الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DR. YAHIA FARES DE MEDEA

FACULTE DE TECHNOLOGIE

Département du Génie Électrique



Mémoire de Fin d'Études

Présenté pour l'obtention du Diplôme
de Master en Génie Éléctrique

Option : Systèmes de télécommunications

THEME

LA ROBUSTESSE DES SYSTEMES DE RECONNAISSANCE DU LOCUTEUR DANS UNE COMMUNICATION MOBILE BASEE SUR L'ESPACE DES I-VECTEURS

Proposé et dirigé par,

Dr. KROBBA Ahmed

Dr. ALIMOHAND Abdennour

Préparé et présenté par,

Melle. BENALI Selma

Melle. BENADDA Widad

Année Universitaire 2016-2017

Remerciements

Avant tout, nous remercions ALLAH tout puissant de nous avoir donné la volonté, le courage et la patience pour terminer ce travail.

Notre profond respect et nos sincères remerciements à Monsieur A.KROBBA. Enseignant à l'université de l'USTHB, pour nous avoir fait confiance, pour avoir accepter de diriger la réalisation de ce mémoire et pour son soutien inconditionnel.

Je tiens à remercier Monsieur A.ALIMOHAND. Enseignant à l'université de Médéa, qui nous a accordé sa confiance et accepter de co-diriger ce travail.

Nous adressons nos respectueux remerciements à tous les membres du jury qui nous ont fait l'honneur de participer à l'évaluation de notre travail.

Nos vifs remerciements s'adressent également à tous les enseignants qui ont contribué à notre formation tout au long de nos études.

Nos sentiments de reconnaissance et nos remerciements vont également à l'encontre de toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédicace

Toutes les lettres ne sauraient trouver les mots qu'il faut…
Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect et la reconnaissance…
Aussi, C'est tout simplement que, je dédie ce mémoire:
A mes très chers Parents…
*Chère **Mama**, tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement, tu n'as pas cessé de m'encourager, toi et mon cher **Papa***
Vos prières et bénédictions m'ont été d'un grand secours pour mener à bien mes études.
Aucune dédicace ne saurait être assez éloquente pour exprimer ce que vous méritez, pour tous les sacrifices que vous n'avez cessé de me donner depuis ma naissance, durant mon enfance et même à l'âge adulte.
Que ce travail, que je vous dédie, en témoignage de mon profond amour, reflète ma profonde affection et ma grande reconnaissance,
Puisse Allah, le tout Puissant, vous préserver et vous accorder santé, longue vie et bonheur, pour que vous demeuriez le flambeau illuminant mon chemin.
*A la mémoire de **mes grandes mères** et mes chers **grands-pères…***
Le destin ne nous a pas laissé le temps pour jouir de ce bonheur ensemble et de vous exprimer tout mon respect,
Puisse Allah tout Puissant vous accorder sa clémence, sa miséricorde et vous accueillir dans son vaste paradis.
*A mes très chères soeurs: **Linda**, **Hayet, Houda, Anissa***
Vous êtes merveilleux des étoiles brillantes, vous êtes la joie et le bonheur, qu'Allah vous garde et vous réalise tous vos rêves et vos souhaits.
*A tous mes oncles, mes tantes, mes cousins et cousines : **Manel, Silya, Sonia et Chahinez***
En témoignage de l'attachement, de l'amour et de l'affection que je porte pour vous. Je vous dédie ce travail avec tous mes voeux de bonheur, de santé et de réussite.
*A tous **mes Professeurs***
Je saisis cette occasion pour vous exprimer mes respects et ma considération.
*A mon cher binôme **Widad***
Je suis très contente de t'avoir connue et travaillée avec toi durant cette année, qu'Allah te réalise tes voeux et te protège…
A mes meilleurs amis : Sarah, Widad, Lina, Imene
A toutes mes amies et collègues de promotion et d'études
A tous ceux ou celles qui me sont chers et que j'ai omis involontairement de citer, je vous dédie ce modeste travail…
A tous ceux qui me connaissent et qui m'aiment de près ou de loin...

Selma

Dédicace

A cœur vaillant rien d'impossible
Quand il y a la soif d'apprendre, tous viennent à point à qui sait attendre
Les études sont avant tout, notre unique et seul atout
Souhaitant que le fruit de nos efforts fournis jour et nuit, nous mène vers le bonheur fleuri
Je dédie ce mémoire à :

A mes très chers Parents

Autant de phrases et d'expressions aussi éloquentes soient-elles ne sauraient exprimer ma gratitude et ma reconnaissance. Vous avez su m'inculquer le sens de la responsabilité, l'optimisme et la confiance en soi face aux difficultés de la vie. Vos conseils ont toujours guidé mes pas vers la réussite. Vos patiences sans fin, vos compréhensions et vos encouragements sont pour moi le soutien dispensable que vous avez toujours su m'apporter. Je vous dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester votre fierté et ne jamais vous décevoir. Qu'Allah, le tout puissant, vous préserve, vous accorde santé, bonheur, quiétude de l'esprit et vous protège de tout mal.

A mes très chers frères : ***Mohamed*** *et* ***Islam***
Ma chère sœur : ***Madina***

Merci d'être toujours à mes côtés, par votre présence, par votre amour dévoué et votre tendresse, pour donner du goût et du sens à ma vie. En témoignage de mon amour et de ma grande affection, je vous prie de trouver dans ce travail l'expression de mon estime et mon sincère attachement. Je prie Allah, le tout puissant, pour qu'il vous donne bonheur et prospérité.

A mes chers grands-parents

Que ce modeste travail, soit l'expression des vœux que vous n'avez cessé de formuler dans vos prières. Qu'Allah vous préserve santé et longue vie.

A tous les membres des familles : ***Benadda*** *et* ***bentchikou****.*
Mes oncles, mes tantes ainsi que mes cousins et mes cousines,
Puisse ce travail témoigner de ma profonde affection et de ma sincère estime.

A tous ***mes enseignants*** *et toutes* ***mes enseignantes***
Je saisis cette occasion pour vous exprimer mes respects et ma considération.

A mon cher binôme ***Selma***
Je suis très contente de t'avoir connue et travaillée avec toi durant cette année, tu étais plus qu'une soeur pour moi, qu'Allah te réalise tes vœux et te protège…

A mes meilleurs amis : Sarah, Selma, Souhila, Iman, Lina.
A toutes mes amies et collègues de promotion et d'études
A tous ceux dont l'oubli du nom n'est pas celui du cœur……

Widad

ملخص

يندرج هذا العمل في ميدان التشخيص الأوتوماتيكي للمتكلم الذي يهدف إلى التعرف على الشخص من خلال تحليل صوته في الاتصالات المتنقلة.
أنظمة التعرف الأوتوماتيكي للمتكلم تستند عادة على المعاملات PLPs و MFCCs للتمثيل الصوتي للإشارة الصوتية.
حساسية هذه المعاملات لتغيرات الاشارة حفز الباحثين على استخدام عدة معايير ونماذج جديدة لجعل النماذج الصوتية أكثر قوة.

في هذا العمل، قمنا باستعمال معامل جديد للتمثيل البيان الصوت وهو "i-vecteurs" والقيام بدراسة مقارنة بين النظامين "PLPs-i vecteur" و "MFCCs-i vecteur" و استنتاج التمثيل النموذجي الأمثل في الاتصالات المتنقلة.

كلمات مفتاحية:

التشخيص الأوتوماتيكي للمتكلم، i-vecteurs, PLPs-i vecteur, MFCCs-i vecteur , البيئة المتنقلة

Résumé:

Ce travail s'inscrit dans le domaine de la reconnaissance automatique du locuteur dont l'objectif est de reconnaitre une personne par l'analyse de sa voix en communication mobile. Les systèmes de reconnaissance automatique du locuteur utilisent généralement les paramètres MFCCs et PLPs comme représentation acoustiques du signal de parole, ces paramètres restent très sensibles aux variations du signal dans les milieux bruits. La sensibilité des paramètres à la variabilité du signal à motiver les chercheurs à utiliser de nouveau paramètres pour rendre les modèles acoustiques plus robustes.
Dans ce travail, nous avons utilisé un nouveau concept de représentation de l'énoncé de voix, désigne sous le terme de "i-vecteurs". Nous faisons une étude comparative sur la performance des systèmes de RAL en "MFCCs-i vecteur" et "PLPs-i vecteur" dans le but de trouver les paramètres les plus robustes en communication mobile.

Mots clés :

Reconnaissance Automatique du Locuteur (RAL), I-vecteur, MFCCs-i vecteur, PLPs-i vecteur, Environnement mobile.

Abstract:

In this work, we focus on automatic speaker recognition for mobile application. The main objective is to recognize a person based on his voice through a mobile communication. Speaker recognition is widely used MFCCs and PLPs as features extractions, these features very sensitive to addition noise. In the fact, we used new a features extraction called "i-vecteurs". We made a comparative study between 'MFCCs-i vector" and "PLPs-i vectors" in the goal of robust features extraction is to improve the performance of speaker recognition system over mobile communication.

Keywords:

Automatic Speaker Recognition, I vectors, MFCCs-I vector, PLPs-I vector, Environnent mobile.

Sommaire

Liste des abréviations

Liste des tableaux

Liste des figures

Introduction générale 1

Chapitre 1 : Généralité sur la reconnaissance automatique du locuteur

1.1. Introduction 3

1.2. L'authentification biométrique 3

1.3. La reconnaissance du locuteur 4

1.3.1. Identification automatique du locuteur 5

1.3.2. Vérification automatique du locuteur 6

1.3.3. L'indexation par locuteur d'un flux audio 7

1.4. Types de la variabilité de la parole 8

1.5. System dépendance et indépendant du texte 9

1.6. L'architecture d'un système RAL 9

1.6.1. La paramétrisation 10

1.6.1.1. Les paramètres prosodiques 11

1.6.1.2. Les paramètres acoustiques 12

1.6.2. Les Approches des modélisations du locuteur 16

1.6.2.1. Modèle de mélange de gaussiennes (GMM) 16

1.6.2.2. Modèle GMM-UBM 17

1.6.2.3. Machine à vecteur de support SVM 18

1.6.2.4. GMM super-vecteur 19

1.6.2.5. Modèle d'analyse conjointe de facteurs (JFA) 19

1.6.2.6. Variabilité total : i-vecteurs 20

1.6.2.7. Analyse discriminante linéaire probabiliste (PLDA) 22

1.6.3. Normalisation des scores 22

1.6.3.1. Z-nrom 23

1.6.3.2. T-norm 23

1.6.3.3. S-norm 23

1.7. Décision et mesure des performances 24

1.8. Conclusion 25

Chapitre 2: L'espace des I-Vecteurs

2.1. Introduction 26

2.2. Développement du système i-vecteurs 26

2.2.1. Extraction i-vecteur 26

2.2.2. Compensation inter-session 30

2.2.3. Méthodes de scoring 33

2.3. Conclusion 40

Chapitre 3: Reconnaissance automatique du locuteur dans la communication mobile

3.1. Introduction 41

3.2. Généralité sur les réseaux mobile 41

3.2.1. La première génération-1G 41

3.2.2. La deuxième génération-GSM 42

3.2.3. La troisième génération-UMTS 43

3.2.4. La quatrième génération-LTE 45

3.3. Codage de la parole 46

3.3.1. Codage de forme d'onde 46

3.3.2. Codage paramétrique 48

3.3.3. Codage hybrides 49

3.4. Les différents codeurs de la parole 50

3.4.1. ACELP 50

3.4.2. Le codeur FR (Full Rate) 51

3.4.3. Le codeur HR (Half Rate) 51

3.4.4. Le codeur EFR 52

3.4.5. Le codeur AMR 52

3.4.6. AMR-WB 53

3.4.7. AMR-NB 53

3.5. Les différents Architecture du système de reconnaissance du locuteur dans la communication mobile 54

3.5.1. La reconnaissance déportée 54

3.5.2. La reconnaissance embarquée 56

3.5.3. La reconnaissance répartie 57

3.6. Conclusion 58

Chapitre 4: Expériences et résultats

4.1. Introduction 60

4.2. Logiciels et langages utilisés dans l'application 60

4.3 Description des bases de données utilisées: 60

4.4. Étapes de la réalisation des systèmes de reconnaissance du locuteur 61

4.5. La structure générale de Système de reconnaissance du locuteur 62

4.6. Extraction des paramètres: 63

4.6.1. Paramétrisation des coefficients MFCCs 63

4.6.2. Paramétrisation des coefficients PLPs 63

4.7 Protocole d'évaluation des résultats 63

4.8. Expériences et résultats: 63

4.8.1. Expériences: 65

Conclusion Générale 71

Références Bibliographiques 72

Liste des abréviations

ADN: Acide Désoxyribonucléique

ADSL: Asymmetric Digital Subscriber Line

AMPS: Advanced Mobile Phone System

ADPCM: Adaptive Differential Pulse Code Modulation

ACELP: Algebraic Code Excited Linear Prediction

ANSI: American National Standards Institute

AMR: Adaptive Multi Rate

AMR-WB: Adaptative Multi-Rate – Wide Band

AMR-NB: Adaptative Multi-Rate – Narrow Band

AWGN: Additive White Gaussian Noise

BSS: Base Station Subsystem

BTS: Base Transceiver Station

BSC: Base Station Controller

CELP: Code Excited Linear Prediction

CVSD: Continuously Variable Slope Delta Modulation

CS-ACELP: Conjugate Structure-Algebraic Code Excited Linear Prediction

CPU: Central Processing Unit

2G: Deuxième Génération-technologies de communication radio mobile

DTW: Dynamic Time Warping

DoD: Department of Defense

DSP: Digital Signal Processor

EM: Expectation Maximisation

ETACS: Extended Total Access communication System

EFR: Enhanced Full Rate

ETSI: European Telecommunication Standards Institute

EER: Equal Error Rate

FDMA: Fréquence Division Multiple Access

FR: Full Rate

GMM-UBM: Gaussian Mixture Model -Universal Background Model

GMM: Gaussian Mixture Model

G-PLDA: Gaussian Probabilistic Linear Discriminant Analysis

GSM: Global System of Mobile communication

GPRS: General Packet Radio Service

GSN: GPRS Support Node

HR: Half Rate

HT-PLDA: heavy-tailed distribution

HSX: Harmonic Stochastic Excitation

HTK: Hidden Markov Model Toolkit

IAL: Identification Automatique du Locuteur

ITU-T: International Telecommunications Union – Telecommunications

JFA: Joint Factor Analysis

LPCC: Linear Prediction Cepstral Coefficient

LDA: Linear Discriminant Analysis

LD-CELP: Low Delay-CELP

LTE: Long Term Evolution

LSP: Line Spectrum Pairs

MFCC: Mel Frequency Cepstral Coefficient

MAP: Maximum A Posteriori

ML: Maximum Likelihood

MSC: Mobile Switching Center

MOS: Mean Opinion Score

MMS: MultiMedia Messaging Service

MBE: Multi Band Excited

MELP: Mixed Excitation Linear Prediction

MPE-LPC: Multi Pulse Excited- Linear Prediction Coefficients

NSS: Network Switching Subsystem

NIST: National Institute of Standards and Technology

OSS: Operating Subsystem

OTAN: Organisation du Traité de l'Atlantique Nord

PLP: Perceptual Linear Prediction

PPCA: Probabilistic Principal Component Analysis

PLDA: Analyse discriminante linéaire probabiliste

PCM: Pulse Code Modulation

PSTN: Public Switched Téléphone Network

RAL: Reconnaissance Automatique du Locuteur

RPE: Regular pulse excitation

RPE-LTP: Regular pulse excitation-long term prediction-linear predictive

RNC: Radio Network Controller

$\boldsymbol{E_b/N_0}$**:** Rapport Signal sur Bruit

SVM: Support Vector Machine

SAC: Similarité Angulaire De Cosinus

SBC: Sub-Band Coding

STC: Sinusoidal Transform Coding

SVQ: Split Vector Quantization

TACS: Total Acces Communication System

TDMA: Time Division Multiple Accès

3GPP: Third Generation Partnership Project

UBM: Universal Background Model

UMTS: Universal Mobile Telephone System

VAL: Vérification Automatique du Locuteur

VoIP: Voice over Internet Protocol

VSELP: Vector Sum Excited Lineary Prédiction

VAD: Voice Activity Detection

WCCN: Within Class Covariance Normalization

W-CDMA: Wideband-Code Division Multiple Access

Liste des tableaux

Tab 4.1: Comparaison des taux d'erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec les différents niveaux bruit du canal (AWGN) 65

