter la lecture et la compréhension d'un texte, les voyelles brèves sont représentées par les signes diacritiques ;
- trois voyelles longues [i:, u:, a:], appelées [huru:f ɛl madd] ;
- le silence, appelé [suku:n].

Les Phonéticiens symbolisent les sons du langage au moyen de signes divers auxquels nous attribuons une valeur conventionnelle. Selon les auteurs la transcription varie beaucoup.

Nous choisissons la transcription de l'Alphabet Phonétique International (API) pour des raisons de simplicité (Tab.1.1).

L'originalité de la phonétique arabe se fonde, sur les consonnes emphatiques, pharyngales et laryngales et huru:f ɛ l mad, car elles donnent une valeur particulière à la langue [29, 30].

Tableau 1.1 : Transcription des phonèmes de l'Arabe Standard en API.

Phonèmes de l'AS (API)	Phonèmes de l'AS (Arabe)	Phonèmes de l'AS (API)	Phonèmes de l'AS (Arabe)
[ʔ]	ء	[ḍ]	ض
[b]	ب	[ṭ]	ط
[t]	ت	[ẓ]	ظ
[θ]	ث	[ɛ]	ع
[ǯ]	ج	[ɣ]	غ
[ħ]	ح	[f]	ف
[x]	خ	[q]	ق
[d]	د	[k]	ك
[ð]	ذ	[l]	ل
[R]	ر	[m]	م
[Z]	ز	[n]	ن
[S]	س	[h]	ه
[ʃ]	ش	[w]	و
[ṣ]	ص	[j]	ى

7. Troubles de la parole

La production d'une voix normale est basée sur les paramètres acoustiques suivants : qualité, intensité, hauteur, débit et résonance. Un trouble de la voix présente une altération d'une ou plusieurs de ces paramètres. Les atteintes peuvent concerner les organes périphériques, qui gênent la production de la parole exemple : bec de lièvre, division palatine, insuffisance vélaire, malformations linguales, labiales ou laryngées. Il s'agit d'anomalies consistant en des erreurs mécaniques et constantes dans l'exécution du mouvement propre à un phonème [28]. L'articulation est la capacité à articuler les sons de façon permanente et systématique, ce qui nécessite des mouvements précis de la mâchoire inférieure, de la langue, des lèvres, des joues, du voile du palais. Le trouble articulatoire isolé est donc l'incapacité à prononcer ou à former un certain phonème correctement. C'est une erreur constante, systématique et mécanique pour un phonème donné. Cette erreur est plutôt de type praxique [32].
La production de la voix normale est basée sur sa qualité, son intensité, son débit. Une voix pathologique présente une altération d'un ou de plusieurs de ces paramètres [33]. On peut d'ores et déjà classifier les troubles du langage en régions pathologiques, c'est ce qui montre que la voix peut être altérée ou modifiée tout le long de sa

production, voire disparaître, phénomène décrit par l'apparition d'une aphonie ou absence de voix complètement, surtout lors du cancer des cordes vocales.

7.1. Classification des troubles

La première partie a concerné les pathologies des organes intervenant dans la production ou l'altération de la voix, celles-ci sont classées comme suit, un dysfonctionnement :

- fonctionnelle : l'organe existe mais, il y a eu soit un mauvais apprentissage, soit une maladie en cours d'évolution ce qui présente un symptôme de pathologie de la parole, si la pathologie n'est pas détectée à temps [34].
- organique : l'organe existe ou est absent, cas du palais en clavé (en forme de massue) ou laryngectomie, mais ne peut pas exécuter la tâche préconçue, soit par atrophie cas de la langue trop courte, soit par surdimensionnement cas du volume du palais démesuré. Les différents défauts émanant de ces pathologies sont classés selon le diagramme de la figure 1.8.

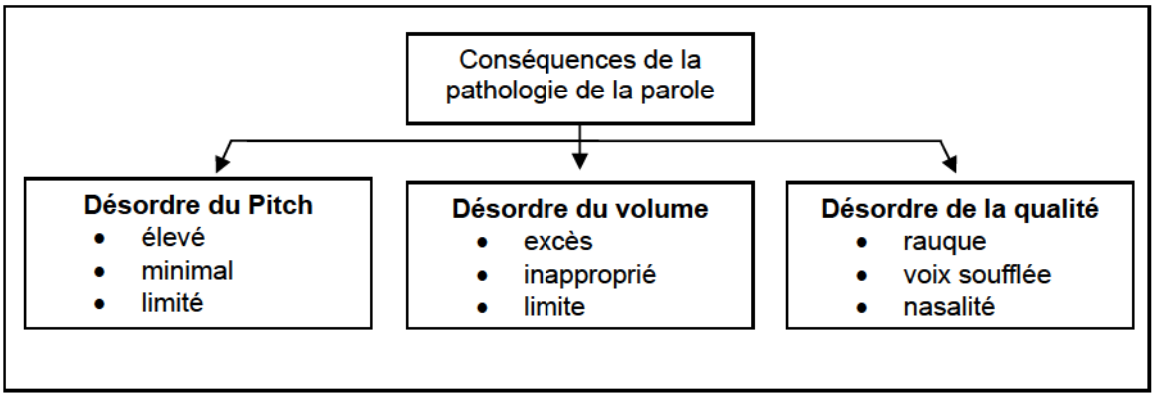

Figure 1.8: Diagramme de classement des pathologies [35]

7.2. Défauts de la voix détectés par l'oreille

Nous présentons les différents défauts de la voix, détectés par le système auditif humain.

7.2.1. Blèsement ou Zézaiement

Le blèsement ou zézaiement est un défaut de prononciation qui consiste en la substitution de [ʃ] (Une consonne chuintante) par [s] (une sifflante) et de [ʒ] ou [j] (Consonnes chuintantes) par [z] (sifflante) [36].

7.2.2. Chuintement

Le chuintement est la prononciation du [s] et du [z] à la manière du [ʃ] et du [j] [36]. Exemple :

- j'ai pris l'auto *bu**s** ju**s**qu'à* la gare ***S**aint-Lazare*
- j'ai pris l'auto *bu**ch** ju**ch**qu'à* la gare ***Ch**aint-Lajare*.

7.2.3. Rhotacisme

Le rhotacisme (terme formé à partir du grec ρ, [r]) est une modification phonétique complexe, consistant en la transformation d'un phonème en [r]; Dans d'autres langues comme le Français c'est avec le [z] que le [r] [34].Pour la langue arabe c'est la confusion entre le [r] et le [ʁ].Donc, au lieu de prononcer ' ريح ' [rihø].La personne atteinte de rhotacisme prononce " غيح " [ʁihø],

7.2.4. Nasonnement

C'est l'altération du son de la voix; le nasonnement provient de la diminution de la résonance nasale par suite de l'obstruction du nez, de la présence de végétations adénoïdes, etc., et produit une déformation des syllabes nasales, [an], [on], [in], et des consonnes nasales, telles que [m], que l'on prononce [b] [37].

7.2.5. Bégaiement

C'est le trouble de la communication affectant le débit et le rythme de la parole se traduisant par :

- une forme clonique : répétition;
- une forme tonique : blocage;
- des troubles associés.

Si rien n'est entrepris, pour un enfant de 2 à 5 ans commençant à bégayer, il restera bègue à l'âge adulte. Il est nécessaire d'intervenir le plus tôt possible pour ne pas prendre le risque de la chronicisation [38], [39].

Les signes d'appel et manifestations du bégaiement se présentent comme suit :

- répétition de sons ou syllabes supérieures à 3 (ex: *toutoutou toupie*) ;
- prolongation de sons ;
- blocage de syllabes ;
- répétitions de mots, de parties de phrases ;
- reprise d'énoncés ;

- hypertonie, blocages respiratoires lors de la prise de la parole ;
- comportement ou modification du comportement : colères, retrait, timidité; énurésie ;
- comportement verbal: refus ou repli ;
- antécédents de bégaiement dans la famille [28].

7.2.6. Clichement

Le clichement est un défaut de prononciation se caractérisant par le fait d'additionner le son [ll] (double L) mouillé, positionné après certaines consonnes. Une consonne mouillée est articulée avec le son [j]. Par exemple [l] dans grisaille.
Un exemple de clichement : prononcer *ch**ill**uchote*r au lieu de *chuchoter* [36].

7.2.7. Gammacisme

Défaut de prononciation se caractérisant par la difficulté voir l'impossibilité de prononcer les consonnes gutturales, [k] à la place de [g].

7.2.8. Retard de parole

Le retard de la parole est l'altération de phonèmes ou de groupes de phonèmes, par leur mise en ordre séquentielle à l'intérieur d'un même mot, le stock phonétique étant acquis.

C'est la forme du mot dans son ensemble qui ne peut être reproduite. Un parler bébé qui perdure au-delà de 4 ans est caractérisé par des:

- omissions mots raccourcis ou élidés : f**l**eur / feur ;
- inversions *br**o**uette* / *bourette* ;
- assimilations : *lavabo la**la**bo* ou ***va**vabo* ;
- interversions : *kiosque* / *kio**kse*** ;
- simplifications : *pa**rapl**uie* / *papui* ... ;
- substitutions : *train* / ***c**rain*, *fleur f**i**eur* ;
- élisions de syllabes finales : *pe**lle** pè*, *assi**ette** assiè*...

Et de façon plus globale :

- problèmes de perception auditive ;
- mauvaise structuration de la perception du temps ;
- mauvaise structuration de la chronologie des sons ;
- difficultés motrices diverses ;

- attention auditive variable ;
- immaturité psychoaffective ;
- un refus de grandir ...

Fréquemment, un retard de langage et/ou un trouble d'articulation peuvent être associés au retard de parole [28].

7.2.9. Facio-Scapulo-Humeral (FSH)

Comme nous l'avons cité précédemment, certaines maladies ne concernent pas les cordes vocales directement, mais affectent les muscles, telle que la FSH qui est la dégénérescence de tous les muscles du corps humain, ou l'évolution du squelette osseux se développe avec quelques anomalies (Courbures de dos, déformation du bassin, etc …), tandis que les muscles d'une partie du corps s'atrophient, ceci donne lieu à une pathologie entre autres langagière qui affecte quelques phonèmes tels que le [b] et le [f] qui sont systématiquement remplacés par le [d] et le [θ].

8. Troubles du Sigmatisme

Nous avons ciblé notre travail sur une pathologie, à savoir le sigmatisme occlusifs et constrictif. Cette dernière nous a permis d'extrapoler notre méthode pour d'autre pathologies. Terme issu de la lettre grecque sigma, c'est la difficulté que présentent certaines personnes à prononcer le phonème [s]. Cette affection ne doit pas être confondue avec le zézaiement qui est un défaut de prononciation d'une personne prononçant le son [s] comme étant [z], le son [žø] comme [s] ou le son [s] comme [sø]. On dit également zozoter Ce défaut est généralement relié à une déviation par la langue, dans le processus d'écoulement d'air. Il y'a deux types de sigmatismes :

- latéral ou schlintement: l'air s'échappe sur le côté de la bouche;
- interdental ou zozotement : la langue vient buter contre les incisives supérieures ou se place entre les dents et produit une interposition linguale lors de l'émission des phonèmes [s] et [z] [36].

Le sigmatisme concerne le remplacement de [ʃ], [j], [s], [z] par [θ] et [d] ou par [f] et [v] [39] (figure 1.9) Il intervient sur les consonnes constrictives et peut avoir plusieurs appellations, suivant son origine.

8.1. Sigmatisme des consonnes constrictives

Le sigmatisme nasal est dû à un positionnement de la langue qui rend impossible le passage de l'air par la cavité buccale ; le sigmatisme dorsal est également dû à un soulèvement de la langue excessif ;

8.2. Sigmatisme des consonnes occlusives

Le sigmatisme occlusif est le remplacement systématique de toute consonne constrictive par la consonne occlusive dont le point d'articulation est le plus proche [40].

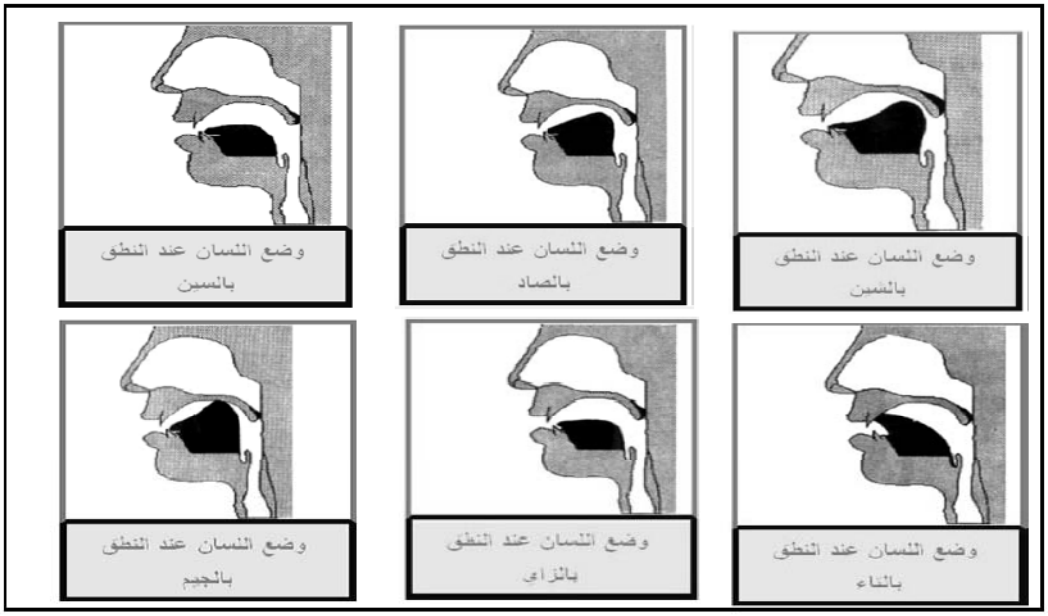

Figure 1.9. : Lieux d'articulation des phonèmes [S, ʃ, Ṣ, Θ, Z, ž]

9. Conclusion

Pour aborder la pathologie de la parole, il était nécessaire d'étudier tous les points qui ont trait à la parole. A travers ce chapitre, nous avons introduit la notion d'anatomie articulatoire, qui est l'une des causes de la pathologie de la parole, par une étude simplifiée, nous avons cité un éventail assez large de pathologies langagières, ayant trait aux variations phonétiques. Le choix du sigmatisme occlusif, pathologie concernant les défauts de prononciation du [ʃ] et [s] prononcés comme [θ] et [z], porte essentiellement sur la disponibilité d'un corpus pathologique ainsi qu'une tendance à mettre une méthodologie d'évaluation par un système d'aide, qui sera bénéfique à l'orthophoniste et au patient. Cette méthodologie sera basée sur la technique de RAL.

CHAPITRE 2 :

ANALYSE PARAMETRIQUE DU SIGNAL VOCAL

1. Introduction

Différentes méthodes de représentation du signal existent. Certaines ont été spécifiquement développées pour l'étude ou la compression des signaux de parole. Elles essaient, soit de résoudre les problèmes posés par les méthodes fondées sur la seule Transformée de Fourier, soit de simuler du mieux possible les caractéristiques du signal vocal [41]. Ces paramètres peuvent être exploités également pour la caractérisation des différentes paroles pathologiques par rapport à la normale.

Connaître d'une manière précise la façon dont est produite la parole permettra de mieux manipuler et traiter celle-ci, afin de parvenir à discriminer efficacement une pathologie vocale, quelle que soit sa nature, par rapport à la normale.

Nous allons maintenant présenter des méthodes adaptées à la Reconnaissance Automatique de la Parole qui sont les plus utilisées actuellement. Ces grandes méthodes sont les techniques cepstrale, le codage par prédiction linéaire.

2. Production de la parole humaine

Comprendre le mécanisme de production de la parole est un aspect d'une grande importance. En effet, c'est l'étude du système de phonation qui va nous permettre d'identifier et de caractériser les grandes classes de sons élémentaires et d'expliquer les variations de ces derniers dans les différents contextes.

De plus, les algorithmes de paramétrisation du signal vocal sont obtenus à partir de modèles du conduit vocal. Les paramètres acoustiques du signal de la parole sont évidemment liés à sa production. L'intensité du son dépend de la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que celle du rythme d'ouverture/fermeture des cordes vocales, induit par la tension de muscles qui les contrôlent. Son spectre résulte du filtrage du signal glottique (impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses [42].

La parole est articulée en interrompant et en modulant le flux d'air à l'aide des lèvres, de la langue, des dents, de la mâchoire inférieure et du palais Les parties différentes de la conduite orale, le voile du palais, la langue et les lèvres servent des articulateurs. Les dimensions différentes de ces articulateurs peuvent causer les coupes transversales diverses de la conduite orale qui sont responsables des

fréquences de résonance possibles multiples du conduit vocal. En revanche, la coupe transversale du conduit nasal est fixée et la quantité d'écoulement d'air au conduit nasal est contrôlée par le voile du palais (Figure. 2.1).

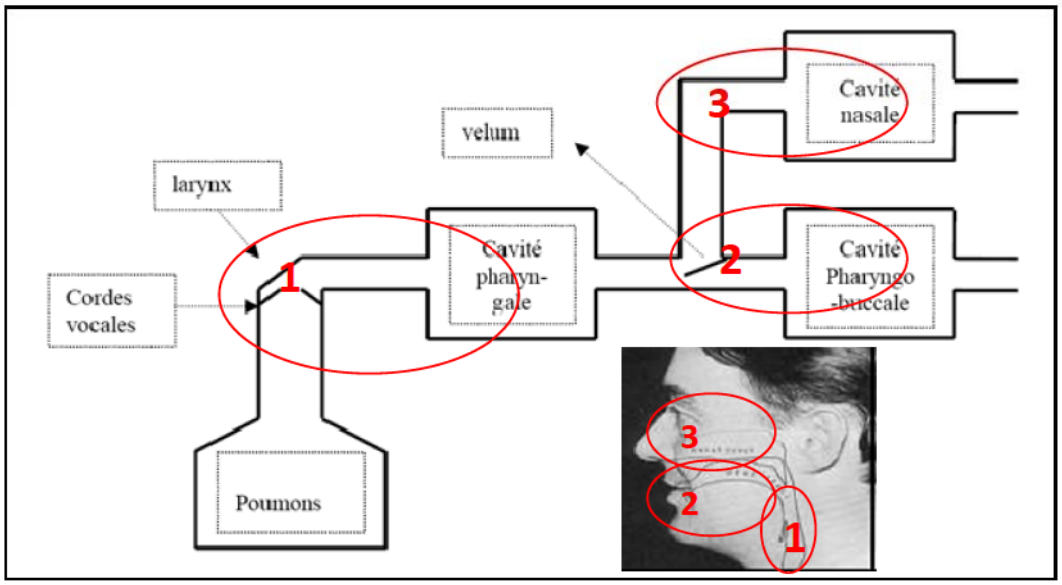

Figure 2.1 : Vue schématique du système de production de la voix humaine

2.1. Production de l'onde glottique

L'air produit par excès de pression dans les poumons rencontre un premier obstacle qui sont les cordes vocales (source d'excitation). Ces dernières accolées, sous l'effet de la pression sub-glottique se mettent à vibrer laissant passer l'air par impulsions. C'est ainsi que se forme l'onde glottique dont la fréquence d'oscillations notée F_0 (fréquence fondamentale ou pitch), est déterminée par la masse et la tension des cordes vocales ainsi que la pression sub-glottique. Quand elles vibrent, il y a émissions de sons dits voisés ou sonores par opposition aux sons non voisés ou sourds qui sont assimilables à un bruit blanc.

2.2. Fonction résonateur du conduit vocal

Le conduit vocal imprime au son émis les caractéristiques spécifiques permettant de distinguer les différents phonèmes et ceci selon deux fonctions, en tant que :

- résonateur de l'onde glottique pour la production des phonèmes sonores ;
- générateur de bruit pour la production des consonnes sourdes.

En effet, l'onde glottique est modifiée lors de son passage à travers le conduit vocal. Les positions de la mâchoire et de la langue déterminent les cavités qui jouent le rôle de caisses de résonance en renforçant certaines régions du spectre acoustique. Les maxima de la courbe de réponse en fréquences du conduit vocal sont appelés formants

2.3. Fonction générateur de bruit du conduit vocal

Le flux d'air créé peut rencontrer soit, un obstacle partiel tel un rétrécissement du conduit vocal pour générer un bruit caractéristique des sons fricatifs ou constrictifs, soit un obstacle total produisant une augmentation de la pression en amont de l'obstacle (lieu d'articulation) suivi d'un relâchement brusque. Ce phénomène engendre la formation des sons occlusifs.

3. Modèle de Source/ Filtre

Le système de production de la parole peut être simplifié au modèle de source- filtre. Le processus de production entier est réduit à deux excitations différentes et un filtre acoustique (figure 2.2).

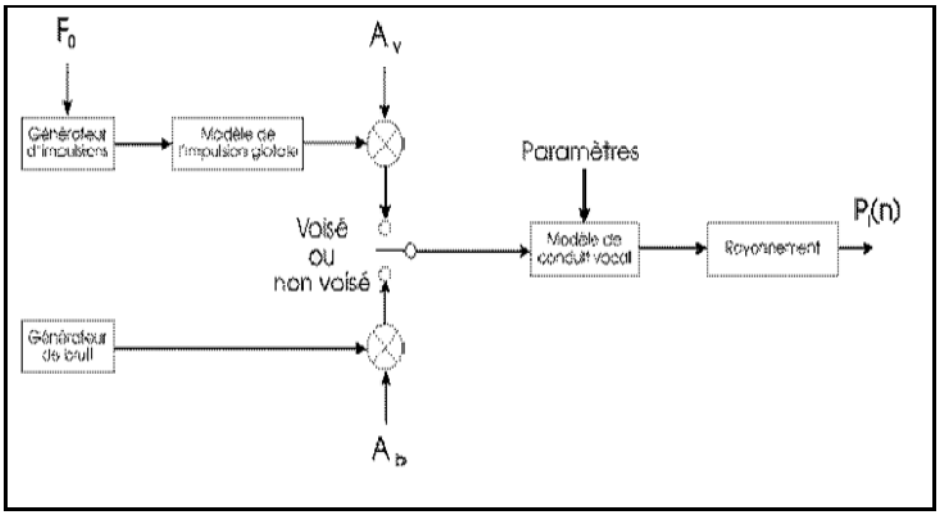

Figure 2.2 : Modèle de production de la parole

Les deux excitations représentent les cordes vocales et le modèle acoustique du conduit vocal. La première excitation du modèle des cordes vocales tendues est réalisée par une source de train d'impulsions caractérisant le voisement ressemble aux voyelles et nasals. Le deuxième est une source de bruit blanc qui est nécessaire pour produire le non voisement exemple sons des fricatives, où les cordes vocales sont ouvertes et la colonne d'air passe par le conduit vocal considérée comme le

filtre acoustique, caractérisé par ces fréquences de résonance, où l'énergie du signal source atteint des maximums locaux. Ces maximums locaux du spectre sont appelés formants et typiquement il y a jusqu'à quatre fréquences de résonance ou formants de signification. La figure 2.3 donne un petit exemple d'une trame de parole courte. Les formants peuvent être observés comme les maximums locaux du signal lissé.

