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

DÉPARTEMENT D'ÉLECTRONIQUE

Mémoire de Fin d'Etudes

En vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

Thème :

Élaboration d'un système d'Identification Automatique du Locuteur par Quantification Vectorielle

Réalisé par :

Mlle F. Bendimerad

Mlle S. Siouane

Devant le jury :

D. Berkani	Professeur	ENP	Président
M. Guerti	Professeur	ENP	Rapporteur
B. Bousseksou	Professeur	ENP	Examinateur

Promotion : Juin 2013

REMERCIEMENTS

En premier lieu nous remercions Dieu le tout puissant de nous avoir donné le courage et la force pour réaliser ce travail.

Notre profonde gratitude et sincère reconnaissance vont tout d'abord à Mme M. Guerti qui a bien voulu nous encadrer. Nous la remercions pour sa disponibilité, son aide, les précieux conseils qu'elle nous a prodigués, ses critiques constructives, ses explications et suggestions pertinentes.

Nous remercions les membres du jury Messieurs D. Berkani et B.Bousseksou Professeurs à l'ENP, pour l'honneur qu'ils nous font de juger notre travail.

Nos remerciements vont également à tous les enseignants de l'Ecole Nationale Polytechnique qui ont contribué à notre formation. Qu'ils trouvent ici l'expression de notre profond respect et notre grande considération.

Nous remercions Mrs M. Kabache de l'ISMAS et F. Ykhlef du CDTA, pour leur aide.

DEDICACES

A la mémoire de mon grand père Si Salah

A mes très chers parents qui ont toujours été là pour moi,

et qui m'ont donnée un magnifique modèle de labeur et de persévérance

A ma chère binôme Saima Siouane - Sofia -

A tous mes amis (ies)

Qui m'ont soutenue et encouragée tout au long de ce projet

J'espère qu'ils trouveront dans ce travail

Toute ma reconnaissance

Et tout mon amour.

Faïza

DEDICACES

A mes très chers parents qui ont toujours été là pour moi,

et qui m'ont donnée un magnifique modèle de labeur et de persévérance

A ma chère binôme Faïza Bendimerad

A ma sœur Nadia et son mari

A mes frères Ilyes, Midou et Raouf

qui m'ont soutenue et encouragée tout au long de ce projet

A mon cher Amine

A tous ceux que j'aime.

Sofia

ملخص

الهدف من عملنا ضمن المشروع النهائي للدراسة هو إنشاء نظام للتحديد الأوتوماتيكي لهوية المتكلم اعتمادا على النص. هذا المجال غني بتطبيقاته الهاتفية وتامين الدخول.

لتحقيق هذا الهدف، قمنا بإنشاء مدونة تحتوي على كلمات منفردة باللغة الفرنسية. تم تسجيل هذه الأخيرة من طرف أربعة متكلمين(رجلان وامرأتان). أجري التحليل الاكستيكي لقاعدة البيانات باستخدام **MFCC**، أما بالنسبة لنمذجة إشارة الصوت طبقنا التكميم الشعاعي، زيادة على خوارزمية **LBG** (**L**inde, **B**uzo et **G**ray) للحد من خطأ النمذجة. اظهرت النتائج التي تحصلنا عليها وجود معدل الاعتراف بنسبة 95٪.

كلمات المفاتيح : التحديد الأوتوماتيكي لهوية المتكلم ،نمذجة اكستيكية، MFCC ، QV ،LBG ، معدل الاعتراف.

RESUME

L'objectif de notre Projet de Fin d'Etude est l'élaboration d'un système d'**I**dentification **A**utomatique du **L**ocuteur (**IAL**) en mode dépendant du texte. Ce domaine est riche en applications téléphoniques, sécurisation d'accès.

Pour atteindre ce but, nous avons construit un corpus constitué de mots isolés en Français. Ce dernier a été enregistré par 4 locuteurs (deux femmes et deux hommes). L'analyse acoustique de cette base de données a été effectuée en utilisant les **MFCC** (**M**el **F**requency **C**epstral **C**oefficients), quant à la modélisation du signal de parole nous avons appliqué la **QV** (**Q**uantification **V**ectorielle), ainsi que l'algorithme **LBG** (**L**inde, **B**uzo et **G**ray) pour minimiser l'erreur de modélisation. Les résultats obtenus donnent un **T**aux de **R**econnaissance (**TR**) qui atteint les 95%.

Mots-clés : Identification Automatique du Locuteur, Modèle Acoustique, MFCC, QV, LBG, Taux de Reconnaissance.

ABSTRACT

The goal of our Final Project Study is to develop an Automatic Speaker Identification system text dependent. This area is rich in mobile applications or securing access.

To achieve this goal, we have constructed a corpus of isolated words in French. The latter was recorded by four speakers (two women and two men). Acoustic analysis of the database was performed using the **MFCC** (**M**el **F**requency **C**epstral **C**oefficients). For modeling the speech signal we applied the **VQ** (**V**ector **Q**uantization) and the **LBG** algorithm (**L**inde, **B**uzo and **G**ray) to minimize the modeling error. The results show a recognition rate reaching 95%.

Keywords : Automatic Speaker Identification, Acoustic Model, MFCC, VQ, LBG, recognition rate.

LISTE DES ABREVIATIONS

API	**A**lphabet **P**honétique **I**nternational
BD	**B**ases de **D**onnées
DE	**D**istance **E**uclidienne
DTW	**D**ynamic **T**ime **W**arping
FFT	**F**ast **F**ourier **T**ransform
GMM	**G**aussien **M**ixture **M**odels
HMM	**H**idden **M**arkov **M**odel
HTK	**H**idden Markov Model **T**ool**K**it
IAL	**I**dentification **A**utomatique du **L**ocuteur
LBG	**L**inde, **B**uzo et **G**ray
LPC	**L**inear **P**redictive **C**oding
LPCC	**L**inear **P**redictive **C**epstral **C**oefficients
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MMC	**M**odèle de **M**arcov **C**aché
OE	**O**reille **E**xterne
OI	**O**reille **I**nterne
OM	**O**reille **M**oyenne
QV	**Q**uantification **V**ectorielle
RAL	**R**econnaissance **A**utomatique du **L**ocuteur
RAP	**R**econnaissance **A**utomatique de la **P**arole
SIAL	**S**ystème d'**I**dentification **A**utomatique du **L**ocuteur
TFD	**T**ransformée de **F**ourier **D**iscrète
VAL	**V**érification **A**utomatique du **L**ocuteur

LISTE DES FIGURES

		Page
Fig.1.1	Processus de production et de perception de la parole chez les êtres humains …..	4
Fig.1.2	Appareil phonatoire humain ……...	4
Fig.1.3	Représentation schématique de la production de la parole …….	5
Fig.1.4	Section du larynx…….	6
Fig.1.5	Système auditif humain……	7
Fig.1.6	Champ auditif humain……..	10
Fig.1.7	Son voisé [a] dans le mot baluchon…….	11
Fig.1.8	Son non voisé [ʃ] dans le mot baluchon……..	13
Fig.2.1	Schéma illustrant les différentes tâches de la RAL…….	19
Fig.2.2	Principe de base de la tâche de Vérification Automatique du Locuteur …….	20
Fig.2.3	Principe de base de la tâche d'Identification Automatique du Locuteur……..	21
Fig.2.4	Principe de base de la tâche d'Indexation par Locuteurs d'un flux audio……	22
Fig.3.1	Analyse numérique du signal parole par FFT …….	26
Fig.3.2	Représentation fonctionnelle du fonctionnement du conduit vocal et des sources d'excitations ……	27
Fig.3.3	Modèle source-filtre de production de la parole …….	27
Fig.3.4	Schéma d'extraction des MFCC …….	30
Fig.3.5	Banc de filtres sur l'échelle Mel …….	30
Fig. 3.6	HMM gauche-droite à trois états ……	33
Fig. 3.7	Modèle de GMM …….	34
Fig. 4.1	Studio d'enregistrement……...	37
Fig. 4.2	Microphone électrodynamique……	38
Fig. 4.3	Station ProTools ……	38
Fig. 4.4	Interface initiale du sytème de l'IAL ……..	40
Fig. 4.5	Interface du sytème de l'IAL avec affichage des résultats …….	40
Fig. 4.6	Etapes suivies lors de la réalisation du système IAL ……..	41
Fig. 4.7	Echantillons du signal de parole à quantifier ……..	42
Fig. 4.8	Quantification uniforme des échantillons ……...	43
Fig. 4.9	Quantification optimisée des échantillons……...	43
Fig. 4.10	Organigramme de l'algorithme LBG……...	46

Table des matières

Tapez le titre du chapitre (niveau 1) **1**

Tapez le titre du chapitre (niveau 2) 2

Tapez le titre du chapitre (niveau 3) 3

Tapez le titre du chapitre (niveau 1) **4**

Tapez le titre du chapitre (niveau 2) 5

Tapez le titre du chapitre (niveau 3) 6

LISTE DES TABLEAUX

		Page
Tableau 1.1	Signes phonétiques de l'API utilisés en Français……..	8
Tableau 1.2	Classification des phonèmes du Français en traits distinctifs ……	9
Tableau 2.1	Exemples d'applications en IAL…….	16
Tableau 4.1	Enregistrement du Corpus des mots isolés …….	36
Tableau 4.2	Paramètres utilisés pour l'extraction des MFCC……	42

TABLE DES MATIERES

Page

INTRODUCTION GENERALE 1

Chapitre 1 GENERALITE SUR LA PAROLE

1.1. INTRODUCTION 3

1.2. PROCESSUS DE PRODUCTION ET DE PERCEPTION DE LA PAROLE

1.3. PRODUCTION DE LA PAROLE 4

1.4. LE SIGNAL DE LA PAROLE 6

1.4.1. Types de sons langagiers produits 7

1.4.2. Notation phonétique du Français 8

1.4.3. Phonétique et phonologie 9

1.4.4. Classification des phonèmes du Français

1.5.SYSTEME AUDITIF ET PERCEPTION DE LA PAROLE 10

1.5.1. Système de transmission

1.5.2. L'aire de l'audition 11

1.6.DESCRIPTION ACOUSTIQUE DE LA PAROLE

1.6.1. La mélodie

1.6.2. La fréquence fondamentale (F0) 12

1.6.3. L'intensité

1.6.4. La durée

1.6.5. Le timbre

1.6.6. Les formants

1.7. CONCLUSION 13

Chapitre 2 BIOMETRIE VOCALE

2.1. INTRODUCTION 15

2.2. DEFINITION DE LA BIOMETRIE

2.3. HISTORIQUE

2.4. DOMAINES D'APPLICATIONS 16

2.5. QU'EST-CE-QU'UNE RECONNAISSANCE DE LA PAROLE ? 17

2.6. NOTIONS SUR LA VARIABILITE DU SIGNAL VOCAL

2.7. RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR 18

2.7.1. Niveau de Dépendance au Texte 19

2.7.2. Vérification et Identification automatique du Locuteur **20**

2.7.2.1. Vérification Automatique du Locuteur

2.7.2.2. Identification Automatique du Locuteur

2.7.2.3. Détection de Locuteurs **21**

2.7.2.4. Indexation par Locuteurs et ses Variantes

2.8. CONCLUSION **22**

Chapitre 3 TECHNIQUES DE RECONNAISSANCE DU LOCUTEUR

3.1. INTRODUCTION **24**

3.2. TECHNIQUES DE TRAITEMENT DU SIGNAL VOCAL

3.2.1. Méthodes non paramétriques

3.2.1.1. Chaîne de prétraitement

3.2.1.2. Analyse temporelle **25**

3.2.1.3. Analyse spectrale **26**

3.2.2. Méthodes paramétriques

3.2.2.1. Codage Prédictif Linéaire et les LPCC **27**

3.2.2.2. Analyse cepstrale **28**

3.3. MODELISATION ACOUSTIQUE DU LOCUTEUR **31**

3.3.1. Approche vectorielle

3.3.1.1. L'Alignement temporel dynamique

3.3.1.2. Quantification Vectorielle **32**

3.3.2. Approche statistique

3.3.2.1. Modèle de Markov Caché

3.3.2.2. Les Mélanges de Gaussiennes **33**

3.4. CONCLUSION **34**

Chapitre 4 EXPERIENCES ET RESULTATS

4.1. INTRODUCTION **36**

4.2. DESCRIPTION DE LA BASE DE DONNEES UTILISEE

4.3. MATERIELS UTILISES **37**

4.4. OUTILS UTILISES **38**

4.5. DESCRIPTION DE L'APPLICATION

4.6. DESCRIPTION DE L'INTERFACE **39**
4.7. ETAPES DE REALISATION DU SYSTEME SIAL **41**
4.7.1. Signal sonore
4.7.2. Paramétrisation
4.7.3. Extraction des MFCC
4.7.4. Quantification Vectorielle **42**
4.7.5. Algorithme LBG **44**
4.7.5.1. Initialisation du dictionnaire pour l'algorithme LBG
4.7.6. La Distance Euclidienne **45**
4.8. ÉVALUATION DES PERFORMANCES DU SIAL **46**
4.9. CONCLUSION **47**
5. CONCLUSION GENERALES ET PERSPECTIVES **49**
6. REFERENCES BIBLIOGRAPHIQUES **51**

Introduction générale

Introduction générale

Nous assistons de nos jours à de grands progrès technologiques qui proposent des outils de plus en plus sophistiqués tout en étant de plus en plus compacts. Le clavier risque de devenir obsolète dans quelques années, c'est pourquoi on cherche de nouveaux moyens de communication plus intuitifs et moins encombrants. Aujourd'hui, après le clavier, la souris et les écrans tactiles, la parole s'impose comme l'alternative la plus directe et la plus naturelle pour communiquer avec les machines.

Les travaux présentés dans notre projet se situent dans le cadre de l'**I**dentification **A**utomatique du **L**ocuteur **IAL**. Un système d'identification a pour objectif la transcription d'un signal de parole. Pour ce faire, deux types de traitements sont nécessaires : un traitement acoustique accompli par un modèle de langage. La conception d'un système d'identification est confrontée à plusieurs difficultés. En effet, la production de la parole est un processus continu et donc l'identification univoque d'unités symboliques dans ce flux n'est pas toujours possible. Par ailleurs, la variabilité du signal de parole inter et intra locuteurs constitue une autre source de complexité.