Tab 4.2: Comparaison des taux d'erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit du canal (Rayleigh) 66

Tab 4.3: Comparaison des taux d'erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit d'environnement (la foule) 68

Tab 4.3: Comparaison des taux d'erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit d'environnement (la voiture) 69

Liste des figures

Fig 1.1: Les principales modalités biométriques 3

Fig 1.2: Différentes tâches du traitement de la parole 4

Fig 1.3: Principe de base de l'identification du locuteur 5

Fig 1.4: Vérification Automatique du locuteur 6

Fig 1.5: Tâche de suivi de locuteurs 7

Fig 1.6: Principe de base de la segmentation en locuteurs 8

Fig 1.7: Structure générale d'un système RAL 10

Fig 1.8: Fréquence fondamentale 12

Fig 1.9: L'extraction des paramètres LPCCs 13

Fig 1.10: Méthode de calcul des coefficients PLPs 14

Fig 1.11: Extraction des paramètres avec MFCCs 15

Fig 1.12: Architecture GMM-UBM d'un système de vérification du locuteur 17

Fig 1.13: concept du GMM super-vecteur 19

Fig 1.14: Schéma approximatif des distributions des scores des locuteurs clients et imposteurs ainsi que le seuil de décision optimale 21

Fig 2.1: Schéma de principe de la reconnaissance du locuteur basé sur i-vecteurs 27

Fig 2.2 : Exemple du calcul de la distance angulaire entre deux points. Les points rouges sont les projections sur la sphère des points bleus 37

Fig.3.1: Structure du réseau GSM simplifiée 42

Fig. 3.2: Structure du réseau UMTS simplifiée 45

Fig 3.3: Schéma simplifié du modèle de synthèse de l'algorithme ACELP 51

Fig. 3.4: L'architecture du système de la reconnaissance déportée 55

Fig. 3.5: L'architecture du système de la reconnaissance embarquée 56

Fig.3.6: L'architecture du système de la reconnaissance répartie 58

Figure 4.1: Structure générale d'un système RAL 62

Figure 4.2: Taux d'erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit du canal AWGN 66

Figure 4.3: Taux d’erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit du canal Rayleigh 67

Figure 4.4: Taux d’erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit d'environnement (la foule) 68

Figure 4.5: Taux d’erreurs (MFCC-I vecteurs et PLP- I vecteurs) avec le bruit d'environnement (la voiture) 69

Introduction générale

Introduction générale

La croissance de l'usage des téléphones mobiles cellulaires a connu un essor remarquable, au cours ces dernières années. L'intégration de technologie vocale dans la communication mobile a créé de nombreux nouveaux défis dans le domaine de reconnaissance du locuteur et la parole telle que : la distorsion de codage de la parole, le canal de transmission, le bruit d'environnement mobile.

La parole est un moyen de communication universel, que la technologie permet aujourd'hui de diffuser, stocker et restituer à une échelle planétaire. La parole peut également servir à caractériser l'identité du locuteur en se basant sur les traits physiologiques et comportementaux véhiculés par le signal vocal. Ceci a donné naissance à une discipline quasi contemporaine dénommée la reconnaissance automatique du locuteur.

La reconnaissance automatique du locuteur (RAL) désigne l'ensemble des procédures automatiques visant à discriminer des locuteurs à partir de leurs énoncés de voix. Elle s'appuie sur la théorie du signal et sur des techniques d'apprentissage automatique. Ces dernières privilégient à la définition de règles la validation statistique. Des représentations numériques du signal de parole, ainsi que des méthodes de détection, sont élaborées afin d'évaluer des hypothèses sur l'identité d'un locuteur présumé. Elle se divise par deux tâches : l'identification des locuteurs et la vérification des locuteurs. L'identification des locuteurs est de trouver l'identité du locuteur à partir d'un énoncé de discours, il existe deux types dans l'identification des locuteurs, l'un dépend du texte et l'autre indépendant du texte. D'autre part, la vérification des locuteurs permet de déterminer si le locuteur est bien celui qu'il ou elle fait valoir.

Plusieurs progrès et méthodes en termes de performances et de robustesse des systèmes de la reconnaissance du locuteur dans une communication mobile. Le tout récent progrès, voire même le plus important, est la proposition de la représentation de la parole par des vecteurs à faible dimension, dénommés les i-vecteurs qui s'appuie sur le paradigme de modélisation par mélange de gaussiennes et présente la particularité de se réduire numériquement à un vecteur de dimension faible, au regard des représentations précédentes, et pourtant très discriminant vis à vis du locuteur. De par sa dimensionnalité modérée (typiquement dans les centaines) et

sa richesse en matière d'information utile, l'espace des i-vecteurs a énormément aidé les chercheurs en reconnaissance du locuteur, et même en d'autres domaines connexes.

Dans le présent travail, nous nous sommes intéressées à la robustesse des systèmes de reconnaissance du locuteur dans une communication mobile basée sur l'espace des i-vecteurs.

Le présent mémoire, se subdivise en quatre chapitre :

- Le premier chapitre est consacré aux principes de la reconnaissance automatique du locuteur (RAL) et des approches de modélisations utilisés dans ce domaine ainsi que l'exploration des différentes tâches de la reconnaissance du locuteur. On explore alors les outils nécessaires à la reconnaissance automatique du locuteur comme paramétrisation, les approches de modélisation du locuteur et la mesure des performances de chaque tâche.

- Le deuxième chapitre, donne la définition de l'espace des i-vecteurs ainsi que la procédure détaillée de leurs extractions seront fournies.

- Le troisième chapitre, est dédié à la présentation d'une façon générale les différentes générations de réseaux mobiles. Ainsi que les différentes techniques de codage du signal, les normes de codeurs de la parole et les différentes architectures du système de reconnaissance automatique du locuteur dans la communication mobile.

- Le quatrième chapitre, présent une étude expérimentale portera les résultats obtenus avec les deux systèmes mis au point, le premier système à basé des coefficients PLPs, le deuxième à basé des coefficients MFCCs.

Finalement on termine par une conclusion générale retraçant les différentes étapes réalisées, résumant les résultats obtenus et soulignant les perspectives qui peuvent envisagées dans la future clôture ce mémoire.

Chapitre I

Généralités sur la reconnaissance automatique du locuteur

1.1. Introduction :

Ce chapitre est consacré aux principes de la reconnaissance automatique du locuteur (RAL) et des approches de modélisations utilisés dans ce domaine. Nous présentons dans un premier temps quelques notions concernant l'authentification biométrique ainsi que les différentes tâches de la reconnaissance du locuteur qui peuvent être définies selon l'application visée. Nous décrivons aussi l'architecture générale d'un système RAL (paramétrisation et les approches de modélisation du locuteur). Enfin, nous donnons la mesure des performances de chaque tâche.

1.2. L'authentification biométrique :

La biométrie est une nouvelle technique, qui consiste à vérifier ou déterminer l'identité d'une personne à partir de ses caractéristiques biologiques (comme l'ADN), comportementales (comme la voix) ou morphologiques (comme l'empreinte digitale) [1].

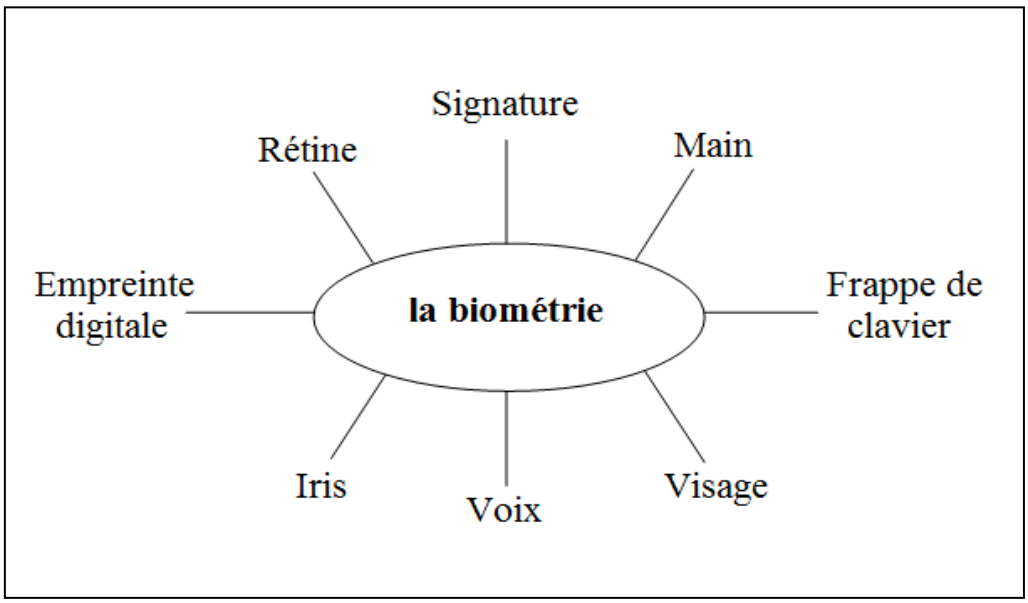


Fig 1.1 : Les principales modalités biométriques.

Les caractéristiques biométriques par lesquelles il est possible de vérifier l'identité d'une personne sont appelées modalités biométriques. La figure 1.1 illustre un exemple de quelques modalités biométriques. Ces modalités sont basées sur l'analyse des données liées à l'individu et sont généralement classées en trois catégories : La biométrie biologique se base sur

l'analyse des données biologiques liées à l'individu (salive, ADN, etc.). La biométrie comportementale se base sur l'analyse de comportements d'un individu (manière de marcher, dynamique de frappe au clavier, la voix, etc.). La biométrie morphologique se base sur les traits physiques particuliers qui, pour toutes personnes, sont permanents et uniques (empreinte digitale, visage, etc.) [1].

1.3. La reconnaissance du locuteur :

La reconnaissance automatique du Locuteur –RAL- s'inscrit dans le domaine du traitement de la parole. Elle consiste à reconnaitre l'identité d'une personne par l'analyse de sa voix [2].

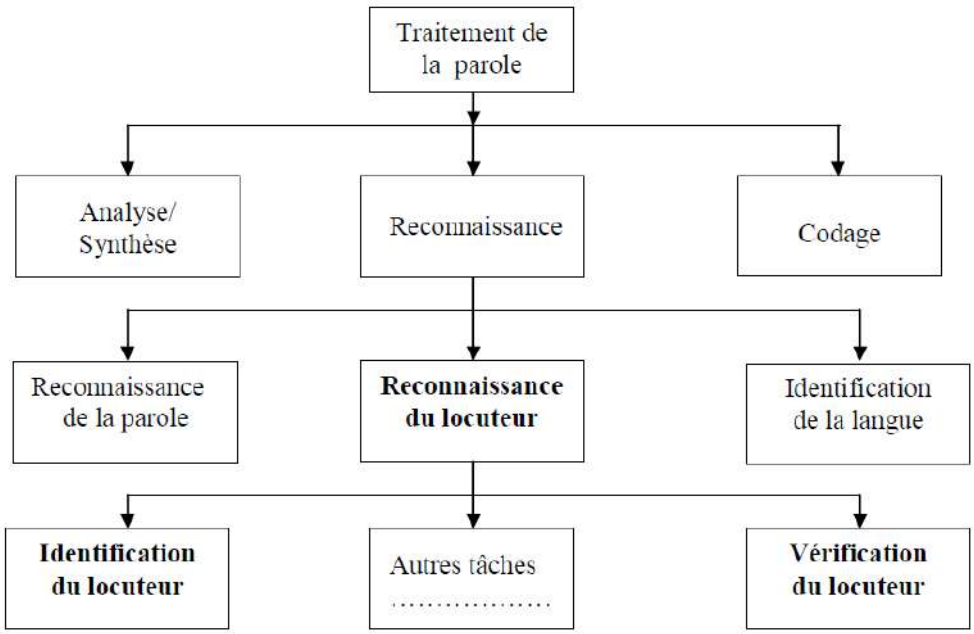


Fig 1.2 : Différentes tâches du traitement de la parole [2].

La parole recouvre également une vaste panoplie d'activités tels que : le codage, la compression, la synthèse et la reconnaissance de la parole, la reconnaissance de locuteur. La reconnaissance du locuteur est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs sont regroupés en trois catégories principales [3] :

- l'identification de locuteurs.

- la vérification du locuteur.
- l'indexation par locuteur d'un flux audio.

1.3.1. Identification automatique du locuteur :

Le principe de l'identification automatique du locuteur l'IAL, consiste à retrouver l'identité du locuteur associé parmi un ensemble de locuteurs connus. D'un point de vue schématique, 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. L'identité du locuteur dont la référence est la plus "proche" de la séquence de parole est donnée en sortie du système d'IAL [4].

Deux modes sont proposés en IAL : l'identification en ensemble fermé, dont on suppose que la séquence de parole est effectivement prononcée par un locuteur connu et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode "ensemble ouvert", le système doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée [4]. De par son principe - déterminer une identité parmi des identités potentielles – les performances des systèmes d'IAL se dégradent à mesure que la population de locuteurs augmente [5].