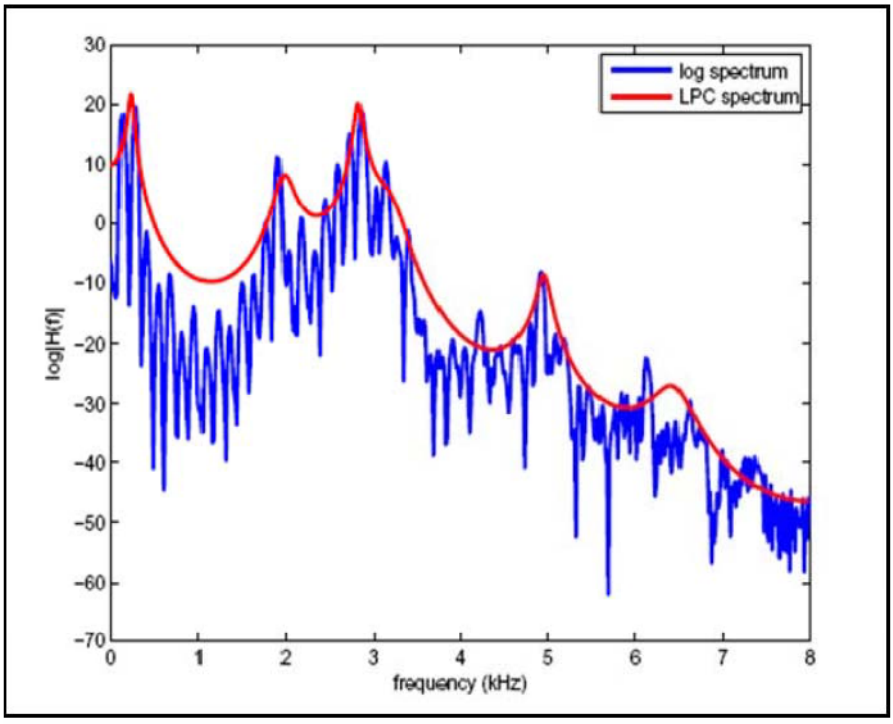

Figure 2.3 : Représentation spectrale des paramètres LPC et des fréquences des formants

4. Traitement du signal vocal

De par la complexité intrinsèque du signal de parole et la quantité d'informations présente dans ce signal, les techniques de la RAP n'utilisent pas ce signal sous sa forme brute. La littérature comporte de nombreux types de traitement de la parole. De par la propriété pseudo-stationnaire de la parole, ce traitement est généralement réalisé périodiquement toutes les 10 à 30 ms, le flux d'information résultant est une suite de vecteurs de paramètres acoustiques.

L'analyse spectrale permet d'extraire des paramètres représentatifs des caractéristiques de l'appareil phonatoire humain. Le calcul des paramètres acoustiques est ainsi réalisé en glissant avec une cadence régulière (ex : 10ms) une fenêtre de pondération d'une longueur bien définie sur tout le signal. On connaît plusieurs type de fenêtrage (ex :Hamming, Hanning, Blackman, etc). En général, le

fenêtrage Hamming est le plus utilisé en traitement de signal de la parole. Chaque fenêtre nous permet d'avoir une trame. Les trames obtenues sur tout le signal de parole sont traitées par la suite afin de produire les vecteurs des paramètres acoustiques. Dans la littérature, il existe trois grandes catégories de paramètres, analyse par :

- bancs de filtres : il s'agit d'une analyse assez simple du système auditif humain qui consiste à calculer l'énergie du signal vocal dans différentes bandes de fréquences ;
- Transformée de Fourier : les coefficients issus de la transformée de Fourier peuvent être utilisés pour une analyse en bancs de filtres de Mel ainsi que pour les calculs des coefficients cepstraux. Parmi les plus connus se trouvent les MFCC (Mel Frequency Cepstrum Coefficient);
- prédiction linéaire : les coefficients issus d'un modèle autorégressif (de l'analyse LPC) peuvent être utilisés comme paramètres caractéristiques. Parmi les plus connus se trouvent les LPCC (Linear Predictive Cepstrum Coefficient).

Les LPC sont censés représenter la forme du conduit vocal, et ainsi présenter une forte variabilité selon le phonème prononcé. Avant de décrire un processeur d'extraction des LPC pour la RAP, il est intéressant d'examiner les raisons pour lesquelles les LPC ont été largement utilisées. Ceux-ci incluent ce qui suit:

- LPC fournit un bon codage de signal de parole, ces coefficients fournissent une bonne approximation de l'enveloppe spectrale de l'appareil vocal ;
- la manière dont les LPC sont appliquées à l'analyse des signaux de la parole conduit à une séparation des voies source-conduit vocal raisonnable. En conséquence, une représentation parcimonieuse des caractéristiques du conduit vocal devient possible.
- l'expérience a montré que la performance de RAP basée sur les LPC, est comparable ou bien meilleures par rapport aux dispositifs de reconnaissance basé sur bancs de filtres [43-45].

4.1. Analyse LPCC

Le calcul des coefficients de la prédiction linéaire (LPCC) est un outil utilisé en traitement du signal audio pour extraire du signal vocal un ensemble de paramètres pertinents dans le but de réduire la redondance du signal vocal pour une tâche de classification de la parole avec des informations de modèle prédictif linéaire.

Le calcul de ces paramètres est réalisé par une chaîne de prétraitement selon les étapes suivantes (Figure 2.4).

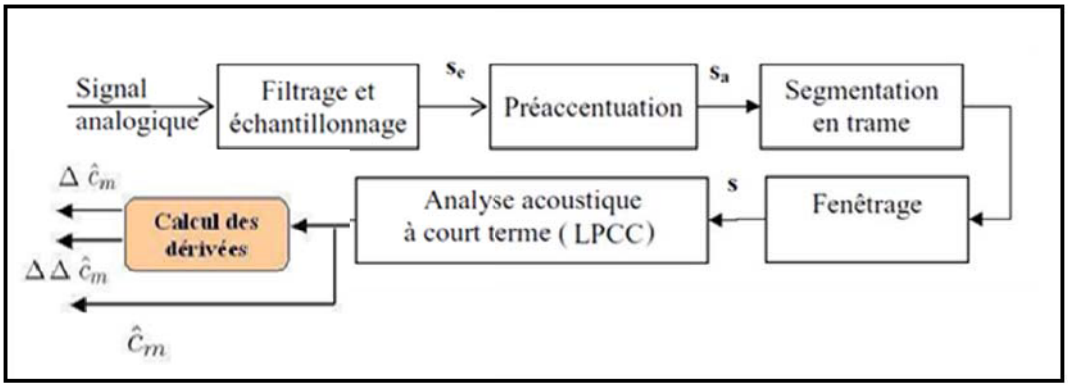

Figure 2.4 : Etapes d'extraction des coefficients LPCC

4.1.1. Échantillonnage du signal vocal

L'échantillonnage est l'opération maîtresse pour la conversion de signaux continus en signaux discrets (Conversion Analogique-Numérique). On peut modéliser l'échantillonnage comme la multiplication du signal de base $x(t)$ par un signal $p(t)$ constitué d'une série d'impulsions unitaires uniformément espacées sur la durée du signal (figure 2.5).

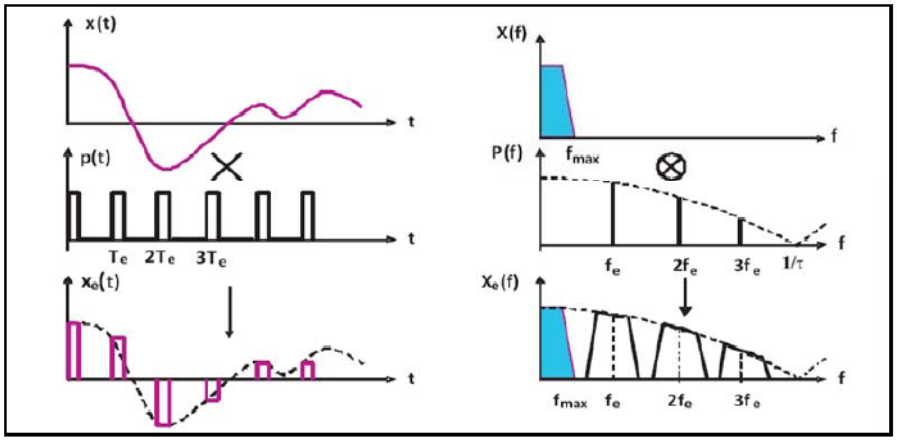

Figure 2.5 : Échantillonnage d'un signal

4.1.2. Préaccentuation

Pour les sons voisés, l'intensité du signal de parole décroît en fonction de la fréquence. Cette décroissance est d'environ 6 dB/octave. En termes absolus, certaines informations pertinentes de hautes fréquences sont donc noyées par les basses fréquences qui occupent une place prépondérante dans l'amplitude du signal. Afin de donner à ces fréquences l'importance qui leur est due et puisque les sons voisés sont plus fréquents que les sons non-voisés, on utilise, dans le domaine temporel, une transformation du signal permettant de redresser le spectre du signal et ainsi de détecter certaines variations aux hautes fréquences. En termes relatifs, les hautes fréquences sont amplifiées par rapport aux basses fréquences. Toutefois, les relations relatives entre fréquences voisines restent pratiquement intactes.

$$\hat{x}[n] = x[n] - a * x[n-1] \tag{2.1}$$

Ceci correspond à la multiplication de la transformée en z du signal par un filtre passe-haut du premier ordre:

$$H[z] = 1 - a * z^{(-1)} \tag{2.2}$$

où, généralement, $0.9 \leq a \leq 1.0$

4.1.3. Segmentation en trames

Le signal de parole est dynamique ou variant dans le temps dans la nature, le signal de parole est considéré comme stationnaire lorsqu'elle est examinée au cours d'une courte période de temps [44].

Afin d'analyser le signal de la parole, il doit être divisé en trames de N échantillons, avec des trames adjacentes étant séparées par M échantillons.

- si M ≤ N, alors les estimations spectrales LPC de trame en trame seront tout à fait lisses.
- si M > N il n'y aura pas de chevauchement entre des trames adjacentes.

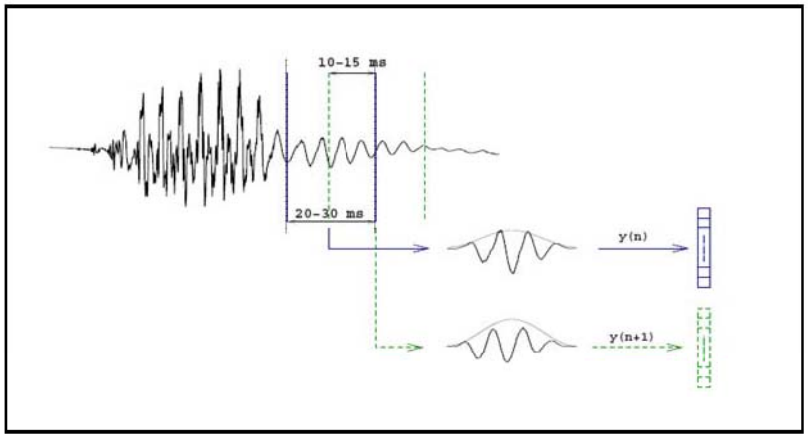

Figure 2.6 : Segmentation en trames

4.1.4. Fenêtrage

L'échantillonnage effectué a permis d'évaluer le signal de parole en un nombre fini de points. Cette propriété ne s'applique pas que pour le cas théorique d'un signal de longueur infinie. De plus, l'information fréquentielle fournie représente une quantité "moyenne" sur la durée totale du signal. Or, la parole étant hautement dynamique, on doit tenter d'extraire les paramètres fréquentiels sur une période beaucoup plus courte pendant laquelle il est raisonnable de supposer un signal stationnaire dans le domaine fréquentiel. Ceci est rendu possible grâce à la lenteur relative de mouvement du conduit vocal. Cette période plus courte est appelée trame. Le terme fenêtrage se réfère à la multiplication d'un signal $x[n]$ par une séquence $\omega[n]$ de durée finie, c'est-à-dire :

$$\omega(n) = \begin{cases} = 0 \text{ pour } n < 0 \\ \neq 0 \text{ pour } 0 \le n < N \\ = 0 \text{ pour } n \ge N \end{cases} \qquad (2.3)$$

Une fois qu'une portion du signal a été fenêtrée, on extrait les fréquences en utilisant la Transformée Discrète de Fourier dont le calcul peut être simplifié à l'aide de l'algorithme de la Transformée Rapide de Fourier (FFT). Une fois que les paramètres ont été extraits sur la portion d'intérêt du signal, on fait "glisser" la fenêtre pour traiter une portion ultérieure et l'on répète le processus d'extraction des paramètres. Ce procédé est communément appelé Transformée de Fourier à fenêtre glissante. Plusieurs modèles de fenêtre ont été proposés. La figure 2.7 présente les valeurs de quelques paramètres qui servent à l'évaluation de trois types de fenêtres.

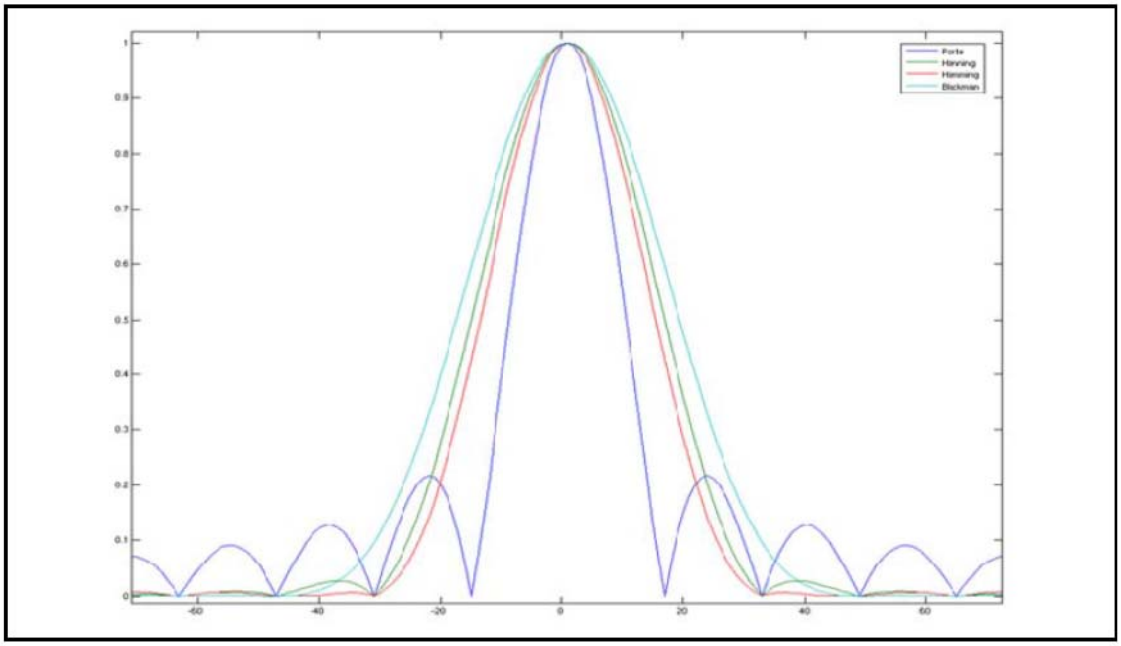

Figure 2.7 : Représentations fréquentielle des fenêtres

L'avantage principal de la fenêtre de Hamming est l'étroitesse du lobe secondaire relativement au lobe principal. Cette fenêtre possède donc une bonne résolution dans le domaine fréquentiel. Cette fenêtre est utilisée dans le domaine de la RAP et la RAL, principalement pour cette raison.

$$\text{Hamming}(n) = \begin{cases} 0.54 - 0.46 \cdot \cos(2\pi n / N) & 0 \le n \le N-1 \\ 0 & \text{ailleurs} \end{cases} \qquad (2.4)$$

4.1.5. Analyse acoustique LPCC

Le signal vocal résulte de l'excitation du conduit vocal par un train d'impulsions ou un bruit blanc produisant respectivement des sons voisés et non voisés. Ainsi, ce signal peut être modélisé par la convolution de la fonction de transfert du conduit vocal (filtre) avec le signal d'excitation (source). Dans l'analyse par prédiction linéaire LPC, la fonction de transfert du conduit vocal peut être modélisée par un filtre linéaire tout-pôles qui produit un signal AutoRégressif (AR). La fonction de transfert est donnée par :

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \qquad (2.5)$$

Où :

- p est le nombre de pôles (l'ordre de prédiction) ,
- a_0=1, {a_i}i=1:p sont les coefficients du filtre.

Chaque échantillon s(n) est prédit comme une combinaison linéaire des p échantillons précédents, à laquelle s'ajoute un bruit blanc gaussien *e* de variance σ^2:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \qquad (2.6)$$

Les coefficients {a_i} sont choisis de telle façon à minimiser l'erreur de prédiction estimée sur la fenêtre d'analyse par la méthode des moindres carrés. Cette minimisation conduit aux équations de Yule-Walker qui expriment le vecteur des coefficients A= $(1,a_1,a_2,\ldots,a_p)^t$ comme :

$$R.A = (\sigma^2,0,\ldots,0)^t \qquad (2.7)$$

où R est une matrice de Toëplitz constituée des p+1 premiers coefficients d'autocorrélation. Levinson en 1947, a développé un algorithme rapide pour résoudre l'équation (2.7) et calculer les coefficients autorégressifs {a_i} avec i=1…p. Cet algorithme a été ensuite modifié par Durbin en 1960. D'autres paramètres peuvent être extraits à partir d'une analyse LPC comme les coefficients de réflexion (ou PARCOR) et les coefficients cepstraux LPCC. Ces derniers coefficients cepstraux peuvent être obtenus à partir des coefficients de la prédiction linéaire. Ainsi, les paramètres LPCC sont calculés à partir d'une analyse par prédiction linéaire décrite en dessus. Si a_0=1, {a_i} avec i=1:p sont les coefficients de cette analyse, estimés sur une trame du signal, les d premiers coefficients cepstraux C_k sont calculés récursivement par :

$$C_k = -a_k - \sum_{i=1}^{k-1} \frac{(k-i)}{k} C_{k-i} a_i \qquad 1 \le k \le d \qquad (2.8)$$

4.1.6. Détection des zones de silences

Cette étape est souvent assimilée à un processus de segmentation silence/parole, ou plus exactement « parole/non-parole ». Une des techniques employées pour cette segmentation consiste à modéliser par une bi-gaussienne l'énergie des trames du signal audio (figure 2.8). Les trames de parole utile sont supposées appartenir à la gaussienne de l'énergie haute, les autres trames considérées comme de la non-

parole (silence, bruit...) appartiennent à la gaussienne de l'énergie basse. L'énergie de trame est lissée par une fonction de fenêtre particulière. Alors, les caractéristiques d'énergie lissées sont utilisées pour construire un modèle de probabilité de parole, qui donne une probabilité qu'une trame contient la parole ou le silence. Seulement les trames que l'on étiquette comme la parole sont utilisés pour le nouveau processus.

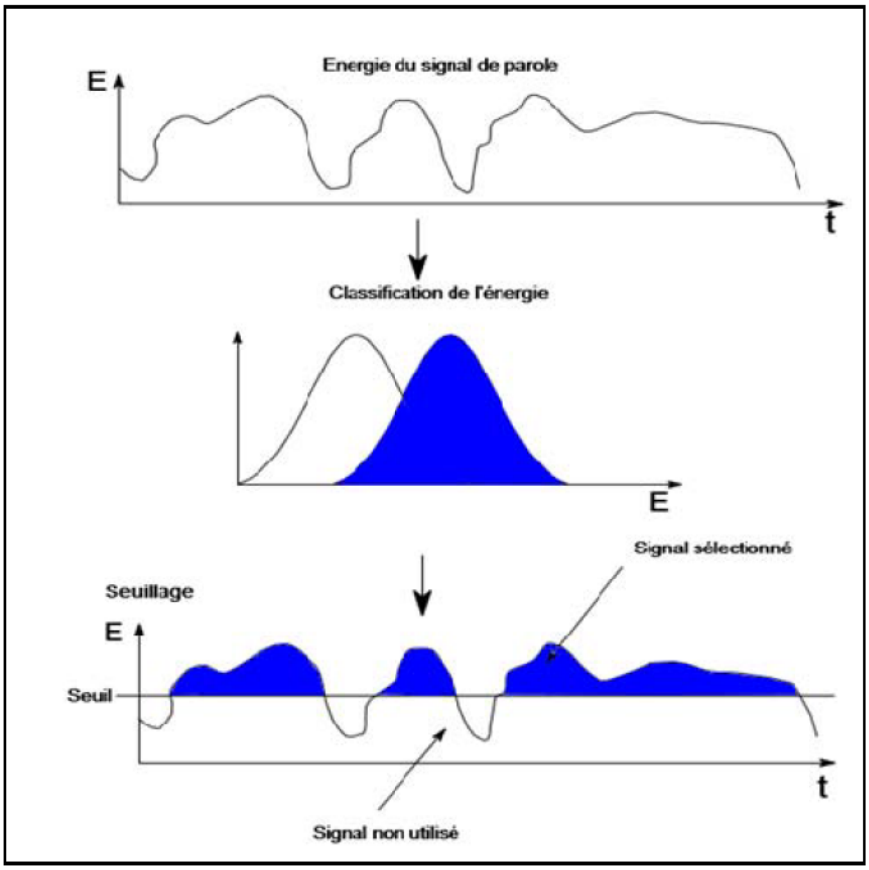

Figure 2.8 : Modélisation de l'énergie pour la Détection des zones de la parole

4.2. Caractéristiques Dynamiques et Mesure d'Énergie

Jusqu'à ce point nous avons seulement considéré les vecteurs de caractéristique statiques calculés pour chaque trame de la parole et nous avons ignoré l'évolution dynamique. L'amélioration supplémentaire des performances des systèmes de la RAP ou la RAL pourrait être obtenue en ajoutant des caractéristiques dynamiques aux vecteurs statiques. Des caractéristiques dynamiques sont les différences de temps des vecteurs de caractéristique statiques donnés par :

$$\Delta_t = \frac{\sum_{i=1}^{N} i(c_{t+1} - c_{t-1})}{2\sum_{i=1}^{N} i^2} \qquad (2.9)$$

Ces caractéristiques sont aussi appelées le delta (Δ) des coefficients. D'une façon semblable la formule (2.9) peut être appliqué pour obtenir les différences de temps d'ordre deux, qui sont appelées le delta de delta ($\Delta\Delta$) des coefficients.

L'énergie des trames de la parole est souvent utilisée dans la RAP. Elle est ajoutée comme composant supplémentaire au vecteur de caractéristique afin de rapporter une amélioration de performance de la RAP. Le logarithme de l'énergie du signal de parole peut être calculé par :

$$E = \log \sum_{i=0}^{d} s(i)^2 \qquad (2.10)$$

5. Conclusion

Nous avons décrit les différentes représentations ainsi que les diverses analyses susceptibles de nous fournir de précieuses informations du signal de parole afin de rendre les données vocales plus facile a traiter et de les rendre moins encombrantes.