Dans ce travail, nous nous intéressons plus particulièrement à la Quantification Vectorielle. Cette approche a fait preuve de simplicité et d'efficacité. Dans cette dernière, le problème d'identification est divisé en phases d'apprentissage et de test. L'identification proprement dite est réalisée au cours de la deuxième étape où nous associons à chaque voix prononcée le locuteur correspondant.

Notre projet est organisé comme suit :

- le premier chapitre présente une étude générale sur les systèmes de la phonation et l'audition chez les êtres humains, ainsi que les mécanismes de production et de perception de la parole et sa description acoustique ;
- le deuxième chapitre est consacré aux principes de la **R**econnaissance **A**utomatique du **L**ocuteur **RAL**. Il présente ses domaines d'application ainsi que ses différentes tâches ;
- dans le troisième chapitre nous exposons les outils de traitement de signal vocal, ainsi que les paramètres les plus efficaces pour le représenter et les différentes approches de reconnaissance;
- le dernier chapitre est le noyau de ce projet, il présente la description de notre application : le principe de fonctionnement des différentes fonctions constituantes

notre **S**ystème d'**I**dentification **A**utomatique du **L**ocuteur **SIAL**, les étapes suivies pendant sa réalisation ainsi que et les résultats obtenus.

Chapitre 1:

Généralités Sur La Parole

1.1. INTRODUCTION

Dans ce chapitre nous présentons le processus de production et de perception de la parole. Nous définissons par la suite le signal de la parole tout en expliquant les types de sons produits, et la classification de phonème. Finalement, nous donnons plus de détails sur la description acoustique de la parole.

1.2. PROCESSUS DE PRODUCTION ET DE PERCEPTION DE LA PAROLE

L'objectif fondamental de la parole est la communication humaine, c'est-à-dire la transmission d'un message entre un locuteur et un auditeur [1] (figure 1.1).

Le processus de production de la parole comporte les étapes suivantes :

- le locuteur formule un message (dans son esprit) qu'il veut transmettre à l'auditeur par la parole ;
- la conversion du message dans un code d'une langue. Une fois la langue est choisie, le locuteur doit exécuter une série de commandes neuromusculaires qui mettent les cordes vocales en vibration, produisant ainsi une onde acoustique propageant vers un auditeur.

Le processus de perception se déroule de la manière suivante :

- le signal acoustique sera traité dans la membrane basilaire au niveau l'oreille interne qui fournit une analyse spectrale ;
- le processus de transduction neuronale convertit le signal spectral à la sortie de la membrane basilaire en potentiel d'action nerveux, correspondant approximativement à un processus d'extraction de paramètres ;
- ce potentiel est converti dans le code de la langue dans le cerveau, et finalement la compréhension du message est atteinte [2].

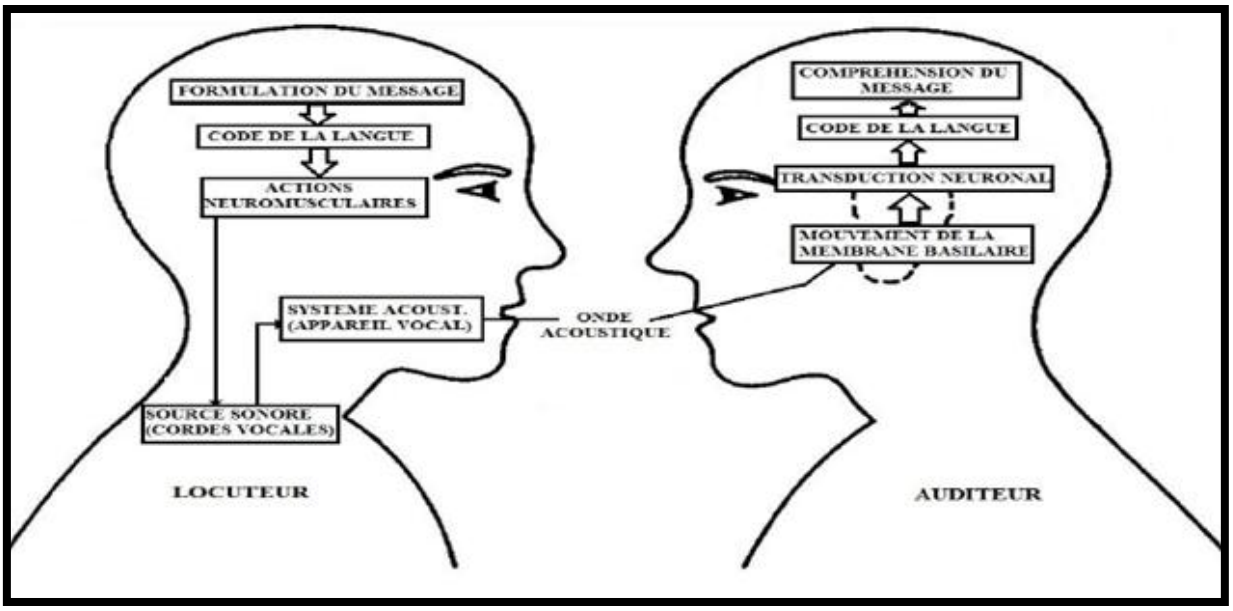

Fig. 1.1 : Processus de production et de perception de la parole chez les êtres humains [2]

1.3. Production de la parole

Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique (figure 1.2).

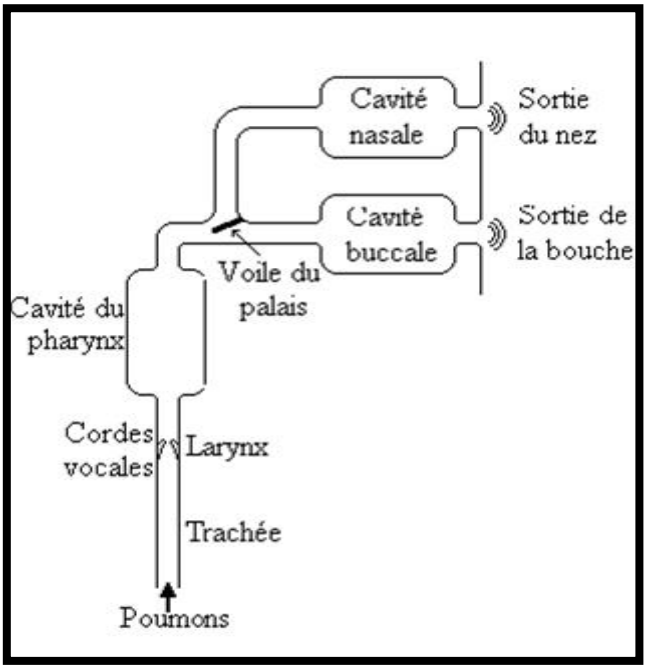

Fig. 1.2: Représentation schématique de la production de la parole [2]

La voix résulte du fonctionnement simultané des organes de l'appareil phonatoire (figure 1.3). Ces derniers modifient sa forme et ses dimensions suivant le son émis [3].

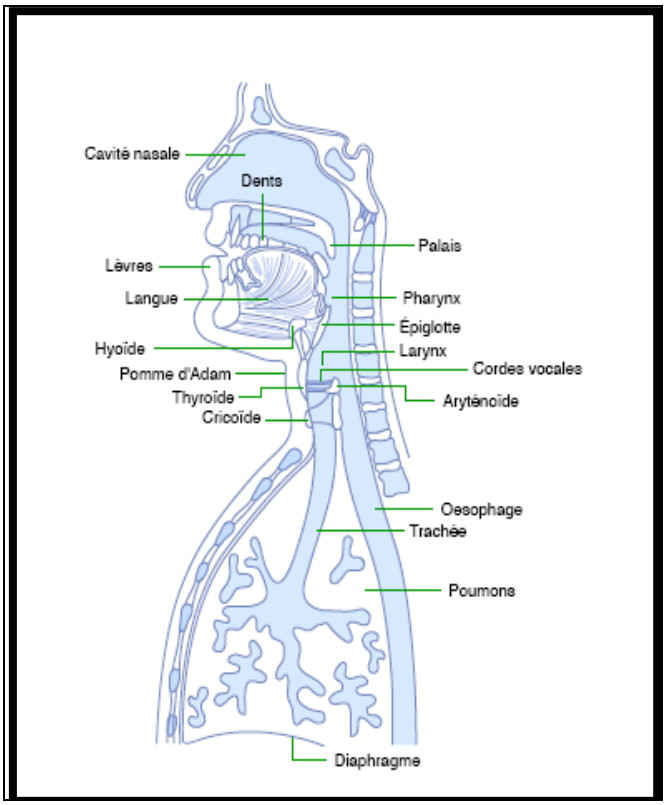

Fig. 1.3 : Appareil phonatoire humain [1]

Des organes et des muscles entrent dans la production des sons :

- les poumons : qui sont une source d'énergie. Ils se comportent comme un générateur d'air qui alimente le larynx. Le processus de génération de l'expiration phonatoire nous permet de disposer d'une énergie ventilatoire qui va pouvoir mettre en mouvement les cordes vocales ;
- le larynx : source vocale (laryngienne ou bruit). C'est est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée artère (figure 1.4).

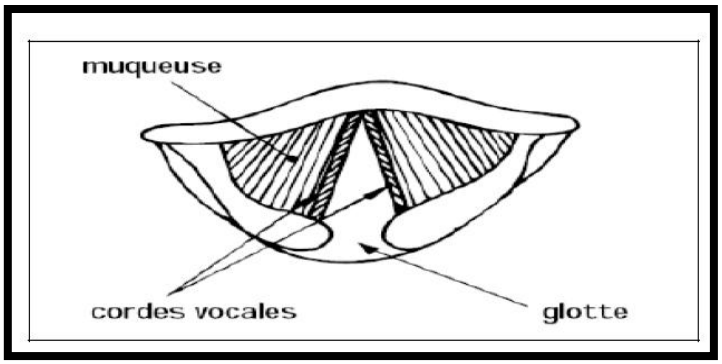

Fig. 1.4 : Section du larynx [2]

- la langue : c'est un articulateur fondamental, sa position est déterminante dans la forme du conduit vocal. La grande mobilité de la langue est due à un système musculaire complexe ;
- les lèvres : elles constituent la terminaison du conduit vocal. Les lèvres sont deux replis musculo-membraneux vers le quel converge un grand nombre de muscles qui permettent une grande mobilité ;
- les fosses nasales et le voile du palais : au cours de la respiration, la bouche est fermée, le voile du palais abaissé et l'air inspiré transite vers les poumons, filtré et réchauffé par les fosses nasales, il refait en sens inverse le même trajet lors de l'expiration. Les cordes vocales vibrent sous l'effet du passage de l'air à travers la glotte, après avoir parcouru le pharynx [4] ;
- les cavités supra-glottiques : renferme les organes qui permettent de modifier le son qui est émis par le travail conjoint des poumons et du larynx [3].

1.4. LE SIGNAL DE LA PAROLE

La parole est un signal réel, continu, d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps : tantôt périodique (plus exactement pseudopériodique) pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt impulsionnelle dans les phases explosives des sons occlusifs.

Les sons voisés résultent donc de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales. Ils ont une structure quasi-périodique.

Pour les sons non voisés les cordes vocales sont relâchées (ne vibrent pas), l'air passe librement au niveau du larynx. Ils ne présentent pas de structure périodique.

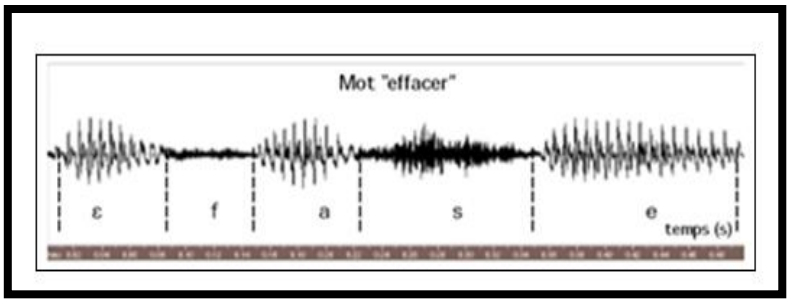

Fig. 1.5 : Audiogramme de signal du mot « effacer » [3]

1.4.1. **Types de sons langagiers produits**

D'un point de vue linguistique, la production des sons ou d'un mot réside dans la production en série de tous les phonèmes constituant ce mot. Ces phonèmes forment les unités phonétiques qui sont classées en voyelles, consonnes et semi-voyelles.

1.4.1.1. Les voyelles

Les voyelles sont caractérisées par le passage de l'air à partir de la cavité supra-glottique, et par le voisement qui crée des formants en bases fréquences. Les voyelles sont [4] :

- Orales : se prononcent avec le voile du palis relevé, ce qui ferme le passage nasal [5] ;
- Nasales : se caractérisent par l'abaissement du voile du palais et donc la mise en communication du conduit nasal avec le conduit oral (pharynx et conduit buccal).

1.4.1.2. Les consonnes

Les consonnes forment, au plan acoustique, une classe très hétérogène que l'on peut décomposer en trois ensembles ayant des caractéristiques distinctes :

- **Les consonnes occlusives :** Une occlusive se compose d'une suite d'événements acoustique : un silence qui correspond à la phase de tenue articulatoire de l'occlusion complète du conduit vocal suivie d'une ouverture expirant brutalement l'air emmagasiné dans le conduit vocal (explosion). On distingue :
  - les occlusives labiales : [p], [b] se caractérisent par une barre d'explosion dont l'enveloppe spectrale a une forme diffuse-descendante. L'énergie, faible et de durée brève, est généralement répartie dans une large bande de fréquence ;
  - les occlusives dentales : [t], [d] dont la barre d'explosion, intense est décrite comme diffuse-montante. L'énergie est répartie dans une large bande de fréquence ;

- les occlusives vélaires : [k], [g] dont l'énergie de la barre d'explosion, intense et de longue durée, est concentrée dans une étroite bande de fréquence.

- **Les consonnes fricatives** : sont des bruits qui résultent d'une turbulence aérodynamique qui prend naissance en un ou plusieurs points du conduit vocal en raison de la présence d'un obstacle placé dans le flot d'air expiratoire.
- **Les consonnes sonantes** : se caractérisent par une structure de formants. Elles ne possèdent que peu ou pas de bruit. On distingue :
  - les consonnes nasales : [m], [n], [ɳ] sont souvent décrites comme les occlusives dans la mesure où l'on ne tient compte que de la partie buccale du conduit vocal, où elles présentent une occlusion ;
  - les liquides [L], [R] : se produisent lorsque la pointe de la langue fait contact avec la zone alvéodentale mais l'air peut passer librement des deux cotés de la pointe ;
  - les semi-consonnes [j], [w] : appelés aussi semi-voyelles. Elles sont produites par le passage de l'air à travers le conduit vocal qui fonctionne en mode résonnant.