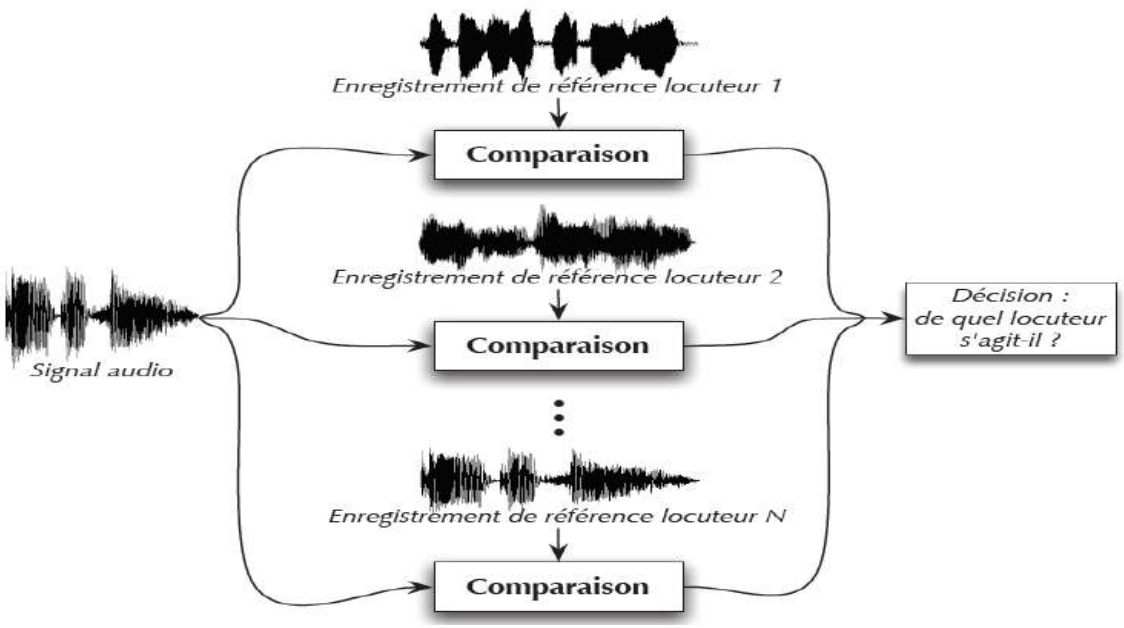


Fig 1.3: Principe de base de l'identification du locuteur [6].

- **Applications :**

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de reconnaissance automatique de la parole (RAP). Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles [4].

1.3.2. Vérification automatique du locuteur :

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu dont représente le principe de VAL. 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 basée sur le rapport de vraisemblance 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 supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et sera rejeté [7].

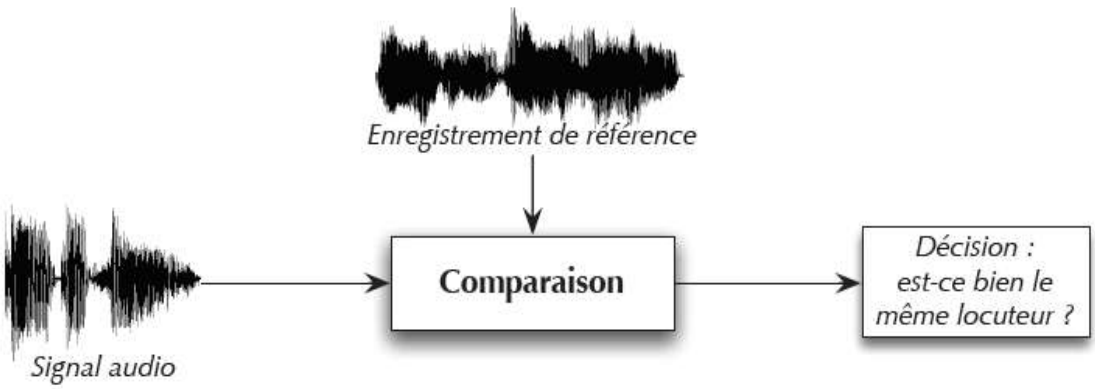


Fig 1.4: Vérification Automatique du locuteur [7].

- Applications :

Les applications de VAL sont multiples et principalement commerciales [7]:

- Serrures vocales pour le contrôle d'accès à des locaux.
- Authentification pour l'accès à distance.
- Protection de matériel contre le vol (téléphones portables, voitures, etc.)

1.3.3. L'indexation par locuteur d'un flux audio:

Le principe est, à partir de l'enregistrement de référence d'un locuteur, de déterminer si ce locuteur est présent au sein d'un enregistrement audio, par exemple une conversation [7].

a) Suivi de locuteur :

Le suivi de locuteur consiste à trouver les limites des interventions du locuteur qu'on a recherché au sein du document audio. Il s'agit donc de déterminer si ce locuteur intervient et si oui, quand. La figure 1.5 donne une illustration de ce principe [6].

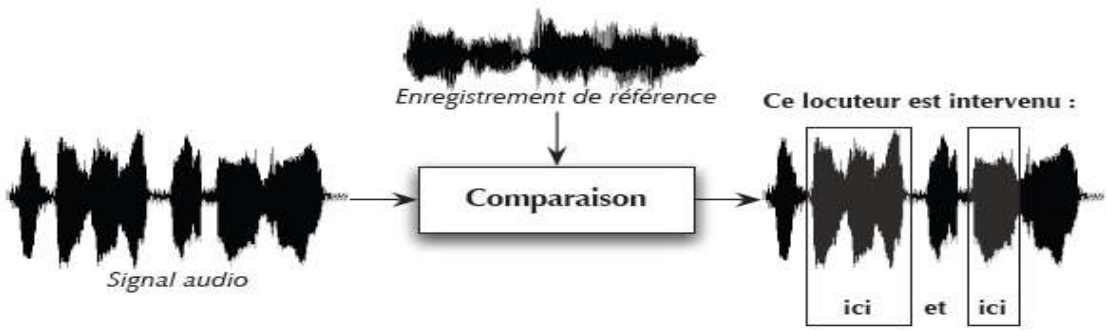


Fig 1.5: Tâche de suivi de locuteurs [6].

b) Segmentation en locuteurs :

C'est la détermination du nombre de locuteurs présents dans un document audio tout en délimitant leurs interventions. La complication de cette tache résulte du traitement de documents pour lesquels peu ou pas d'informations sont connues a priori. Notamment, pas

d'information n'est disponible à la primitive concernant les locuteurs participant dans le document : ni leur nombre, ni leur identité, ni aucun échantillon de leur voix permettant d'avoir une référence. Toutes ces informations doivent être extraites du document étudié [6].

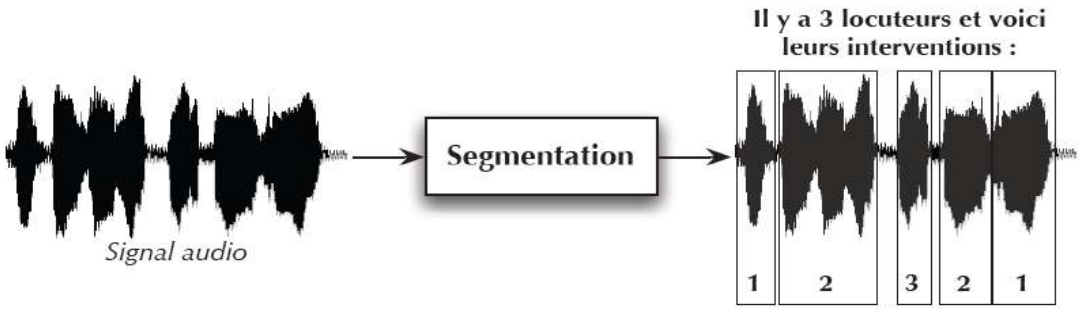


Fig 1.6: Principe de base de la segmentation en locuteurs [6].

- Applications:

Le domaine d'application est la segmentation automatique d'échanges radio entre pilotes et contrôleurs aériens. Depuis, le champ d'application de la segmentation en locuteurs s'est étendu et cette tache se retrouve intégrée dans le cadre plus vaste de l'indexation en locuteurs de bases de données de documents multimédia [6].

1.4. Types de la variabilité de la parole:

Il est important d'introduire les types de variabilités de la parole. En plus du contenu linguistique, le signal vocal véhicule aussi des informations sur : l'identité du locuteur, son genre, son état émotionnel, son état de santé, sa tranche d'âge, etc. Dans une base de données de plusieurs locuteurs où chaque locuteur a au moins un enregistrement, les enregistrements d'un même locuteur peuvent être effectués dans différents lieux (par exemple : au bureau, à la maison, à l'extérieur, dans une chambre insonorisée, etc.) et en utilisant différents types de canaux (par exemple : téléphone fixe, téléphone portable, microphone, etc.). Usuellement, nous pouvons limiter toute cette gamme d'informations ou de variabilités dans ces trois catégories suivantes [8] :

- **La variabilité intra-session** : il s'agit de la variabilité perçue au cours d'un même enregistrement et qui peut être causée par un changement de l'état émotionnel, de lieu, d'emplacement par rapport au microphone ou bien du microphone lui-même, etc.

- **La variabilité inter-session**: il s'agit de la variabilité perçue entre l'ensemble des enregistrements d'un même locuteur, elle est ainsi due aux mêmes causes que l'intra-session, mais elle est susceptible d'être plus importante vu les différences possibles entre les dates et les circonstances d'enregistrement.

- **La variabilité interlocuteurs** : il s'agit de la variabilité perçue entre les distincts locuteurs et due principalement à la différence d'anatomie des cordes vocales, ça pourrait être aussi dû au genre, à l'âge, à la langue.

1.5. System dépendance et indépendant du texte :

Diverses applications reposant sur une même tache peuvent se différencier entre autres par leur degré de dépendance au texte. Les systèmes de RAL dits indépendants du texte si le texte prononcé n'est pas connu à priori. À l'opposé, les systèmes dits dépendants du texte lorsque le texte prononcé par le locuteur est fixé et connu à l'avance. En mode indépendant du texte, le locuteur peut prononcer n'importe quelle phrase pour être reconnu. Les systèmes dépendants du texte donnent généralement de meilleures performances de reconnaissance que les systèmes indépendants du texte car la variabilité due au contenu linguistique de la phrase prononcée est alors neutralisée. Cependant, dans le cas des systèmes basés sur des techniques de reconnaissance de la parole à large vocabulaire, les performances d'identification en mode indépendant du texte ou en mode dépendant du texte sont pratiquement les même [5].

1.6. L'architecture d'un système RAL:

Un système reconnaissance automatique du locuteur tel qu'il illustré par la figure 1.7, se compose de trois processus principaux, qui sont : la paramétrisation (analyse acoustique), la modélisation et la décision. En plus, un système de RAL possède deux modes [9]:

- Apprentissage, où un modèle est estimé pour chaque locuteur « client » du système puis servira de référence pour les tâches de reconnaissance à venir.

- Test, où l'étape de reconnaissance (vérification, identification...) est effectuée. En sortie de cette phase, le système émet une réponse : une identité pour la tâche d'identification ou une décision accès/rejet pour la vérification.

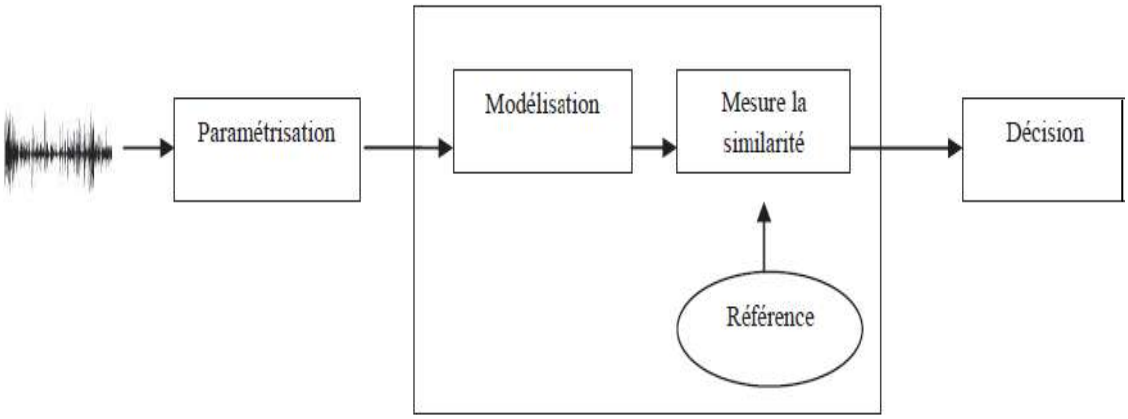


Fig 1.7: Structure générale d'un système RAL [9].

1.6.1. La paramétrisation :

Dans cette étape en décompose le signal parole en trames de longueur variantes de (10 à30ms).car le signal vocal est supposé stationnaire dans cette durée (invariance par changement d'observation). Pour chaque fenêtre on fait L'extraction des paramètres acoustique qui peuvent fournir une représentation plus simple du signal parole. Il existe trois grandes catégories de paramètres [10] :

- **Les paramètres prosodiques:** ces paramètres illustrent le débit, la durée et les fréquences de pauses, pitch et l'énergie [10].

- **Paramètres acoustique :** Ces paramètres sont utilisés en RAL. Ils représentent l'influence des caractéristiques physiques de l'appareil phonatoire de chaque individu. Ces paramètres sont les LPCCs, PLPs et MFCCs [11].

- **Les paramètres dynamiques :** Le vecteur des paramètres issus des paramétrisations précédentes peut être complété par le vecteur correspondant aux dérivées du premier et second ordre de ces paramètres qui sont calculées à partir de plusieurs trames adjacentes. Elles permettent d'introduire une information concernant le contexte temporel d'une trame courante [11].

1.6.1.1. Les paramètres prosodiques :

Les traits ou indices prosodiques d'un signal de parole sont sa fréquence fondamentale (ou pitch), son énergie et son timbre. Le pitch représente la fréquence de vibration des cordes vocales. Cet élément est différent pour la voix d'un homme (entre 60 Hz et 150Hz), la voix d'une femme (aux alentours de 250Hz) ou celle d'un enfant (entre 300Hz et 400Hz). L'énergie d'un son est liée à la pression de l'air en amont du larynx et caractérise son intensité. Le timbre est la caractéristique d'un son permettant de le différencier d'un autre son (c'est la mélodie d'un mot ou d'une phrase) [10].

a) Energie totale :

L'énergie à court terme du signal de la parole fournit une représentation convenable qui reflète ces variations d'amplitude. Elle est calculée à partir de la relation suivante [5]:

$$\overline{E}(m) = \sum_{n=0}^{N-1} x^2(n).w(m-n) \qquad (1.1)$$

Avec : $\overline{E}(m)$: L'énergie moyennes d'une trame "m".