CHAPITRE 3 :

APPLICATION DES TECHNIQUES DE LA RAL À LA CLASSIFICATION DES TROUBLES DE LA PAROLE

1. Introduction

Dans ce chapitre nous présentons la chaîne de reconnaissance en développant l'étape de l'extraction des caractéristiques du signal de la parole ainsi que les techniques de classification des formes qui sont utilisées en RAL. La majorité des systèmes actuels de la RAL, sont basés sur l'utilisation de modèles de mélange de Gaussiennes (GMM). Cependant, l'apprentissage génératif ne s'attaque pas directement au problème de classification étant donné qu'il fournit un modèle de la distribution jointe. Ceci a conduit récemment à l'émergence d'approches discriminantes qui tentent de résoudre directement le problème de classification, et qui donnent généralement de bien meilleurs résultats. Par exemple, les Machines à Vecteurs de Support (SVM), combinées avec les supervecteurs GMM sont parmi les techniques les plus performantes. Ces techniques serviront, dans l'élaboration de notre système de détection des troubles articulatoires de la parole, en vue de l'évaluation des prononciations phonémiques pathologiques [10].

2. Reconnaissance Automatique du Locuteur

Contrairement à la Reconnaissance Automatique de la Parole (RAP), la Reconnaissance Automatique du Locuteur (RAL) s'intéresse tout particulièrement aux informations extralinguistiques véhiculées par un signal de parole, informations porteuses de renseignements sur les spécificités d'un individu (identité, émotivité, caractéristiques physiques, particularités régionales, etc.). Son objectif est d'identifier une personne à l'aide de sa voix grâce à la variabilité interlocuteur qui permet de reconnaître une voix parmi plusieurs voix possibles. L'état de l'art des systèmes actuels de la RAL utilise une approche statistique fondée sur les théories de la détection, de la décision bayésienne et de l'information. Dans le domaine de la RAL, on distingue différentes tâches :

- l'Identification Automatique du Locuteur : qui consiste à déterminer la personne ayant prononcé un message donné, parmi un ensemble de locuteurs connus. On distingue deux modes en ensemble :
    - fermé : le locuteur à identifier est connu du système ;
    - ouvert : le locuteur à identifier peut ne pas être connu du système.

Ces applications sont peu nombreuses. En ensemble ouvert et dépendant du texte (par exemple, un même mot de passe pour les employés d'une même société), certaines applications d'IAL peuvent permettre le contrôle d'accès à un bâtiment, à un réseau ;

- Vérification Automatique du Locuteur (VAL) : consiste à déterminer la véracité de l'identité revendiquée par un individu, au moyen d'un message vocal. Ces applications sont multiples comme les serrures vocales pour le contrôle d'accès aux locaux, l'accès par le téléphone à des services distants sécurisés, la protection de matériel contre le vol (téléphones portables, voitures…) ;
- détection/suivi de locuteurs : se rapproche de la VAL. Sa tâche consiste à déterminer si un locuteur donné intervient ou non dans un document audio (conférences, débats, conversations,…). Ces applications sont principalement judiciaires et militaires. Cependant, en indexation de documents audio, elle peut faciliter la recherche d'un document audio particulier par la détection d'un locuteur connu (émission de télévision, de radio) ;
- indexation de locuteurs : consiste à cibler les interventions de locuteurs dans un document audio (conférences, débats, conversations). Ces applications sont principalement orientées sur le traitement de bases de données audio, comme par exemple la recherche de séquences d'émissions télévisées pour un locuteur particulier.

Le processus de RAL comprend 3 phases : la paramétrisation, l'apprentissage et la phase de test.

2.1. Phase de paramétrisation

La paramétrisation permet de réduire la redondance du signal de parole et d'en extraire les informations pertinentes en vue de la reconnaissance. Elle fournit ainsi une représentation simplifiée du signal nécessaire avant les phases d'apprentissage et de test. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques correspondant à des trames de signal (Généralement, la longueur varie de 20 à 31,5 ms) calculées périodiquement sur le signal de parole (par exemple, toutes les 10 ms). Suivant la nature des informations que l'on souhaite extraire du signal de parole, différentes représentations sont proposées. Celles-ci peuvent être classées en quatre grandes classes.

2.1.1. Analyse spectrale

L'analyse spectrale met en évidence les caractéristiques physiques de l'appareil phonatoire (forme du conduit vocal et nasal) de chaque individu, à travers des vecteurs de paramètres qui en sont déduits. Les paramètres les plus pertinents en RAL sont les :

- LPC obtenus à partir de la prédiction linéaire ;
- LFC et MFC (Linear/Mel Frequency Coefficients) obtenus par analyse en banc de filtres.

Pour plus de détails, nous nous référons aux travaux [46,47].

2.1.2. Analyse cepstrale

L'analyse cepstrale est une méthode qui vise à séparer la contribution de la source et du conduit vocal par déconvolution, en prenant comme hypothèse que le signal vocal est produit par un signal excitateur (source glottique) traversant le conduit vocal. Le spectre ainsi débarrassé de la contribution de la source ne contient que des informations sur le conduit vocal. Les paramètres les plus pertinents en RAL sont les :

- LPCC (Linear Predictive Cepstral Coefficients) obtenus par prédiction linéaire ;
- LFCC et MFCC (Linear/Mel Frequency Cepstral Coefficients) obtenus par analyse en banc de filtres.

2.1.3. Paramètres dynamiques

Souvent les paramètres dynamiques constituent un facteur d'amélioration des performances. Ils reflètent les phénomènes de coarticulation, les trajectoires formantiques ainsi que les informations temporelles (vitesse d'élocution, distribution des pauses). Un exemple d'exploitation des informations dynamiques a trait aux coefficients dérivés des spectres instantanés, appelés communément les coefficients Delta (ou Δ) pour la 1ère dérivée et Delta-Delta (ou ΔΔ) pour la 2ème dérivée.

static	E	delta	dE	delta delta	ddE

Figure 3.1 : Paramètres acoustiques statiques et dynamiques

2.2. Phase d'apprentissage des modèles

La phase d'apprentissage consiste à construire un modèle à partir des paramètres extraits du signal d'apprentissage afin d'obtenir le modèle du client.

2.2.1. Modélisation par GMM

En RAL, les systèmes GMM de l'état de l'art utilisent des matrices de covariance diagonales et sont appris par adaptation MAP (Maximum a Posteriori) des vecteurs moyens d'un modèle du monde. Les GMM reposent sur une modélisation statistique qui représente les valeurs des vecteurs acoustiques des caractérisations d'un locuteur, ou bien d'un ensemble de locuteurs [48].

Un GMM ″X″ est une somme pondérée de ″M″ distributions gaussiennes multidimensionnelles, chacune d'elle est caractérisée par un vecteur moyen ″x″ et une matrice de covariance ″Σ″ et un poids ″p″. Durant la phase d'apprentissage, les paramètres des GMM (le vecteur moyen ″x″ de dimension ″d″, la matrice de covariance ″Σ″ de dimension d×d, la pondération p de chaque distribution gaussienne) sont estimés par l'algorithme EM [49].

La densité de probabilité d'un vecteur y_t de dimension d s'écrit :

$$p(y_t|X) = \sum_{i=1}^{M} p_i N(y_t, \overline{x}_i, \Sigma_i) \qquad (3.1)$$

Classiquement, deux phases d'apprentissage sont nécessaires en RAL [52] (figure 3.2), apprentissage d'un modèle :

- générique de parole (aussi appelé modèle du monde) estimé par l'algorithme EM/ML sur une grande quantité de données (population de locuteurs) ;
- locuteur dérivé du modèle du monde par application des techniques d'adaptation (MAP) [53].

Cet apprentissage fait souvent appel à la technique d'Estimation par Maximum de vraisemblance (Maximum Likelihood Estimation) MLE. On utilise souvent l'algorithme Espérance-Maximisation (Expectation-maximisation) EM.

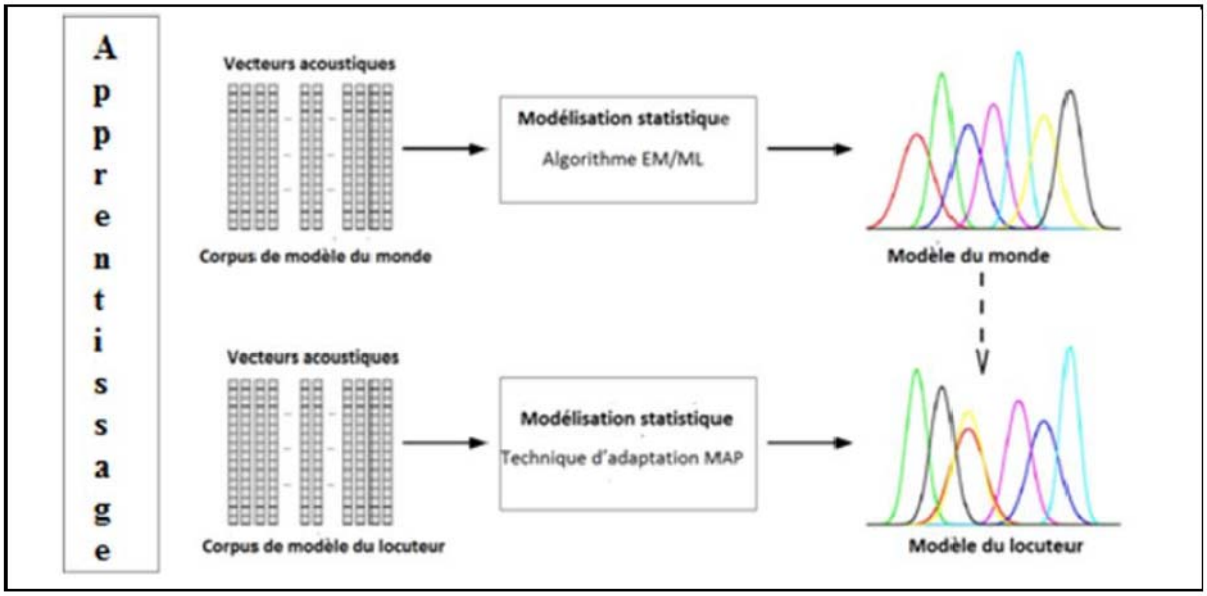

Figure 3.2 : Phase d'apprentissage d'un système de RAL

2.2.1.1. L'Algorithme EM

L'algorithme EM se compose de deux paliers. Le premier est une initialisation du modèle par Quantification Vectorielle (par exemple). Le second palier est une optimisation des paramètres du mélange par l'algorithme classique Expectation-Maximization (figure 3.3).

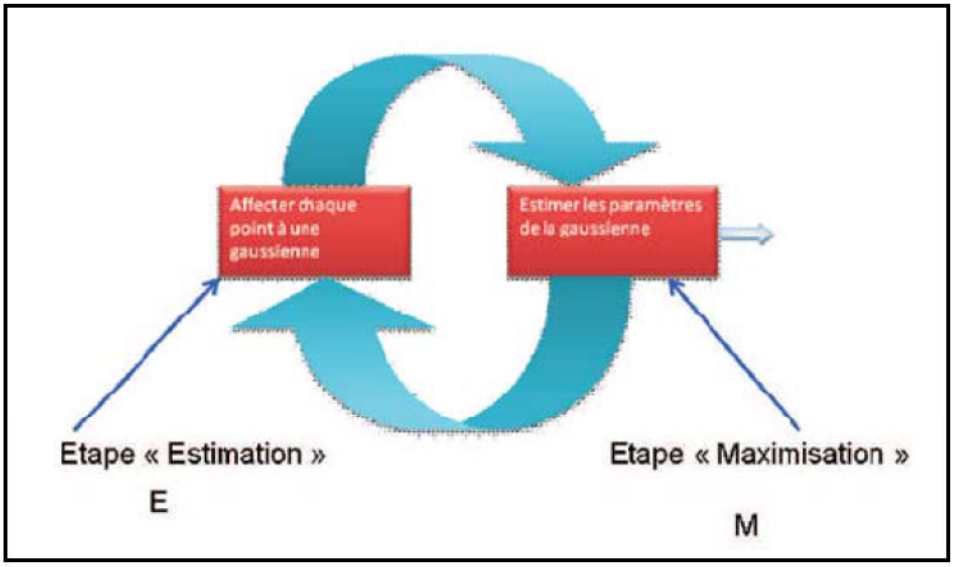

Figure 3.3 : Etapes de l'algorithme itératif EM,

La partie optimisation est un algorithme itératif qui comporte deux étapes : estimation et maximisation. Voici le détail de l'algorithme EM [49] :

- initialisation

  Utilisation de l'algorithme Llyod (Quantification Vectorielle) pour l'initialisation des moyennes des M gaussiennes du modèle.

  Initialisation de toutes les matrices de covariance $\sum_{i=1}^{N}$ à la matrice unité I.

  Initialisation équiprobable des poids des composantes : $\omega_i = 1/M$.

- itération

  pour $i = 1, \ldots, N-1$

- phase d'Estimation

  Pour tous les vecteurs acoustiques $n = 1, \ldots, T$ Calcul de la probabilité P_{ni}, probabilité que le vecteur x_n soit généré par la loi gaussienne i.

$$P_{ni} = \frac{\frac{\omega_i}{(2/pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp\left[-\frac{1}{2}\left(x_n - \mu_i\right)^T \Sigma_i^{-1}\left(x_n - \mu_i\right)\right]}{\sum_{i'=0}^{N-1} \frac{\omega_{i'}}{(2/pi)^{d/2} \mid \Sigma_{i'} \mid^{1/2}} \exp\left[-\frac{1}{2}\left(x_n - \mu_{i'}\right)^T \Sigma^{-1}\left(x_n - \mu_{i'}\right)\right]} \quad (3.2)$$

  Cette étape est équivalente à avoir un ensemble Q de variables continues cachées, prenant des valeurs dans l'intervalle [0; 1], qui donnent un étiquetage des données (vecteurs acoustiques) en indiquant dans quelle proportion un vecteur x_n appartient à la gaussienne i.

- phase de Maximisation

  Réestimation des paramètres à partir des probabilités P_{ni}.

$$\omega_i^* = \frac{1}{T}\sum_{n=1}^{T} P_{ni} \quad (3.3)$$

$$\mu_i^* = \frac{\sum_{n=1}^{T} P_{ni} x_n}{\sum_{n=1}^{T} P_{ni}} \quad (3.4)$$

$$\Sigma_i^* = \frac{\sum_{n=1}^{T} P_{ni}(x_n - \mu_i^*)(x_n - \mu_i^*)'}{\sum_{n=1}^{T} P_{ni}} \quad (3.5)$$

  Dans le cas présent, tous les vecteurs de données participent à la mise à jour du modèle, mais leur participation est proportionnelle à la valeur P_{ni}

  Incrémentation de i à $i+1$ et retour à la phase d'estimation

- arrêt de l'algorithme

  Tout d'abord, nous procédons au calcul de la vraisemblance par rapport à tous les vecteurs. Si la variation de la vraisemblance descend en dessous d'un seuil fixé alors l'estimation est terminée sinon l'estimation est reprise à l'itération suivante. La variation de la vraisemblance est négligeable à partir de 15 itérations lorsque l'initialisation est basée sur l'algorithme K-means.

2.2.1.2. Apprentissage du modèle du Monde

La notion de modèle du monde (il s'appelle aussi modèle générique ou modèle universel UBM) est introduite la première fois dans les travaux de Carey et Parris [52], pour estimer le modèle d'un ensemble non locuteur, l'objectif de ce modèle est de représenter une population générique des locuteurs. Le principal avantage de cette approche est de considérer un modèle générique indépendant des locuteurs clients. Pour la construction du modèle UBM, plusieurs approches peuvent être employées. L'approche la plus simple est de collecter toutes les données d'apprentissage pour former un seul modèle (UBM) à l'aide de l'algorithme EM [49]. Mais il faut faire un équilibrage entre les sous populations pendant le choix des données. Par exemple, si on emploie des données indépendantes du genre, on devrait être sûr qu'il y a un équilibre des discours masculins et féminins. Autrement, le modèle final sera décentré vers la sous population dominante. Le modèle du monde représente les conditions d'enregistrement, l'environnement, le type et la qualité de parole, produits dans la phase d'apprentissage.

2.2.1.3. Maximum a Posteriori

La méthode d'adaptation la plus utilisée en RAL est celle du maximum a posteriori (figure 3.4). Elle consiste à définir des distributions a priori $p(\Theta)$ pour les paramètres du modèle et à maximiser leurs probabilités a posteriori $p(\Theta|X)$ sur un signal d'apprentissage X. Le critère d'adaptation pour l'estimation des nouveaux paramètres s'écrit comme suit :

$$\hat{\Theta} = \arg\max_{\Theta} p(\Theta \mid X) = \arg\max_{\Theta} p(X \mid \Theta) p(\Theta) \tag{3.6}$$

Des formules adaptées à la modélisation GMM ont été développées en proposant un choix spécifique des densités a priori sur les paramètres [50]. Ce choix s'oriente vers les distributions a priori conjuguées permettant aux distributions a posteriori

d'appartenir à la même famille qu'aux distributions a priori. L'adoption de ces distributions permet de conserver l'utilisation de l'algorithme EM pour l'implémentation du MAP. Dans le cas des GMMs, ce choix s'oriente vers une distribution Gaussienne comme a priori pour les paramètres moyenne/variance et une distribution de Dirichlet pour les paramètres de poids. En pratique, dans un système de RAL indépendant du texte, seuls les paramètres de moyenne sont modifiés. Les moyennes du modèle du monde sont les a priori pour celles du locuteur [53]. Dans ce cas, l'estimation de la moyenne pour une composante est obtenue par une combinaison linéaire des moyennes a priori μ_k et empiriques $\bar{y}_k$, issues des données d'apprentissage.

$$\hat{\mu}_k = \frac{\eta_k}{\eta_k + \tau - k}\bar{y}_k + \frac{\tau_k}{\eta_k + \tau_k}\mu_k \qquad (3.7)$$

$$avec\ \eta_k = N * \gamma_k$$

Où γ_k est le vecteur des variables cachées d'EM et N le nombre de trames d'apprentissage. Le facteur τ, appelé facteur de relevance, permet de contrôler l'adaptation du modèle aux données en modifiant la confiance sur la distribution a priori des paramètres de moyenne. Cette formule d'adaptation pose la distribution a priori sur les moyennes comme une gaussienne de moyenne μ_k et de variance $\frac{\sigma_k^2}{\tau_k}$.

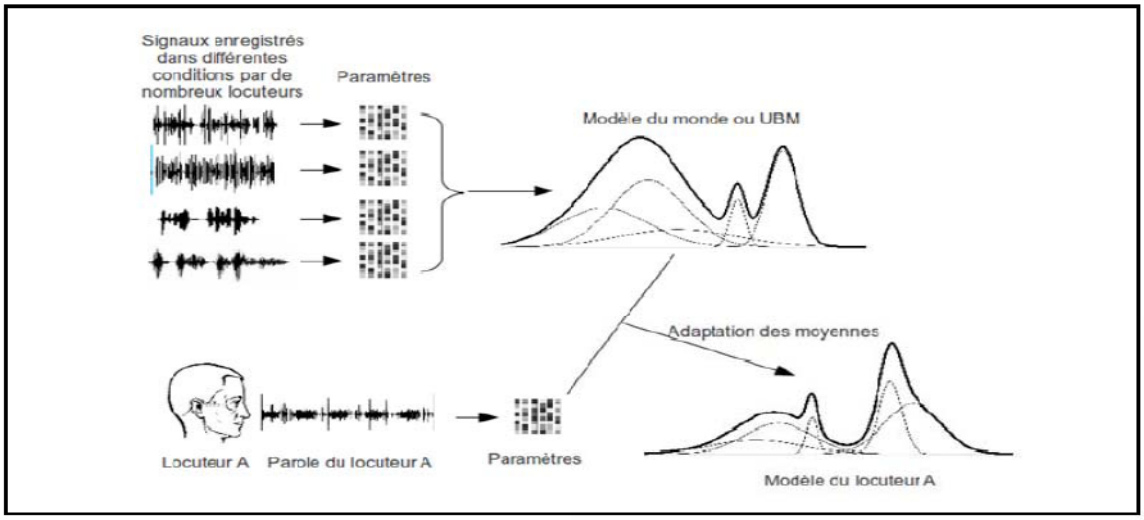

Figure 3.4 : Apprentissage du modèle de locuteur : Adaptation du modèle de monde selon les paramètres extraits du signal d'apprentissage

2.2.3. Machines à Vecteurs de Support

La majorité des systèmes de la RAL sont basés sur une modélisation générative des vecteurs cepstraux issus du signal vocal d'un locuteur. L'utilisation du paradigme GMM-UBM présenté dans la section précédente est dorénavant une étape indispensable pour obtenir des performances proches de l'état de l'art, mais ces dernières années ont vu l'apparition d'approches discriminantes présentant des performances proches des méthodes génératives.

Les méthodes à base de machines à vecteurs supports (SVM) présentent des intérêts particuliers : leur capacité à traiter des problèmes de grande dimension et la bonne réalisation du compromis complexité/généralisation. De plus, leur mise en œuvre est aisée au vu des multiples logiciels de grande qualité disponibles en licence libre pour la communauté des chercheurs. Cette section présente les Machines à Vecteurs de Support (SVM), qui regroupent une catégorie de méthodes qui ont montré de bonnes performances pour de nombreux problèmes de classification. Le fondement théorique des SVM vient de la théorie de Vapnik. A la base les SVM ont été formulés pour des problèmes de classification binaire. Leur puissance vient du critère de "marge" et de "l'astuce du noyau".

2.2.3.1. Théorie d'apprentissage de Vapnik

La théorie fondatrice des SVM repose sur deux critères qui permettent de juger de l'adéquation d'une méthode de classification pour un problème donné. Les SVM constituent une classe d'algorithmes basée sur le principe de minimisation du du risque structurel décrit par la Théorie de l'Apprentissage Statistique de Vapnik et Chervonenkis qui utilise la séparation linéaire [56].

Cela consiste à séparer par un hyperplan des individus représentés dans un espace de dimension égale au nombre de caractéristiques, les individus sont alors séparés en deux classes. Cela est possible quand les données à classer sont linéairement séparables. Dans le cas contraire, les données seront projetées sur un espace de plus grande dimension afin quelles deviennent linéairement séparables comme le montre la figure 3.5.