1.4.2. Notation phonétique du Français

Transcrire phonétiquement un énoncé oral, c'est noter à l'aide d'un alphabet conventionnel, en général l'alphabet phonétique international **API**, (**IPA,** 1982), la séquence des sons phonétiques qui composent cet énoncé. Les symboles utilisés pour transcrire le français sont donc un sous-ensemble des symboles de l'**API** définis par une description articulatoire et un exemple choisi dans une langue particulière (tableau 1.1).

Tableau 1.1 : Des signes phonétiques de l'API utilisés en Français [4]

Consonnes			
	[p] paie	[t] taie	[k] quai
	[b] baie	[d] dais	[g] gai
	[m] mais	[n] nez	[ɲ] gagner
	[f] fait	[s] sait	[ʃ] chez
	[v] vais	[z] zéro	[ʒ] geai
	[w] ouais	[ɥ] huer	[j] yéyé
		[l] lait	[R] raie
Voyelles			
	[i] lit	[y] lu	[u] loup
	[e] les	[ø] leu	[o] lot
	[ɛ] lait	[œ] leur	[ɔ] lotte
	[a] là	[ə] le	
	[ɛ̃] lin	[ɑ̃] lent	[ɔ̃] long

1.4.3. Phonétique et phonologie

Phonétique et phonologie sont deux sciences qui ont le même objet : les sons du langage. Ce qui les différencie, c'est le point de vue, adopté dans la description, c'est à dire ce que l'on veut faire découvrir dans l'objet étudié :

- la phonétique s'intéresse à la manière dont les sons du langage sont produits, transmis, perçus par les sujets parlants ;
- la phonologie s'efforce de découvrir comment ces sons participent au fonctionnement de la langue dans l'acte de la parole et en assument le codage.

1.4.4. Classification des phonèmes du Français

Les phonèmes sont les éléments sonores les plus brefs qui permettent de distinguer différents mots.

Le mode d'articulation définit le degré de contact entre les articulations qui existe durant la prononciation d'une consonne.

Le point d'articulation définit l'endroit où se produit une consonne.

Diverses transcriptions phonétiques sont utilisées [4] (tableau 1.2).

Tableau 1. 2 : Classification des phonèmes du francais en traits distinctifs [4]

CONSONNES Mode d'articulation ↓	Labiales	Dentales	Vélo-palatales ←	Lieu d'articulation
Occlusives				
non voisées	[p]	[t]	[k]	
voisées	[b]	[d]	[g]	
Nasales	[m]	[n]	[ɲ]	
Fricatives				
non voisées	[f]	[s]	[ʃ]	
voisées	[v]	[z]	[ʒ]	
Glissantes	[w]	[ɥ]	[j]	
Liquides		[l]	[R]	
VOYELLES				
Orales	Antérieures		Postérieures	
	Non arrondies		Arrondies	
Fermées	[i]	[y]	[u]	
	[e]	[ø]	[o]	
	[ɛ]	[œ]	[ɔ]	
Ouvertes	[a]			
Nasales	Antérieures		Postérieures	
Fermées	[ɛ̃]		[ɔ̃]	
Ouvertes		[ɑ̃]		

1.5. Système auditif et perception de la parole

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production.

Les processus complexes par lesquels un auditeur comprend un message oral émis par un locuteur peuvent être fonctionnellement décomposés en deux grandes phases :

- l'oreille transforme l'information contenue dans le signal acoustique et le transmet ensuite au cerveau par l'intermédiaire du nerf auditif ;
- la reconnaissance du message linguistique par l'interprétation d'indices fournis à l'issu de prétraitement auditif sans référence à la signification puis la réalisation de l'accès au sens.

1.5.1. Système de transmission

L'appareil auditif comprend l'**O**reille **E**xterne (**OE**), l'**O**reille **M**oyenne (**OM**) et l'**O**reille **I**nterne (**OI**) (Figure 1.5).

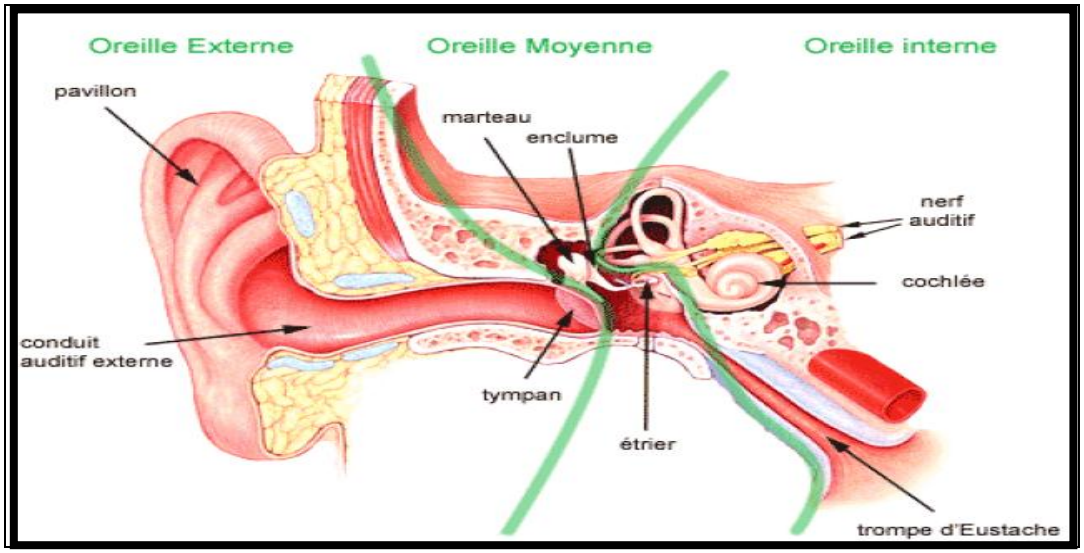

Fig. 1.6 : Le système auditif [6]

- l'**OE** relie le pavillon et conduit auditif permet de recueillir les sons et de les orienter vers l'oreille moyenne ;
- l'**OM** comprend le tympan et les osselets. Elle assure la fonction de transmission, qui inclut une transformation d'ondes sonores aériennes en ondes liquidiennes. Elle

joue aussi un rôle d'accommodation auditive. La position des osselets les uns par rapports aux autres assure l'amplification des sons ;

- le mécanisme composé de marteau, étrier et enclume permet une adaptation d'impédance entre l'air et le milieu liquide de l'**OI**. Les vibrations de l'étrier sont transmises au liquide de la cochlée qui avec le nerf auditif assurent la fonction de réception [4,7].

1.5.2. L'aire de l'audition

Le système auditif ne répond pas également à toutes les fréquences. Le champ auditif humain est délimité par la courbe de ***seuil de l'audition*** et celle du ***seuil de la douleur*** (figure 1.6). Sa limite supérieure en fréquences (≈ 16 000-20 000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (≈ 32000 Hz) [7].

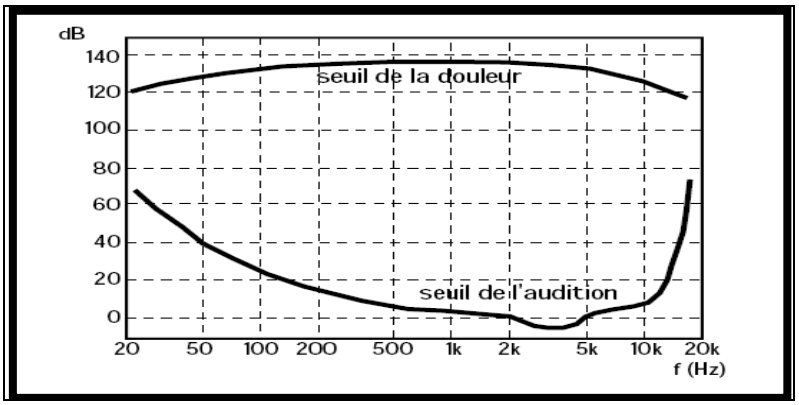

Fig. 1.7 : Le champ auditif humain [8]

1.6. DESCRIPTION ACOUSTIQUE DE LA PAROLE

Du point de vue phonétique acoustique, les sons du langage humain sont constitués par des ondes en mouvement. Il s'agit d'un mouvement vibratoire régulier ou irrégulier par les articulateurs et les cordes vocales.

Ces vibrations sont ensuite transmises par les milieux matériels (l'air, l'eau, le bois, le métal,…). Contrairement à la lumière, le son ne se propage pas dans le vide, il est à signaler que le son se propage dans l'air à une vitesse de 340 m/ s [9].

1.6.1. **La mélodie**

La mélodie de la voix résulte de la vibration des cordes vocales, et se traduit phonétiquement par l'évolution de la fréquence de vibration laryngienne, on utilise plutôt le terme fréquence fondamentale, qui correspond à une estimation de fréquence laryngienne réalisée à partir du signal de parole [4]

1.6.2. La fréquence fondamentale (F0)

La fréquence fondamentale ou F_0 est également appelée pitch. Elle représente le nombre de vibrations par seconde des cordes vocales. La F0 n'est calculée que sur des parties voisées de la parole. La gamme de variation moyenne de la fréquence fondamentale dépend, essentiellement, de l'âge, de l'état et du sexe du locuteur. Elle peut varier de :

- 70 à 250 Hz chez l'homme ;
- 150 à 400 Hz chez la femme ;
- 200 à 600 Hz chez l'enfant [9].

1.6.3. L'intensité

Par l'intensité, nous désignons la qualité qui nous fait distinguer un son fort d'un son faible. L'intensité augmente avec l'amplitude des vibrations sonores. On utilise une unité de mesure relative, le décibel (dB), pour rendre compte de l'intensité d'un son.

1.6.4. La durée

La durée est le paramètre acoustique le plus délicat à évaluer. La difficulté de mesure réside dans sa grande variabilité qui est due au contrôle quasi impossible du système phonatoire.

Chaque phonème se caractérise par ses propres durés intrinsèques et extrinsèques.

1.6.5. Le timbre

Il représente la qualité particulière du son, indépendante de son intensité ou de sa hauteur, mais spécifique de l'instrument ou de la voix qui l'émet.

Le timbre est la qualité qui nous permet de distinguer les différents instruments de musique ou de reconnaître une voix familière, par exemple.

1.6.6. Les formants

Les formants sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces

formants représentent les maxima de la courbe de réponse en fréquences du conduit vocal. Chaque son à ses formants caractéristiques (figure 1.6 et 1.7) [5].

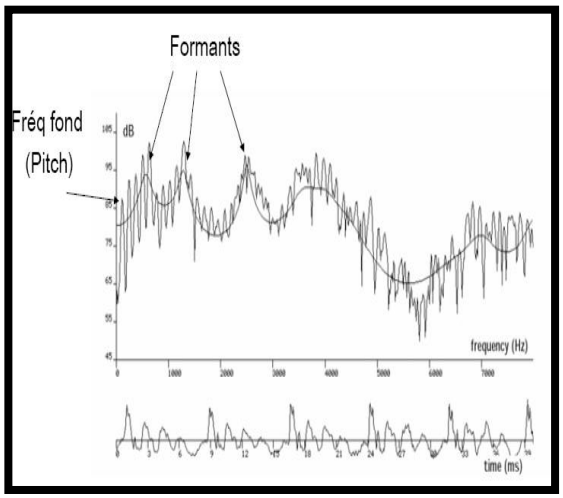

Fig. 1.8 : Le son voisé [a] de baluchon[2]

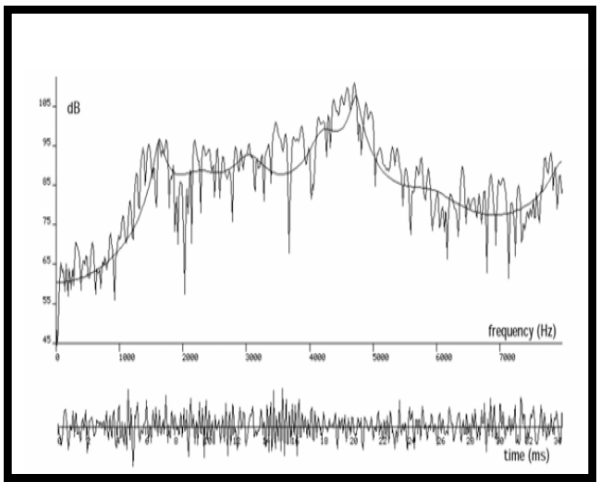

Fig. 1.9 : Le son non voisé [ ] de baluchon[2]

1.7. CONCLUSION

Dans ce chapitre, nous avons expliqué les deux approches : la production de la parole chez les êtres humains et sa perception par le système auditif. Nous avons exposé aussi les types de sons langagiers et la description acoustique de la parole.

Chapitre 2 :
Biométrie Vocale

2.1. INTRODUCTION

Ce chapitre est consacré aux principes de la RAL, il présente ses domaines d'application ainsi que ses différentes tâches, telles que l'**I**dentification et la **V**érification **A**utomatique du **L**ocuteur (IAL et VAL).

2.2. DEFINITION DE LA BIOMETRIE

Le mot **biométrie** signifie « mesure du vivant », et désigne dans un sens très large l'étude quantitative des êtres vivants. L'usage de ce terme se rapporte de plus en plus aux techniques de reconnaissance, d'authentification et d'identification d'un être vivant [11].

Il existe différents systèmes de biométrie, tels que les empreintes, la reconnaissance d'iris ou la reconnaissance de la parole. La biométrie vocale est une technologie non intrusive qui présente de nombreux avantages, de fiabilité et de sécurité, ce qui lui permet d'être perçue comme outil de haute technologie.

C'est dans les années 40 à l'USA, que les premières tentatives de création d'une machine capable de comprendre le discours humain eurent lieu. Leurs principaux objectifs étaient d'interpréter les messages russes interceptés.

2.3. HISTORIQUE

Le système de reconnaissance de la parole a connu une évolution rapide, voici quelques dates qui ont marqué l'histoire de la biométrie vocale :

1952 : reconnaissance des 10 chiffres par un dispositif électronique câblé.

1960 : utilisation des méthodes numériques.