N: La largeur de la fenêtre d'analyse.

$x(n)$: C'est le signal de la trame avec une longueur "N".

$w(m-n)$**:** La fenêtre

b) Fréquence fondamental :

La période du fondamental est par définition la fréquence de vibration des cordes vocales, elle est appelée aussi le pitch. La figure 1.8 représente la fréquence fondamentale [5].

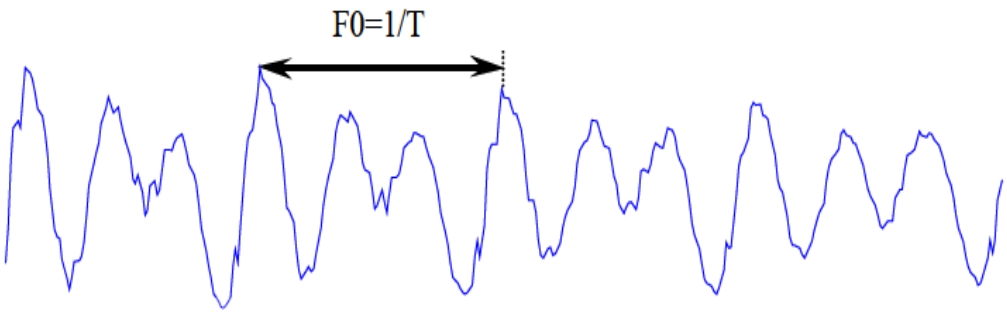


Fig 1.8: Fréquence fondamentale [5].

L'extraction du pitch est une tâche particulièrement difficile pour trois raisons [5]:

- La vibration des cordes vocales n'a pas nécessairement une périodicité complète.
- Il est difficile de séparer le pitch des effets du trait vocal.
- La plage de dynamique de la fréquence fondamentale est très grande.

1.6.1.2. Les Paramètres acoustique :

L'analyse spectrale de la parole présente des avantages au niveau de la perception, car l'oreille humaine effectue ce genre d'analyse. On introduit en traitement du signal parole les outils suivants [11]:

a) Les coefficients LPCCs :

L'extraction des paramètres LPCC se base sur l'analyse LPC qui permet d'extraire les coefficients de prédiction. Les coefficients cepstraux peuvent être calculés à partir de la sortie d'un banc de filtres ou des coefficients de la prédiction linéaire. Ainsi les paramètres LPCC sont calculés à partir d'une modélisation auto régressive du signal [12].

Chaque trame est représentée par des coefficients statiques. Nous avons choisi d'utiliser 12 coefficients LPCC. Le principe de calcul de ces coefficients est présenté par la figure 1.9 [11].

Après l'analyse du signal de parole, la procédure utilisée pour la paramétrisation par la technique LPCC est la suivante [11]:

- Préaccentuation pour aplatir le spectre moyen du signal qui décroît de 6 dB par octave.
- Fenêtrage : Chaque trame est multipliée par une fenêtre de Hamming.
- Extraction des coefficients d'autocorrélation, on note que la fonction d'autocorrélation s'avère pour traîner M, où M est l'ordre de l'analyse désirée de LPC.
- les coefficients de prédiction linéaire calculés par l'algorithme de Levinson-Durbin.

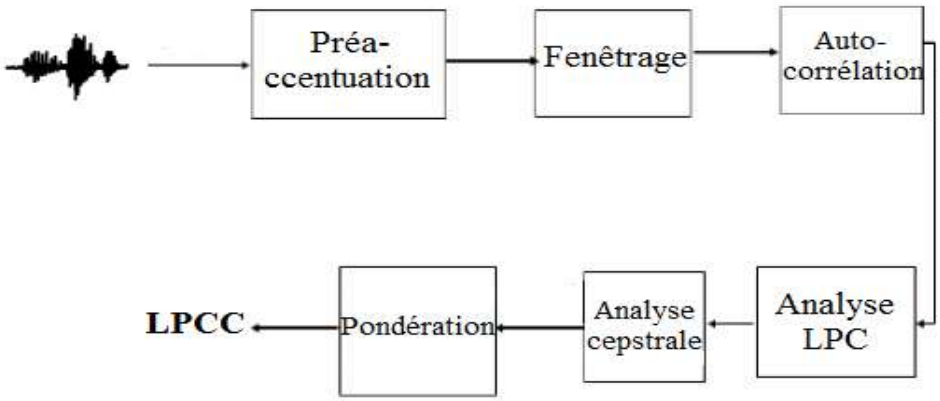


Fig 1.9 : L'extraction des paramètres LPCCs [5].

b) Les coefficients PLPs:

La méthode de coefficients de prédiction à base de notions psycho acoustiques connue sous le nom PLP (Perceptual Linear Prediction ou Perceptually based Linear Prediction), est une méthode inspirée du principe de prédiction linéaire (LPCs). La méthode de calcul de ces paramètres est détaillée dans la Figure 1.10. Tout d'abord, un fenêtrage de Hamming est effectué. Ensuite, l'amplitude spectrale du signal est calculée et suivie d'un filtrage selon

l'échelle Bark. L'échelle Bark, comme l'échelle Mel, reproduit approximativement la sensibilité du système auditif humain. Puis, une pré-accentuation des composantes fréquentielles est appliquée. Ensuite, une accentuation du volume d'intensité sonore est appliquée par élévation à la puissance de 0,33. Un traitement par prédiction linéaire est réalisé sur le spectre obtenu. Enfin, les coefficients cepstraux sont obtenus par récurrence à partir des coefficients de prédiction, ce qui est l'équivalent du log du spectre suivi d'un transformé de Fourier inverse [5].

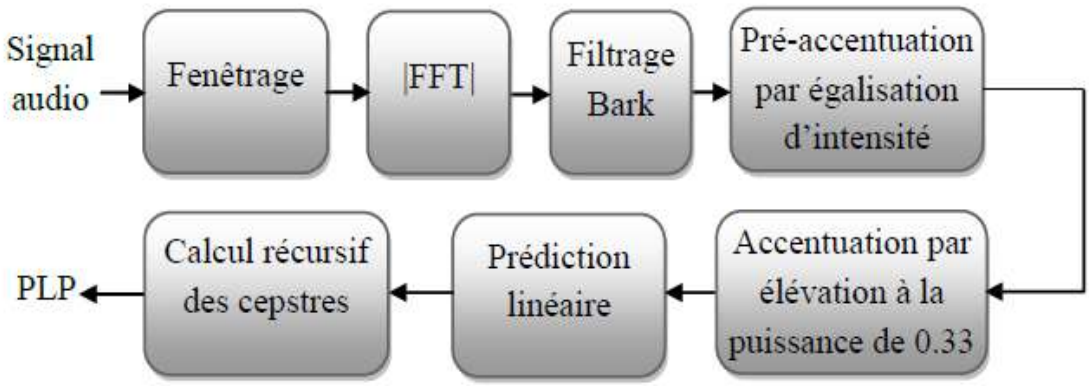


Fig 1.10 : Méthode de calcul des coefficients PLPs [5].

Dans la première étape on a [5]:

- une analyse en bandes critiques selon une échelle de Bark par un banc de filtres.
- une préaccentuation des valeurs obtenues selon une courbe suivant approximativement les mêmes principes que les traitements effectués par l'oreille, avec accentuation des basses fréquences et atténuation des hautes fréquences.

Dans la deuxième étape on a [5]:

- une interpolation des sorties de banc de filtre pour obtenir un spectre sur une échelle fréquentielle auditive.
- une transformée de Fourier inverse.

- une résolution d'un ensemble d'équations linéaires pour obtenir les coefficients issus d'un filtre tout pôle d'ordre 5 ce qui permet d'obtenir au moins deux sommets caractéristiques selon les autours.

Dans cette méthode a pour avantage de permettre une analyse ou un codage de la parole qui respecte le principe de la prédiction linéaire, qui suit l'échelle fréquentielle observable dans l'oreille et enfin, qui réduit l'espace de représentation [5].

c) Les coefficients MFCCs:

Le coefficient MFCC (Mel Frequency Cepstral Coding) est une technique très utilisée en traitement de la parole. Il est basé sur la variation des bandes critiques de l'oreille humaine avec la fréquence, les filtres espacés linéairement aux basses fréquences et logarithmiquement à hautes fréquences. Ces filtres sont modélisés par une échelle non-linéaire issue de connaissances sur la perception humaine : l'échelle Mel [13].

Pour les MFCCs, on utilise la fenêtre de Hamming durant la transformation du domaine temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier. Un filtrage, est appliqué ensuite, par banc de filtres triangulaires qui vont chacun couvrir une fréquence, permettant de mieux simuler le fonctionnement de l'oreille humaine. Par la suite les filtres sont espacés selon l'échelle de Mel. Cette échelle reproduit la sélectivité de l'oreille qui diminue avec l'accroissement de la fréquence. Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel [13].

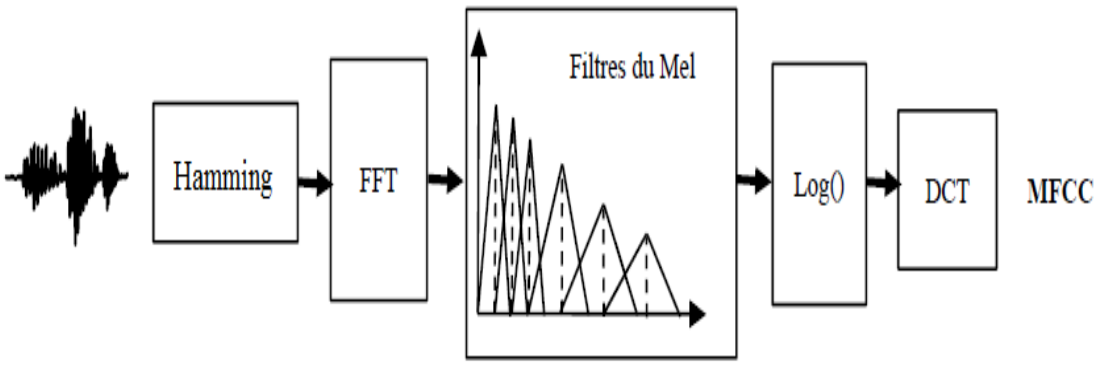


Fig 1.11: Extraction des paramètres avec MFCCs [8].

1.6.2. Les Approches des modélisations du locuteur :

L'étape de modélisation exploite les données fournis dans l'étape de la paramétrisation afin de créer la représentation d'un individu qui servira, par la suite, à l'authentifier. Le modèle utilisé est généralement une représentation statistique des données acquises [9].

Différentes approches ont été développées, néanmoins on peut les classer en quatre grande familles [9] :

- **l'approche vectorielle :** le locuteur est représenté par un ensemble de vecteurs de paramètre dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW (dynamique time warping) et par quantification vectorielles [2].

- **L'approche statistique :** consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustique. Elle couvre les techniques de modélisation par les modèles de Markov cachés, par les mélanges de gaussiennes et par des mesures statistiques du second ordre [2].

- **L'approche connexionniste :** consiste principalement à modéliser les locuteurs par des réseaux de neurones [2].

- **L'approche discriminante :** les modèles des différentes classes sont appris simultanément et l'apprentissage vise à séparer le plus possible les classes entre elles (minimisation du taux d'erreur de classification) [14].

1.6.2.1. Modèle de mélange de gaussiennes (GMM) :

Un modèle de mélange de gaussiennes (GMM, Gaussian Mixture Model) représente la distribution d'une variable aléatoire *X* par [15] :

$$p(X/\Theta) = \sum_{i=1}^{M} w_i N(X; \mu_i; \sum_i) \qquad (1.2)$$

Où w_i est le poids de la i-ème gaussienne tel que $\sum_i w_i = 1, N()$ est la fonction de densité de probabilité de la loi normale et μ_i et $\sum_i$ sont le vecteur moyen et la matrice de covariance

de la gaussienne i .soit $X = (x_1,....x_t,....x_T)$ un ensemble de vecteurs de paramètres acoustiques calculées sur un segment de parole de durée T. Un algorithme itératif de type Expectation-Maximisation (EM) permet l'estimation des paramètres du modèle Θ en maximisant la vraisemblance des échantillons d'apprentissage X [15].

Les GMM sont utilisés d'une manière ou d'une autre dans la plupart des systèmes de reconnaissance de parole et du locuteur. En augmentant le nombre de composants du mélange, il est théoriquement possible de s'approcher d'une densité de probabilité quelconque, mais en pratique le nombre de paramètres du modèle est limité par la quantité des données d'apprentissage disponible. On travaille le plus souvent avec des GMM à matrice de covariance diagonale. D'une part, les différentes dimensions des coefficients cepstraux habituellement utilisés sont faiblement corrélées entre elles. D'autre part, à nombre de paramètres du modèle constant, l'estimation de GMM à matrices de covariance diagonales est plus robuste et les modèles obtenus sont plus performants que dans le cas de matrices de covariance pleines [15].

1.6.2.2. Modèle GMM-UBM :

Un modèle GMM générique de référence, adopte l'architecture GMM-UBM présentée dans la figure 1.12.

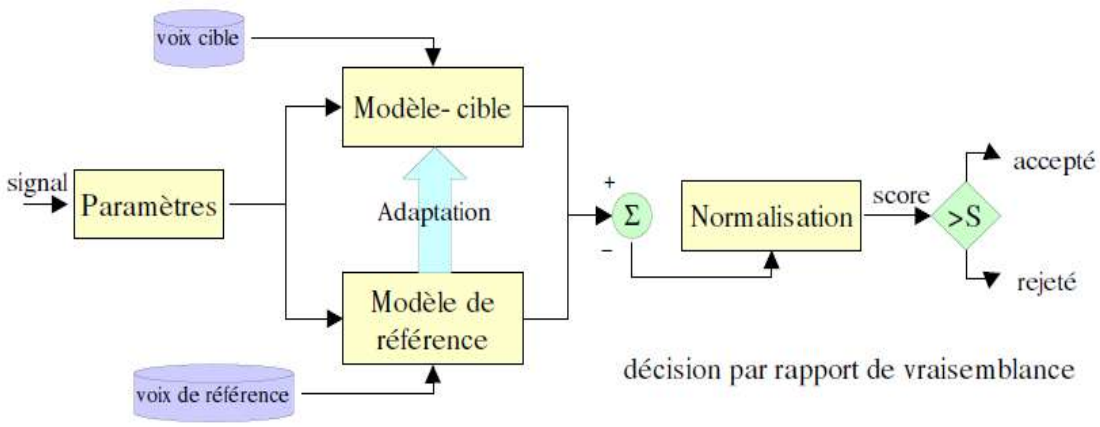


Fig 1.12 : Architecture GMM-UBM d'un système de vérification du locuteur [15].