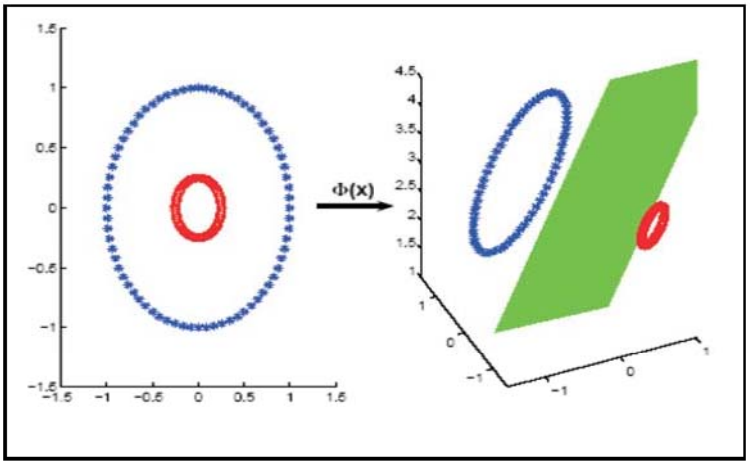

Figure 3.5 : Exemple de projection des données dans un espace plus grand

2.2.3.2. Classification binaire par hyperplan

Considérons maintenant l points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$

Avec $x_i : \epsilon \Re^N$ .et $i = 1, \ldots, l$ et $y_i \varepsilon \{\pm 1\}$.

Classons ces points en utilisant une famille de fonctions linéaires définie par :

$$wx + b = 0$$

Avec $w \epsilon \Re^N$ et $b \epsilon \Re$ de telle sorte que la fonction de décision concernant l'appartenance d'un point à l'une des deux classes soit donnée par :

$$f(x) = \text{sgn}(wx + b) \qquad (3.8)$$

2.2.3.3. Cas de données linéairement séparables

Nous présentons brièvement le problème de la classification linéaire à 2 classes, reliées étroitement avec le formalisme des SVM. Un SVM peut être exprimé comme un classifieur bi-classe [55], les stratégies d'extensions multi-classes étant souvent exprimées comme des extensions du modèle binaire. Considérons un jeu de données d'apprentissage $(x_t, y_t)_{t=1\ldots T}$

- $H_1 : wx + b = +1;$
- $H_2 : wx + b = -1.$

telle que les deux conditions suivantes soient respectées :

- il n'y a aucun point qui se situe entre H_1 et H_2. Cette contrainte se traduit par les inégalités :

- $wx_i + b \geq +1$ pour $y_i = +1$
- et $wx_i + b \leq -1$ pour $y_i = -1$
- la distance ou la marge entre H_1 et H_2 est maximale.

Dans ce cas, la distance entre H_1 et H_2 est donnée par: $M = \frac{2}{|w|}$ Maximiser M revient à minimiser |w| ou à minimiser $|w|^2$ avec $|w|^2 = w^T w$ (carré de la norme euclidienne du vecteur w). Le problème de séparation par hyper plan optimal peut être formulé comme suit :

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ \text{sous contraintes} \\ y_i (wx_i + b) \geq +1 \text{ avec } i = 1...l \end{cases} \quad (3.9)$$

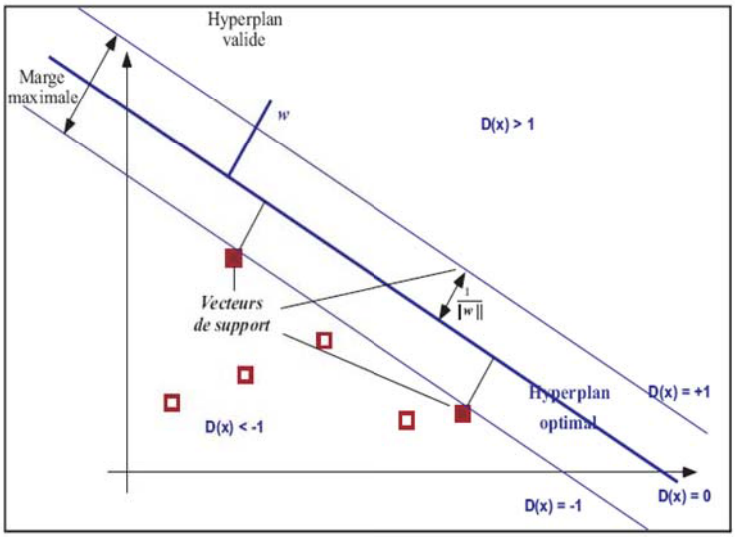

Figure 3.6. Données linéairement séparables

Ce problème d'optimisation quadratique peut être résolu en introduisant des multiplicateurs de Lagrange $\alpha_i \geq 0$. Le lagrangien associé au problème précédent d'optimisation est :

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i (wx_i + b) - 1\right) \quad (3.10)$$

Le lagrangien doit être minimisé par rapport à w et b et maximisé par rapport à α.

$$\frac{\partial L}{\partial w} = 0 \tag{3.11}$$

$$\frac{\partial L}{\partial b} = 0 \tag{3.12}$$

et les $\alpha_i \geq 0$ à partir des relations (3.11) et (3.12) nous pouvons déduire :

$$w = \sum_{i=1}^{I} \alpha_i y_i x_i \text{ et } \sum_{i=1}^{I} \alpha_i y_i = 0 \tag{3.13}$$

En les remplaçant dans $L(w, b, \alpha)$, on obtient le problème dual :

$$\begin{cases} L_D = \sum_{i=1}^{I} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{I} \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{sous contraintes} \\ \sum_{i=1}^{I} \alpha_i y_i = 0 \\ \text{et } \alpha_i \geq 0 \\ \text{et } i = 1,\ldots,I; \end{cases} \tag{3.15}$$

La fonction de décision est alors :

$$f(x) = \text{sgn}\left(\sum_{i=1}^{I} \alpha_i y_i (x_i x) + b \right) \tag{3.16}$$

Cette fonction de décision est donc seulement influencée par les points correspondants à des α_i non nuls. Ces points sont appelés les Vecteurs de Support. Ils correspondent, dans un cas linéairement séparable, aux points les plus proches de la limite de décision, c'est-à-dire aux points se trouvant exactement à une distance égale à la marge.

Il s'agit là d'une propriété très intéressante des SVM: seuls les Vecteurs de Support sont nécessaires pour décrire cette limite de décision, et le nombre de Vecteurs de Support pour le modèle optimal est généralement petit devant le nombre de données d'entraînement.

2.2.3.4. Cas des données non-linéairement séparables

En pratique, il est assez rare d'avoir des données linéairement séparables. Afin de traiter également des données bruitées ou non linéairement séparables, les SVM ont été généralisées grâce à deux outils : la marge souple (soft margin) et les fonctions noyau (kernel functions). Le principe de la marge souple est d'autoriser des erreurs de classification. Le nouveau problème de séparation optimale est reformulé comme suit : L'hyperplan optimal séparant les deux classes est celui qui sépare les données avec le minimum d'erreurs, et satisfait donc les deux conditions suivantes :

- la distance entre les vecteurs bien classés et l'hyperplan doit être maximale.
- la distance entre les vecteurs mal classées et l'hyperplan doit être minimale.

Pour formaliser cela, on introduit des variables de pénalité non-négatives, ε_i pour $i = 1, \ldots, l$ appelées variables d'écart. Le principe de la marge souple se traduit par la transformation des contraintes (III.2) qui deviennent :

$$y_i\left(wx_i + b\right) \geq +1 - \varepsilon_i \text{ Pour } i = 1,\ldots,l \tag{3.17}$$

Avec l'introduction d'un terme de pénalité, la fonction objective devient :

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \varepsilon_i \qquad C \geq 0. \tag{3.18}$$

Le paramètre C est défini par l'utilisateur. Il peut être interprété comme une tolérance au bruit du classificateur. C'est aussi la pénalité associée à toute violation des contraintes du cas linéairement séparable. Pour de grandes valeurs de C, seules de très faibles valeurs de ε sont autorisées et, par conséquent, le nombre de points mal classés sera trés faible (données faiblement bruitées). Cependant, si C est petit, f peut devenir assez grand et on autorise alors bien plus d'erreurs de classification (données fortement bruitées). La nouvelle formulation du problème d'optimisation est alors :

$$\begin{cases} \min_{w,b,\varepsilon} \frac{1}{2} w^T w + C\sum_{i=1}^{I} \varepsilon_i \; C \geq 0 \\ \quad \text{sous contraintes} \\ y_i \left(w x_i + b \right) \geq +1 - \varepsilon_i \\ \text{et} \;\; \varepsilon_i \geq 0 \quad \text{pour } i = 1,\ldots,I; \end{cases} \tag{3.19}$$

En introduisant les multiplicateurs de Lagrange, le lagrangien associé au nouveau problème d'optimisation devient :

$$\begin{aligned} L(w,b,\varepsilon_i,\alpha,\mu) &= \frac{1}{2} w^T w + C\sum_{i=1}^{I} \varepsilon_i - \sum_{i=1}^{I} \alpha_i \left[y_i \left(w^T x_i - b \right) + \varepsilon_i - 1 \right] - \sum_{I}^{i=1} \mu_i \varepsilon_i \\ &= \frac{1}{2} w^T w + \sum_{i=1}^{I} \left(C - \alpha_i - \mu_i \right) \varepsilon_i - \left(\sum_{i=1}^{I} \alpha_i y_i x_i \right) w - \left(\sum_{i=1}^{I} \alpha_i y_i \right) b + \sum_{i=1}^{I} \alpha_i \end{aligned} \tag{3.20}$$

Le lagrangien doit être minimisé par rapport à w, b, ε_i et maximisé par rapport à α et μ.

$$\frac{\partial L}{\partial w} = 0 \tag{3.21}$$

$$\frac{\partial L}{\partial b} = 0 \tag{3.22}$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \tag{3.23}$$

De ces dernières relations, nous tirons les trois égalités suivantes :

$$w = \sum_{i=1}^{1} \alpha_i y_i x_i \tag{3.24}$$

$$\sum_{i=1}^{I} \alpha_i y_i = 0 \quad \text{et } \alpha_i = C - \mu_i \tag{3.25}$$

Ce qui conduit à un problème dual légèrement différent de celui du cas séparable :

$$\begin{cases} L_D = \sum_{i=1}^{I} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{I} \alpha_i\alpha_j y_i y_j x_i x_j \\ \sum_{i=1}^{I} \alpha_i y_i = 0 \quad \text{et } \alpha_i \geq 0 \\ \alpha_i \leq C \text{ et } i = 1,\ldots,I \end{cases} \tag{3.26}$$

La seule différence avec le cas linéairement séparable est donc l'introduction d'une borne supérieure pour les paramètres α_i. Il est également interessant de noter que les points se trouvant sur la limite de decision sont tous des vecteurs de support, quelle que soit leur distance à cette limite, ce qui signifie qu'ils exercent une influence sur le calcul de cette limite. Pour le cas des données qui ne sont pas linéairement séparables, L'idée est de projeter l'espace d'entrée (espace des données) dans un espace de plus grande dimension appelée espace des caractéristiques (feature space) afin d'obtenir une configuration linéairement séparable (à l'approximation de la marge souple) de nos données, et d'appliquer alors l'algorithme des SVM. Cette projection est équivalente à l'application d'une transformation sur les données initiales par l'intermédiaire d'une fonction ϕ. Le nouvel algorithme peut donc être écrit ainsi :

Soit la fonction $\phi: \Re^N \rightarrow \Re^M, M > N : x \rightarrow \phi(x)$

$$L_D = \sum_{i=1}^{I} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{I} \alpha_i\alpha_j y_i y_j \left(\phi(x_i)\phi(x_j)\right) \tag{3.27}$$

$$\sum_{i=1}^{I} \alpha_i y_i = 0 \quad \text{et } \alpha_i \geq 0 \quad \alpha_i \leq C \text{ et } i = 1,\ldots,I \tag{3.28}$$

2.2.3.5. Fonctions noyaux

Afin d'effectuer des décisions non linéaires en utilisant le SVM, il n'est pas nécessaire de définir une transformation explicite car ce genre de transformation peut devenir très coûteux du point de vue calcul pour de grandes valeurs. En analysant les formules (3.27) et (3.28), on remarque que les vecteurs d'entrée se présentent dans les fonctions objectives sous formes de produits scalaires entre les paires de vecteurs. L'astuce est de calculer le produit scalaire dans l'espace des caractéristiques en fonction des vecteurs de l'espace d'entrée directement [57]:

$$u' \cdot v' = u \cdot v + u_1 \cdot u_2 \cdot v_1 \cdot v_2 \tag{3.29}$$

Donc on peut définir le noyau :

$$K(u,v) = u \cdot v + u_1 \cdot u_2 \cdot v_1 \cdot v_2 \tag{3.30}$$

Les produits scalaires dans les formules (3.29) et (3.30), peuvent être remplacés par une fonction noyau. On peut utiliser n'importe quelle fonction noyau valide (satisfaisant la condition de Mercer) sans avoir besoin de connaître des informations sur la transformation linéaire qui lui a donné lieu [58]. C'est également plus efficace que d'effectuer des transformations non-linéaires sur les données puis calculer leurs produits scalaires séparément.

- Condition de Mercer

Théorème : (condition de Mercer) La fonction $K(u,v): X \times X \to \Re$ est un noyau si est seulement si :

$$G = K(x_i, x_j), \quad i,j = 1,\ldots,n \tag{3.31}$$

est définie positive. Notons qu'une fonction $K: X \times X \to \Re$ générant une matrice définie positive possède les trois propriétés fondamentales du produit scalaire :

- Positive : $K(x_i, x_j) > 0$
- Symétrie : $K(x_i, x_j) = K(x_j, x_i)$
- Inégalité de Cauchy-Shwartz : $\left|K(x_i, x_j)\right| \leq \| x_i \| \cdot \| x_j \|$

La condition de Mercer nous indique si une fonction est un noyau mais ne fournit aucune information sur la fonction ϕ (et donc sur l'espace des caractéristiques) induit par ce noyau.

- Exemple de noyaux
  - ✓ Le noyau Linéaire

$$K(x,y) = \langle x, y \rangle \tag{3.32}$$

✓ Le noyau Polynomial

$$K(x,y) = (a.\langle x,y\rangle + b)^d \tag{3.33}$$

Prenons une instance simple de cette fonction : $K(x,y) = (\langle x,y\rangle)^2$ et essayons de trouver un candidat Φ tel que:

$$(\langle x,y\rangle)^2 = \langle \Phi(x), \Phi(y)\rangle \tag{3.34}$$

En supposant que l'espace d'entrée est de dimension 2, on peut utiliser les projections suivantes :

$$\Phi_1 : \Re^2 \to \Re^3 \; x \to \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right) \tag{3.35}$$

$$\Phi_1 : \Re^2 \to \Re^4 \; x \to \left(x_1^2, x_1x_2, x_1x_2, x_2^2\right) \tag{3.36}$$

Cet exemple montre que la fonction de projection Φ et l'espace des caractéristiques ne sont pas uniques. En général, quand on utilise un noyau polynomial, on prend des paramètres a et b égaux à 1.

✓ Le noyau Polynomial réel de Vovk

La forme générique de ce noyau est:

$$K(x,y) = \frac{1-(x,y)^d}{1-(x,y)} \tag{3.37}$$

Avec : $-1 < x, y < 1$.

✓ Le noyau Polynomial réel infini de Vovk

Sa forme générique est de la forme:

$$K(x,y) = \frac{1}{1-(x,y)} \tag{3.38}$$

avec $-1 < x, y < 1$.

✓ Le noyau RBF (Radial Basis Function)

La forme générique de ce noyau « kernel » est:

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{3.39}$$

2.2.4. Système hybride GMM-SVM proposé

L'approche majoritairement utilisée en RAL est basée sur les modèles génératifs pour représenter le locuteur. L'utilisation du paradigme GMM-UBM [51-54] apparaît maintenant comme une étape indispensable pour obtenir des performances proches de l'état de l'art dans des campagnes d'évaluation internationales telles que les campagnes NIST-SRE. Ces dernières années ont vu l'apparition d'approches discriminantes basées sur l'utilisation des (SVM). Nous présentons une méthode simple et peu coûteuse permettant de combiner les approches génératives et discriminantes (figure 3.7).

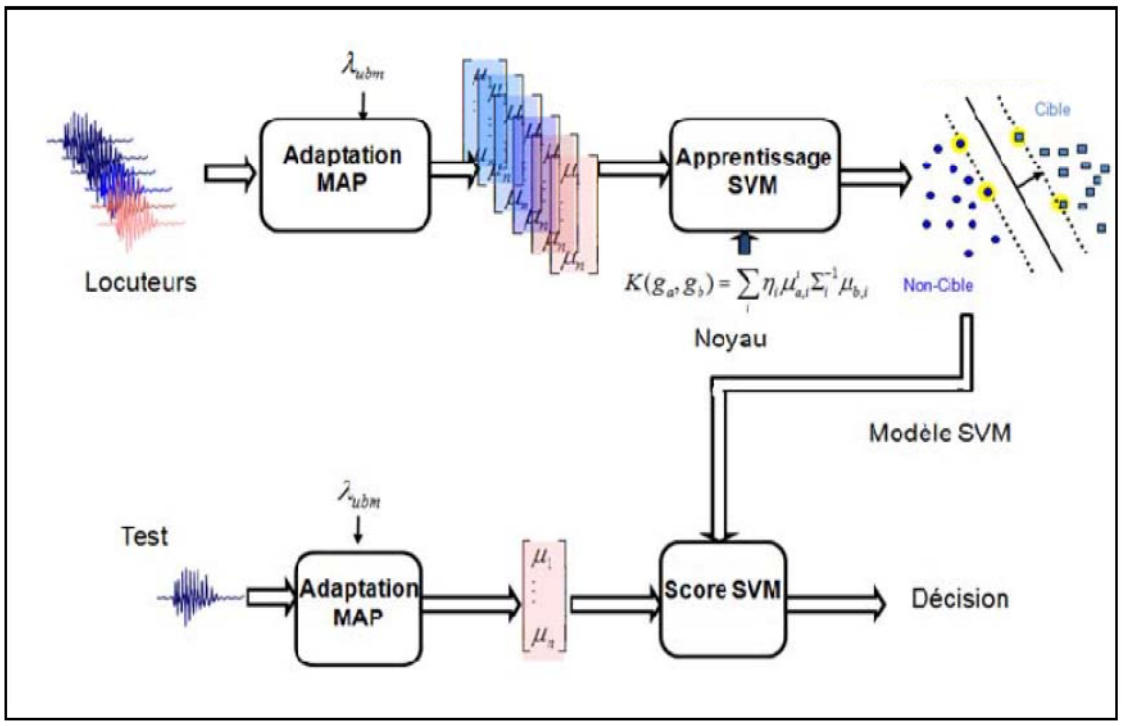

Figure 3.7 : Structure générale d'un système SVM-GMM

Le vecteur d'entrée pour le noyau d'apprentissage SVM est les moyennes des GMM de chaque locuteur, Ces modèles SVM sont enregistrés dans une base de données, pour les comparer avec les modèles GMM des signaux tests, en utilisant un noyau SVM de tests.

Dans ce cadre, la combinaison des méthodes discriminantes et génératives est particulièrement intéressante. De par leur capacité à bien représenter les données, les modèles (GMM) sont souvent employées comme modèle génératif pour ces techniques.

2.2.4.1. Supervecteur GMM

Soit un modèle GMM-UBM

$$p(y_t|X) = \sum_{i=1}^{M} p_i N(y_t, \overline{x}_i, \Sigma_i) \qquad (3.40)$$

Où ″p_i″ sont les poids de mélange, ″ N″ est une gaussienne, ″$\overline{x}_i$″ et ″Σ_i″ sont la moyenne et la covariance des gaussiennes, respectivement. Le GMM supervecteurs comprend les moyennes des modèles aprés l'adaptation (figure 3.8).

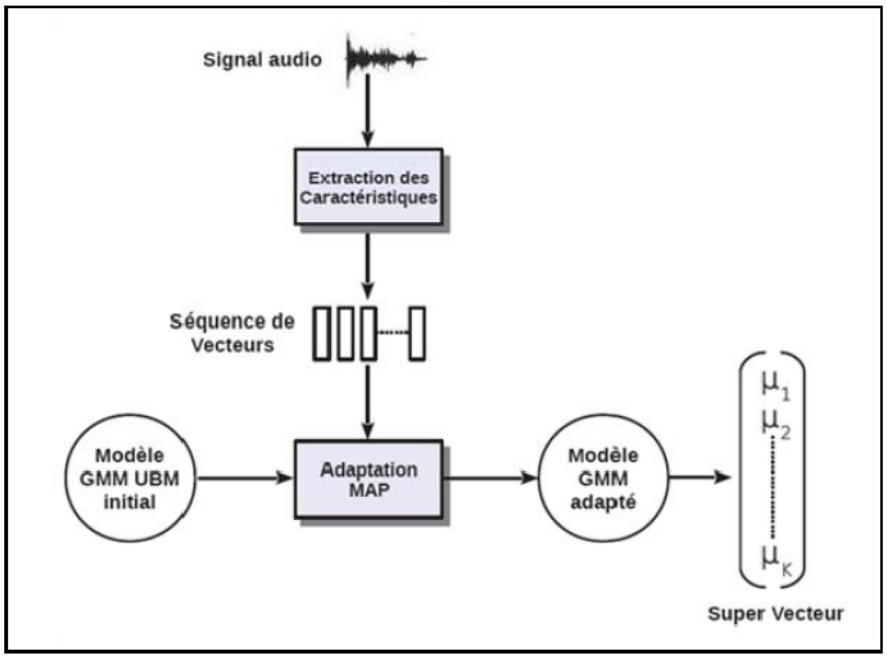

Figure 3.8 : Supervecteur des moyennes GMM

2.3. Phase de test et mesure de ressemblance

Dans la phase de test, pour chaque segment de test x et une identité x_i, un vecteur d'entrée est construit de la même façon que dans l'apprentissage. Un score de décision est obtenu par la fonction de classement des SVM suivante :

$$f(x) = \sum_i \alpha_i \, y_i K(x, x_i) + b \qquad (3.42)$$

Où x_i est un vecteur support du modèle SVM du client, α_i est le coefficient de Lagrange correspondant au vecteur support x_i, y_i est la classe de x_i, $K(x, x_i)$ représente le noyau utilisé pour apprendre le modèle SVM du client, et b est le biais du modèle SVM du client.

Supposons que dans l'apprentissage des modèles SVM des clients, les vecteurs d'entrée représentant la classe client ont été étiquetés par (1) et les vecteurs représentant la classe non-clients ont été étiquetés par (-1). Si la classe ˝X_0˝ est positive alors le système décide que le segment X_0 est prononcé par le client, sinon le système décide que le segment X_0 provient d'un imposteur.

3. Adaptation des techniques de RAL pour la reconnaissance des troubles de la parole

L'approche statistique à base de GMM-SVM a été mise à l'épreuve dans le cadre de la RAL lors de la campagne d'évaluation NIST (National Institute of Standards and Technologies) [54] des systèmes. Elle sera adaptée à la tâche de reconnaissance des troubles de la parole à différents niveaux comme le montrent les sections suivantes.