1965 : reconnaissance de phonèmes en parole continue.

1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots).

1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables.

1972 : premier appareil commercialisé de reconnaissance de mots.

1978 : commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés.

1983 : première mondiale de commande vocale à bord d'un avion de chasse en France.

1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots.

1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réels.

1988 : apparition des premières machines à dicter par mots isolés.

1990 : premières véritables applications de dialogue oral homme-machine.

1994 : IBM lance son premier système de reconnaissance vocale sur PC.

1997 : lancement de la dictée vocale en continu par IBM.

2.4. DOMAINES D'APPLICATIONS

Nous donnons quelques exemples d'applications en RAL que nous pouvons regrouper en 3 catégories principales : applications sur sites géographiques, téléphoniques et juridiques [12] :

Tableau 2. 1 : Exemples d'applications en RAL [12]

Applications sur sites géographiques Cette catégorie concerne les applications qui se trouvent sur site géographique particulier, elles sont utilisées principalement pour limiter l'accès à des lieux privés. Par exemple la protection de domicile, garage, bâtiment, etc. l'intérêt de ce type d'application est :	• l'environnement est facilement contrôlable ; • la vérification du locuteur un effet dissuasif ; • la reconnaissance vocale peut être associée à d'autres techniques de reconnaissance d'identité (ex : analyse du visage, des empreintes digitales, etc.) ; • l'utilisateur peut avoir son propre modèle sur lui (ex : sur la puce d'une carte).
Applications téléphoniques Ce type d'applications utilise le téléphone comme un moyen matériel de Communication entre l'Homme et la Machine. C'est la catégorie la plus importante parce qu'elle permet de vérifier ou identifier le locuteur à longue distance. Il existe plusieurs applications dans cette catégorie parmi elles, nous citons :	• la validation de transactions bancaires par téléphone ; • l'accès à des **B**ases de **D**onnées **(BD)** pour plus de sécurité et pour plus de protection (ex : consultation de répondeur).

Applications juridiques Enfin nous trouvons que le domaine d'application qui propose actuellement le plus de problèmes, est le domaine juridique et criminalistique. La reconnaissance du locuteur est utilisée par exemple pour :	• l'orientation des enquêtes ; • la constitution des éléments de preuves au cours d'un procès.

2.5. QU'EST-CE-QU'UNE RECONNAISSANCE DE LA PAROLE ?

Le terme "reconnaissance" est défini comme étant l'identification de quelque chose, sachant qu'on doit connaître au préalable son modèle de référence.

La reconnaissance automatique d'un individu consiste à utiliser des caractéristiques physiques dans le but de faire une discrimination entre les différents individus. Pour ce faire, plusieurs caractéristiques sont proposées dans la littérature : la photographie du visage, les empreintes digitales, les traits génétiques ou encore le signal de parole.

L'authentification par la voix est appelée "Reconnaissance Automatique du Locuteur". Il convient dans ce domaine de rechercher à connaître non pas ce qui a été dit, mais l'identité de la personne qui parle, à partir de son « empreinte » vocale [4]. Cependant, ici nous rencontrons un problème majeur, qui est défini par la difficulté de trouver des caractéristiques pertinentes en discrimination : ce problème implique la nécessité de trouver des caractéristiques possédant une grande variabilité inter-locuteur et une faible variabilité intra-locuteur.

2.6. NOTIONS SUR LA VARIABILITE DU SIGNAL VOCAL

Il existe deux types de variabilités (pour une caractéristique acoustique donnée), la variabilité :

- **intra-locuteur** : elle identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un

être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie ;

- **inter-locuteur** : la voix de chaque individu possède des qualités qui lui sont propres, l'âge, le sexe, le tempérament du locuteur (et bien d'autres facteurs encore) lui confèrent une identité vocale originale qui est la combinaison de multiples paramètres dont la hauteur (pitch), l'intensité et le timbre de sa voix, la qualité de son articulation, ou encore son accent (national et régional).

Le Coefficient de Wolf "CW" est le rapport de la variabilité inter-locuteur sur la variabilité intra-locuteur :

- une grande variabilité inter-locuteur : indique qu'on peut séparer, facilement, les locuteurs par leurs caractéristiques ;
- une petite variabilité intra-locuteur : indique que chaque locuteur peut être représenté par une référence qui le représente très bien.

Par conséquent, un grand coefficient de Wolf indique, alors, que le paramètre choisi est pertinent en identification du locuteur et devrait donner un bon score de reconnaissance.

2.7. RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR

La caractérisation automatique du locuteur est un vaste domaine dans lequel la "machine" a pour tâche d'extraire du signal de parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc. Elle s'applique à différents thèmes de recherche traitant des informations véhiculées par la voix tels que la classification d'individus, ou l'étude psychique ou physiologique d'une personne.

La **R**econnaissance **A**utomatique du **L**ocuteur (RAL) est un sous-problème de la caractérisation automatique du locuteur. Son objectif est de reconnaître l'identité d'une personne à l'aide de sa voix. La variabilité de la parole entre locuteurs (variabilité inter-locuteur) est l'essence même de la RAL. Sans cette variabilité, il serait impossible de reconnaître une voix parmi plusieurs voix possibles [9].

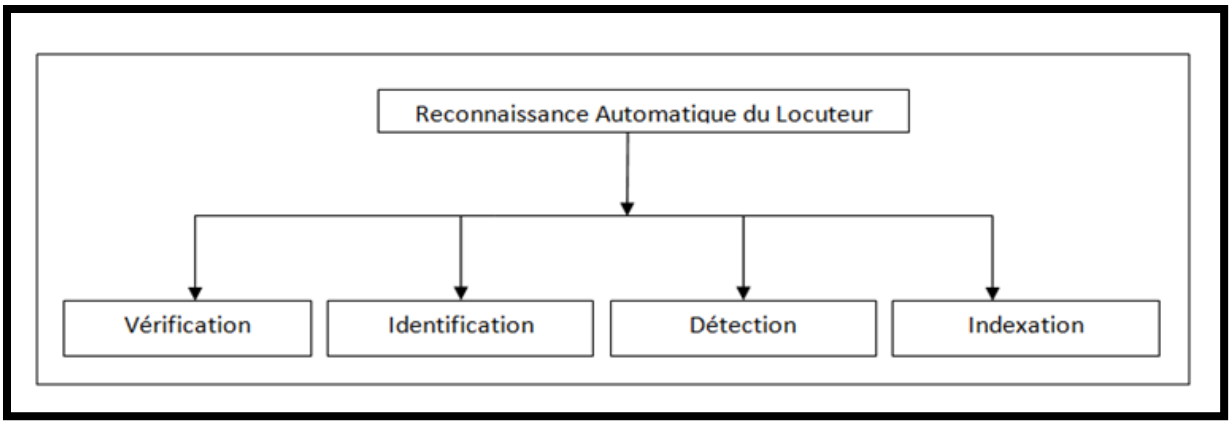

Fig. 2.1 : Schéma illustrant les différentes taches de la RAL

2.7.1. Niveau de Dépendance au Texte

Les systèmes de RAL peuvent nécessiter de la part du locuteur qu'il prononce un texte déterminé, ou lui laisser prononcer ce qu'il veut.

Dans les systèmes de VAL, souvent réalisés pour permettre l'accès physique à des zones surveillées, ou autoriser une transaction bancaire par téléphone par exemple, on suppose le locuteur coopératif, et il sera donc possible d'utiliser des systèmes dépendants du texte, plus simples à mettre en œuvre, et aux performances convenables. De fait, la nature du texte à prononcer elle-même, telle un mot de passe, ajoute un degré de sécurité supplémentaire.

Par contre, dans les systèmes d'IAL, souvent utilisés pour déterminer l'identité d'un locuteur à son insu, à partir d'une conversation téléphonique par exemple, l'indépendance de la méthode au texte prononcé peut être une nécessité [4]. En résumé :

- en mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne ;
- en mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système, une terminologie plus fine peut être donnée à un système suivant l'application visée, systèmes à :
  - messages fixés : la personne est contrainte de prononcer un message, qu'elle aurait au préalable (mots de passe personnalisés) ou qui sera imposé par le système ;
  - messages prompts : un message, différent à chaque nouvelle session de reconnaissance, est imposé par le système sous forme visuelle ou auditive. Ces systèmes ont pour première motivation de se protéger des attaques de personnes

malveillantes (imposteurs) qui disposeraient d'un enregistrement de la voix d'une personne ;

- unités segmentales : la personne doit prononcer un message comportant soit une séquence de mots (séquence de chiffres), soit des traits phonétiques (séquence de phonèmes) connus du système [9].

2.7.2. Vérification et Identification automatique du Locuteur

La reconnaissance du locuteur regroupe en fait deux taches distinctes :

2.7.2.1. Vérification Automatique du Locuteur

Il s'agit ici, après que le locuteur ait décliné son identité, à l'aide d'un badge magnétique, par la frappe d'un code numérique, ou même vocalement, de vérifier l'adéquation du message vocale qu'on lui prétend être. C'est donc une décision par tout ou rien (et un système opérant une décision aléatoire aurait un taux de réussite de 50%) [4] (Figure 2.2). L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans le cas où la mesure de similarité est inférieure au seuil, l'individu est accepté, dans le cas contraire, l'individu est considéré comme un imposteur, est rejeté [9].

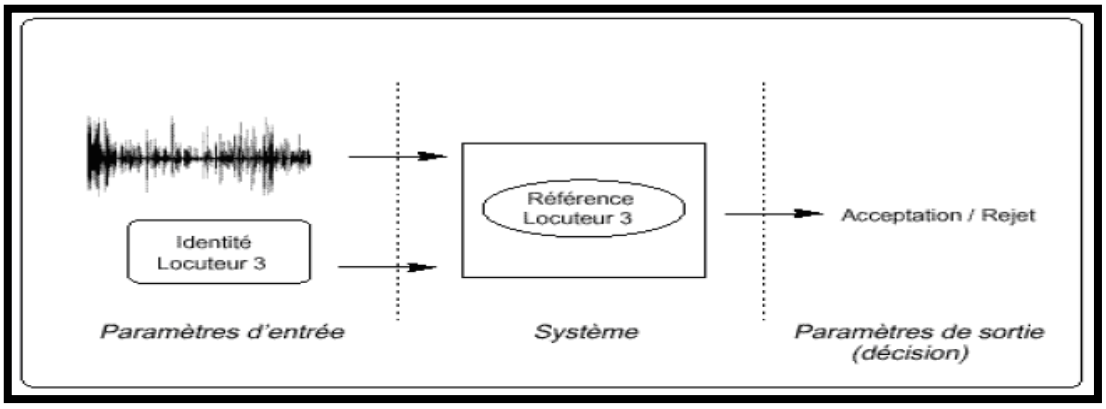

Fig. 2.2 : Principe de base de la tâche de Vérification Automatique du Locuteur [9]

2.7.2.2. Identification Automatique du Locuteur

Il convient de comparer un message vocal avec un ensemble de références acoustiques correspondant à plusieurs personnes, et de déterminer par cet examen quelle est la personne qui a parlé. Le choix est donc multiple, et une décision aléatoire conduirait à un taux de réussite de 1/N %, ou N est le nombre de personnes à reconnaître [4].

D'un point de vue schématique (figure 2.3), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL.

Deux modes sont proposés en IAL, l'identification en ensemble :

- fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système ;
- ouvert pour lequel le locuteur peut ne pas être connu.

En mode "ensemble ouvert", le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée. De par son principe - déterminer une identité parmi les identités potentielles - les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

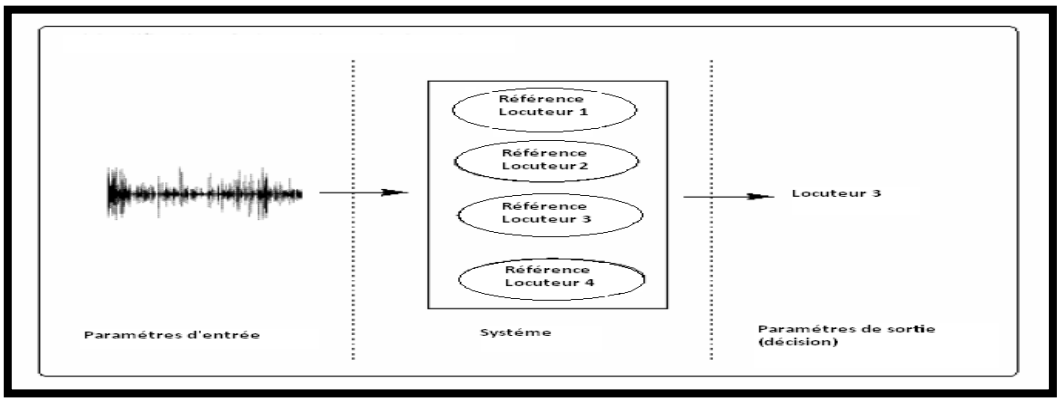

Fig. 2.3 : Principe de base de la tâche d'Identification Automatique du Locuteur [9]

2.7.2.3. Détection de Locuteurs

La détection de locuteurs dans un flux audio est une variante de la VAL. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux audio monolocuteur, la tâche de détection se résume à la tâche de vérification.

2.7.2.4. Indexation par Locuteurs et ses Variantes

La tâche d'Indexation Automatique par Locuteurs consiste à cibler les interventions des locuteurs dans un flux audio (figure 2.4). En d'autres termes, indexer un document audio en

locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu.

La seule entrée d'un système d'indexation est le document audio à indexer. Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité [9].