Un modèle GMM générique de référence, encore appelé UBM (Universal Background Model) ou modèle du monde, est préalablement appris sur un ensemble données d'échantillons de voix le plus représentatif possible. Lors de la phase d'enrôlement, ce modèle UBM est adapté par MAP à la voix du locuteur cible en utilisant les données d'enrôlement. Lors de la phase d'exploitation, les deux hypothèses que l'échantillon *X* ait été généré par le modèle cible *M* ou par le modèle de référence *B* sont comparées au moyen du logarithme du rapport de vraisemblance normalisé par la durée du signal [15]:

$$S(X) = \frac{1}{T}(\log f(X/M) - \log f(X/B)) \tag{1.3}$$

et le score obtenu est comparé à un seuil de décision pour reconnaître l'échantillon présenté comme appartenant au locuteur cible (acceptation) ou le considérer comme inconnu (rejet) [15].

1.6.2.3. Machine à vecteur de support SVM :

Les machines μa support de vecteurs est une nouvelle technique discriminante dans la Théorie de l'apprentissage statistique. Elle permet d'aborder des problèmes très divers comme la classification, la régression, la fusion, etc. [16].

Le principe de cette technique est de projeter les données de l'espace d'entrée (appartenant à deux classes différentes non-linéairement séparables) dans un espace de plus grande dimension appelé espace de caractéristiques. Dans cet espace, on construit un hyperplan optimal séparant les deux classes tel que [16]:

- Les vecteurs appartenant aux différentes classes se trouvent de différentes cotes de l'hyperplan.
- la plus petite distance entre les points et l'hyperplan (la marge) est maximale.

1.6.2.4. GMM super-vecteur :

Un super-vecteur GMM est défini comme la concaténation des composantes des moyennes des gaussiennes du GMM. Si D est la dimension de l'espace du paramètre, la dimension du super vecteur est M*D où M est le nombre de gaussienne dans le GMM [17].

La notion de super-vecteur implique une transformation d'un signal de parole vers un vecteur de plus grande dimension [17].

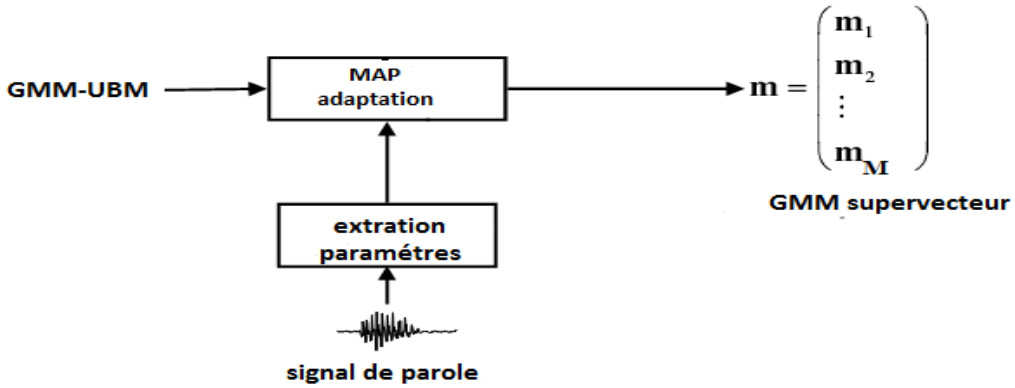


Fig 1.13 : concept du GMM super-vecteur [17].

Étant donné un énoncé de locuteur, la formation GMM-UBM est effectuée par MAP adaptation des moyens, M partir de ce modèle adapté, le GMM super-vecteur se forme. Le processus est illustré à la figure 1.13 [17].

1.6.2.5. Modèle d'analyse conjointe de facteurs (JFA) :

Le modèle d'analyse conjointe de facteurs est un modèle génératif des variables cachées qui tente principalement de s'attaquer au problème de la disparité entre les données d'enrôlement et celles de test dans l'espace des super-vecteurs des GMM. L'hypothèse de base de ce modèle stipule qu'un super-vecteur *S* de dimension *CF* dépendant du locuteur et du canal (c.-à-d. un super-vecteur estimé d'un segment donné d'un locuteur donné) peut se diviser en deux super-vecteurs statistiquement indépendants comme suit [8]:

$$S = s + c \tag{1.4}$$

Où ***s*** est le super-vecteur qui dépend uniquement du locuteur et ***c*** est celui qui dépend du canal. Le super-vecteur ***s*** dépendant du locuteur est donné par la formule suivante [2]:

$$S = m + Vy + Dz \tag{1.5}$$

Où les super-vecteurs moyens m représentés souvent par le super-vecteur du modèle du monde UBM, *V* connue par la matrice des voix propres, c'est une matrice rectangulaire de dimension *CF* × *Rs (avec Rs ≪CF),* elle peut aussi être vue comme une matrice de projection en dimension réduite, elle modélise principalement la variabilité interlocuteur. *D* est une matrice diagonale de dimension *CF* × *CF* qui joue un rôle similaire à celui de l'information a priori dans l'adaptation MAP du modèle GMM-UBM. Les vecteurs cachés *y* et *z* sont respectivement les facteurs du locuteur (speaker factors) et les facteurs communs (common factors). Les vecteurs cachés suivent une distribution normale standard $N = (.,0,1)$ où 0 est un vecteur moyen dont toutes ses composantes sont égales à zéro et *I* est une matrice-identité. Donc, le super-vecteur **s** suit une distribution gaussienne d'un vecteur moyen *m* et d'une matrice de covariance qui est égale à $VV' + D^2$ [8].

Le super-vecteur **c** dépendant du canal est donné par la formule suivante [8]:

$$C = Ux \tag{1.6}$$

où la matrice *U*, dite matrice des canaux propres (eigenchannels), est une matrice rectangulaire de dimension *CF* × *Rc (avec Rc≪CF),* ses colonnes constituent les axes de l'espace du canal de dimension réduite. Le vecteur caché *x* de dimension *Rc* distribué selon une loi normale standard est appelé le vecteur des facteurs du canal (channel factors). Ainsi, le super-vecteur *c* suit une loi normale d'un vecteur moyen nul et d'une matrice de covariance égale à UU' [8].

1.6.2.6. Variabilité total : i-vecteurs:

Le succès de l'architecture GMM-UBM dans le domaine de la reconnaissance du locuteur ainsi que la proposition des super-vecteurs ont donné naissance au modèle JFA. Ce modèle tente de remédier au problème des disparités causées par les effets du canal dans l'espace de très grande dimension des super-vecteurs. À son tour, le modèle JFA a largement préparé le terrain à l'arrivée de la représentation de la parole par des vecteurs à faible dimension dénommés par le terme d'i-vecteurs. La définition la plus élémentaire d'un i-vecteur est de le considérer comme étant la projection d'un super-vecteur dans un espace de dimension réduite. Cet espace, à l'inverse des deux sous-espaces de JFA, ne fait aucune distinction entre les

différents types de variabilité (c.-à-d. la variabilité interlocuteur et celle due au canal), ainsi, il est nommé l'espace de la variabilité totale [8].

Mathématiquement, un super-vecteur *S* d'un segment audio dépendant du locuteur et du canal peut se réécrire sous la forme suivante [8] :

$$S = m + Tx \tag{1.7}$$

où le super-vecteur moyen *m* de dimension *CF* est le super-vecteur de modèle du monde UBM, la matrice rectangulaire *T* de dimension $CF \times D$ *(avec* $D \ll CF$) est nommée la matrice de la variabilité totale (c.-à-d. la variabilité interlocuteur et celle due au canal) et *x* est un vecteur caché de dimension *D* qui suit une distribution normale standard. Ainsi, le super-vecteur *S* suit une distribution normale du vecteur moyen *m* et de la matrice de covariance qui est égale à TT'. Le modèle génératif proposé à Figure 1.14 peut être estimé via un modèle d'analyse en facteurs ou bien via une analyse en composantes principales probabiliste (Probalistic Principal Component Analysis, PPCA) [8].

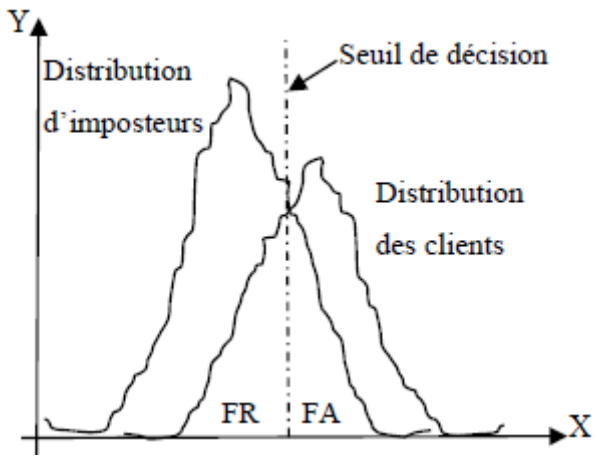


Fig 1.14 : Schéma approximatif des distributions des scores des locuteurs clients et imposteurs ainsi que le seuil de décision optimale [8].

Un système de vérification d'identité biométrique peut faire deux types d'erreur. Une fausse acceptation (FA) se produit lorsqu'un imposteur n'est pas rejeté par le système et un faux rejet (FR) se produit lorsqu'un client est rejeté par le système. Ces deux types d'erreur dépendent du seuil de décision auquel est comparé le score issu du processus de comparaison [8].

Une valeur élevée du seuil de décision tendra à rendre l'accès plus difficile, en augmentant le nombre de faux rejets et diminuant le nombre de fausses acceptations. Inversement, une valeur faible du seuil de décision tendra à faciliter l'accès et donc à augmenter le nombre de fausses acceptations [8].

Il est à noter qu'une estimation ponctuelle (point estimate) du vecteur caché **x** nous donne le i-vecteur [8].

De par sa nature, l'espace de la variabilité totale (c.-à-d. l'espace des i-vecteurs) contient simultanément la variabilité désirable et indésirable à la distinction des locuteurs. Ainsi un mécanisme de filtration est nécessaire pour avoir un système de reconnaissance du locuteur robuste. Dès l'apparition de l'espace des i-vecteurs dans le domaine de la vérification du locuteur, deux méthodes de classification sont émergées dans ce domaine, à savoir l'analyse discriminante linéaire probabiliste (Probabilistic Linear Discriminant Analysis, PLDA) comme un modèle génératif et la similarité angulaire du cosinus (SAC) comme un modèle à base de distance/similarité [8].

1.6.2.7. Analyse discriminante linéaire probabiliste (PLDA) :

Le PLDA est un modèle purement génératif, il n'est en réalité qu'une version miniature du modèle JFA opérant dans le sous-espace des i-vecteurs au lieu de celui des super-vecteurs. Le modèle de base de PLDA est apparu pour la première fois dans le domaine de l'imagerie où il est utilisé pour la reconnaissance faciale. Il est ensuite introduit dans la vérification du locuteur avec deux variantes et quelques modifications dans l'algorithme d'apprentissage [8].

1.6.3. Normalisation des scores :

En observant le processus de vérification, nous constatons que ce dernier est composé d'une suite d'opérations de réduction de dimensionnalité (c.-à-d. de super-vecteurs de grande dimension jusqu'au score scalaire). En fait, chaque étape réduit principalement les dimensions qui représentent la variabilité indésirable. En dépit de la multitude des étapes filtrantes, la variabilité indésirable peut se propager jusqu'aux scores. En conséquence, plusieurs méthodes ont été proposées dans la littérature pour compenser ces variabilités. L'hypothèse principale de toutes les méthodes de normalisation stipule que la distribution des scores des clients et celle des scores des imposteurs suivent des lois normales unidimensionnelles. Dans le but

d'atténuer la variabilité au niveau des scores, les méthodes de normalisation visent à standardiser la distribution des imposteurs de la manière suivante [8] :

$$score^{(norm)} = \frac{score - \mu_{imp}}{\sigma_{imp}} \tag{1.8}$$

Où $score^{(norm)}$ est le score obtenu par la normalisation du score brut en utilisant la moyenne μ_{imp} et la déviation standard σ_{imp} estimées à partir des scores d'une cohorte des modèles d'imposteurs [8].

1.6.3.1. Z-nrom :

Le but de la méthode Z-norm est d'atténuer les biais causés par les différentes qualités d'estimations des modèles de locuteurs. En effet, la qualité d'estimation d'un modèle du locuteur est essentiellement liée à la longueur, au nombre et à la qualité audio des segments d'enrôlement de ce locuteur. Les paramètres de la normalisation dans ce cas sont estimés en dehors de l'étape du test (c.-à-d. un calcul hors-ligne) et ce, en calculant les scores d'un ensemble d'énoncés des imposteurs vis-à-vis le modèle du locuteur cible [8].

1.6.3.2. T-norm :

La normalisation T-norm, quant à elle, vise à réduire la variabilité intersessions propagée jusqu'aux scores. Le mécanisme derrière cette méthode consiste à estimer les paramètres de la normalisation (c.-à-d. la moyenne et la déviation standard) en calculant le score du segment du test vis-à-vis un ensemble de modèles des imposteurs [8].

1.6.3.3. S-norm :

Un des aboutissements dans le domaine de la vérification du locuteur après l'introduction d'espace d'i-vecteur est de rendre symétrique le calcul du score entre un modèle du locuteur cible et un segment de test. De ce fait, une méthode tenant en compte cet aspect est proposée dans la littérature sous l'appellation S-norm. La formule de la normalisation symétrique est donnée comme suit [8] :

$$score^{(norm)} = \frac{score - \mu_e}{\sigma_e} + \frac{score - \mu_t}{\sigma_t} \quad (1.9)$$

Où $\sigma_e, \sigma_t, \mu_t$ et μ_e sont les moyennes et les déviations standards estimées à partir d'un ensemble de scores des segments des imposteurs calculés vis-à-vis le segment d'enrôlement et celui du test [8].

1.7. Décision et mesure des performances:

Un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Un module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système de reconnaissance [9].

Le module de décision décrit dans le paragraphe précédent reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté [9].