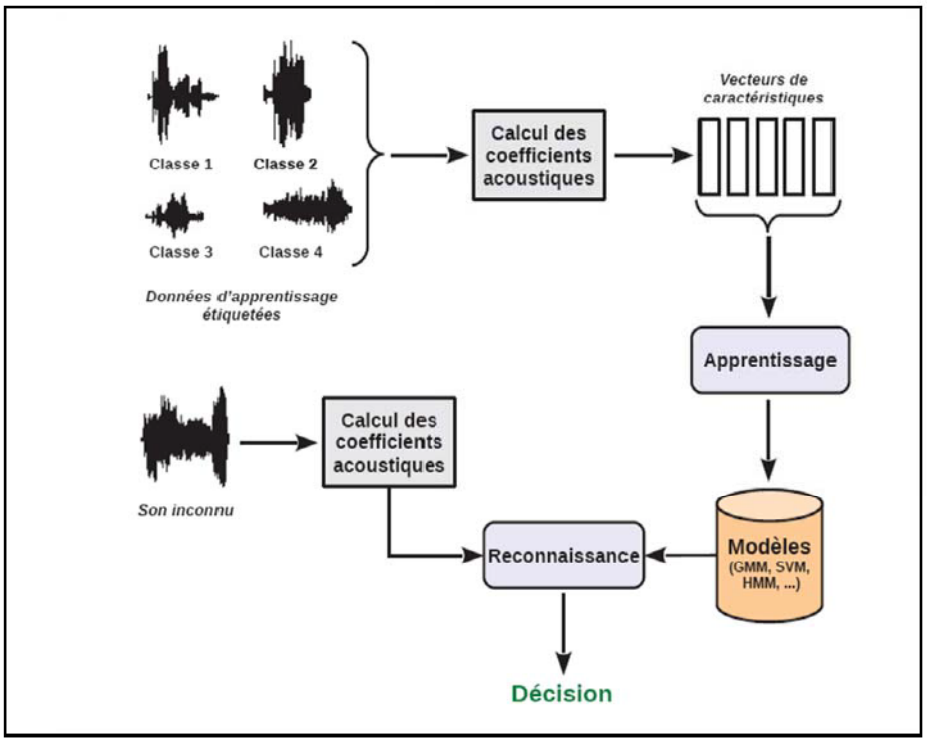

Figure 3.9 : Architecture de base d'un système de reconnaissance d'un son

3.1. Modèle des clients, de monde et prise de décisions

Dans le contexte pathologique, un modèle ne correspond plus ici à un locuteur donné mais à modèle des troubles de la parole. Le modèle est appris en utilisant l'ensemble des locuteurs de même classe. On s'assurera que les voix utilisées pour l'apprentissage des modèles des troubles de la parole, sont exclues des jeux de tests afin de différencier la détection des troubles de la parole, de la reconnaissance du locuteur.

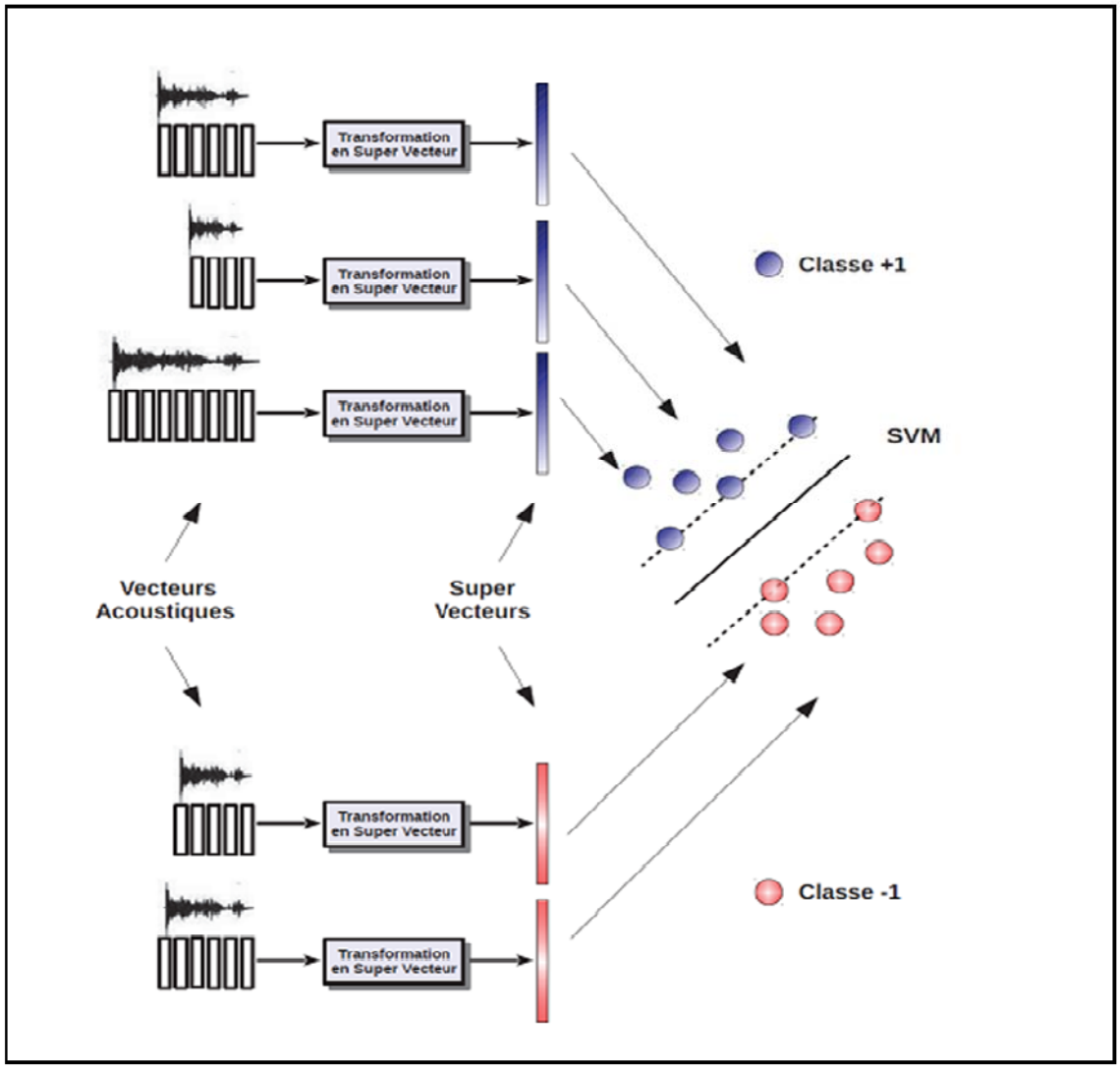

Figure 3.10 : Des vecteurs acoustiques au super vecteurs SVM

3.2. Tâches de classification

Il existe deux tâches de classification :

- la classification Contrôle/Pathologique : la première action consistera à observer si un système de RAL, adapté à notre sujet, réagit favorablement à

la classification binaire c'est-à-dire détecte si une voix donnée est reconnue en tant que voix pathologique ou voix normale (= contrôle) ;

- la classification des phonèmes : la seconde phase analysera le comportement du système à une classification par phonème. L'ensemble des voix sera testé à travers le système.

3.3. Prise de décision

Un supervecteur GMM est constitué par les moyennes des composantes du GMM. Soit deux énoncés utt_m (pour un L_P) et utt_n (pour un L_N), avec deux modéles GMM apprentis respectivement g_m et g_n.

On peut utiliser comme noyaux la distance entre les deux modèles en calculant la divergence de KL (Kullback-Leibler) :

$$D(g_m \| g_n) = \int g_m(x) \log\left[\frac{g_m(x)}{g_n(x)}\right] dx \qquad (3.43)$$

Cependant, la divergence KL ne satisfait pas à la condition de Mercer, qui est la condition optimale, et n'est donc pas susceptible d'être mise en œuvre dans le système SVM [55-58]. Une autre approche se fait par approximation:

$$0 \le D(g_m \| g_n) \le d(m^m, m^n),$$

$$\text{Où } d(m^m, m^n) = \frac{1}{2}\sum_{i=1}^{N} \lambda_i (m_i^m - m_i^n)^t \Sigma_i^{-1} (m_i^m - m_i^n) \qquad (3.44)$$

L'inégalité signifie que la divergence correspondant serait petite si la distance entre mm et mn est petite. Selon cette approche, le noyau résultant est :

$$K(utt_m, utt_n) = \sum_{i=1}^{N} \lambda_i \, (m_i^m)^t \Sigma_i^{-1} m_i^n$$

$$= \sum_{i=1}^{N} (\sqrt{\lambda_i}\Sigma_i^{-\frac{1}{2}} m_i^m)^t \left(\sqrt{\lambda_i}\Sigma_i^{-\frac{1}{2}} m_i^n\right). \qquad (3.45)$$

La décision correspond à une Acceptation ou Rejet suivant que le score est supérieur à un seuil, ce seuil peut être calculé tel que le plus petit chevauchement entre les deux histogrammes présente la plus petite probabilité d'erreur (figure 3.8).

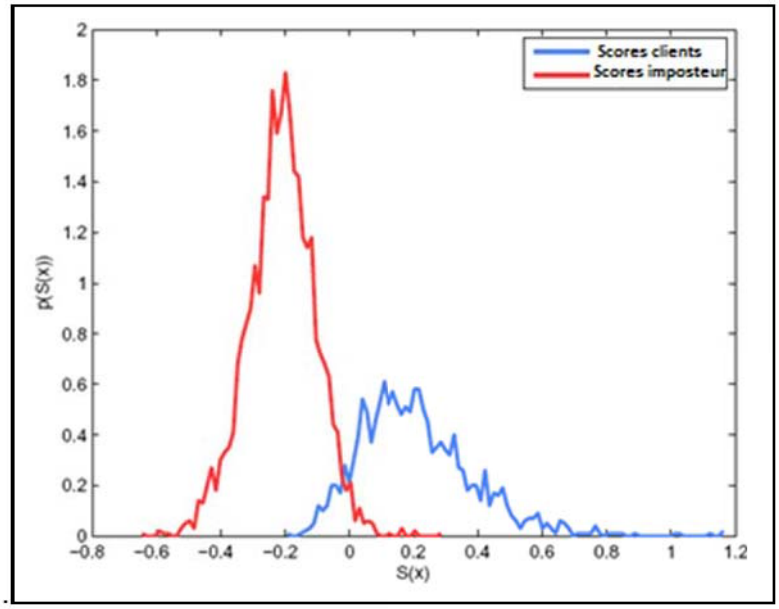

Figure 3.11 : Distributions de score pour un locuteur cible et des imposteurs

4. Conclusion

Dans ce chapitre nous avons présenté les principales techniques d'extraction des caractéristiques du signal vocal, ainsi que les techniques de RAL: GMM-UBM, les systèmes hybrides SVM-GMM. Nous avons aussi présenté les principaux détails de chaque technique utile en RAL en citant les algorithmes utilisés, ceci étant un préambule sur le background mathématique nécessaire pour la réalisation du travail que nous nous sommes fixé, à savoir la réalisation d'un système d'aide aux orthophonistes où aux malades eux- mêmes ayant des pathologies de la parole.

CHAPITRE 4 :

IMPLEMENTATION DU SYSTEME DE CLASSIFICATION DES TROUBLES DE LA PAROLE (SCTP)

1. Introduction

Le traitement du signal vocal est un processus très complexe en termes d'extraction de l'information utile et sa reconnaissance, car il concerne l'une des problématiques les plus influentes dans cette phase. Le corpus sur lequel s'effectuent les tests de performance est fondamental étant donné que l'extraction des paramètres acoustiques est sujette aux différentes influences qui sont dues à l'enregistrement, au réglage du niveau sonore du microphone, au bruit environnant, les défauts de prononciation non pathologiques, la prononciation incorrecte d'un mot et lors de la segmentation manuelle, etc. Tous ces défauts influent directement, d'une façon décisive, sur la phase de la reconnaissance.

Dans ce chapitre, nous présentons dans un premier temps les modules de l'application inspirés des classes de la plateforme Alize pour les deux systèmes GMM-UBM et GMM-SVM, dans un second temps, Les performances et les résultats obtenus sont discutés [60].

2. Architecture de l'implémentation du SCTP

Les expériences ont été réalisées après l'adaptation du système de RAL du LIA-ALIZE à la classification des troubles articulatoires de la parole Ce système de RAL (appelé LIA_SpkDet) repose sur la plateforme libre Alize [59,60] conçue et réalisée dans le cadre du programme Technolangue. LIA_RAL est une plate-forme biométrique contenant tout le code spécifique à la reconnaissance du locuteur, bien que la plupart de ces codes soient également utiles dans la reconnaissance de la parole. LIA_RAL est mise en œuvre en C ++ sous le système d'exploitation LINUX. LIA_RAL est subdivisée en sous parties pour la vérification du locuteur. Cette dernière constitue le système de base que le LIA présente aux évaluations NIST-SRE chaque année. Ce « package » est diffusé à la communauté chaque année, permettant aux autres laboratoires de reproduire les expériences.

D'un point de vue de l'architecture de base, l'application est construite autour de plusieurs serveurs de données et de calcul :

- le serveur de données audio qui va stocker les données issues soit d'un microphone, soit d'un fichier, soit d'autres sources ;

- le serveur de paramètres qui va stocker les paramètres issues soit d'un fichier, soit d'un calcul réalisé à partir des données audio ;
- le serveur de mélanges/distributions sert à stocker les modèles de parole (mixtures de gaussiennes) calculés à partir des paramètres ou chargés à partir de fichiers ;
- le serveur de statistiques regroupe les algorithmes de base les plus courants (calcul de vraisemblance, EM) et permet de conserver et d'accumuler les résultats de calculs pour réaliser des moyennes sur un ensemble de paramètres (Figure 4.1).

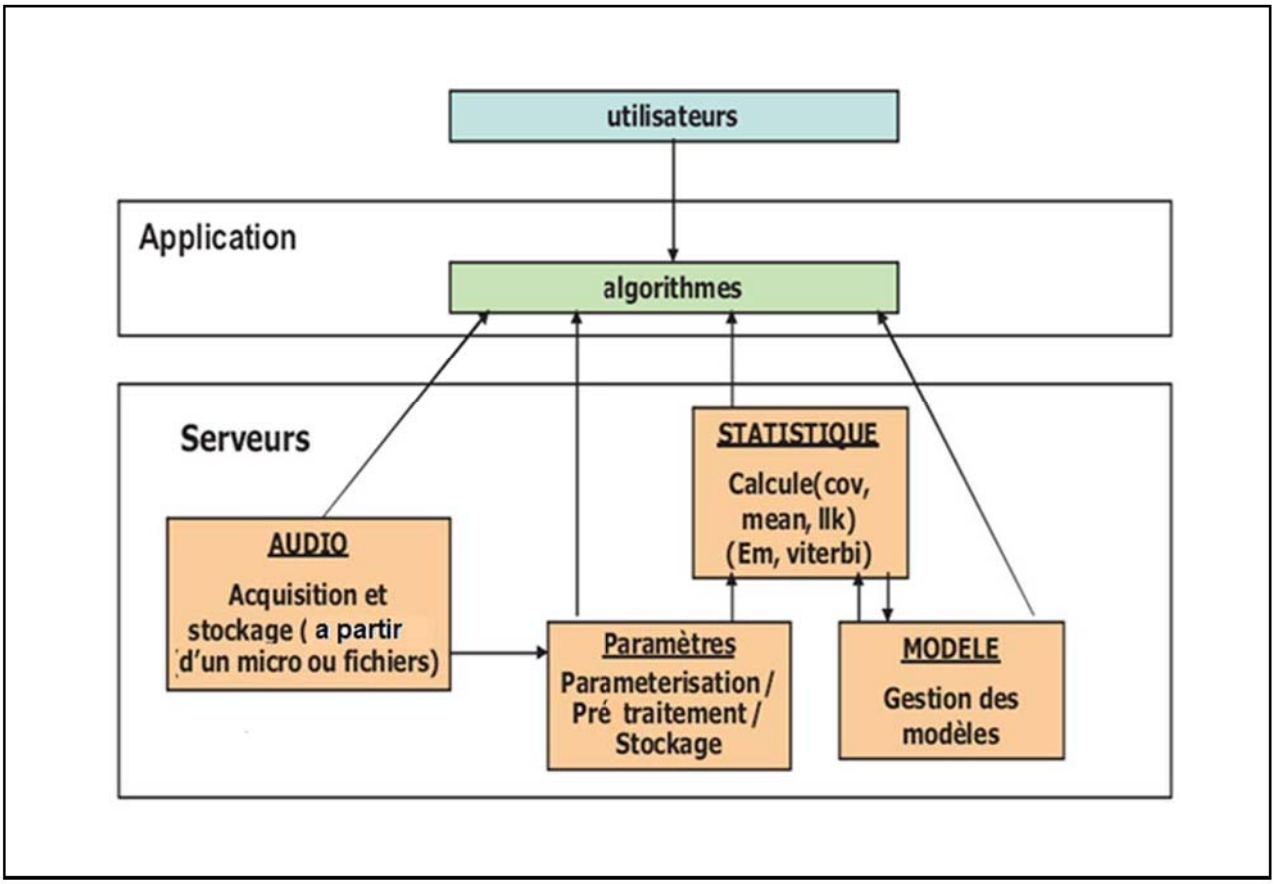

Figure 4.1 : Architecture de l'application SCTP inspirée de la plateforme ALIZE [59]

2.1. Paramétrisation

Le bruit dans l'environnement est le problème le plus évident qui peut modifier la forme du signal audio. Une représentation plus robuste tels que les coefficients Cepstrum est nécessaire. Pour la phase de paramétrisation, le module standard Spro [61] du consortium ELISA est utilisé (figure 4 2).

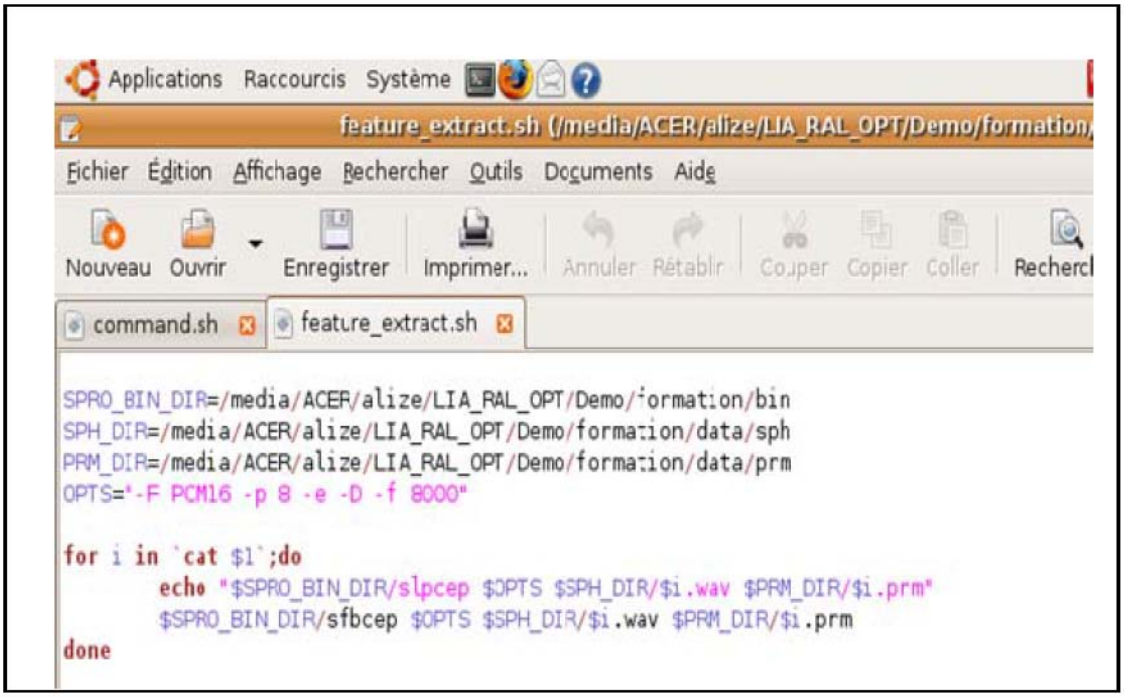

Figure 4.2 : Fichier de paramétrisation des corpus

2.2. Détection d'énergie

Dans le monde réel, il n'est souvent pas possible à assumer chaque fichier de caractéristiques pour être emballé avec des données de parole. Plusieurs fois, il y a du silence et du bruit ambiant qui amène les données de parole à être floues. Pour éviter l'apprentissage et les tests sur des vecteurs de caractéristiques qui ne sont pas la parole, la plupart des systèmes font la détection de l'énergie.

Le module de *energydetector*, lit une liste des fichiers des paramètres (ou un fichier de paramètres) et les segmentés suivant l'énergie de la parole ou non parole. Le processus de sélection des trames est fait par l'intermédiaire d'un analyseur GMM (utilisation 2 ou 3 gaussiennes) appris sur chaque vecteur de paramètres.

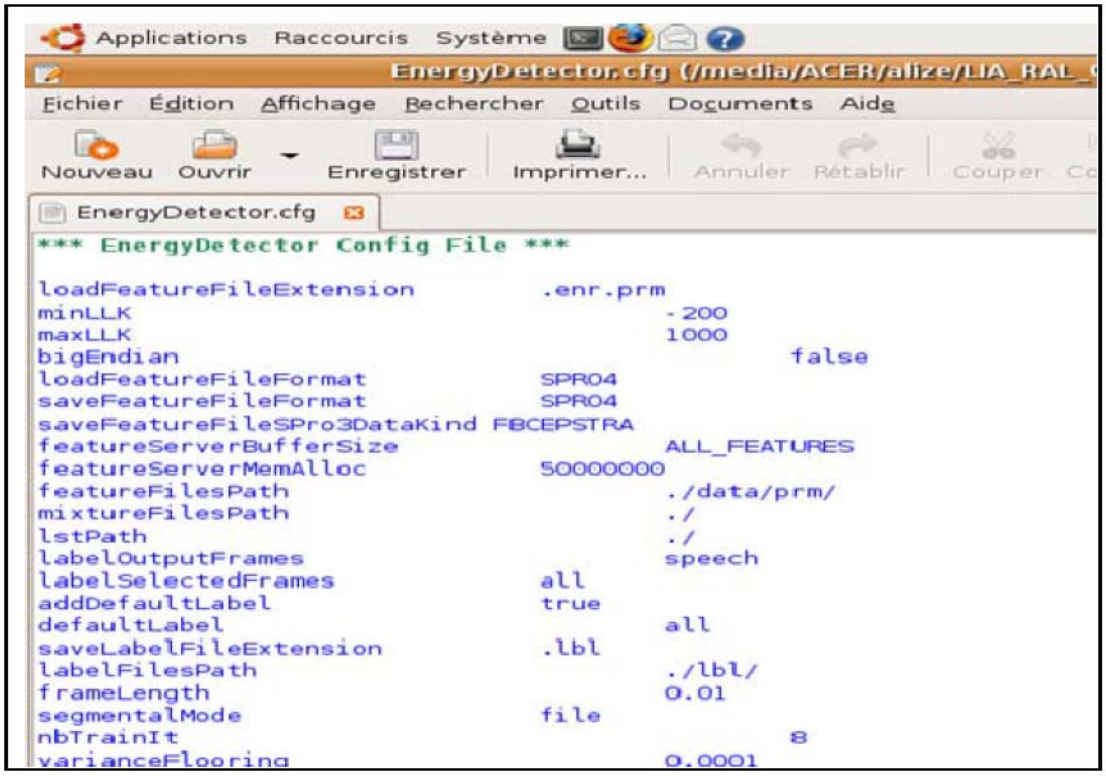

Figure 4.3 : Fichier de configuration de détection d'énergie

2.3. Normalisation des paramètres

La normalisation moyenne variance cepstrale est une technique très simple et très répandue en RAL. Elle consiste à retirer la moyenne de la distribution de chacun des paramètres cepstraux (la composante continue), et à ramener la variance à une variance unitaire en les divisant par l'écart type global des paramètres acoustiques. Le passage du domaine temporel aux domaines log-spectral et cepstral, transforme les bruits convolutifs en des bruits additifs. Des bruits convolutifs variant lentement dans le temps seront alors représentés par une composante additive presque constante tout au long de l'enregistrement de parole. Par conséquent, la suppression de la composante continue permet de réduire l'effet de ces bruits. L'estimation de la moyenne et de la variance est réalisée sur l'intégralité de la séquence de parole. Le module *Normfeat* lit une liste des fichiers des paramètres (ou un fichier de paramètres) et les normalisés sur la base des segments (donnés par les fichiers Label et générés par le module *energydetector*).