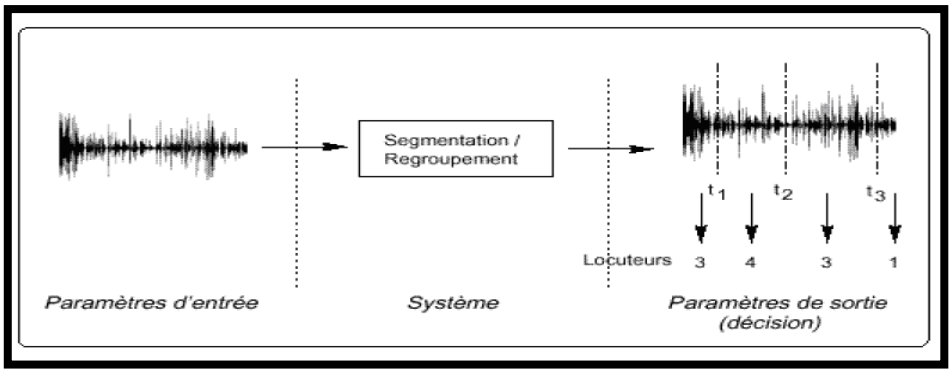

Fig. 2.4 : Principe de base de la tâche d'Indexation par Locuteurs d'un flux audio [9]

2.8. CONCLUSION

Dans ce chapitre, nous avons exposé les différents types de systèmes de RAL, qui sont l'IAL et VAL. La biométrie vocale est une des technologies de contrôle d'identité par des attributs physiques les plus répandus, car la voix est celle qui peut se déployer le plus rapidement. Nous étudierons dans les chapitres suivants l'analyse acoustique du signal parole, ainsi que les différentes techniques de modélisation de ce dernier.

chapitre 3 :
TECHNIQUES DE RECONNAISSANCE DE LOCUTEUR

3.1. INTRODUCTION

Dans ce chapitre, nous allons décrire les outils de traitement du signal vocal avec des méthodes non paramétriques et paramétriques, puis la modélisation acoustique du locuteur à savoir les approches vectorielle et statistique.

3.2. TECHNIQUES DE TRAITEMENT DU SIGNAL VOCAL

Le signal de la parole est un signal très complexe. Il contient une quantité importante d'informations imbriquées entre elles.

Le traitement du signal vocal a pour but de fournir une représentation moins redondante de la parole, tout en permettant une extraction précise des paramètres significatifs.

Les principales classifications des méthodes de traitement du signal vocal sont:

- les transformées usuelles comme la Transformée Discrète de Fourier qui ne se réfère pas à un modèle de production ni de perception ;
- les méthodes fondées sur la déconvolution « source - conduit vocal » cepstre et codage prédictif linéaire qui s'appuient sur le modèle de production de la parole [4].

3.2.1. Méthodes non paramétriques

Le signal de la parole peut être analysé dans les domaines spectral ou temporel par des méthodes non paramétriques, sans faire l'hypothèse d'un modèle pour rendre compte du signal observé.

3.2.1.1. Chaîne de prétraitement

Le calcul de la représentation du signal est réalisé par une chaîne d'analyse numérique en suivant ces étapes :

- l'échantillonnage transforme le signal à temps continu $s\ (t)$ en signal à temps discret $s\ (nT_e)$ défini aux instants d'échantillonnage T_e. Celui-ci doit se faire selon le théorème de Shannon (la fréquence d'échantillonnage doit être supérieure ou égale à deux fois leur plus haute composante fréquentielle) ;
- la préaccentuation : Le filtre de préaccentuation qui est souvent non récursif du premier ordre, permet d'égaliser les aigus toujours plus faibles que les graves. la préaccentuation s'_n de l'échantillon n est calculé pour une valeur α comprise entre 0,9 et 1 comme

$$S'_n = S_n - \alpha . S_{n-1} \quad (3.1)$$

- le fenêtrage : Vu que le signal vocal est non stationnaire, nous utilisons une fenêtre glissante; chaque trame couvrant une durée de 20 à 30 ms sur laquelle le signal est supposé quasi-stationnaire. Le pas d'analyse entre deux trames successives est de l'ordre de quelques dizaines de ms. Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre. Pour compenser ces effets de bord, nous multiplions en général préalablement chaque tranche d'analyse par une fenêtre de pondération de type fenêtre de Hamming [13]:

$$S''_n = w_n . S'_n \quad (3.2)$$

avec

$$w(n) = 0.54 - 0.46 cos\left(2\pi \frac{n}{N}\right), \qquad 0 \leq n \ll N \quad (3.3)$$

N : nombre d'échantillons dans une fenêtre

L = N+1 : longueur de la fenêtre

3.2.1.2. Analyse temporelle

Dans l'analyse temporelle du signal, on peut extraire des paramètres tels que l'énergie et la fréquence fondamentale :

- L'énergie du signal est un indice qui peut, par exemple contribuer à la détection du voisement d'un segment de parole. L'énergie totale E_0 est calculée directement dans le domaine temporel sur une trame de signal S_n, $0 \leq n \ll N-1$ par :

$$E_0 = \sum_{n=0}^{N-1} S_n^2 \quad (3.4)$$

- fréquence fondamentale : La parole est obtenue à partir de la vibration des cordes vocales. Des harmoniques sont filtrées par la cavité buccale, ce qui permet d'en extraire les sons voisés. Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale [14].

3.2.1.3. Analyse spectrale

Parmi les techniques utilisées, la **FFT** (**F**ast **F**ourier **T**ransform) joue un rôle de premier plan puisqu'elle permet d'obtenir des spectres en temps réel. Elle exprime la répartition fréquentielle de l'amplitude, de la phase et de l'énergie (ou de la puissance) des signaux considérés [4].

En **RAP**, il est important de connaître l'évolution de ce spectre dans le temps.

Soit *s (t)* un signal déterministe. Sa transformée de Fourier est une fonction, généralement complexe, de la variable *f* et définie par:

$$S(f) = TF[S(t)] = \int_{-\infty}^{+\infty} S(t)\ e^{-j2\pi ft}\ dt \qquad (3.5)$$

Le calcul de l'intensité s'effectue en décibels (figure 3.1).

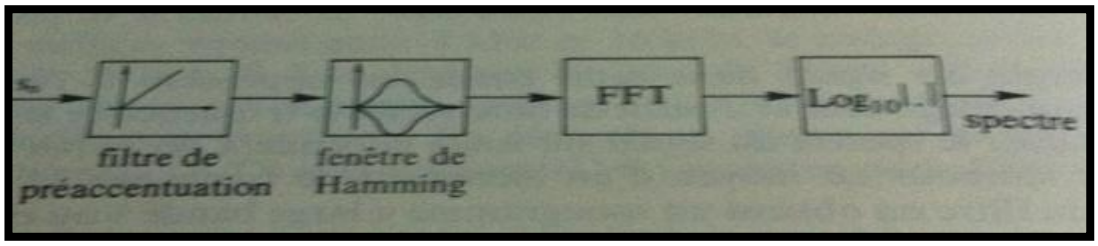

Fig. 3.1 : Analyse numérique du signal parole par FFT [4]

3.2.2. Méthodes paramétriques

Les méthodes paramétriques appelées aussi méthodes d'identification sont fondées sur une connaissance des mécanismes de production de la parole. Les plus utilisées sont celles basées sur :

- l'analyse prédictive linéaire ;
- l'analyse cepstrale.

Les avantages de cette approche sont :

- la souplesse de l'analyse et une meilleure caractérisation du signal ;
- l'introduction naturelle de l'information et les choix variés des espaces de représentations paramétriques [15].

3.2.2.1. Codage prédictif Linéaire et LPCC

Le **C**odage **P**rédictive **L**inéaire **LPC (L**inear **P**redictive **C**oding) est basé sur le modèle de production de la parole, qui considère que l'appareil de production de la parole est constitué d'une source (source pseudopériodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal) (figure 3.2).

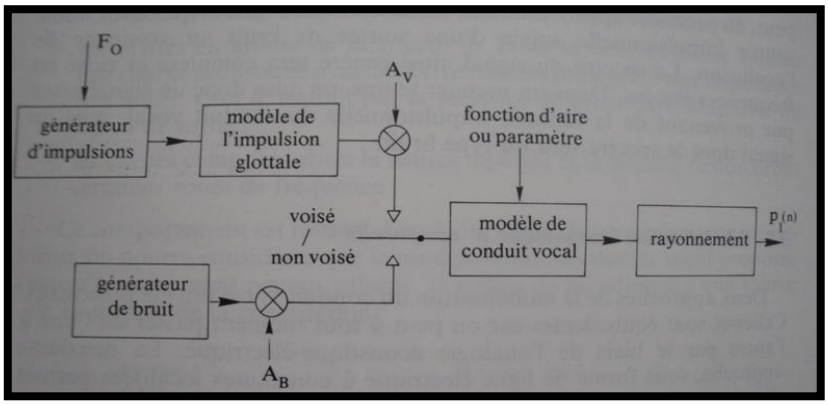

Fig. 3.2 : Représentation fonctionnelle du fonctionnement du conduit vocal et des sources d'excitations [4]

Le signal de parole peut être ainsi modélisé comme étant le signal en sortie d'un filtre H(z) dont la source d'excitation à l'entrée du filtre u(t) est soit une source de série d'impulsions quasi-périodiques, soit un bruit blanc (figure 3.3) [4].

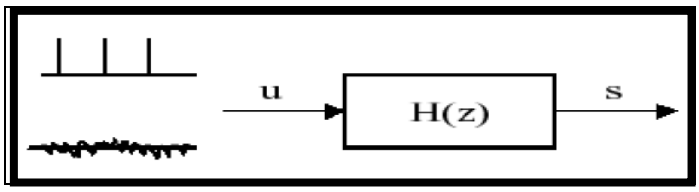

Fig. 3.3 : Modèle source-filtre de production de la parole [4]

L'analyse LPC repose sur l'hypothèse que le filtre est tous-pôles, avec cette hypothèse, le signal de la parole peut être considéré comme un signal auto régressif :

$$S(n) = \sum_{k=1}^{p} a_k . S(n-k) + G.u(n) \quad (3.6)$$

$$H(z) = \frac{S(z)}{G.\ U(z)} = \frac{1}{1-\sum_{k=1}^{p} Z^{-k}\, a_k} = \frac{1}{A(z)} \quad (3.7)$$

Où :

G est le coefficient de gain du modèle source-filtre;

a_k sont les coefficients LPC ;

p est l'ordre du filtre.

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance ou la matrice d'autocorrélation.

Les **L**inear **P**redictive **C**epstral **C**oefficients **LPCC** (c_n) sont dérivés directement des coefficients LPC à travers le système d'équations suivant [15]:

$$\begin{cases} C0 = Ln\ G & (3.8) \\ \mathrm{Cm} = \mathrm{am} + \sum_{k=1}^{m-1} (\frac{k}{m})\ \mathrm{ck\,am-k} \qquad 1 \leq \mathrm{m} \leq \mathrm{p} & (3.9) \\ \mathrm{Cm} = \sum_{k=1}^{m-1} (\frac{k}{m})\ \mathrm{ck\,am-k} \qquad \mathrm{m} > \mathrm{p} & (3.10) \end{cases}$$

3.2.2.2. Analyse cepstrale

Le défaut majeur des méthodes d'analyse, comme la FFT, pour le calcul du spectre réside dans l'intermodulation source - conduit vocal, qui rend difficile la mesure du fondamental F_0 et des formants [4].

Les coefficients cepstraux les plus répandus sont les **MFCC** (**M**el **F**requency **C**epstral **C**oefficients). Ils présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale.

Les MFCC sont une extension des cepstres qui sont utilisés pour mieux représenter les modèles de l'audition humaine. Le principe de calcul des MFCC est issu des recherches psychoacoustiques sur la tonie et la perception des bandes de fréquences par l'oreille humaine.

Les MFCC assurent une séparation entre les deux composantes:

- la fonction d'excitation glottique qui est caractérisée par le pitch ou la contribution de l'excitation se localise dans les quéfrences élevées ;
- la fonction de transfert du conduit vocal ou la contribution se retrouve dans les faibles quéfrences (premiers coefficients cepstraux) [16].

Le cepstre du signal de parole est défini comme étant la Transformée de Fourier Inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal [17] :

$$S(t) = e(t) * h(t) \quad (3.11)$$

Où *s(t)* est le signal de parole, *e(t)* est la source d'excitation glottique et *h(t)* est la réponse impulsionnelle du conduit vocal.

L'application du logarithme sur le module de la Transformée de Fourier dans l'équation donne :

$$Log\ |S(f)| = Log\ |E(f)| + Log\ |H(f)| \quad (3.12)$$

Par une transformée de Fourier inverse on obtient :

$$S'(\ cef\ ) = e'\ (\ cef\ ) + h'\ (\ cef\ ) \quad (3.13)$$

La dimension du nouveau domaine est homogène à un temps et s'appel la *quéfrence* (*cef*), le nouveau domaine s'appelle donc le domaine *quéfrentiel*. Un filtrage dans ce domaine s'appelle *liftrage* [18].

Ce domaine est intéressant pour faire la séparation des contributions du conduit vocal et de la source d'excitation dans le signal de parole. En effet, si les contributions relevant du conduit vocal et les contributions de la source d'excitation évoluent avec des vitesses différentes dans le temps, alors il est possible de les séparer par l'application d'un simple fenêtrage dans le domaine quéfrentiel (liftrage passe-bas) pour le conduit vocal.

Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'informations utiles dans le signal de parole. Les échelles perceptives les plus utilisées sont [17] :

- l'échelle Mel qui est linéaire en basses fréquences et logarithmique en hautes fréquences;

$$Mel\ (f) = 2595\ log\ (1 + \frac{f}{700}) \quad (3.14)$$

- l'échelle Bark

$$Bark\ (f) = 6\ arcsinh\ (\frac{f}{1000}) \quad (3.15)$$

f représente la fréquence (Hz).

La procédure de calcul des MFCC est illustrée comme suit (figure 3.4) :

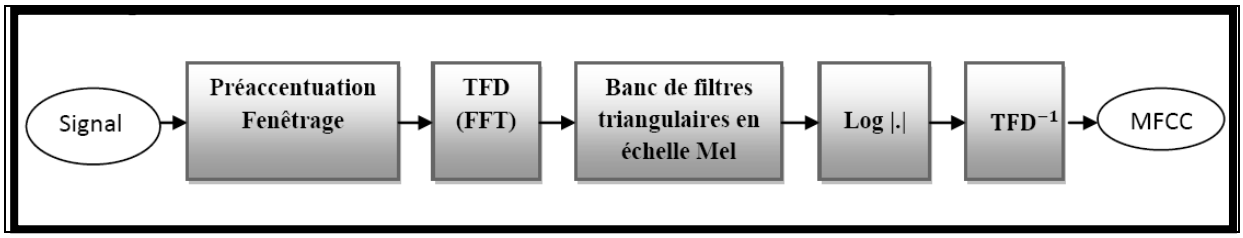

Fig. 3.4 : Schéma d'extraction des MFCC

Soit un signal discret *s(n)* avec, $0 \leq n \ll N-1$, N est le nombre d'échantillons d'une fenêtre d'analyse, F_e est la fréquence d'échantillonnage, la Transformée de Fourier Discrète à court terme *s(k)* est obtenue avec la formule :

$$S(k) = \sum_{n=0}^{N-1} S(n)\, exp^{\left(\frac{-j2\pi nk}{N}\right)} \quad , \quad 0 \leq k \leq N\text{-}1 \qquad (3.16)$$

Le spectre du signal est filtré par un banc de filtres triangulaires, dont les bandes passantes sont de même largeur dans le domaine des fréquences Mel (figure 3.5). Les points de frontières B_m des filtres en échelle de fréquence Mel sont calculés à partir de la formule :

$$B_m = B_b + m\,\frac{Bh-Bb}{M+1} \quad , \quad 0 \leq m \leq M+1 \qquad (3.17)$$

M: Le nombre de filtres.