- **L'identification du locuteur :** Consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues**.** les performances du système d'identification sont données en termes de taux d'identification correcte Ic ou incorrecte I_i , soit [2] :

$$Ic = \frac{\text{Nombre de tests correctement identifiés}}{\text{Nombre total de tentatives}} \quad (1.10)$$

Et

$$I_i = \frac{\text{Nombre de tests mal identifiés}}{\text{Nombre total de tentatives}} \quad (1.11)$$

Avec $I_C + I_i = 100\%$

- **Vérification du locuteur** : C'est une décision en tout ou rien. les performances de vérification de locuteur sont données en termes des faux rejets f_r et de fausses acceptations f_a [2]:

$$f_r = \frac{\text{Nombre de tentatives d'abonnés rejetées}}{\text{Nombre total de tentatives d'abonnés}} \tag{1.12}$$

$$f_a = \frac{\text{Nombre de tentatives d'imposteurs acceptées}}{\text{Nombre total de tentatives d'imposteurs}} \tag{1.13}$$

1.8. Conclusion :

Dans ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes du système. La reconnaissance automatique du locuteur est probablement la méthode la plus ergonomique pour résoudre les problèmes d'accès. Cependant, la voix ne peut être considérer comme une caractéristique biométrique d'une personne compte tenu de la variabilité intralocuteur. Un système de reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume en trois étapes principales: l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision qui se base sur les deux processus d'identification et /ou de vérification de locuteur, et cela quelle que soit l'application ou la tache visée.

Chapitre II

L'éspace d'i-vecteurs

2.1. Introduction :

Dans ce chapitre, nous nous focalisons sur la représentation de la voix, une étape très importante pour le processus de la reconnaissance du locuteur. En effet, durant ces quatre dernières années, la représentation de la voix par des i-vecteurs est devenue dominante dans le domaine de la reconnaissance du locuteur. Ce chapitre détaille le processus d'extraction des i-vecteurs et les méthodes de scoring.

2.2. Développement du système i-vecteurs :

Des développements récents dans la technologie de reconnaissance de locuteurs ont vu le succès de systèmes basés sur une représentation à faible dimension d'un segment de parole appelé « i-vecteurs » [18].

Un i-vecteur est une représentation compacte d'un super-vecteur du modèle de mélange gaussien (GMM), qui capture la majeure partie de variabilité des vecteurs GMM. La disponibilité de caractéristiques de faible dimension a renforcé l'intérêt de la recherche pour les modèles probabilistes génératifs. Ces techniques visent à décomposer les composantes de variabilité inter-session et de locuteur des i-vecteurs, à estimer leurs distributions et à effectuer une induction sur l'identité du locuteur dans un cadre bayésien [18].

L'approche i-vecteur dans un système de reconnaissance du locuteur peut être divisée en trois étapes: extraction i-vecteur, compensation inter-session et méthodes de scoring [18].

2.2.1. Extraction i-vecteur:

La distribution complexe des vecteurs acoustiques nécessite une modélisation multimodale. En fait, la modélisation par les mélanges de gaussiennes (GMM) est devenue le paradigme le plus répandu dans le domaine de la reconnaissance du locuteur. Le modèle du monde (UBM pour Universal Background Model) est le terme utilisé pour référer au GMM de M composantes gaussiennes entrainé à partir d'une quantité importante de vecteurs acoustiques de la parole de plusieurs locuteurs [8].

Etant donné un modèle GMM-UBM, la représentation d'un segment de voix s'obtient par adaptation de ce modèle aux données contingentes du segment. La méthode utilisée s'appuie sur la notion statistique de Maximum à Posteriori (MAP) pour son application en reconnaissance du locuteur). Le paramètre $\Theta_{UBM} = \{\gamma, \mu, \Sigma\}$ du GMM-UBM est adapté aux nouvelles données contingentes aux trames du segment à représenter, formant un paramètre adapté $\hat{\Theta} = \{\hat{\gamma}, \hat{\mu}, \hat{\Sigma}\}$ [9].

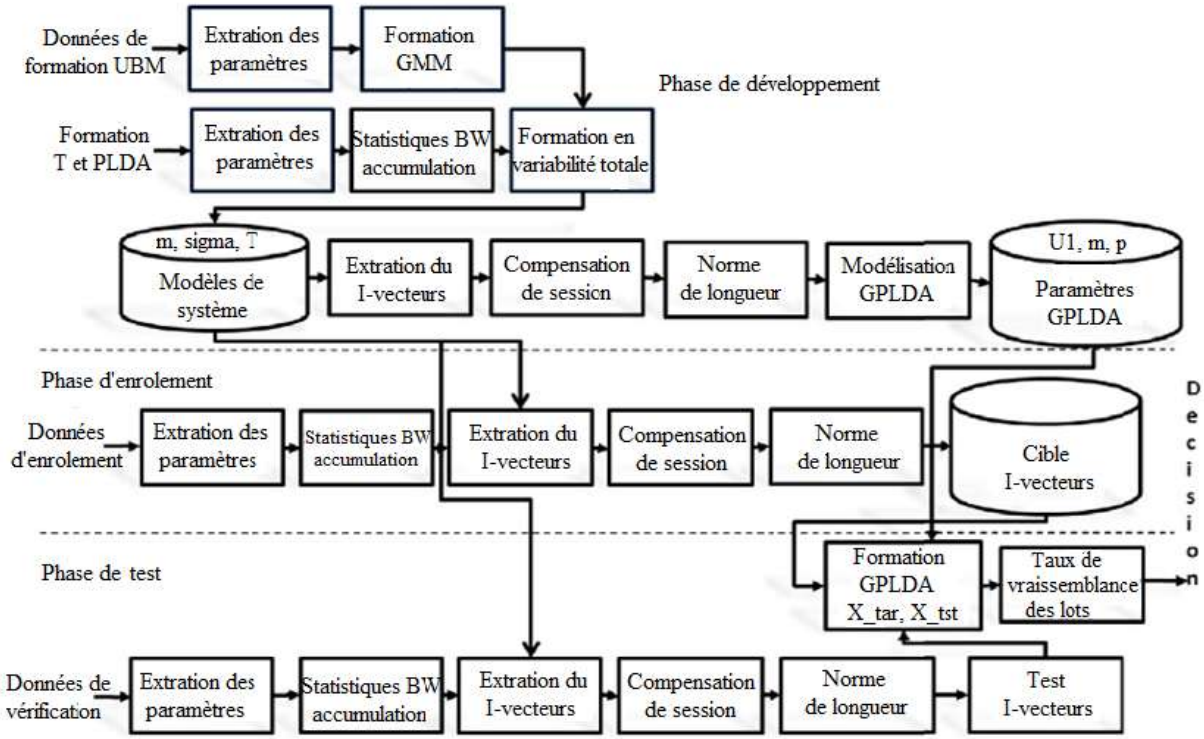


Fig 2.1 : Schéma de principe de la reconnaissance du locuteur basé sur i-vecteurs [19].

L'adaptation selon le critère MAP s'effectue par algorithme EM. A chaque itération *t*, la phase d'expectation met à jour la variable latente *y* par son espérance [9] :

$$\forall x \in X, y_m^{(t)} = \frac{\gamma_m^t f(x|\mu_m^{(t)}, \Sigma_m^t)}{\sum_{m'=1}^{M} \gamma_{m'}^t f(x|\mu_{m'}^{(t)}, \Sigma_{m'}^t)} \quad (2.1)$$

Et celle de maximisation évalue alors (itération t+1) les paramètres $\hat{\gamma}, \hat{\mu}, \hat{\Sigma}$ par maximum de vraisemblance [9]:

$$\gamma_m^{(t+1)} = \frac{n_m}{n_m + \tau_m^{[w]}} \left(\frac{1}{n} \sum y_m^{(t)} \right) + \frac{\tau_m^{[w]}}{n_m + \tau_m^{[w]}} \gamma_m \tag{2.2}$$

Où *n* taille de la collection *X* et $n_m = \sum y_m^{(t)}$

$$\mu_m^{(t+1)} = \frac{n_m}{n_m + \tau_m^{[\mu]}} \left(\frac{\sum y_m^{(t)} x}{\sum y_m^{(t)} x} \right) + \frac{\tau_m^{[\mu]}}{n_m + \tau_m^{[\mu]}} \mu_m \tag{2.3}$$

$$\sum_m^{(t+1)} = \frac{n_m}{n_m + \tau_m^{[\mu]}} \left(\frac{\sum y_m^{(t)} (x - \mu_m^{(t+1)})(x - \mu_m^{(t+1)})}{\sum y_m^{(t)}} \right) + \frac{\tau_m^{[\Sigma]}}{n_m + \tau_m^{[\Sigma]}} \sum_m \tag{2.4}$$

Nous détaillons ici ces formules pour y souligner le rôle des facteurs τ .Qualifiables de facteurs de confiance (relevant factor : facteur latent) ou de pertinence, ils contrôlent le degré d'adaptation de chaque paramètre UBM aux données. En théorie, ils dépendent du paramètre à estimer. En pratique, ils sont confondu, qu'il s'agisse de poids, moyenne ou matrice de covariance, en un seul vecteur $(\tau_m)_{m=1}^M$.Lorsque le nombre τ tend vers l'infini, les estimations $\gamma_m^{(t+1)}, \mu_m^{(t+1)}, \sum_m^{(t+1)}$ tendent vers les paramètres $\gamma_m, \mu_m, \sum_m$ de l'UBM. Lorsque ce nombre tend vers 0, ces estimations sont seulement déterminées par la collection de trames du segment à représenter (algorithme EM-ML).Le problème du manque de données dans la seule collection de trames d'un segment de voix est donc évité, par balance entre un modèle du monde robuste et les nouvelles observations [9].

Comme le montrent les formules précédentes, un facteur τ_m indique une quantité d'informations associée à chaque composante du GMM-UBM. Le prolongement empirique de l'adaptation MAP consiste à ramener τ_m ces valeurs à des échelles qui les rendent comparables aux effectifs n_m des segments utilisés et à conserver les valeurs fournissant les meilleures performances [9].

Il s'est avéré que la seule adaptation des moyennes suffisait à atteindre les performances optimales et cette technique s'est généralisée dans les systèmes état de l'art. L'adaptation MAP fournit alors en sortie une batterie de vecteurs de moyennes adaptées $\hat{\mu}_m \in R^F$ (un vecteur par composante gaussienne du GMM).Ces M vecteurs peuvent être alors exprimés, par concaténation, sous forme d'un super-vecteurs de dimension $M \times F$ [9]:

$$s = [\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_M] \in R^{MF} \tag{2.5}$$

Ce super-vecteur constitue la statistique d'ordre 1 issue du GMM-UBM. Les procédures succédant à la production de modèles adaptés nécessitent d'ajouter à la représentation la statistique d'ordre 0 des trames du segment considéré. Il s'agit du vecteur n des effectifs de trames associées (probabilistiquement) à chaque gaussienne [9]:

$$n = (n_m)_{m=1}^{M} \in R^M \tag{2.6}$$

Où n_m effectif de trames affectées à la gaussienne g, est égal à la somme sur toutes les trames de la collection des probabilités d'occupation de cette gaussienne [9].

La représentation par adaptation du GMM-UBM fournit donc en sortie, dans sa version la plus courante, une numérisation du signal de parole sous forme vectorielle de dimension $M\,(F+1)$ [9].

Dans le domaine de la reconnaissance du locuteur, les i-vecteurs sont devenus état de l'art. Ils fournissent un cadre de travail intéressant, permettant de réduire la taille d'un vecteur de très grande dimension en un vecteur plus compact, où toute l'information importante du locuteur est conservée [9].

a) outils mathématiques :

L'approche i-vecteur est basée sur l'idée de l'analyse factorielle conjointe (JFA) que les facteurs du canal dans JFA contiennent également des informations dépendantes des locuteurs. Mathématiquement, un super-vecteur *S* d'un segment audio dépendant du locuteur et du canal peut se réécrire sous la forme suivante [2] :

$$S = m + Tx \tag{2.7}$$

Où m est le vecteur moyen de dimension *F*, la matrice ***T*** de dimension $F \times D$ [2].

2.2.2. Compensation inter-session:

Parce que l'espace des i-vecteurs contient à la fois une variation de locuteur et canal. La compensation inter-session joue un rôle important dans l'espace des i-vecteurs.

a) Matrice de covariance :

Soit $\{w_i\}_{i=1}^{n}$ un jeu de *n* i-vecteurs de moyenne μ, appartenant à un ensemble de *s* locuteur distinct. On notera par commodité $\{w_i^s\}_{i=1}^{n_s}$ le sous-ensemble des *ns* i-vecteurs d'un locuteur *s*.

La matrice de covariance totale $\sum$ de ce jeu peut s'écrire [20]:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(w_i - \mu)(w_i - \mu)^t \tag{2.8}$$

Pour tout locuteur s, la matrice de covariance w_s de ses observations est estimée par [20]:

$$w_s = \frac{1}{n}\sum_{i=1}^{n}(w_i^s - \overline{w}_s)(w_i^s - \overline{w}_s)^t \tag{2.9}$$

Où $\overline{w}_s$ est la moyenne des i-vecteurs de *s*.

b) Normalisation de la covariance intra-classe (WCCN) :

Pour composer les variations indésirables intra-classe dans l'espace de variabilité totale, la normalisation de la covariance intra-classe et appliquée pour l'extraction d'i-vecteurs. À cette fin, une matrice de covariance intra-classe *w*, est calculé en utilisant [21]:

$$w = \frac{1}{S}\sum_{S=1}^{S}\sum_{i=1}^{n_S}(w_i^s - \overline{w}_s)(w_i^s - \overline{w}_s)^T \tag{2.10}$$

Où w_i^S est la représentation d'une session *i* du locuteur *s*, $\overline{w}_S$ est le vecteur moyen du locuteur *s* qui est égale à $\frac{1}{n_s}\sum_{i=1}^{n_S} w_i^S$, *s* est le nombre totale des locuteurs, n_s est le nombre d'enregistrement du locuteur *s* [22].

La matrice de covariance intra-classe B_1 peut être calculé en utilisant la décomposition de Cholesky de : $B_1 B_1 = w^{-1}$ [22].

L'i-vecteur compensé par le canal WCCN peut être calculé comme suit [22] :

$$\hat{w}_{WCCN} = B_1^T w \tag{2.11}$$

c) Analyse discriminante linéaire (LDA) :

L'Analyse discriminante linéaire est une méthode principalement destinée aux problèmes de la réduction de dimensionnalité. L'avantage principal de cette méthode est le fait de tenir compte des étiquettes des classes afin de former un sous-espace (c.-à-d. de dimension réduite) optimal [8]. Ce sous-espace minimise la variabilité intra-classe et maximise la variabilité inter-classes d'où :

$$S_b v = \lambda S_w v \tag{2.12}$$

Où S_b est la dispersion inter-classe, S_w est la dispersion intra-classe, v est un vecteur propre et λ est sa valeur propre associé.

Les matrices de dispersions, inter-classe et intra-classe, sont respectivement données par [22] :

$$S_b = \sum_{S=1}^{S} n_S (\overline{w}_S - \overline{w})(\overline{w}_S - \overline{w})^T \tag{2.13}$$

$$S_w = \sum_{S=1}^{S} \sum_{i=1}^{n_S} n_S (w_S^i - \overline{w}_S)(w_S^i - \overline{w}_S)^T \tag{2.14}$$

Où n_s est le nombre d'enregistrement du locuteur *s*, $\overline{w} = \frac{1}{N} \sum_{S=1}^{S} \sum_{i=1}^{n_S} w_i^S$ est le vecteur moyen de toutes les données d'apprentissage (*N* est le nombre totale de session), $\overline{w}_S = \frac{1}{n_S} \sum_{i=1}^{n_S} w_i^S$ est le vecteur moyen de toute les données du locuteur *s*, *s* est le nombre totale des locuteurs [22].