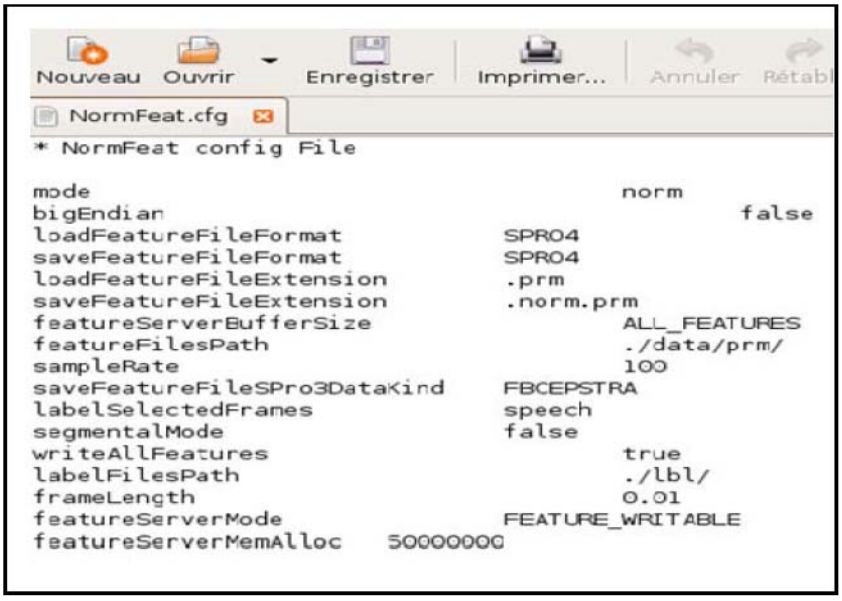

Figure 4.4 : Fichier de configuration pour la normalisation ds paramètres

2.4. Modèle de monde

L'apprentissage du modèle de monde avec les fichiers d'apprentissage de monde est exécutée par le module *Trainworld*. L'entrée est une liste des segments des paramètres, la sortie est un GMM. Initialement, toutes les distributions sont placées à variance et moyenne globale, et tous les poids sont égaux. Alors le modèle est itérativement apprentis avec tous les vecteurs des paramètres en utilisant l'algorithme EM avec un critère ML.

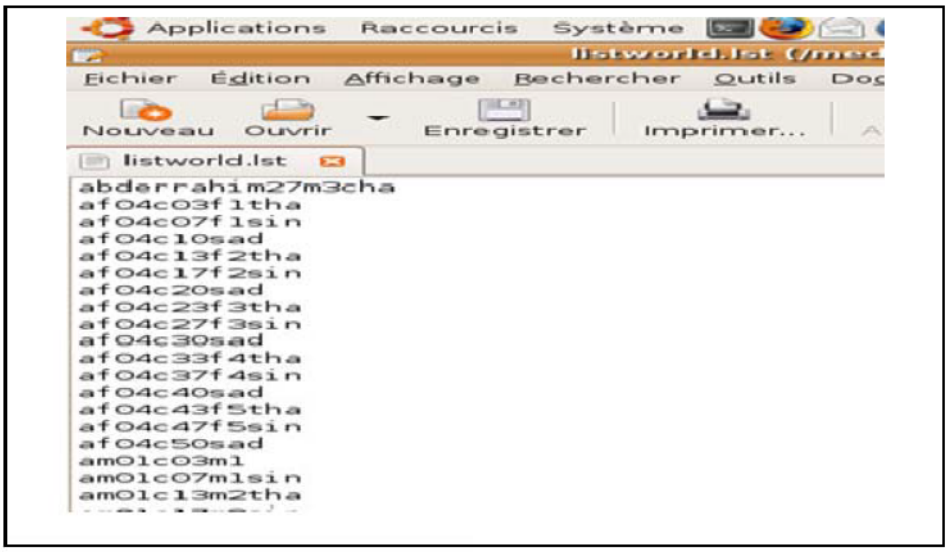

Figure 4.5 : Génération de la liste des corpus utilisée pour l'apprentissage du modèle de monde

2.5. Modèle locuteur

Une fois qu'un modèle de monde a été créé, le module *Traintarget* est exécuté pour adapter le modèle de monde pour chaque corpus (en employant l'algorithme EM modifié avec un critère MAP). L'entrée au programme est une liste, avec des lignes tel que la première colonne indique l'identité de locuteur, les autres colonnes sont les noms des fichiers des paramètres.

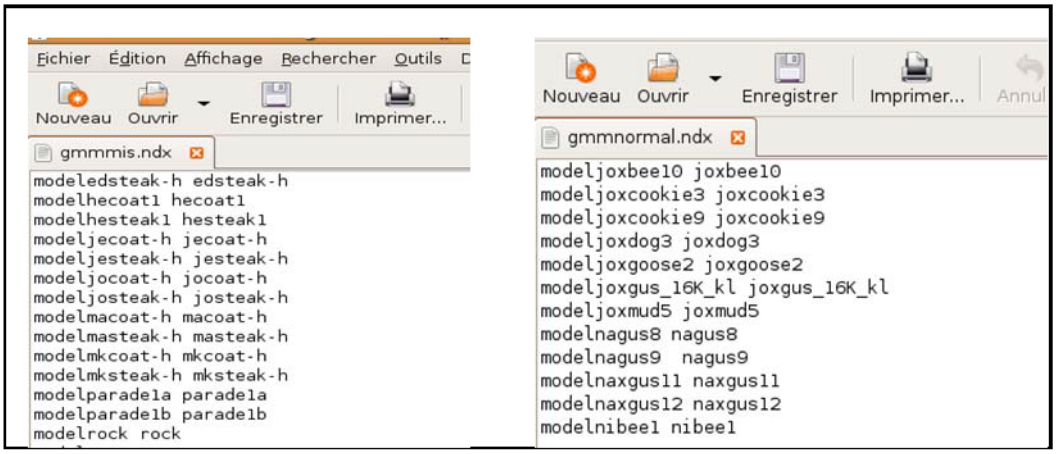

Figure 4.6 : Génération des listes utilisées pour l'apprentissage des modèles de la parole

2.6 Tests et scores

Le module *ComputeTest* produit des fichiers des scores contient une ligne par test (le modèle et le vecteur paramètres de test), Le programme est une mesure de LLR pour le calcul des scores.

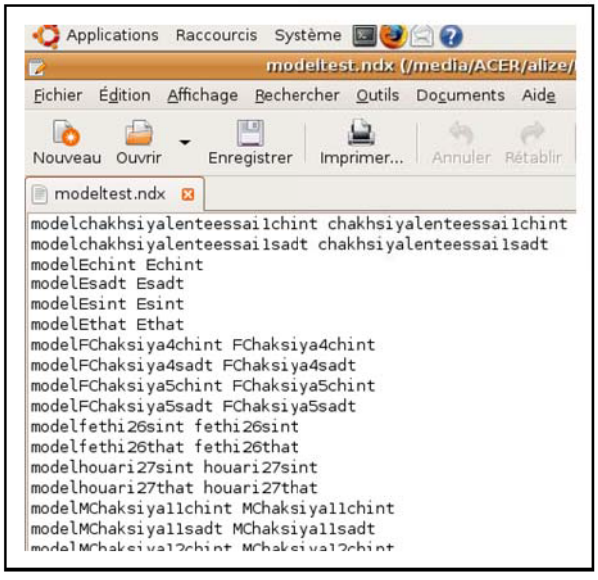

Figure 4.7 : Génération des fichiers pour la pahase de test

3.8. Vecteurs des moyennes des GMM

Ce module sert à extraire les moyennes de toutes les gaussiennes dans un vecteur sauvegardé dans un fichier d'extension .vect. Ce vecteur représente l'entrée de module SVM.

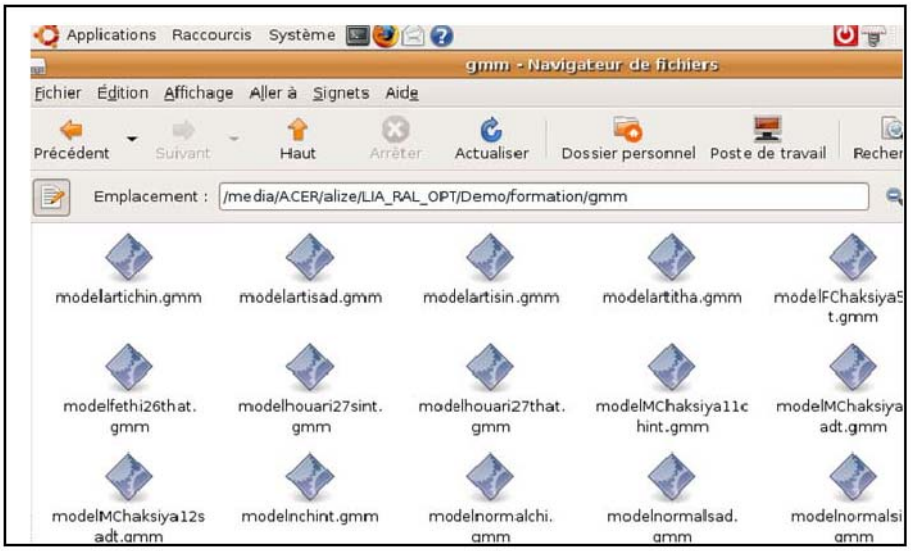

Figure 4.8 : Création des vecteurs des GMM

2.7. Modèles SVM

LIBSVM est une bibliothèque des SVM. Son but est de faciliter l'utilisation SVM comme outil. Elle contient deux modes, le premier mode pour l'apprentissage des modèles locuteurs et le deuxième mode pour le test, l'entrée de ce module est les vecteurs des moyennes des GMM, la sortie correspond aux fichiers de scores [62].

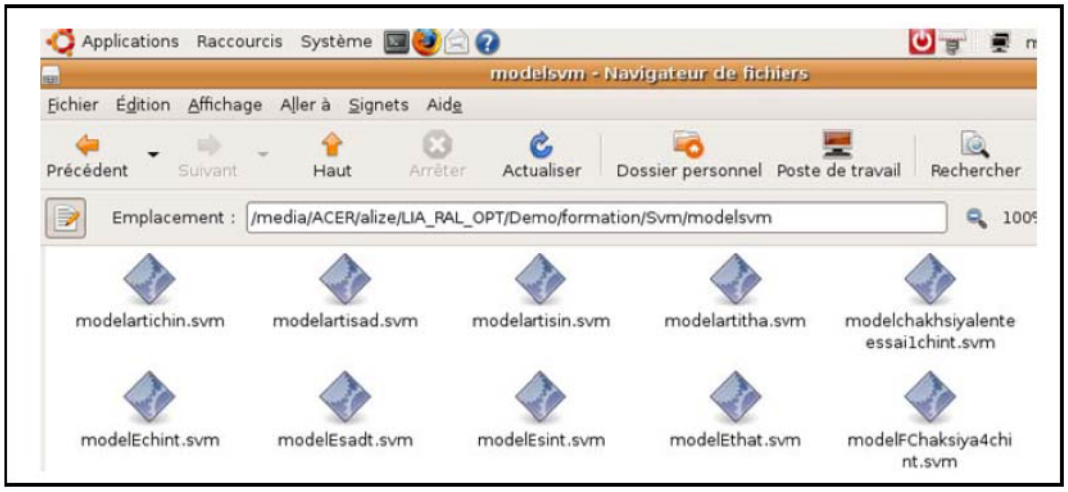

Figure 4.9 : Apprentissage des modèles SVM

3. Classification Contrôle/Pathologique

Une classification ou système de classification est un système organisé et hiérarchisé de catégorisation d'objets appelées classes. Dans notre cas nous nous limiterons à deux classes pour organiser les signaux sonores : la première classe devra contenir les signaux de voix normales, et la seconde classe les signaux de voix pathologiques. Par conséquent, deux modèles doivent être estimés, contrôle G et patho G correspondant respectivement au modèle des voix normales et au modèle des voix ayant des troubles articulatoires.

L'absence d'une grande base de données sonore Arabe nous a conduit à faire les tests de reconnaissance sur une base préenregistrée ainsi que sur une base contenant des phonèmes présentant certaines caractéristique proches des sons de l'Arabe développé par le MIT Open Course Ware **M**assachusetts **I**nstitute of **T**echnology, afin de valider nos résultats.

3.1. Description de la Base de Données WMIT

Dans cette section, nous allons examiner le corpus de certains enfants dont la plage d'âge est 3-5 ans. Dans cet âge, les enfants continuent acquérir le contrôle et la coordination des organes phonatoires pour produire les consonnes et les voyelles nécessaires pour générer des mots et des séquences de mots.

La base de données utilisée dans notre étude est une introduction aux troubles articulatoires de la parole et à l'origine, développée par le MIT Open Course Ware Massachusetts Institute of Technology sous licence Creative Commons. Il y a plusieurs corpus produits par des enfants normaux et des enfants qui ont des troubles articulatoires de la parole. Dans notre étude, l'échantillon de spectrogramme a été obtenu à partir de corpus du mot «sève» prononcé par des enfants enregistrés dans le laboratoire de WMIT [63]. Ces corpus sont destinés à illustrer certains troubles articulatoires du son [s]. Les troubles articulatoires pourraient donner lieu à des attributs acoustiques observées

En regardant les propriétés acoustiques de la parole pathologique, nous allons essayer de tirer des conclusions sur les caractéristiques LPC et des configurations articulatoires qui donnent lieu à la parole. (La première ligne indique la représentation du temps, la deuxième indique l'intensité spectrale, la troisième représente la LPC) (Figure 4.10).

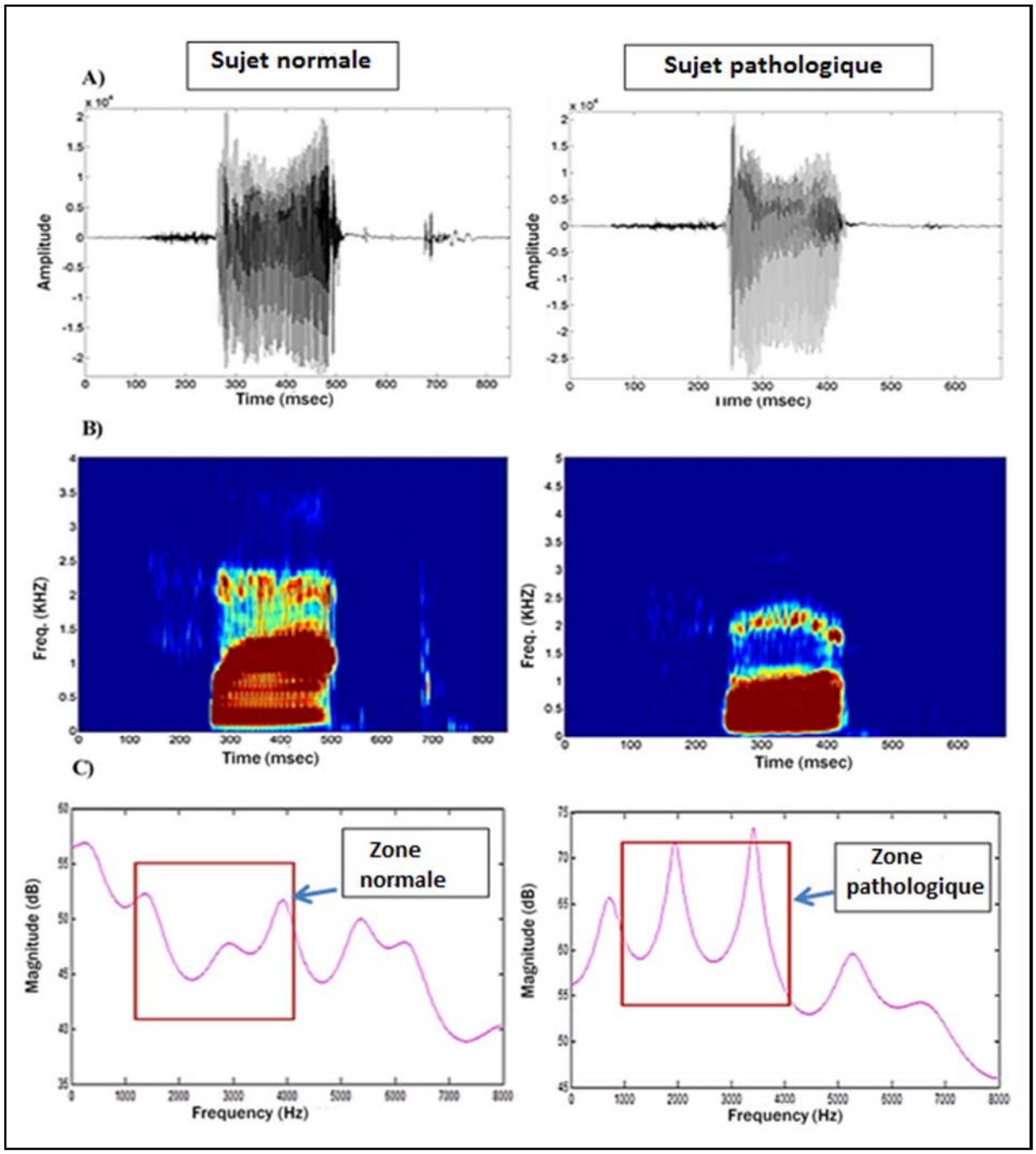

Figure 4.10 : Caractéristiques acoustiques observées de la consonne [s] dans le mot "sève" pour un sujet normal (à gauche) et pathologique (à droite) : (a) la représentation temporelle, (b) l'intensité spectrale, (c) les LPC.

3.2. Paramétrisation des corpus

Le signal de parole préaccentué (un filtre IIR d'ordre inférieur avec un coefficient $0,9 \leq a \leq 1$ est utilisé à cet effet, dans notre cas, nous avons choisi a = 0.95) est subdivisé en trames de 20 ms, extraites toutes les 10 ms, sur laquelle une fenêtre de

Hamming est appliquée. Pour chaque trame, la corrélation est calculée et utilisée pour estimer LPCC. Enfin, les dérivées première et seconde des coefficients cepstraux et des log-énergies peuvent être ajoutés aux vecteurs de caractéristiques (42) LPC. Les vecteurs de paramètres sont ensuite normalisés pour obtenir une distribution de moyenne 0 et une variance 1.

3.3. Modélisation des données

Le modèle du monde est appris par l'algorithme EM, deux autres modèles doivent être estimés pour des LS_P et des LS_N. Les deux modèles sont composés de 64 composantes gaussiennes avec des matrices de covariance diagonale et sont obtenus par l'adaptation de modèle du monde. Nous concaténons toutes les moyennes des composantes de GMM pour former un super-vecteur qui peut être utilisé pour former à la fois des modèles SVM des LS_P et des LS_N. Les corpus est subdivisé en trois parties :

- 20 fichiers des LS_N ont été utilisés pour former UBM ;
- 6 fichiers de parole normale ont été utilisés pour former un modèle normal ;
- 6 fichiers de parole pathologique ont été utilisés pour former le modèle pathologique.

3.4. Décision et calcul des performances

La décision est la mesure de similarité, elle est calculée pour un corpus de test (24 fichiers LS_P et LS_N) par rapport au modèle de troubles articulatoires.
A cause du manque des données pour calculer les performances de notre système, nous avons utilisé une simulation de la courbe des performances empiriques basées sur un ensemble de valeurs de décision [64-66]. Les performances obtenues par les deux systèmes GMM-UBM et GMM-SVM sont présentés dans le tableau 4.1.

Tableau 4.1. Performance du système de la classification Contrôle/Pathologique

Systèmes	Modèle	
	Normal	Pathologique
GMM-UBM	79.1%	87%
GMM-SVM	91.6%	95.8%

Le tableau 4.1 des résultats de la classification Contrôle/Pathologique donne un résultat prometteur et encourageant. Les systèmes GMM-UBM donne 79,1% pour le modèle de L_N et 87% pour le modèle de L_P en termes de classification correcte.

Les résultats montrent clairement qu'un système à base des SVM-GMM est plus performant qu'un système implémenté par des GMM-UBM. En effet, un gain absolu de 11% et 9% est observé.

Nous avons réalisé une expérience présentant la sensibilité du classificateur au nombre de coefficients cepstraux. La figure 4.11 illustre ces différences en présentant les gains de performance des systèmes où le vecteur de paramètres cepstraux augmente de 9 à 13 coefficients. L'ajout des dérivées secondes aux vecteurs des caractéristiques ainsi que le log-énergie est également considérable. Dans le cas des 16 coefficients nous remarquons une dégradation des performances que nous pouvons l'expliquer par un sur échantillonnage.