B_h: La fréquence la plus haute du signal.

B_b: La fréquence la plus basse du signal.

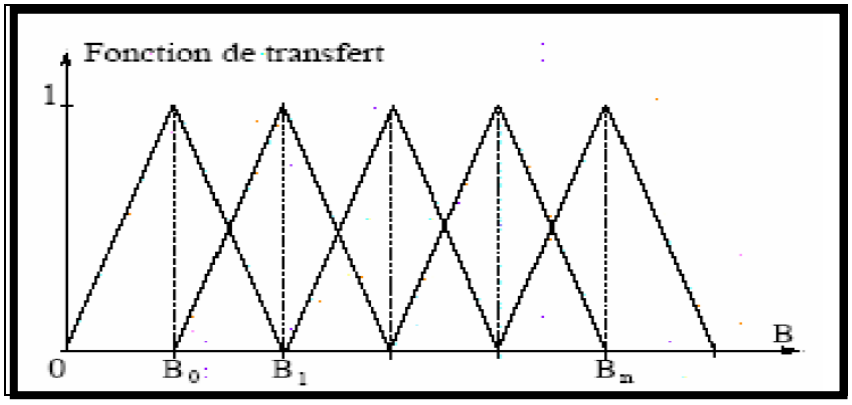

Fig. 3.5 : Banc de filtre sur l'échelle Mel [11]

Les coefficients cepstraux peuvent être calculés directement à partir du logarithme des énergies E_i issues d'un banc de M filtres par la transformée en cosinus discrète inverse définie par :

$$C_k = \sum_{i=1}^{M} log\ E_i \cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2}\right)\right] \quad , \quad 1 \leq k \leq d \tag{3.18}$$

et qui permet d'obtenir des coefficients peu corrélés.

Le coefficient C_0 qui est la somme des énergies n'est pas utilisé ; il est éventuellement remplacé par le logarithme de l'énergie totale E calculée dans le domaine temporel et normalisée [19].

$$E = \text{Log} \sum_{n=1}^{N} x_n^2 \tag{3.19}$$

$$E_{NOR} = E - E_{max} \tag{3.20}$$

3.3. MODELISATION ACOUSTIQUE DU LOCUTEUR

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage, ou de reconnaissance. Elle repose en général sur un modèle. Celui-ci possède un ensemble de paramètres numériques dans le but de lui faire correspondre le signal analysé. Pour ce faire, on met en œuvre un algorithme d'analyse, qui cherche généralement à minimiser la différence, appelée erreur de modélisation, entre le signal original et celui qui serait produit par le modèle [20].

3.3.1. Approche vectorielle

Elle consiste à représenter un locuteur par un ensemble de vecteurs issus directement de la phase de paramétrisation. Cette approche comporte deux techniques principales [21].

3.3.1.1. L'Alignement Temporel Dynamique

L'alignement temporel dynamique **DTW** (Dynamic Time Warping) est un modèle basé sur le calcul d'une distance entre deux vecteurs. Principalement, il fait la comparaison d'une séquence de vecteurs avec une autre séquence de vecteurs par le calcul de la distance accumulée entre ces deux séquences. Si les deux séquences sont identiques alors le chemin entre eux est diagonal, et par conséquent, la distance qui les sépare est minimale. Cette méthode est utilisée souvent dans les systèmes de reconnaissance automatique du locuteur

dépendant du texte. Elle est efficace pour la reconnaissance mono-locuteur à petit vocabulaire et en mots isolés [22].

3.3.1.2. Quantification Vectorielle

La **Q**uantification **V**ectorielle **QV** s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques. Cela consiste à faire un partitionnement de cet espace en régions, qui seront représentées par leur vecteur centroïde. Pour déterminer la distance d'un vecteur acoustique à cet espace, on effectue une mesure de distance avec chacun des centroïdes des régions et on retient la distance minimale. Si le vecteur acoustique provient du même locuteur pour lequel on a établi le dictionnaire de quantification, la distorsion sera en général moins grande que si ce vecteur provient d'un autre locuteur. Ainsi, on va représenter un locuteur par son dictionnaire de quantification [23].

3.3.2. Approche statistique

Une approche qui repose essentiellement sur des fondements mathématiques (probabilité et statistique). L'objet de cette approche est de décrire les formes à partir d'un modèle probabiliste simple à utiliser et de regrouper les formes dans des classes.

3.3.2.1. Modèle de Markov Caché

Le **M**odèle de **M**arkov **C**aché **MMC** permet de modéliser l'évolution temporelle du signal vocal. Ils sont basés sur une théorie probabiliste. Un **MMC** est caractérisé par :

- les états initiaux du système c'est-a-dire la probabilité d'être dans un état particulier du système au temps 0 ;
- la matrice de transition entre états qui représente la probabilité de transition pour aller d'un état à un autre ;
- la distribution de probabilité d'émission représente la probabilité qu'un état du système ait génère une observation particulière

Le **MMC** donne pendant la phase de reconnaissance, le chemin le plus probable des états et donc des classes sonores à chaque instant. Un algorithme efficace, largement utilisé, permet la détermination de la séquence d'état. Il est basé sur la programmation dynamique et est appelé algorithme de Viterbi.

Dans l'exemple de la figure 3.6, le modèle de Markov à trois états possède des transitions gauche-droite, sans retour arrière possible, afin de représenter l'évolution acoustique d'un phonème au cours du temps : le début du phonème, sa partie centrale et la fin du phonème.

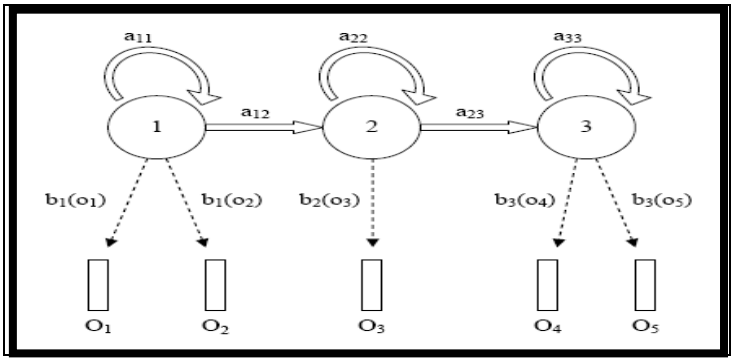

Fig. 3.6 : HMM gauche-droite à trois états [24]

La réalisation d'un processus de Markov se traduit par l'existence d'une séquence Q = (q_0, …, q_T) d'états. Le processus d'émission de ce modèle associe a Q une séquence de T observations O = (o_1, …, o_T). Avant le début du processus, le système se trouve dans un état initial q_0 sans émettre d'observations. Au temps t, le HMM effectue une transition vers l'état q_t et émet l'observation o_t [24].

3.3.2.2. Les Mélanges de Gaussiennes

La reconnaissance du locuteur par mélange de lois gaussiennes **GMM** (**G**aussien **M**ixture **M**odels) consiste à modéliser le signal d'un locuteur par une somme pondérée de composantes gaussiennes. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèles modélise bien les caractéristiques spectrales des voix des locuteurs, et il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes sont considérés comme un cas particulier des **HMM** et une extension de la quantification vectorielle soit un locuteur s et un vecteur acoustique x de dimension D, le mélange de gaussienne est définit comme suit :

$$P(x|\lambda_s) = \sum_{m=1}^{M} \pi_m^s \, b_m^s(x) \qquad (3.21)$$

Où les

$b_m^s(x)$ représentent des densités gaussiennes paramétrées par un vecteur de moyenne μ_m^s et $\sum_m^s$ une matrice de covariance ;

π_m^s représentent les poids du mélange, avec $\sum_m^s \pi_m^s = 1$ (3.22)

un locuteur est donc modélisé par un ensemble de paramètres noté $\lambda_s = \{ \pi_m^s, \pi_m^s, \sum_m^s \}$;

pour m= 1 ,2,……,M (figure 3.7).

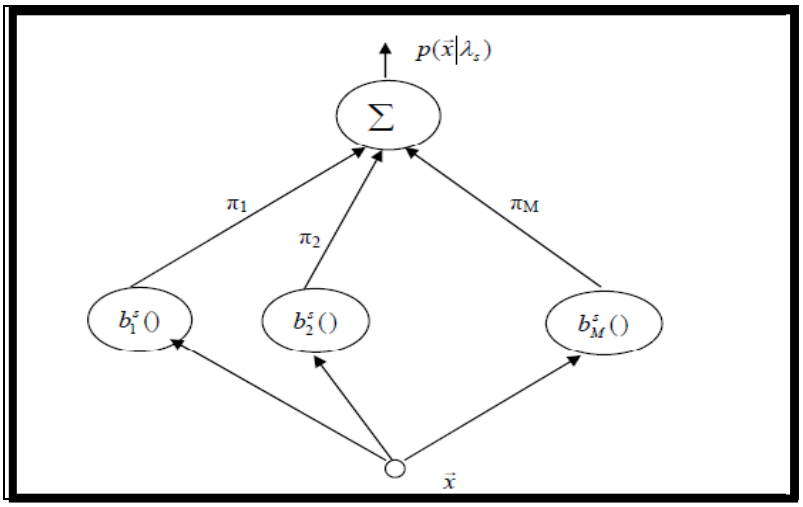

Fig. 3.7 : Modèle de GMM [25]

Le modèle GMM peut prendre pusieurs formes, notament en ce qu concerne les matrices de covariances. On peut utiliser une matrice de covariance poue chaque gaussienne, ou bien une matrice de covariance globale commune à toytes les gaussinnes [25].

3.4. CONCLUSION

Dans ce chapitre, nous avons, présenté les différentes méthodes d'analyse acoustique du signal de parole ainsi que les deux approches de modélisation du locuteur.

Chapitre 4 :
Expériences et Résultats

4. 1. INTRODUCTION

Dans ce chapitre, nous présentons les étapes suivies dans l'élaboration du **SIAL** dépendant du texte, par mots isolés, en utilisant la quantification vectorielle et l'algorithme LBG. Ainsi nous décrirons la BD et l'outil, utilisés ainsi que les résultats obtenus.

Pour pouvoir effectuer les expériences citées dans le paragraphe précèdent, nous avons utilisé le logiciel MATLAB, ainsi que l'outil **HTK** (**H**idden Markov Model **T**ool**K**it). Ces derniers nous ont permis de tester le **SIAL** que nous avons étudié.

4.2. DESCRIPTION DE LA BASE DE DONNEES

La **BD** utilisée dans notre travail est constituée par des enregistrements de trois corpus différents (tableau 4.1), prononcés par quatre locuteurs algériens : deux hommes et deux femmes, âgés en moyenne de 22 ans.

Chaque locuteur a prononcé les différents mots des corpus, constituant la base d'apprentissage. Pour la phase test nous avons choisi d'enregistrer la voix du locuteur en temps réel, et ceci en utilisant uniquement le corpus numéro 3 où chaque locuteur doit prononcer son prénom durant cette phase.

Les enregistrements ont été effectués sous le format WAV (dans un studio de l'Institut Supérieur des Métiers des Arts du Spectacle et de l'Audio Visuel Bordj-El-Kiffan - Alger), avec une fréquence d'échantillonnage de 48 KHz.

Tableau 4.1 : Corpus des mots isolés enregistrés

Corpus	**Mots isolés prononcés**
Corpus 1	Avance
	Recule
	Tourne à droite
	Tourne à gauche
Corpus 2	Un
	Deux
	Trois
	Quatre

Corpus 2 (suite)	Cinq
	Six
	Sept
	Huit
	Neuf
	dix
Corpus 3	Amine
	Faïza
	Mossaab
	Sofia

4.3. MATERIELS UTILISES

Nous avons procédé à l'enregistrement dans un studio de l'ISMAS. Celui-ci est constitué de deux cabines : cabine speaker et cabine technique (figure 4.1).

Fig. 4.1 : Studio d'enregistrement

Nous avons utilisé comme matériel hardware un **microphone électrodynamique** de marque **Beyerdynamic M69TG** (figure 4.2), et software, la station ProTools LE version 8 (figure 4.3), qui est une station audionumérique (en Anglais : **DAW**, pour **D**igital **A**udio

Workstation). Pro Tools est utilisée par une grande partie de l'industrie de la production sonore. On la trouve dans des domaines aussi variés que l'enregistrement et le mixage musical, la post production audio film et télévision, le montage son, la création et l'illustration sonore, la création et la composition musicale, etc.

Fig. 4.2 : Microphone électrodynamique

Fig. 4.3 : Station ProTools

4.4. OUTILS UTILISES

Le SIAL est construit par deux outils : MATLAB et la plate-forme HTK avec une paramétrisation du signal par les coefficients cepstraux de type MFCC.

- MATLAB: version 7.8.0.347(R2009a);
- Le Hidden Markov Model Toolkit (HTK) est un outil gratuit pour l'analyse des séries temporelles. Il utilise la modélisation par Chaînes de Markov Cachées, et est essentiellement - mais non exclusivement - employé pour la RAP.

4.5. DESCRIPTION DE L'APPLICATION SIAL

Concernant l'outil MATLAB, nous avons utilisé 4 fonctions réparties en 4 fichiers.m : train.m, test.m, vqlbg.m, disteu.m. Nous décrivons les deux fonctions principales qui sont :

- le fichier train.m : est une fonction d'apprentissage dans laquelle on :

  - charge les fichiers sonores des 4 locuteurs en utilisant la fonction wavread ;
  - paramétrise le signal parole de chaque locuteur a l'aide de l'outil HTK, en initialisant préalablement les paramètres de ce dernier pour le calcul des MFCC ;

- calcule la Quantification Vectorielle par l'algorithme LBG, avec comme entrées les MFCC et le nombre de centroïdes. Cette tâche est effectuée en appelant la fonction vqlbg.m ;

- le fichier test.m : est une fonction qui teste le système de Reconnaissance Automatique du Locuteur. Elle est organisée comme suit :
  - chargement du fichier sonore du locuteur inconnu a l'aide de la fonction wavrecord ;
  - paramétrisation du signal parole du locuteur a l'aide de l'outil HTK, en initialisant préalablement les paramètres de ce dernier pour le calcul des MFCC ;
  - calcul de la QV comme décrit précédemment dans la dernière étape du fichier train.m.
  - calcul de la distance euclidienne pour la mesure de ressemblance entre les paramètres du locuteur inconnu et celle de la BD, cette tache est réalisée par fonction disteu.m.