I-vecteurs $\hat{w}_{LDA}$ compensé par le canal LDA peut alors être calculé comme suit [22]:

$$\hat{w}_{LDA} = A^T w \tag{2.15}$$

2.2.3. Méthodes de scoring:

a) Modèle à deux covariances :

Le modèle à double-covariance (Two-Covariance Model) est un modèle génératif simple, linéaire et Gaussien, basé sur la décomposition présumée de tout i-vecteurs *w* d'un locuteur *s* en [20]:

$$w = y_s + \varepsilon \tag{2.16}$$

Où le modèle locuteur y_s, de même dimension que *w*, est la part propre au locuteur de *w* et ε est un bruit gaussien, avec les hypothèses (en dénotant par commodité $N(\,)$ la densité d'une loi normale multivariée) [20]:

$$P(y_s) = N(\mu, B) \tag{2.17}$$

$$P(w|y_s) = N(y_s, w) \tag{2.18}$$

La densité de y_s suit une loi normale de moyenne μ et la matrice de covariance inter-locuteur *B* et la densité conjointe de *w* sachant y_s suit une loi normale de moyenne y_s et de covariance la matrice intra-locuteur *w* [20].

Les deux hypothèses précédentes peuvent s'écrire, pour tout i-vecteur *w* d'un locuteur *s* [20]:

$$y_s = \mu + B^{\frac{1}{2}}\eta \text{ où } \eta \approx N(0, I) \tag{2.19}$$

$$w = y_s + w^{\frac{1}{2}}\varepsilon \text{ où } \approx N(0, I) \tag{2.20}$$

Le vecteur *w* se décompose en [20]:

$$w = \mu + B^{\frac{1}{2}}\eta + w^{\frac{1}{2}}\varepsilon \tag{2.21}$$

Où η et ε suivent des lois normales standards. Ce modèle constitue donc une décomposition en facteurs locuteur et session dans l'espace des i-vecteurs [20].

b) Score à deux covariances :

La formule de score appliquée aux i-vecteurs, si elle est dépendante du modèle considéré, a pris ces dernières années une allure (semi-)standardisée. Le score par « log-ratio de vraisemblance d'hypothèses complémentaires» dans le cas des i-vecteurs est présenté au paragraphe suivant [20].

Etant donnés deux i-vecteurs w_1, w_2 à comparer dans un problème de discrimination du locuteur, deux hypothèses peuvent être définies [20]:

- H_S : $''w_{1,}\, w_2$ appartiennent au même locuter$''$.
- H_d : $''w_{1,}\, w_2$ n'appartiennent pas au même locute$''$.

$$score(w_1, w_2) = \log \frac{P(w_{1,}\, w_2 | H_S)}{P(w_{1,}\, w_2 | H_d)} \tag{2.22}$$

Où *P* ( ) est une probabilité, au sens le plus large du terme. Un tel type de score peut être appliqué en général, quelles que soient les hypothèses probabilistes. Notamment la probabilité *P* ( ) peut s'appuyer sur différentes lois théoriques, mais aussi sur des estimations non paramétriques de densité : locales, par voisinage, etc... [20].

Lorsque w_1 et w_2 sont deux observations indépendantes d'un locuteur s connu, il est possible d'écrire [20] :

$$p(w_1, w_2 \left| H_s, s) = P(w_1 \right| s) P(w_2 \left| s) \tag{2.23}$$

En vérification du locuteur, la classe est inconnue. La probabilité conjointe peut alors être calculée par [20]:

$$p(w_1, w_2 \left| H_s, s) = \int P(w_1, w_2 \right| s)\, dP(s) \tag{2.24}$$

De même, la probabilité conjointe de w_1 et w_2 sous l'hypothèse H_s peut être décomposée en [20]:

$$p(w_1, w_2 \left| H_d \right.) = \int P(w_1 \left| s) \right.\, dP(s) \int P(w_2 \left| s) \right.\, dP(s) \tag{2.25}$$

et le score peut s'écrire [20]:

$$score(w_1, w_2) = \log \frac{\int P(w_1 \left| s) \right.\, P(w_2 \left| s) \right.\, dP(s)}{\int P(w_1 \left| s) \right.\, \mathrm{dP(s)} \int P(w_2 \left| s) \right.\, dP(s)} \tag{2.26}$$

L'expression finale de ce score dépendra donc des modélisations des variabilités intra-locuteur et inter-locuteur qui ont été proposées [20].

c) Modèle LDA et WCCN-cosine-scoring :

N. Dehak propose un certain nombre de méthodes de scoring. La plus efficace est une méthode enchaînant une réduction de dimension par LDA et un scoring de type cosinus entre les i-vecteurs issu de travaux initiaux autour des SVM [20].

Etant donnés deux i-vecteurs w_1 et w_2 (ou leur projection par LDA), ce score s'écrit [20]:

$$score(w_1, w_2) = \cos_{W_{WCCN}^{-1}}(w_1, w_2) \tag{2.27}$$

Le cosinus s'entend ici au sens de la métrique de W_{WCCN}^{-1} . Ce score s'écrit analytiquement [20]:

$$score(w_1, w_2) = \frac{w_1^t W_{WCCN}^{-1} w_2}{\sqrt{w_1^t W_{WCCN}^{-1} w_1}\sqrt{w_2^t W_{WCCN}^{-1} w_2}} = \frac{\left(W_{WCCN}^{-\frac{1}{2}} w_1\right)^t W_{WCCN}^{-\frac{1}{2}} w_2}{\left\|W_{WCCN}^{-\frac{1}{2}} w_1\right\|\left\|W_{WCCN}^{-\frac{1}{2}} w_2\right\|} \tag{2.28}$$

Ou plus brièvement [20]:

$$score(w_1, w_2) = \frac{(Bw_1)^t Bw_2}{\|Bw_1\|\|Bw_2\|} \tag{2.29}$$

La matrice *B* est obtenue par décomposition de Cholesky de la matrice $W_{WCCN}^{-1} = B_1 B_1^T$. Cette opération justifie le terme WCCN (le N pour Normalisation). Partant d'un cosinus euclidien, celui-ci prend en compte la nécessité de réduire la variabilité inter-session. Après avoir réduit la dimension par LDA, l'application de la matrice $W_{[WCCN]}$ dans la formule de

score permet d'estimer une proximité angulaire, suivant l'orientation de la covariance intra-classe [20].

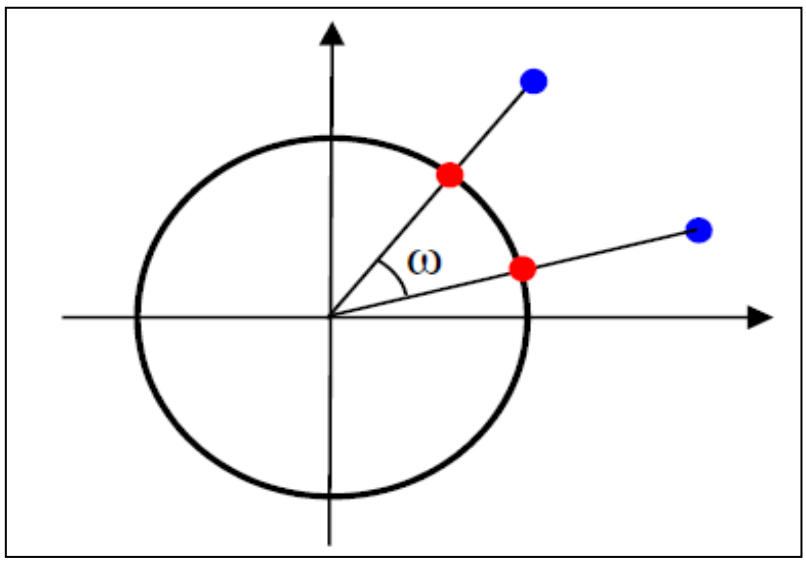


Fig 2.2 : Exemple du calcul de la distance angulaire entre deux points. Les points rouges sont les projections sur la sphère des points bleus [8].

d) Analyse discriminante linéaire probabiliste (PLDA) :

L'analyse discriminante linéaire probabiliste (Probabilistic Linear Discriminant Analysis, PLDA) élargit le modèle à deux covariances, qui en est un cas particulier, au cas probabiliste et à celui de la décomposition en facteurs. Proposée par P. Kenny en reconnaissance du locuteur, elle peut être vue comme un équivalent de la Joint Factor Analysis dans l'espace des i-vecteurs, c'est à dire dans un espace mono-gaussien [20].

Ce modèle est génératif. Il ignore le mécanisme d'extraction des i-vecteurs pour les considérer comme des observations d'un modèle génératif probabiliste. Le modèle PLDA est défini comme [20]:

$$w_{S,h} = m + Ux_S + Vy_{S,h} + \varepsilon_{S,h} \tag{2.30}$$

Où $m+Ux_S$ la composante ne dépend que du locuteur, la composante $Vy_{S,h}+\varepsilon_{S,h}$ diffère pour chaque session de voix et représente la "variabilité" intra-locuteur, *m* représente la moyenne globale, à la fois indépendante du locuteur et de la session, des modèles i-vecteurs du corpus d'apprentissage. Les espaces de variabilité du locuteur et de la session sont respectivement représentés par les colonnes de la matrice *U* et de la matrice *V*. $\varepsilon_{S,h}$ représente la variabilité résiduelle modélisée par une matrice de covariance pleine. Le terme x_S correspond au facteur locuteur, et le terme $y_{S,h}$ représente le facteur canal de la session *h* pour le locuteur *S*. Ces deux facteurs sont supposés indépendants et suivent une distribution normale sur *N (0, I)* [23].

On fait deux suppositions sur les distributions de probabilité a priori de ces variables. Soit on considère qu'elles suivent une distribution gaussienne, et on parle dans ce cas là d'une PLDA gaussienne (G-PLDA), ou soient qu'elles suivent une distribution à queue lourde (heavy-tailed distribution), une loi de Student (Student's t-distribution), et on a alors une PLDA à queue lourde (heavy-tailed PLDA), que nous allons noter HT-PLDA. En supportant de grandes déviations de la moyenne, la distribution à queue lourde permet de mieux traiter les données aberrantes qui occurrent souvent en reconnaissance automatique du locuteur [24].

Notons par $N(\mu,\sum)$ une distribution gaussienne de moyenne μ et de matrice de covariance $\sum$, et par $\Gamma(k,\theta)$, une distribution Gamma de paramètres k et θ. Et considérons les paramètres scalaires n_1, n_2, v, appelés degrés de liberté, et les variables scalaires cachées $u_S, u_{S,h}, v_{S,h}$. On assume dans la HT-PLDA que [24]:

$$x_S \approx N(0, u_S^{-1} I) \quad \text{où} \quad u_S \approx \Gamma(n_1/2, \mathrm{n}_1/2)$$

$$y_{S,h} \approx N(0, u_{S,h}^{-1} I) \quad \text{où} \quad u_{S,h} \approx \Gamma(n_2/2, \mathrm{n}_2/2) \tag{2.31}$$

$$\varepsilon_{S,h} \approx N(0, u_{S,h}^{-1} I) \quad \text{où} \quad v_{S,h} \approx \Gamma(v/2, v/2)$$

La PLDA permet de réduire la dimension d'i-vecteurs téléphone/microphone tout en transformant ces i-vecteurs (appartenant initialement à un espace regroupé téléphone/microphone) en un espace homogène. Cette transformation élimine certaines informations redondantes et nuisances du canal, ce qui permet d'améliorer les performances par rapport aux i-vecteurs originaux [24].

Notons par w_1 (pour l'enrôlement), w_2 (pour le test) les deux i-vecteurs définissant l'unité d'évaluation. Dans une application de vérification de locuteurs, on évalue les hypothèses de production [24]:

- H_S : Les deux modèles i-vecteur w_1 et w_2 sont produits par un même locuteur, auquel cas ces modèles partagent le même facteur locuteur x_S .

- H_d : Les deux modèles i-vecteur w_1 et w_2 sont produits par deux locuteurs différents, auquel cas ces modèles partagent le même facteur locuteur x_S de ces deux modèles est différent.

Le score PLDA par le score log-ratio de vraisemblance s'exprime sous forme de [24]:

$$score_{PLDA}(w_1, w_2) = \log \frac{P(w_{1,} w_2 | H_S)}{P(w_1 | H_d) P(w_2 | H_d)} \tag{2.32}$$

Ce logarithme du rapport des vraisemblances se calcule facilement dans le cas d'une PLDA gaussienne (les vraisemblances marginales sont gaussiennes).

2.3. Conclusion :

Dans ce chapitre, nous avons présenté l'espace des i-vecteurs, qui est une technique introduite dans le domaine de la reconnaissance automatique du locuteur. Cette technique vise à réduire la grande quantité de données représentant un locuteur en un vecteur discriminant de dimension réduite.

i-vecteurs est une représentation compacte et normalisée des énoncés complets, issue d'une concaténation vectorielle d'informations locales du GMM-UBM. Le résultat est un vecteur de variables acoustiques de très haut niveau d'informations.

Chapitre III

Reconnaissance automatique du locuteur dans la communication mobile

3.1. Introduction :

Dans ce chapitre, nous nous intéressons aux bases de la reconnaissance automatique du locuteur dans la communication mobile, au début nous présentons d'une façon générale les différentes générations de réseaux mobiles. Nous décrivons aussi un aperçu sur les codages de la parole à savoir le codage de forme d'onde, codage paramétrique, et codage hybride, ensuite les normes de codeurs de la parole: (FR, HR, EFR, AMR-NB et AMR-WB). Enfin, nous citons les différents architectures du système de reconnaissance automatique du locuteur dans la communication mobile telles que : la reconnaissance déportée, la reconnaissance embarquée, la reconnaissance répartie.

3.2. Généralité sur les réseaux mobile :

Les réseaux mobiles sont des réseaux sans fil (en anglais Wireless network), est comme son nom indique, un réseau dans lequel au moins deux terminaux peuvent communiquer sans liaisons filaires. Grace à ce type de réseau, un utilisateur (abonné) à la possibilité de rester connecté au réseau tout en se déplaçant dans un périmètre géographique plus ou moins étendu, c'est la raison pour laquelle on entend parfois parler de ‹‹ mobilité » [25].