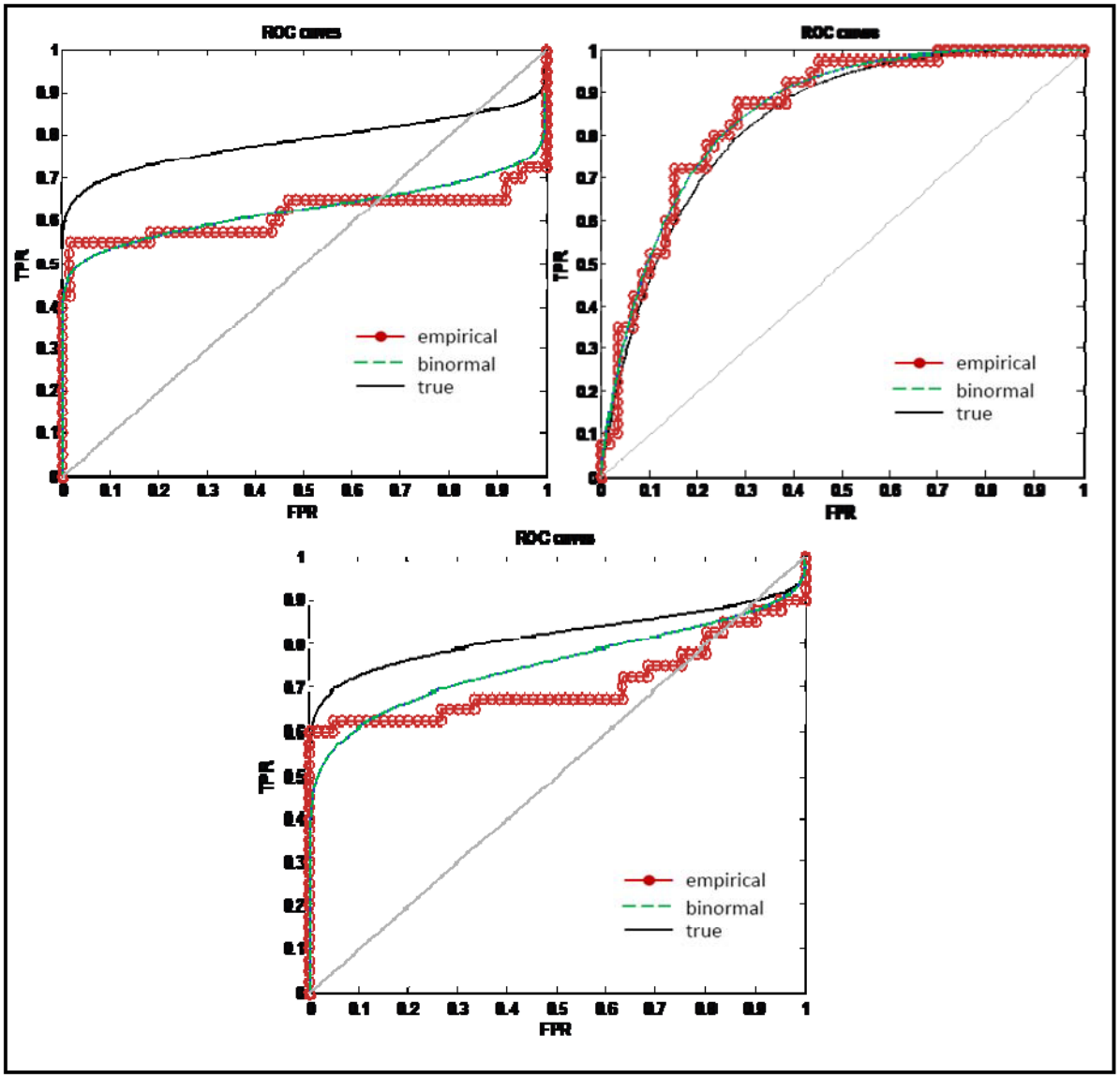

Figure 4.11 : Influence de coefficients cepstraux sur les performances du SCTP

L'effet désiré de l'application du GMM-SVM est illustré dans les figures 4.11-4.14. Ces dernières montrent les distributions des scores pour les deux modèles. Nous pouvons dire que les distributions des scores (client et imposteur pour les figures. 4.12 et 4.13) sont plus séparées. L'utilisation des supervecteurs GMM présentant de l'information utile se révèle être un enjeu important pour améliorer les performances du SCTP.

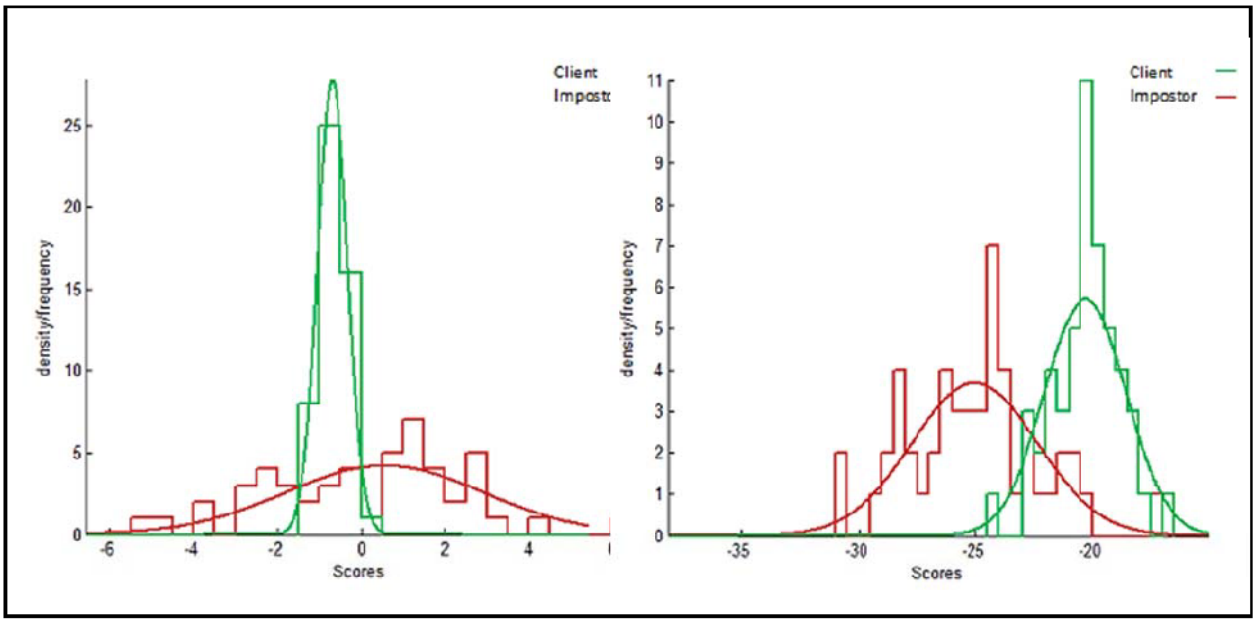

Figure 4.12 : Distribution des scores pour le modèle du L_P en utilisant les : a)GMM-UBM, b)GMM-SVM.

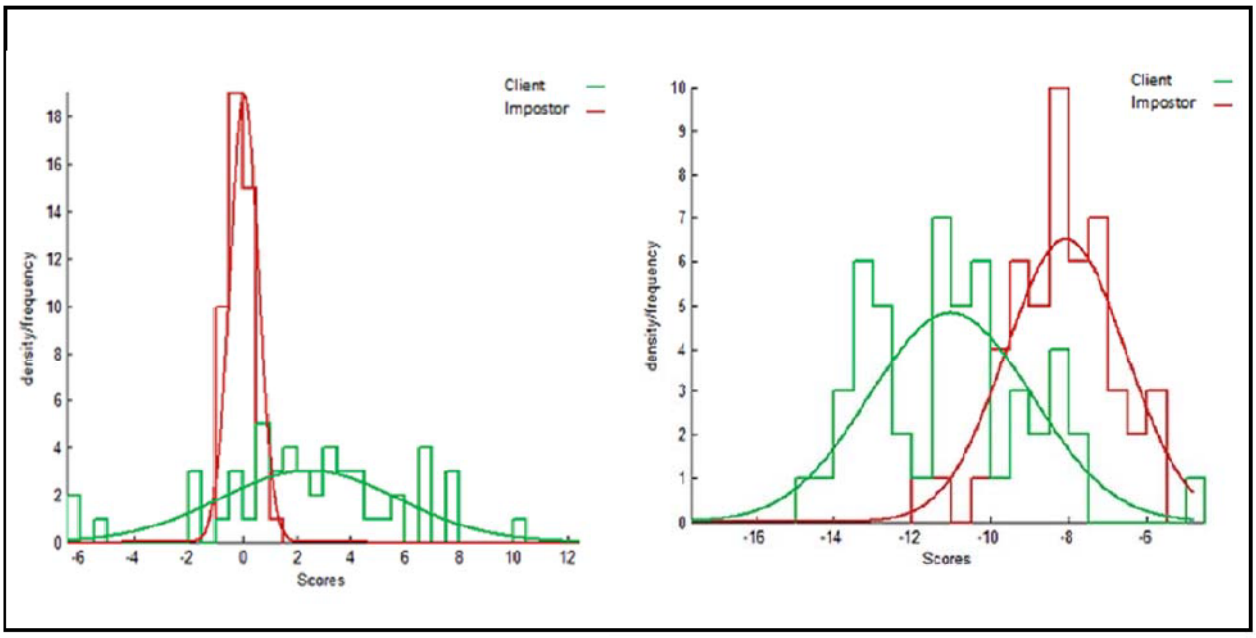

Figure 4.13 : Distribution des scores pour le modèle du L_N en utilisant les : a)GMM-UBM, b) GMM-SVM.

La figure 4.14 compare les performances du SCTP en terme de courbe ROC entre le classificateur hybride GMM-SVM et le classificateur de référence GMM-UBM.

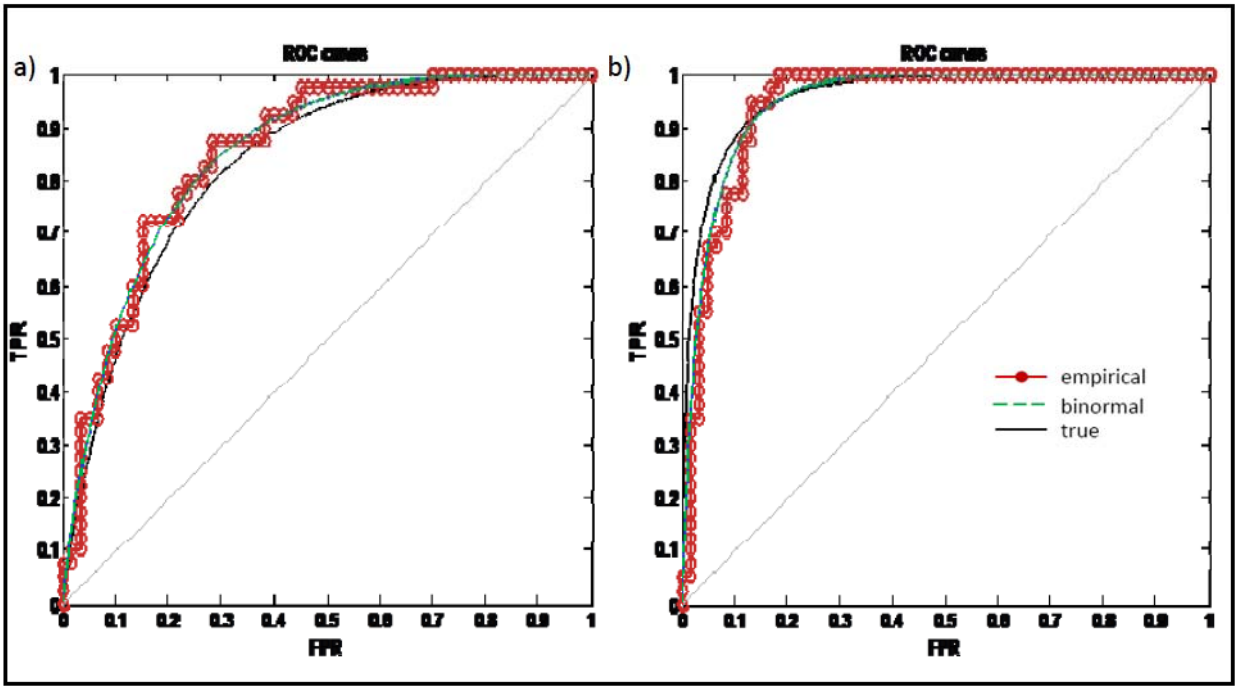

Figure 4.14 : Courbe ROC pour le modèle du L_P en utilisant les : a) GMM-UBM ; b) GMM-SVM

4. la classification des phonèmes étudiés

Il s'agit de classer un corpus du test par rapport aux 4 phonèmes ([S] [ʃ] [ʂ] [Θ].). Par conséquent, les 4 modèles des phonèmes sont à estimer.

4.1. Description de la Base de Données enregistrée

Le corpus utilisé dans cette étude est constitué de parole, enregistrée par des LS_P et LS_N, dans un milieu ambiant sous la supervision d'un orthophoniste du CHU Lamine Debaghine-Alger.

Tableau 4.2 : Locuteurs ayant enregistré le corpus des phonèmes [S] [ʃ] [ʂ] [Θ].

	Personnes (Age 8-14 ans)	
	Normal	**pathologique**
Masculin	11	8
Féminin	9	7

Les enregistrements du corpus ont été faits dans une salle ordinaire, ceci afin d'inclure le bruit environnant, car nous préconisons d'utiliser notre SCTP dans un environnement en conditions normales (dans un cabinet de médecin, à la maison, dans une salle de classe, etc.).

Le corpus est composé de 104 fichiers. Ces derniers ont été enregistrés par une trentaine de locuteurs prononçant quatre phonèmes, qui sont divisés en deux parties entre les phases de test et d'apprentissage. La classification des troubles de la parole traite essentiellement la détection du défaut de prononciation afin d'appliquer un traitement adéquat.

La pathologie, que nous avons étudiée, concerne l'interposition linguale dans la production des phonèmes arabes comme [S, ʃ, Ṣ, Ɵ]. En raison de l'absence des bases de données de la parole pathologique, nous avons choisi les mots [ʃaχsija], [ʃamsø] et [ƟalaƟa] où la pathologie de l'occlusion est intensive [67], les sons entendus transcrits sont présentés dans le tableau 4.2.

Tableau 4.2 : Transcription phonétique des mots prononcés par un L_P

Mots écrits en AS	**Prononciation**		**Mots incorrects écrits en AS**
	Correcte	**incorrecte**	
شخصية	[ʃaχṢija]	[ƟaχƟija]	ثخثية
		[ƟacƟʃija]	ثخشثية
شمس	[ʃamsø]	[ʃamʃø]	شمش
ثلاثة	[ƟalaƟa]	[salasa]	سلاسة

Avant d'entamer le traitement des données par une modélisation hybride entre deux classificateurs GMM et SVM, nous allons voir quelques aspects visuels des signaux sonores et faire des comparaisons, dans le but de comprendre le fait de prononcer le [θ] au lieu du [ʃ] et [Ṣ] figures 4.15-4.17.

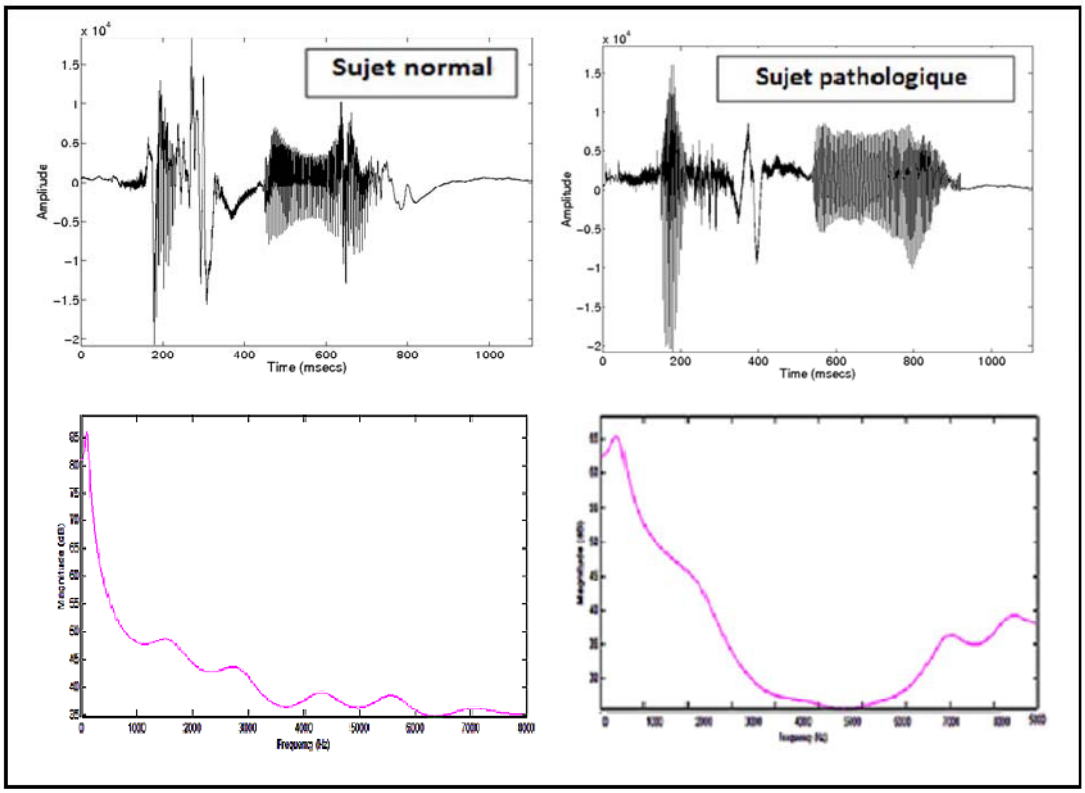

Figure 4.15 : Comparaisons visuelles des représentations LPC des mots [ʃaxsija] et [ӨaxӨija]

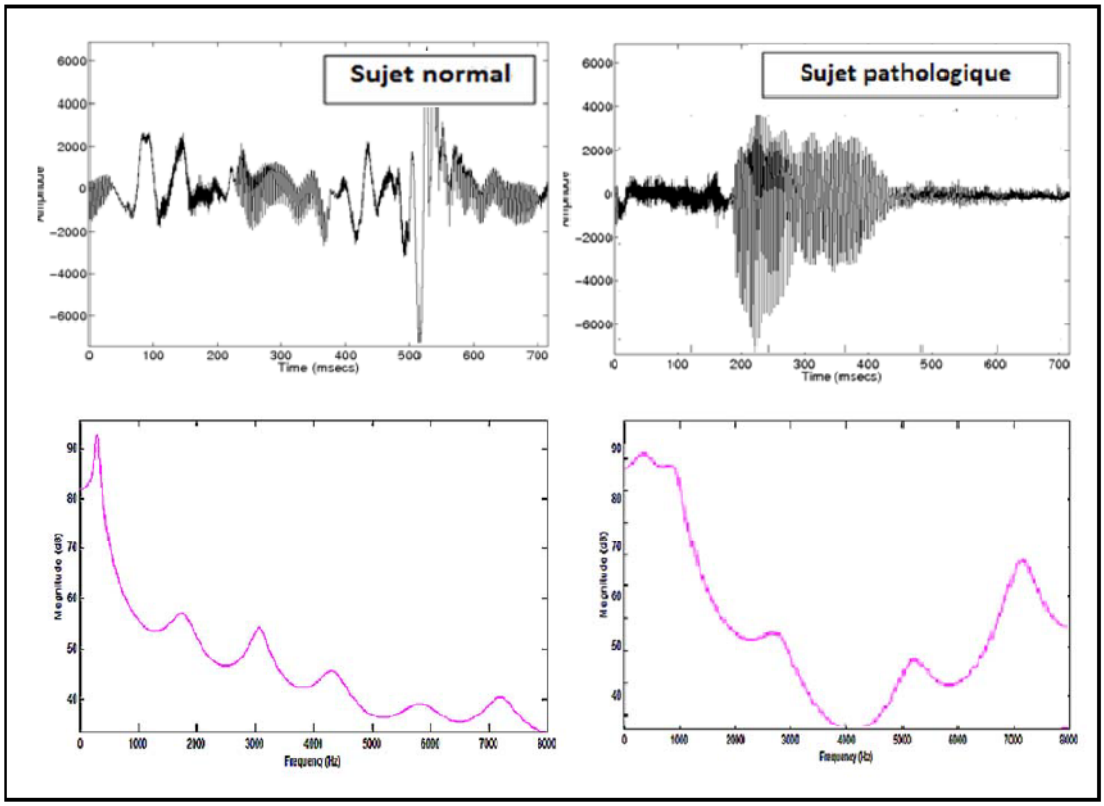

Figure 4.16 : Comparaisons visuelles des représentations LPC des mots [ʃamsø] et [ʃamʃø]

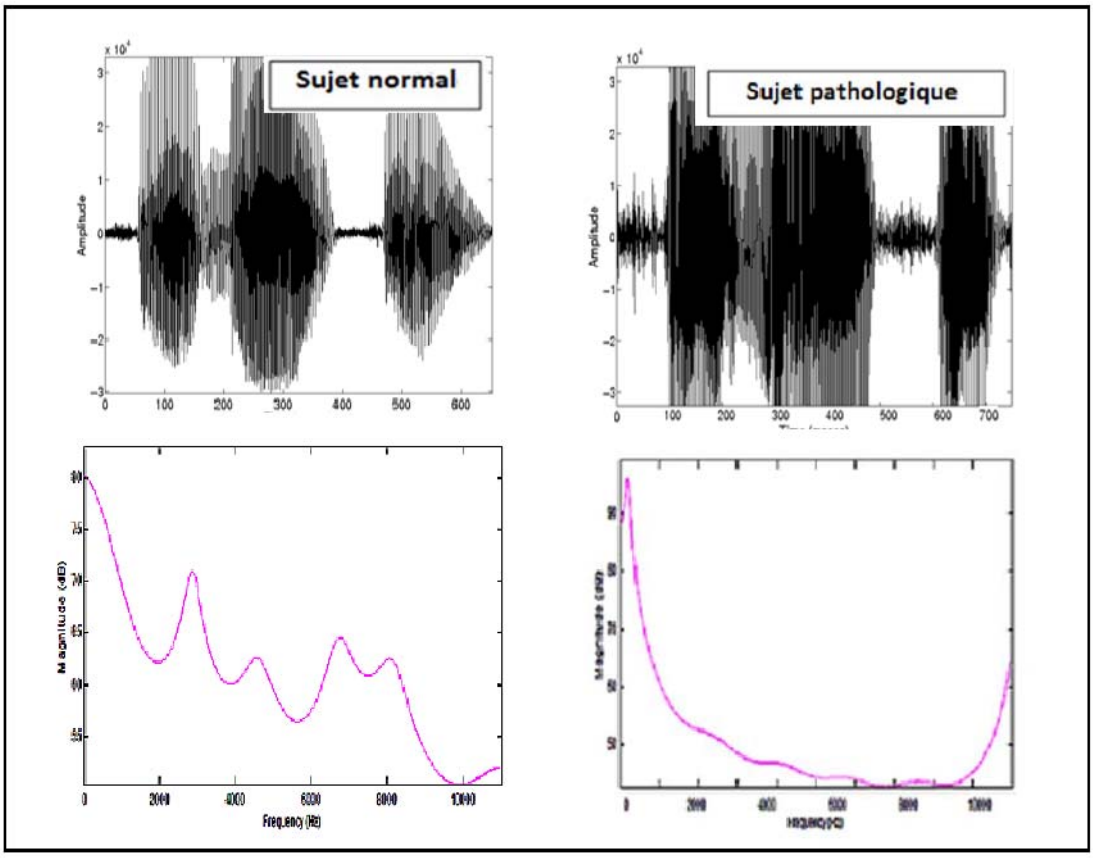

Figure 4.17 : Comparaisons visuelles des représentations LPC des mots [ΘalaΘa] et [salasa]

Toutes les comparaisons entre les mots prononcés par des LS_N et LS_P présentées précédemment se basent sur une connaissance a priori. Toutes les données graphiques précédentes sont inexploitables par le thérapeute ainsi que le patient, pour cela, il y a lieu d'automatiser les actions et de donner des courbes et/ou graphes d'évolution ou même des scores obtenus à partir du SCTP.