4.6. DESCRIPTION DE L'INTERFACE

Pour une meilleure présentation des résultats obtenus, nous avons réalisé une interface (figure 4.4) avec pour principaux composants :

- la barre de titre ;
- le menu : avec deux boutons quitter et recommencer ;
- la zone de dessin : où s'affiche la photo du locuteur identifié ;
- les zones de textes : pour aider le locuteur durant l'enregistrement de sa voix et se termine par l'affichage du résultat désiré (figure 4.5).

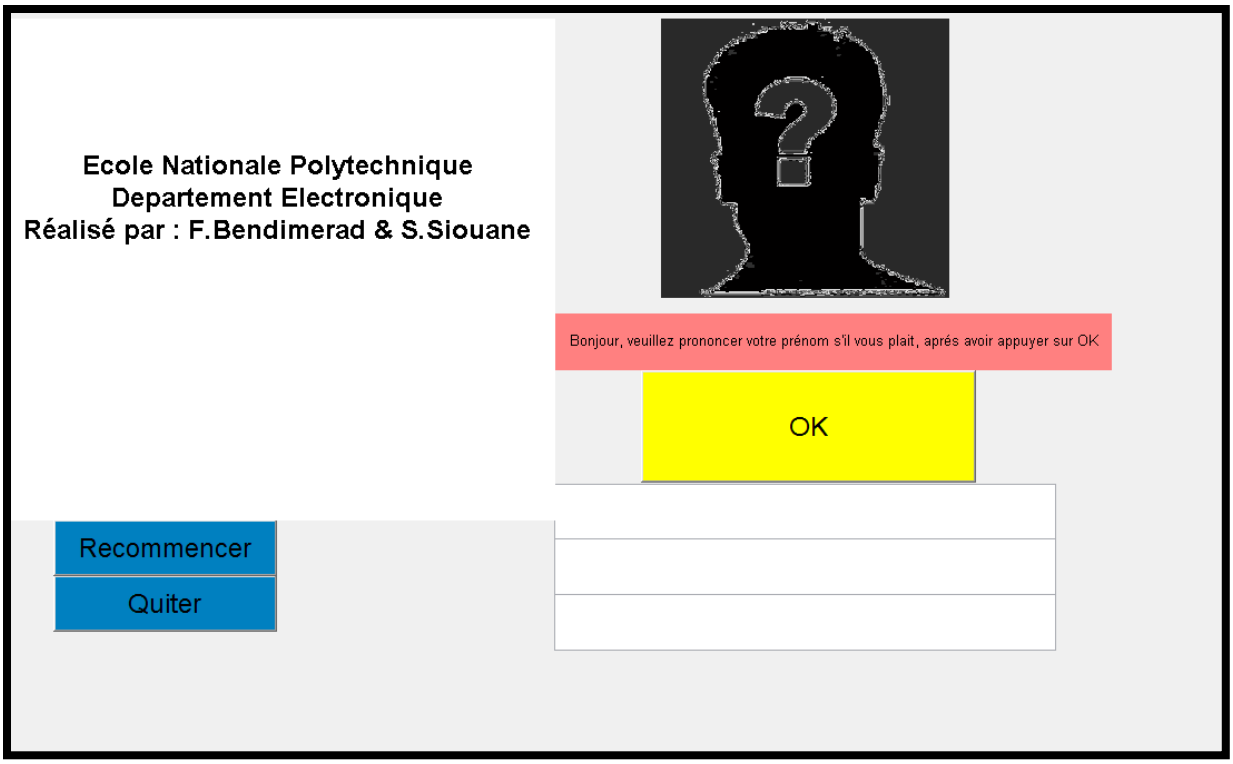

Fig. 4.4 : Interface initiale du SIAL

.

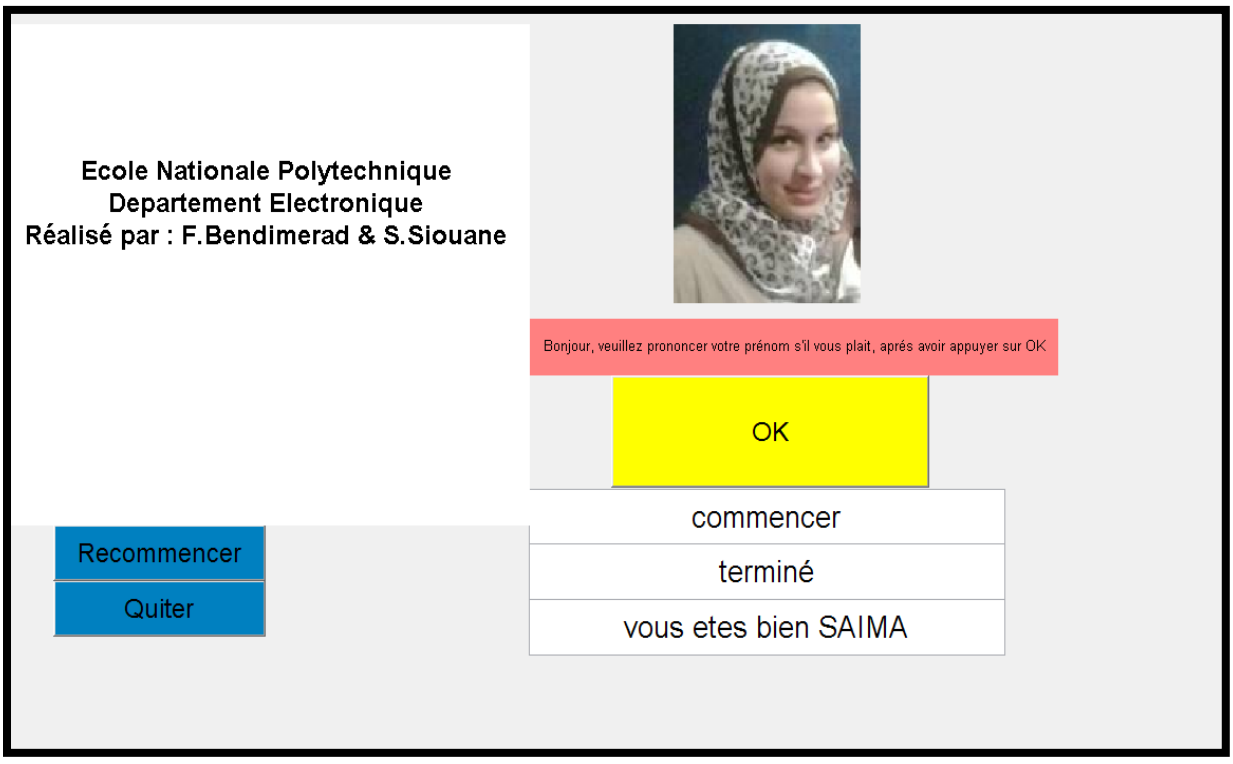

Fig. 4.5 : Interface du SIAL avec l'affichage des résultats

4.7. ETAPES DE REALISATION DU SYSTEME

Les étapes que nous avons suivies pour aboutir à un système de reconnaissance de locuteur sont les suivantes (figure 4.6) :

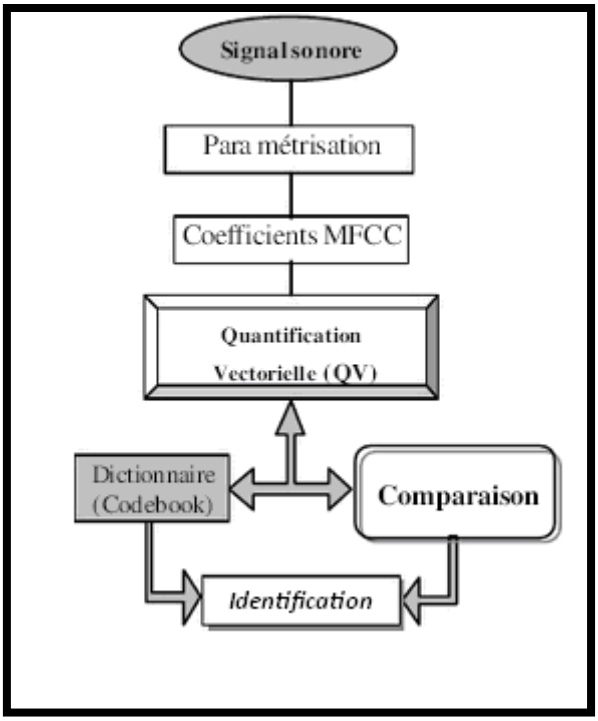

Fig. 4.6 : Etapes suivies lors de la réalisation du SIAL

4.7.1. Signal sonore

Les enregistrements ont été faits au sein de l'institut SMAS à Alger.

4.7.2. Paramétrisation

La paramétrisation du signal aux vecteurs acoustiques, le signal audio est analysé par fenêtre glissante. Les fenêtres d'analyse sont espacées régulièrement et se chevauchent. Leur durée est d'environ 20 ms, ce qui fournit une résolution fréquentielle et une résolution temporelle suffisantes pour la parole. Leur analyse fournit des paramètres statiques, qui sont des analyses temps - fréquence fenêtre par fenêtre, et des paramètres dynamiques, qui sont le plus souvent leurs dérivées temporelles et font intervenir des trames successives. Les paramètres de chaque fenêtre sont ensuite regroupés en un vecteur acoustique, aussi appelé trame.

4.7.3. Extraction des MFCC

Comme nous l'avons exposé dans le chapitre 3, les MFCC sont les paramètres pertinents qui différencient nos locuteurs (tableau 4.2), cette tâche a été réalisée, par l'outil HTK.

Tableau 4.2 : Paramètres utilisés pour l'extraction des MFCC

paramètres	valeurs
Tw : durée d'analyse par trame	20 (ms)
Ts : décalage par trame	10 (ms)
Alpha : coefficient de préaccentuation	0.97
M : nombre de banc de filtre sur l'échelle de Mel	20
C : nombres des MFCC	12
LF : fréquence minimale	0 (Hz)
HF : fréquence maximale	Fs/2 (Hz)

4.7.4. Quantification Vectorielle

La quantification scalaire consiste à coder des échantillons qui sont représentés par une valeur. La QV code de manière efficace des échantillons représentés par plusieurs valeurs (ou vecteurs). Imaginons qu'on ait un ensemble d'échantillons, chacun représenté par un couple de valeur (x_1, x_2).Ces valeurs sont réelles et nous cherchons à les quantifier (figure 4.7).

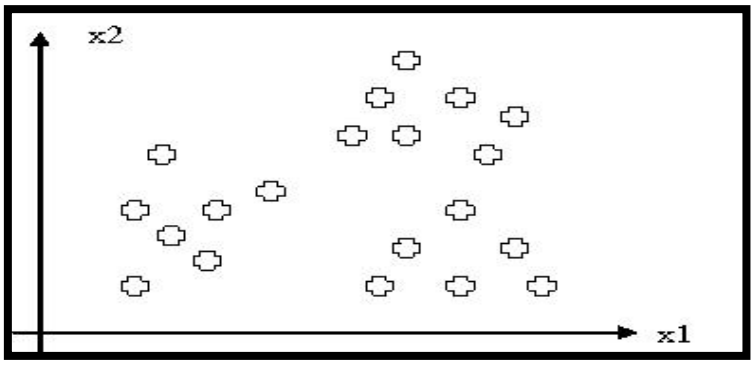

Fig. 4.7 : Echantillons du signal de parole à quantifier [26]

On peut faire une quantification uniforme sur chaque dimension (figure 7.8), mais cela risque de ne pas être optimisé.

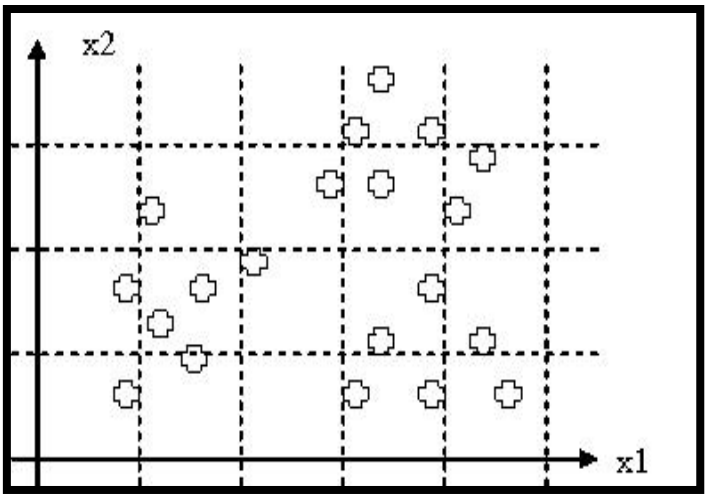

Fig. 4.8 : Quantification uniforme des échantillons [26]

La QV propose une quantification optimisée : l'espace est divisé en classes adaptées à l'ensemble des échantillons et on calcule un représentant pour chaque classe (élément rouge sur la figure 4.9). L'ensemble des représentants est appelé dictionnaire. Pour quantifier un échantillon (x_1, x_2), on lui attribue les valeurs du représentant le plus proche.

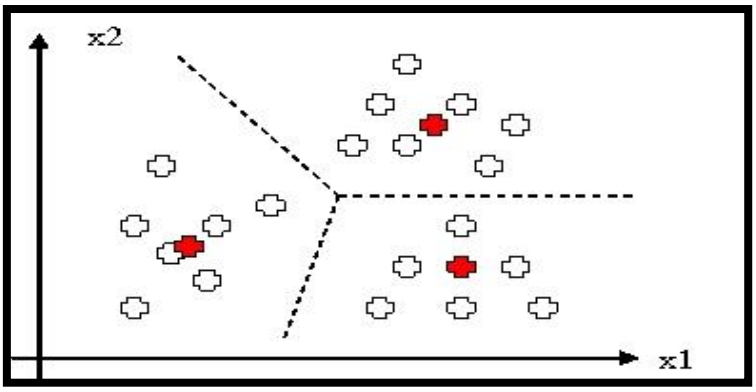

Fig. 4.9 : Quantification optimisée des échantillons [26]

La quantification n'est pas qu'une simple généralisation du cas scalaire. Elle permet de prendre en compte la corrélation entre les échantillons. De plus, il a été montré par Shannon qu'aucune autre méthode ne pouvait surpasser les performances de la QV. C'est une technique très utilisée en reconnaissance et compression de la parole, codage d'image.

La QV consiste donc à coder non plus un échantillon, mais un groupe d'échantillons, ou vecteur. Notons K sa dimension (K=2 dans l'exemple précédent). A partir du dictionnaire, composé des représentants notés $\hat{X}$, on choisira le meilleur représentant de chaque vecteur à coder au sens d'une certaine distance. Le vecteur X est alors remplacé par son représentant $\hat{X}$.

$$d(X, \hat{X}) = \frac{1}{K}\sum_{k=1}^{K}(x_k - \hat{x}_k)^2 \quad (4.1)$$

Si le codage par QV est simple, la conception d'un dictionnaire est plus compliquée et a donné lieu à de nombreux algorithmes. Un des plus utilisés est l'algorithme LBG (Linde, Buzo et Gray) ou algorithme de Lloyd généralisé. L'algorithme LBG part d'un dictionnaire (que nous verrons plus loin comment initialiser), qu'il cherche à améliorer.

4.7.5. Algorithme LBG

Soit les données suivantes :

C_0 : un dictionnaire initial à M éléments ;

X : l'ensemble des L vecteurs d'apprentissage ;

d : une mesure de distorsion ;

n = 0

L'algorithme LBG consiste à :

- trouver une partition de X à partir du dictionnaire C_n. Chaque partition est une cellule de Voronoï : à chaque vecteur de X, on associe le meilleur représentant dans le dictionnaire ;
- calculer la distorsion moyenne pour tous les éléments de X, notée d_n

$$d_n = \frac{1}{L}\sum_{l=1}^{L} d(X_l, \hat{X}_l) \quad (4.2)$$

- si la distorsion est telle que $(d_{n-1} - d_n)/ d_n < \varepsilon$, alors C_n est le dictionnaire souhaité ;
- recalculer les centres de chaque cellule de Voronoï pour obtenir C_{n+1} ;
- n = n+1, retour en 1.
- on prend pour d la **MSE** (**M**ean **S**quare **E**rror) ou moyenne des écarts au carré.

4.7.5.1. Initialisation du dictionnaire pour l'algorithme LBG

L'algorithme LBG repose sur une bonne initialisation du dictionnaire. Différentes méthodes sont possibles :

- *codes aléatoires*: les M premiers vecteurs d'apprentissage forment le dictionnaire initial ;

- *division récursive des données* : il s'agit de prendre le barycentre des vecteurs, de diviser l'ensemble, et ainsi itérativement jusqu'à ce que le nombre de vecteurs désiré (M) soit atteint;
- *assemblage des paires des plus proches voisins (Pairwise Nearest Neighbour (PNN) Clustering)* : il s'agit de fusionner les vecteurs d'apprentissage les plus proches jusqu'à ce que le nombre de vecteurs désiré (M) soit atteint (voir détails ci-dessous) [26].

4.7.6. La Distance Euclidienne

Pour analyser des données, nous commençons par choisir les caractéristiques des objets que nous voulons analyser en les plaçant dans un espace de représentation. Ensuite, nous devons nous doter d'outils métriques permettant de mesurer des distances (des ressemblances, des dissemblances, etc.) entre les dits objets. Ainsi, dans un espace métrique, la distance entre deux vecteurs notés x et y doit satisfaire à l'expression suivante :

$$\begin{cases} d(x,y) \geq 0 \\ d(x,y) = d(y,x) \\ d(x,u) \leq d(x,u) + d(u,y) \end{cases} \quad (4.3)$$

u : vecteur

La distance la plus connue est la **D**istance **E**uclidienne **DE**. La mesure de cette distance dans un espace métrique est très utilisée pour les données quantitatives. Dans l'espace défini par j variables, la distance Euclidienne entre les vecteurs x et y est [11]:

$$d^2(x,y) = \sum_j (x_j - y_j)^2 \quad (4.4)$$

Dans ce travail, nous faisons une comparaison entre le vecteur du signal de parole prononcé par le locuteur et ceux enregistrés dans la BD. On dit qu'un vecteur x est plus proche d'un vecteur y, si la distance euclidienne entre x et y est la plus petite comparée à celle des autres vecteurs. On prend donc la distance minimale calculée.

Notre programme considère x et y comme étant deux matrices dont chaque colonne est une donnée vectorielle, le calcul de la distance euclidienne se fait entre les colonnes des matrices deux à deux.

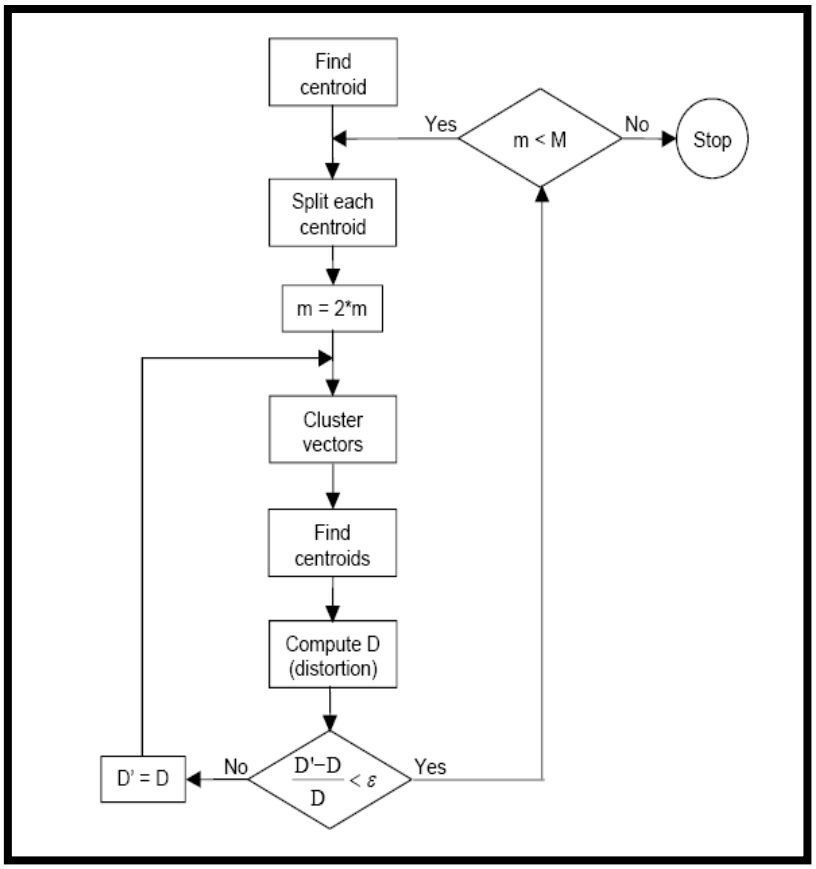

Fig. 4.10 : Organigramme de l'algorithme LBG [27]

4.8. EVALUATION DES PERFORMANCES

L'identification qui est un procédé permettant de déterminer l'identité d'une personne, ne comprend qu'une étape. Son évaluation se fait donc uniquement par calcul du **T**aux de **F**ausse **A**cceptation (TFA), contrairement à un système de vérification qui nécessite le calcul du **T**aux de **F**aux **R**ejet (TFR).

- **Fausse Acceptation** : Événement ayant lieu lorsqu'un système biométrique accepte une personne alors qu'elle n'est pas dans sa base d'utilisateurs. Cet événement doit être le plus rare possible pour assurer la sécurité d'un système biométrique.
- **Faux Rejet** : Événement ayant lieu lorsqu'un système biométrique refuse une personne alors qu'elle est dans sa base d'utilisateurs. Cet événement est souvent dû à une mauvaise acquisition des données biométriques et est perçu comme une gène par l'utilisateur.
- **TFA** - Taux de fausse acceptation : Indique la probabilité qu'un utilisateur inconnu soit identifié comme étant un utilisateur connu. Ce taux définit la sécurité du système biométrique.

Le Taux de fausse acceptation est égale au nombre de fausses acceptations divisé par le nombre de tests imposteur de la base (N_i).

$$TFA(\tau) = \frac{FA(\tau)}{N_i} \quad (4.5)$$

- **TFR -** Taux de faux rejet : Indique la probabilité qu'un utilisateur connu soit rejeté par le système biométrique. Ce taux définit en partie le confort d'utilisation du système biométrique.

$$TFR(\tau) = \frac{FR(\tau)}{N_c} \quad (4.6)$$

Le Taux de faux rejet est égale au nombre de faux rejets divisé par le nombre de tests cible dans la base (N_c) [28]

Le taux d'erreur de décision sont dépendant su seuil de décision fixé dans le module de décision et sont en générale en fonction du seuil (τ).

Lors des essaies effectué, nous avons obtenu un taux de reconnaissance de 95%.

4.9. CONCLUSION

Dans ce chapitre nous avons présenté les expériences et les résultats du système de reconnaissance automatique du locuteur avec la BD et ses performances.

Conclusions générales et perspectives

5. Conclusions générales et perspectives

Dans notre travail, nous avons traité le problème de l'IAL en mode dépendant du texte avec des mots isolés. Il s'agit d'extraire les vecteurs acoustiques, à partir des signaux de paroles enregistrés par les locuteurs de la BD, qui servent à la phase d'apprentissage des modèles représentants chaque locuteur. Pour l'analyse paramétrique nous avons utilisé les MFCC avec comme approche de modélisation la Quantification Vectorielle.

La première partie de ce travail a été consacrée à l'extraction des MFCC, réalisée avec l'outil HTK, la deuxième partie pour l'IAL réalisée avec MATLAB.

Les résultats obtenus indiquent un taux de reconnaissance de 95%. Ce taux de réussite élevé est lié au nombre restreint de locuteur dans la BD.

Comme perspectives à ce travail nous proposons :

- d'utiliser une paramétrisation acoustique basé sur les MFCC, constitué cette fois-ci des premières et deuxièmes dérivées ΔMFCC et ΔΔMFCC, pour améliorer la robustesse du système ;
- la mise au point d'une BD plus riche à enregistrer en milieu ambiant et par une centaine de locuteur.

Références Bibliographiques

6. Références Bibliographiques

[1] O. Godin, Chapitre 5-Analyse de la parole IMN317, Université de Sherbrooke/Canada, 23 Novembre 2011.

[2] Traitement de la Parole Cours 2: Signal de parole, Université de Fribourg suisse, 27/03/2006.

[3] T.En-najjary, conversion de la voix pour la synthèse de la parole, thèse de doctorat de l'université de Rennes 1, France, Mars 2005.

[4] Calliope. La parole et son traitement automatique. Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pages, 1989.

[5] S.Djeghiour, application des réseaux de neurones à la synthèse de la parole en arabe standard, Ecole Normale Supérieure des Sciences Humaines, Alger/Algérie

[6] http://www.iframsurdite.com/thematique.html

[7] L.Buniet, traitement automatique de la parole en milieu bruité : étude des modèles connexionnistes statiques et dynamiques, thèse de doctorat de l'université Henri Poincaré, Nancy 1, France, Février 1997.

[8] T. Dutoit, Introduction au traitement automatique de la parole, faculté polytechnique de Mons, 2000.

[9] S. Ouamour, Indexation automatique des documents audio en vue d'une classification par locuteurs-Application a l'archivage des émissions TV et Radio-, thèse de doctorat, Ecole Nationale Polytechnique, Alger/Algérie, 2009.

[11] WWW.Wikipedia.com

[12] O. KENAI, Authentification Vocale Criminalistique d'un locuteur arabophone.

Mémoire de magister, USTHB, Alger, 2011. 71 pages.

[13] L.R. Rabiner & R.W. Shafer, "Digital Processing of Speech Signal"», Printice Hall, Ing, Englewoodchiffes, New Jersey, 1978.

[14] J.P.Haton, J.M.Pierrel, G.Perennou, J.Caelen, J.L.Gauvain, 1991, Reconnaissance automatique de la parole, DUNOD.

[**15]** J.Makhoul, Linear prediction: A tutorial review, Proc, IEEE, vol. 63, pp. 561-580.

[16] B.S. Atal & S.L. Hanauer, "Speech analysis and synthesis by linear prediction of speech wave", J. Acous. Soc. Am. VOL. 50(2), pp. 637-655. 1971.

[17] M.Bouchamekh, Identification du locuteur indépendante du contexte, mémoire de Magister, ENP.

[18] J.M.Pierrel, 1982, utilisation de contraintes linguistiques en compréhension automatique de la parole continue : Le système MYRTILLE II.

[19] H.Takhedmit & N.Ait Saadi, 2005, « Identification du locuteur en mode indépendant du texte », Projet de fin d'étude à l'ENP, Dép. d'Electronique.

[20] http://tcts.fpms.ac.be/cours/1005-07-08/speech/parole.pdf.

[21] B. Tounsi, Inférence d"identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte algérien, Thèse de Magister, Institut National de Formation en Informatique (i.n.i) oued-smar, Alger, Algérie, 2008

[22]T. Matsui, S. Furui, Comparaison of tex-independent speaker recognition methods using VQ- distorsion and discrete /contiuous hmms, IEE international conference on acoustics, speech and signal processing, ICASSP, vol 2, pp. 157-160, San Francisco, CA, USA, 1992

[23] Décodage acoustico-phonétique et applications à l'indexation audio automatique, En vue de l'obtention du DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE Délivré par l'Université Toulouse III – Paul Sabatier Discipline ou spécialité : Informatique Présentée et soutenue par Olivier Le Blouch, Le 12 Juin 2009.

[24] D.A REYNOLDS, « An Overview of Automatic Speaker Recognition Technology », IEEE 2002.

[25] D.A REYNOLDS, T.F QUAIERI, and R.B DUNN « Speaker Verification Using Adapted Gaussian Mixture Models » M.I.T. Lincolon Laboratory 2000.

[26] http://carolinepetitjean.free.fr

[27] http://www.ifp.illinois.edu

[28] http://www.biometrix21.com/content/7-biometrie