Dans cette section, nous nous concentrons sur la reconnaissance au niveau des phonèmes. Nous avons utilisé la segmentation phonémique d'un signal de parole afin d'identifier les trames voisées et non-voisées (figure 4.16).

Le **T**aux de **P**assage par **Z**éro (TPZ) représente le nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la valeur centrale de l'amplitude (généralement zéro). Il est fréquemment employé pour des algorithmes de détection de section voisée / non voisée dans un signal. Un seuil d'amplitude ˝S˝ permet de définir une zone autour du zéro de largeur ˝2xS˝ au sein de laquelle les oscillations ne sont pas prises en compte.

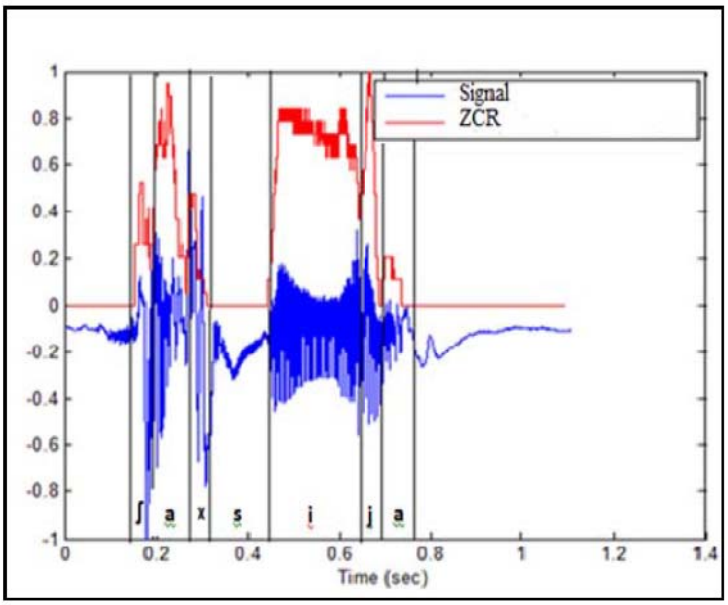

Figure 4.18 : Segmentation semi automatique du mot [ʃaxsija]

Cette segmentation est basée sur le TPZ pour obtenir le corpus des phonèmes [S, ʃ], Ṣ, Θ]. Le SCTP est basé sur une hybridation des GMM-UBM et les SVM. Trois phases sont nécessaires.

4.2. Paramétrage des corpus

L'analyse spectrale est un procédé qui a pour but de séparer la contribution de la source et du conduit vocal. Les paramètres de RAL les plus pertinents sont les LPC. Le signal de parole pré-accentué (a = 0.95.) est divisé en trames de 20 ms toutes les 10 ms, sur lequel une fenêtre de Hamming est appliquée. Les dérivées première et seconde des coefficients cepstraux ainsi que des log-énergies sont ajoutés aux vecteurs des LPCC (Nombre des LPCC=13). Les vecteurs de paramètres sont ensuite normalisés.

4.3. Modélisation des données

Dans le cadre de notre système, un modèle correspond à un phonème prononcé par le locuteur dans la même catégorie des LS_P ou LS_N.

Tous les modèles sont composés de 64 gaussiennes avec des matrices de covariance diagonales. La même procédure a été appliquée pour avoir le super-vecteur utilisé afin de former à la fois des modèles SVM pour des LS_P ou des LS_N pour chaque phonème. L'ensemble des corpus est divisé en deux parties:

- le modèle du monde (32 fichiers, 8 fichiers équilibrés entre hommes et femmes pour chaque phonème) ;
- le modèle des LS_P ou des LS_N (16 fichiers, 2 hommes et 2 femmes pour chaque phonème).

4.4. Decision et calcul des performances

La décision dépend du modèle de phonème prononcé par un locuteur (L_P ou L_N) sur lequel la plus grande mesure de similarité est calculée pour un corpus de test.
Les performances obtenues par le système GMM -SVM sont présentées dans les tableaux 4.3- 4.5.

Les résultats sont donnés en termes de matrice de confusion. Le comportement de classificateur est évalué en termes de **T**aux de **R**econnaissance correcte dans l'ensemble de test. La méthode de calcul du **T**aux de **R**econnaissance **G**lobale (TRG) est donnée par :

$$TR = \left(\text{Cas corrects} \ / \text{Nb Total}\right) \quad (4.1)$$

$$TRG(\%) = \left((\Sigma TR)/M\right) \times 100 \quad (4.2)$$

Où M=4 est le nombre des phonèmes utilisés.

Tableau 4.3 Matrice de confusion pour des LS_N

	[S]	[ʃ]	[ʂ]	[Θ]
[S]	75%	50%	50%	50%
[ʃ]	25%	100%	50%	25%
[ʂ]	50%	50%	100%	25%
[Θ]	25%	0%	25%	75%
		TRG		87.5%

Tableau 4.4 Matrice de confusion pour des LS_P

	[S]	[ʃ]	[ʂ]	[Θ]
[S]	50%	50%	75%	100%
[ʃ]	75%	50%	50%	75%
[ʂ]	75%	50%	75%	100%
[Θ]	100%	50%	25%	50%
			TRG	56.25%

Tableau 4.5 Matrice de confusion entre des LS_N et LS_P

	[S]	[ʃ]	[Ṣ]	[Θ]
[S]	100%	50%	50%	75%
[ʃ]	25%	75%	50%	25%
[Ṣ]	50%	50%	100%	25%
[Θ]	25%	25%	0%	100%
			TRG	93,75%

A partir de la comparaison des performances du SCTP, nous pouvons remarquer que:

- les valeurs de la diagonale de la matrice de confusion sont grandes par rapport à l'ensemble. Nous obtenons un TRG = 87,5%. Ceci montre que le SCTP a reconnu les phonèmes prononcés par les sujets normaux (tab 4.3) ;
- à cause de l'utilisation des modèles d'apprentissage des phonèmes prononcés par des sujets ayant des troubles articulatoires, le SCTP détecte les phonèmes mal prononcés (tab 4.5). Dans ce cas, nous obtenons un TRG = 93,75%. La figure 4.17 confirme les conclusions de l'expérience menée sur les corpus en termes de densités de distribution des scores et de courbe ROC ;
- le tableau 4.4 montre que dans cette situation le SCTP ne tient pas compte des phonèmes qui sont mal prononcés, nous obtenons un TRG = 56,25%.

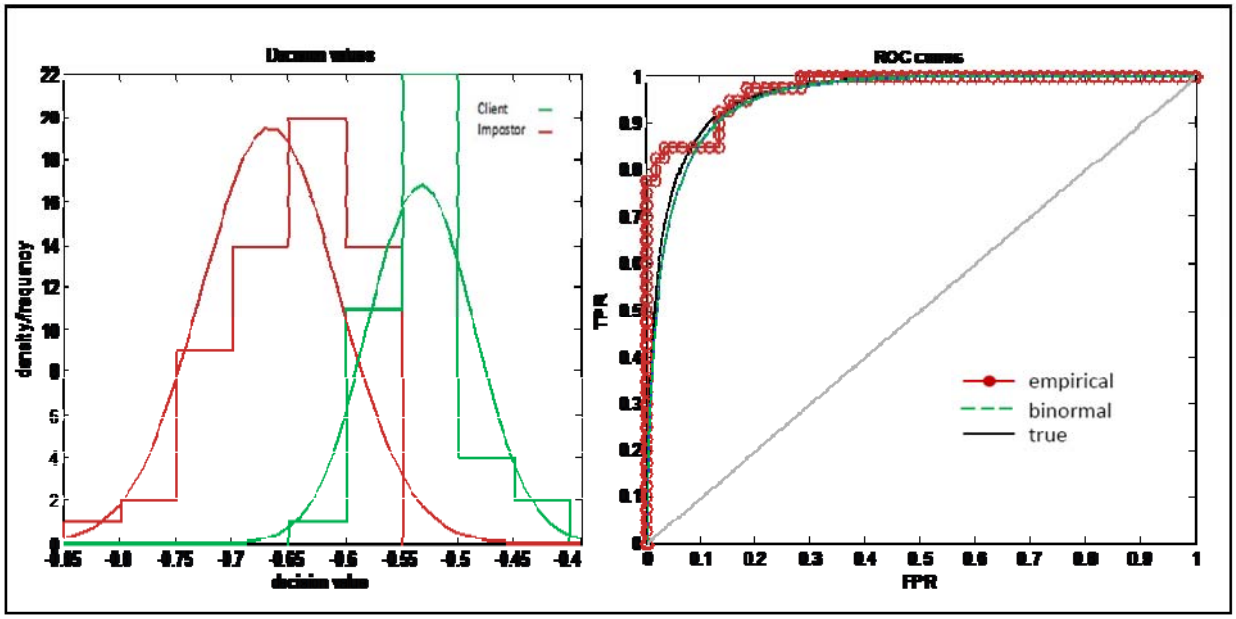

Figure 4.19 : Performances du SCTP dans le cas GMM-UBM pour des LS_N et LS_P

5. Conclusion

Dans ce chapitre, nous avons adopté une méthodologie de détection du sigmatisme par l'utilisation des LPCC connus pour leur robustesse au bruit et une modélisation stochastique basée sur les SVM en utilisant les GMM. Différents tests et comparaisons ont été réalisés sur un corpus enregistré. Ceci nous a permis de déduire des règles de travail relatives aux paramètres de l'enregistrement et au choix de la configuration adéquate. Toutefois l'utilisation de l'hybridation des classificateurs nous a permis de résoudre une grande partie du problème posé.

CONCLUSIONS GENERALES ET PERSPECTIVES

Dans ce travail de thèse nous avons présenté un système de RAL adapté à la classification des voix pathologiques. Dans un premier temps, une classification basique *Contrôle/Pathogique* a permis de constater que le système répond plutôt favorablement à la classification des voix des sujets normaux et des voix des patients ayant des troubles articulatoires de la parole.

La deuxième phase expérimentale *la classification 4-phonèmes* a permis d'évaluer les différents phonèmes prononcés par des sujets dans le cas pathologique et le cas normal. Il est à noter que le manque manifeste de données peut influencer fortement la qualité.

D'autres types d'analyses instrumentales permettent de meilleurs résultats pour le moment. En ce qui nous concerne un TRG de 93,75% avec 35 locuteurs et 4 classes de phonèmes est considéré encourageant. L'originalité et l'intérêt d'une telle approche sont les suivants :

- une capacité à analyser la parole continue proche de l'élocution naturelle ;
- une capacité à traiter de vastes bases de données, autorisant des études à grande échelle et des résultats statistiques significatifs ;
- une analyse acoustique simple et automatique permettant une simplicité d'utilisation.

Nous avons présenté quelques expériences comparatives permettant d'illustrer la sensibilité des systèmes à certaines techniques (nombres des coefficients spectraux, hybridation des systèmes de classification).

Ces résultats sont le reflet du travail que nous avons effectué pour proposer un SCTP basé sur les GMM-SVM correspondant à l'état de l'art.

En effet, les performances du SCTP peuvent être améliorées, en :

- augmentant le corpus d'apprentissage (élément très important dans les systèmes de la RAL) ;
- extrayant les informations acoustiques mieux adaptées à l'analyse des troubles de la prononciation ;
- utilisant une fusion des données au niveau des paramètres acoustiques et des scores.

Nous préconisons de modéliser d'autres sigmatismes ainsi que les schlintements, le zézaiement ainsi que d'autres défauts pathologiques détectables par le système auditif humain, selon la méthodologie mentionnée dans l'organigramme que nous avons proposé. L'intégration d'autres maladies s'effectuera par l'addition du mot pathologique, des différentes occurrences du phonème en question, ainsi que la Base de Données Correspondante.

Enfin, nous avons conscience que l'intérêt majeur de ce type d'outil de classification automatique est un certain déterminisme qui fait actuellement défaut à l'analyse perceptive. Cet outil restera un instrument d'évaluation et non un outil de décision qui reste clairement entre les mains du clinicien.

REFERENCES BIBLIOGRAPHIQUES

[1] L. Ben moussa, Les troubles de la voix dans le milieu clinique algérien, thèse de doctorat à l'université d'Alger, 2009.

[2] P. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman; G. Friedrich, P. Van De Heyning, M. Remacle, & V. Woisard, A basic protocolfor functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques, Eur. Arch. Otorhinolar., 258, pp. 77-82, 2001.

[3] R. Bruce Gerratt, and Mika Ito, When and why listeners disagree in voice quality assessment tasks,J. Acoust. Soc. Am., 122, pp. 2354–2364, 2007.

[4] A. Giovanni, Is the Perception of Voice Quality Language-Dependant?, inetrspeech, 2011.

[5] A. Giovanni, V. Molines, N. Nguyen & B. Teston, L'évaluation objective de la dysphonie : une méthode multiparamétrique, Proceedings of International Congress of Phonetic Sciences(ICPhS) (12 : août 19-24 : Aix-en-Provence, France), pp. 274-277, 1991.

[6] C. Fredouille, Application of Automatic Speaker Recognition techniques to pathological voice assessment, Proc. European Conference on Speech Communication and Technology (Eurospeech), 2005.

[7] A. Giovanni, D. Robert, B. Teston, M.D. Guarella & M. Zanaret, Etude préliminaire des paramètres acoustiques et aérodynamiques après laryngectomies frontales antérieures de Tucker, Ann. Otolaryngol. Chir. Cervicofac., 113, pp. 277-284, 1996.

[8] A. Ghio & B. Teston, Evaluation of the acoustic and aerodynamic constraints of apneumotachograph for speech and voice studies, Proceedings of International Conference on VoicePhysiology and Biomechanics (août 18-20 : Marseille, France), Marseille : Univ. Méditerranée, pp. 55-58, 2004.

[9] V. Parsa & D.G. Jamieson, Acoustic discrimination of pathological voice: sustained vowels versus continuous speech, J. Speech Hear. Res., 44, pp. 327-339, 2001.

[10] C. Fredouille, G. Pouchoulin, J.F. Bonastre, M. Azzarello, A. Giovanni & A. Ghio, Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia), Proc. Eurospeech, Lisboa, ISCA, pp. 149-152, 2005.

[11] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia & D.A. Reynolds, A tutorial on text independent speaker verification, EURASIP Journal on Applied Signal Processing, 4, pp. 430-451, 2004.

[12] L. Chen, J.-L. Gauvain, and L. Lamel ad G. Adda. Dynamic Language Modeling for Broadcast News In Proceedings of ICSLP, Jeju Island, October 2004.

[13] N. Ramou, M. Djeddou and M. Guerti, Two Classifiers Score Fusion for Text Independent Speaker Verification, 11th International Conference on Intelligent Systems Design and Applications(IEEE Conference), pp. 937:940, 2011.

[14] N. Ramou and M. Guerti, Automatic Detection of Articulations Disordersfrom Children's Speech Preliminary Study,Journal of Communications Technology and Electronics, Vol. 59, No. 11, pp. 1274–1279, 2014.

[15] L. Crevier-buchman, M.-C. Monfrais-pfauwadel, O. Laccourreye, V. Jouffre, D. Brasnu, et H. Laccourreye, La Laryngostroboscopie, Ann. Otolaryngol. Chir.Cervicofac., 110, p. 355-357, 1993.

[16] A. Giovanni, Ch. Assaiante, A. Galmiche, M. Vaugoyeau, M. Ouaknine et F. Le huche, forçage vocal et posture : etudes experimentales chez le sujet sain, revue de laryngologie, d'otologie et de rhinologie, vol. 127, no 5, p. 285-291, 2006.

[17] D. Bernard Bleicher « Anatomie de l'appareil respiratoire et desmécanismes Phonatoires » URL ; http://www.operalab.org/fr/4bib/1art/1gd

[18] D. Bernard Bleicher « Anatomie de l'appareil respiratoire et desmécanismes Phonatoires » URL ; http://www.operalab.org/fr/4bib/1art/1gd

[19] « Encarta encyclopédie » International programme manager interactifmedia group, USA. http://www.les-encyclopedies.com/encarta.html

[20]J. Domat et J.Bournef, Nouveau larousse médicale.
http://www.larousse.fr/archives/medical

[21] Lolkj. Vandervan « L'apparail phonatoire » CM phonétique
URL :http://www.lesla.univ-lyon2.fr/IMG/pdf/doc-284.pdf

[22] M. Hinich « Detecting a transient signal by bispectral analysis ». IEEEtransation on acoustic speech and signal processing, Vol. 38, N° 7,pp. 1257-1265, March 1990.

[23] T. Dutoit, Introduction au Traitement Automatique de la Parole Faculté Polytechnique de Mons, 2000.

[24] A.Malraux. http://andremalrauxtpeson.e-monsite.com/pages/la-physique-du-son/l-intensite-du-son.html.

[25] P.N. Garner , A Simple Continuous Pitch Estimation Algorithm, Signal Processing Letters, IEEE, Vol:20 , N: 1,pp : 102 – 105, Jan. 2013

[26] K. Ferrat and M. Guerti, Synthèse de la parole en Arabe Standard. Cas des phénomènes spécifiques à la langue, Colloque International en Traductologie et TAL, Université d'Oran, Algérie, 9-11 avril 2007.

[27] http://www.geneva-link.ch/ceppim/final/ORL/Laphonation.htm, Université de Genève.

[28] E.Nemer, R.Goubran Automatic voice activity detection in different speech applications, Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop, 2008.

[29] D.Laurent, Les troubles du langage et de la communication chez l'enfant, Collection 'Que sais je ?', Presses Universitaires de France, 2013.

[30] Fédération nationale des orthophonistes, Semaine Nationale de Prévention des Troubles du langage 27 au 31 Mai 2002, Fédération Nationale des Orthophonistes, Paris, France.

[31] M. Guerti, Contribution à la synthèse de la parole par diphones en Arabe Standard, Thèse de Magister en Electronique Acoustique et Physiologique de la Parole. Université d'Alger, Algérie, 1983.

[32] Larousse encyclopédie du corps humain. http://www.larousse.fr/encyclopedie/ animations/Corps_humain_squelette_et_organes/1100545

[33] H. S. Venkatagiri, Voice is one aspect of speech production, Department of Psychologylowa State University, 2003.

[34] www.vulgaris-medical.com.

[35] http://www.cairn.info/zen.php?ID_ARTICLE=RFLA_132_0045

[36] http://dictionnaire.metronimo.com/term/53aa5da457a7acaaa2,xhtml

[37] http://www.begaiement.org/spip.php?article19

[38] http://www.geneva-link.ch/ceppim/final/ORL/La_phonation.htm#trouble

[39] M. Rokibul A Kotwal, Recurrent Neural Network Based Phoneme Recognition Incorporating Articulatory Dynamic Parameters, Communications in Computer and Information Science Vol : 192, pp 349-356, 2011.

[40] http://tecfa.unige.ch/tecfa/teaching/UVLibre/ tp-iish/ex-9798/ logopedie/ trouble1 .html.

[41] M. Zbancioc, M. Costin, Using neural networks and LPCC to improve speech recognition, International Symposium on Signals, Circuits and Systems. SCS , Vol.2 , pp : 445 - 448, 2003.

[42] K. Nishikawa, and all., Speech production of an advanced talking robot based on human acoustic theory, IEEE International Conference on Robotics and Automation. Proceedings. ICRA '04, Vol.4, pp : 3213 - 3219, 2004

[43] L.Rabiner, and B. -H. Juang, Fundamentals of Speech Recognition, PTR Prentice Hall,San Francisco, NJ. pp.no 507,1993:

[44] B.A. Dautrich, L.R. Rabiner, and T.B. Martin, "On the effects of Varying Filter bank Parameters on isolated Word Recognition," IEEE Trans. Acoustics, Speech, Signal Proc. ASSP-31 Vol :4, pp :793-807, August 1983.

[45] G.M. white and R.B. Neely, Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming, IEEE Trans. Acoustic, Speech, Signal Proc., ASSP-24 Vol :2, pp :183-188,1976.

[46] J. Kahn, Parole de locuteur : performance et confiance en identification biométrique vocale, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 2011.

[47] M. Homayounpour & G.Chollet, Performance comparison of some relevant spectral representations for speaker verification, Workshop on Automatic Speaker Recognition, Identification, Verification, Martigny, Suisse, pp. 27-30, 1994.

[48] D.A. Reynolds, A Gaussian Mixture Modeling approach to text-independent speaker identification, PhD, Georgia Institute of Technology, 1992.

[49] M. R. Gupta, Y. Chen, Theory and Use of the EM Algorithm, Journal of Foundations and Trends in Signal Processing, Vol : 4 N 3, pp 223-296, March 2011 .

[50] F. Bimbot, J-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia & D.A. Reynolds, A tutorial on text independent speaker verification, EURASIP Journal on Applied Signal Processing, Vol : 4, pp. 430-451, 2004.

[51] M.Ferras, C. Leung, Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition , IEEE Transactions On Audio, Speech, And Language Processing, Vol. 18, n :. 6, august 2010.

[52] M. J. Carey et E. S. Parris. Speaker verification using connected words. Dans Proceedings of Institute of Acoustics, 1992.

[53] D. A. Reynolds, Speaker identification and verification using gaussian mixture speaker models. Dans Speech Communication, 1995.

[54] A. Martin and M. Przybocki, "The NIST speaker recognition evaluation series, National Institute of Standards and Technology's in *Odyssey*, 2004. website, http://www.nist.gov/speech/tests/spk," **.**

[55] D. A. Reynolds, T.F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models , Digital Signal Processing Journal, 2000.

[56] V. N. Vapnik. Statistical Learning Theory.Wiley, 1998.

[57] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition.Data Mining and Knowledge Discovery , 1998.

[58] N. Cristianiniet J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.Cambridge University Press, 2000.

[59] ALIZE: open tool for speaker recognition, Software available at http://www.lia.univ-avignon.fr/heberges/ALIZE/.2006

[60] J.-F. Bonastre, F. Wils, and S. Meignier, Alize, a free toolkit for speaker recognition, in ICASSP, pp : 737 - 740, 2005.

[61] G.Gravier.spro: a free speech signal processing toolkit : http://www.irisa.fr /metiss /guig/spro/.

[62] Site web, http ://www.csie.ntu.edu.tw/ cjlin.

[63]http://ocw.mit.edu/courses/electrical_engineering_and_computer_science/6_542j_laboratory_on_the_physiology_acoustics_and_perception_of_speech_fall_2005/lab_data_base/.

[64] M. A. Aizerman, E. M. Braverman, and L. I. Rozomer, Theoretical foundations of The potentiel fonction method in pattern recognition learning, In Automation and Remote Contol,

[65] M. Stone, .Methods of Mathematical Physics, Inter-science, 2003.

[66] Kay H. Brodersen_y, Cheng Soon Ong_, Klaas E. Stephany and Joachim M. Buhmann, The binormal assumption on precision-recall curves, International Conference on Pattern Recognition, pp : 4247 - 4266, 2010

[67] Z.A. Benselama, M. Guerti and M.A. Bencherif.; Arabic Speech Pathology Therapy Computer Aided System, Journal of Computer Science Vol. 3, Issue 9, pp. 685-692, 2007.