Le développement des réseaux mobiles n'a pas cessé d'accroitre, plusieurs générations ont vues le jour (1G, 2G, 3G, 4G) et connues une évolution remarquable, en apportant un débit exceptionnel et qui ne cesse d'augmenter, une bande passante de plus en plus large et un des avantages d'une telle bande passante est le nombre d'utilisateur pouvant être supportés [26].

3.2.1. La première génération-1G :

La première génération de téléphonie mobile (noté 1G), Elle est basée sur une transmission analogique avec une modulation de fréquence. Elle est constituée d'appareils relativement volumineux, utilisant une faible bande passante. Cette norme ne permet que la transmission de la parole. La zone de couverture est divisée en cellules de tailles différentes. On y retrouve essentiellement les standards suivants [25]:

- **AMPS** (Advanced Mobile Phone System) apparu aux USA, constitue le premier standard des réseaux de 1G, il possède de faibles mécanismes de sécurité rendant possible le piratage des lignes téléphonie.

- **TACS** (Total Acces Communication System) est la version européenne du standard AMPS utilisant une bande de fréquence de 900 MHz.

- **ETACS** (Extended Total Access communication System) est une version amélioré du standard TACS, il a été développé au Royaume-Uni et permet l'utilisation d'un nombre plus important de canaux de communications.

3.2.2. La deuxième génération-GSM :

Le GSM (Global Système for mobile communication) est un réseau de 2ème génération qui a marqué une rupture avec la 1G grâce au passage de l'analogique au numérique. La norme permet la transmission de la parole et des données simultanément. Elle offre la possibilité aux utilisateurs de partager un même canal de transmission, ceci est possible grâce à l'utilisation du mécanisme de division de fréquence FDMA (Fréquence Division Multiple Access) et le mécanisme de division de temps TDMA (Time Division Multiple Access) [25].

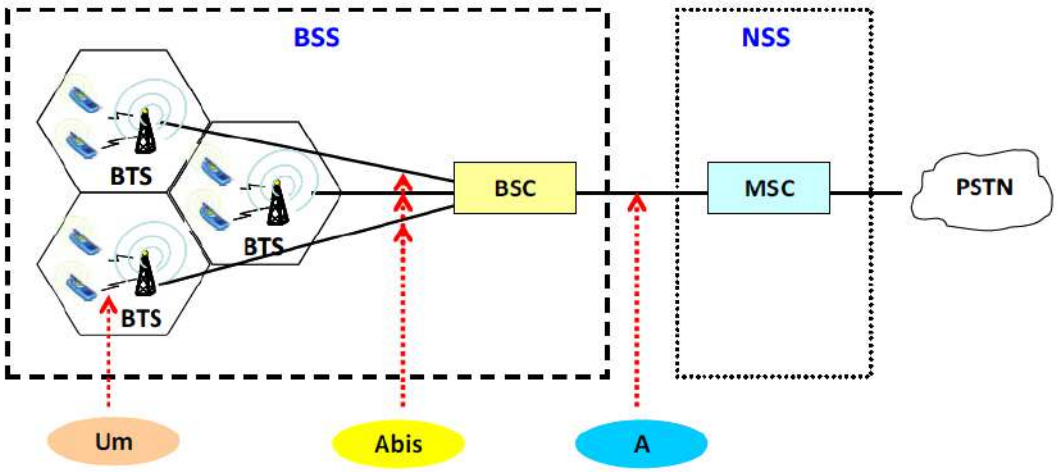


Fig.3.1 : Structure du réseau GSM simplifiée [27].

Typiquement le réseau GSM est composé de 3 sous-ensembles [27]:

- **Le sous-système radio (BSS) (**Base Station Sub-system**):** Il gère la partie radio des communications et se compose d'émetteurs-récepteurs radio (BTS) contrôlés par une BSC.

- **Sous-système réseau (NSS)** (Network Station Sub-system**):** comprend l'ensemble des fonctions nécessaires de commutation et de routage.

- **Sous-système opérationnel OSS (**Operating Sub-System**):** Il contrôle les droits d'accès au réseau, les droits des usagers et assure l'interface homme-machine d'exploitation. Il gère aussi le maintien en conditions opérationnelles du réseau et la remontée des alarmes.

Les composants principaux de cette structure du réseau GSM sont [27]:

- **BTS (Base Transceiver Station) :** permet de communiquer avec les mobiles sur l'interface radio (interface Um). Elle est comme un émetteur-récepteur central pour chaque zone cellule qui comprend un ou plusieurs TRX (Transmission/Réception), chaque TRX peut supporter 8 communications simultanées.

- **BSC (Base Station Controller):** permet de contrôler un ensemble de BTS sur l'interface Abis. Il prend en charge l'allocation de la ressource radio et la décision de l'activation/désactivation d'un canal vers un mobile. Le BSC s'occupe aussi de la gestion du (handover) quand le mobile passe d'une BTS à l'autre.

- **MSC (Mobile Switching Center):** est un commutateur téléphonique qui permet de gérer les appels. Il multiplexe les trafics en provenance de plusieurs BSC sur l'interface A et les relie au réseau téléphonique commute public (PSTN : Public Switched Téléphone Network).

3.2.3. La troisième génération-UMTS :

Abréviation d'Universal Mobile Télécommunications System, l'UMTS désigne une nouvelle norme de téléphonie mobile. On parle plus généralement de téléphonie de troisième génération ou 3G. Les puristes préfèrent utiliser le terme W-CDMA (Wideband Code Division Multiple Access) qui reprend le nom de la technologie déployée en Europe et par certains opérateurs asiatiques [28].

Son principe : exploiter une bande de fréquences plus large pour faire transiter d'avantage de données et donc obtenir un débit plus important. En théorie, il peut atteindre 2 Mbit/s, soit une vitesse de transmission équivalente à celle proposée pour l'internet "très haut débit" permis par l'ADSL ou le câble [28].

Après de nombreux atermoiements, le déploiement de la téléphonie de troisième génération en Europe a enfin débuté. La norme UMTS/W-CDMA a été retenue sur le Vieux Continent par l'association 3GPP (Third Génération Partnership Project) regroupant les principaux acteurs des télécommunications. Elle implique, pour chaque opérateur mobile qui souhaite se lancer dans l'aventure, l'achat d'une licence émise par l'État et la mise en place d'infrastructures en parallèle du réseau GSM/GPRS [28].

La norme UMTS exploite le nouveau protocole de communication W-CDMA et de nouvelles bandes de fréquences situées entre 1900 et 2200 MHz. À la différence du GSM qui fait passer les données par une cellule (antenne) divisée en canaux de fréquences différentes, elles-mêmes réparties selon des créneaux de temps, le W-CDMA permet d'envoyer simultanément toutes les données, par paquets et dans le désordre (sur n'importe quelle fréquence), reste au téléphone à réceptionner les paquets de données et les rassembler [28].

Cette technologie permet de faire transiter d'avantage de données simultanément et offre un débit bien supérieur à ceux permis par les GSM et GPRS. En théorie, il peut atteindre 2 Mbit/s à partir d'un lieu fixe et 384 kbit/s en mouvement. Bien qu'on soit encore loin de la promesse initiale, le maximum en Europe a été fixé à 384 kbit/s : soit 64 à 128 kbit/s en émission et 128 à384 kbit/s en réception, même en mouvement (train, voiture). Ce "haut débit" mobile n'est pas si éloigné des vitesses de transmission proposées dans le cadre des premières offres d'accès à l'internet par l'ADSL ou le câble [28].

L'UMTS présente des avantages qui s'appliquent autant aux communications vocales qu'aux transferts de données. Comme la technologie exploite une bande de fréquences plus large, elle permet de passer trois fois plus d'appels. En théorie, l'UMTS devrait donc remédier à la saturation des réseaux existants et proposer des services de meilleure qualité. Le débit cinq à dix fois plus rapide a permis le développement de nouvelles applications, notamment dans le domaine du multimédia (visiophonie, diffusion de contenu vidéo et audio, MMS vidéo ou audio, etc.). Le haut débit mobile facilite aussi l'accès aux données, web et e-mails, en situation de mobilité [28].

La topologie du réseau UMTS avec la simplification pour cette étude est montrée par la figure. 3.18. Elle est composée de deux parties principales : le réseau d'accès (similaire au sous-système BSS dans le réseau GSM) et le réseau cœur. Comme un réseau GSM/GPRS, le réseau cœur de l'UMTS est compose de deux domaines : d'une part, le domaine de commutation de paquets (PS : Paquet Switched) permet l'accès à Internet, d'autre part, le domaine de commutation de circuits (CS : Circuit Switched), qui est très similaire à celui du NSS dans le réseau GSM est utilisé principalement pour la téléphonie [27].

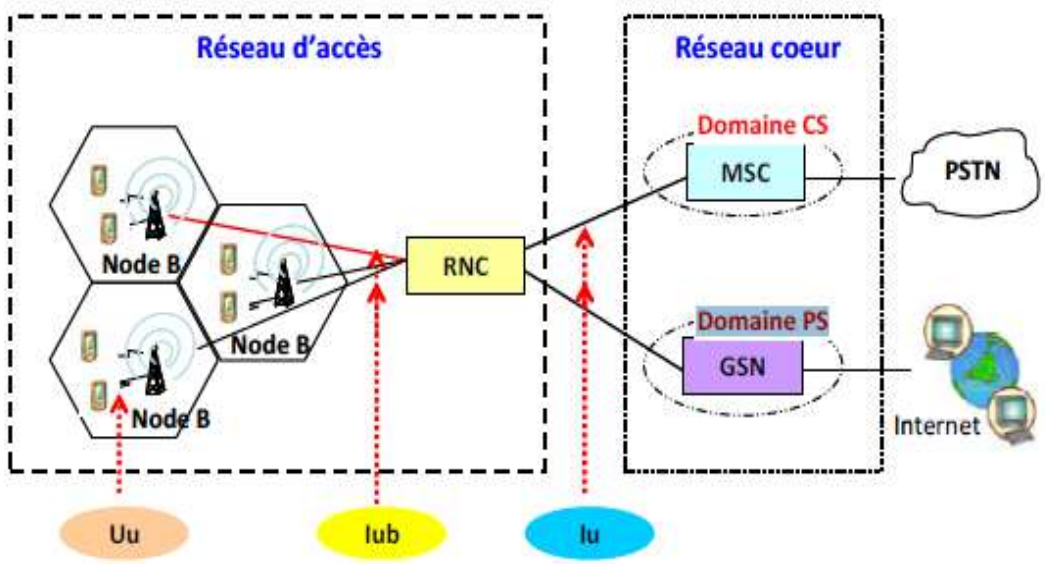


Fig. 3.2 : Structure du réseau UMTS simplifiée [27].

3.2.4. La quatrième génération-LTE:

La technologie LTE (Long Term Evolution) ou la 4G s'appuie sur un réseau de transport à commutation de paquet IP. Elle n'a pas prévu de mode d'acheminement pour la voix, autre que la VoIP, contrairement à la 3G qui transporte la voix en mode circuit [26].

Les services de communications mobiles sont en train de suivre la même évolution que celle des services fixes, c'est-à-dire une transition accélérée vers l'accès à très haut débit. Ce sont les réseaux 4G qui permettent de répondre aux demandes croissantes des usages mobiles,

tant en termes de qualité des services offerts que de capacité d'écoulement du trafic par les réseaux [26].

Ces fréquences sont destinées au déploiement de réseaux mobiles à très haut débit, pour apporter au consommateur une capacité et une qualité de services supérieures aux offres actuelles d'internet mobile. La technologie LTE « Long Term Evolution » offre aux utilisateurs des débits de plusieurs dizaines de Mbit/s, largement supérieurs aux performances des technologies 3G et 3G+ actuellement déployées, ainsi que des latences plus faibles favorisant une meilleure interactivité [26].

Avec le 4G, on se dirige vers la transmission de toutes les informations. Voix et données. par IP, le même protocole qu'on utilise sur Internet. Pour les fournisseurs, c'est plus facile et moins cher à gérer. Ça facilite aussi le développement d'applications multimédias. Cette génération permet des vitesses de téléchargement plus rapides et des temps de latence plus courts [26].

Selon les critères de l'Union internationale des télécommunications (UIT), qui établit les normes pour les réseaux cellulaires, le vrai 4G devrait offrir des vitesses de téléchargement de 100 Mbit/s pour un utilisateur en mouvement et de 1 Gbit/s en mode stationnaire [26].

3.3. Codage de la parole:

Le but de toute technique de codage est de réduire la quantité de données binaires et son attribut fondamental est la qualité de la parole restituée : typiquement, plus le débit de codage est faible, moins la qualité du signal restitué est bonne. Les différentes normes couvrent diverses applications avec des débits variant de 128 Kb/s jusqu'à 600 b/s. Les algorithmes de codage de parole utilisés dans les normes peuvent être groupés en 3 catégories majeures qui sont : le codage de formes d'onde, le codage paramétrique et le codage hybride [29].

3.3.1. Codage de forme d'onde :

Les codeurs temporels codent directement la forme d'onde du signal, sans connaissance apriori de ses caractéristiques. Ce genre de codage est utilisé dans plusieurs standards destinés Principalement aux réseaux téléphoniques filaires. Il conduit à un débit de codage élevé,

typiquement compris entre 16 et 64 Kb/s, et permet de préserver une bonne qualité perceptuelle pour le signal vocal. La complexité de ces codeurs est minimale ainsi que le délai de codage [29].

Le codage porte directement sur les échantillons de la forme d'onde du signal de la parole qui sont quantifiés et transmis. Le signal variant peu d'un échantillon à un autre, le codage différentiel est introduit afin de réduire le débit. Dans ce cas, c'est la différence entre la valeur d'un échantillon et sa valeur prédite qui est quantifiée puis transmise. La prédiction est réalisée par une combinaison linéaire à partir des valeurs précédemment quantifiées [29].

La technique de modulation par impulsions et codage PCM (Pulse Code Modulation), adoptée par le standard G.711, effectue une quantification pseudo-logarithmique qui permet, à débit identique, d'obtenir une meilleure qualité perceptuelle qu'une quantification linéaire et ceci compte-tenu des propriétés de la perception auditive. Deux lois de compression pseudo-logarithmiques sont utilisées : la loi A et la loiμ. Chaque échantillon est codé sur 8 bits et la fréquence d'échantillonnage est de 8 KHz, ce qui donne un débit de codage de 64 Kb/s. Le rapport signal sur bruit de ce codeur est de l'ordre de 35 dB sur une large plage d'amplitudes [29].

La modulation par impulsion et codage différentiel ADPCM (Adaptive Differential Pulse Code Modulation) est une technique de codage différentiel qui utilise un pas de quantification Adaptatif. Le pas est proportionnel à l'énergie à court terme normalisée du signal. La technique de codage permet d'obtenir des débits réduits allant de 16 à 40 Kb/s avec une faible dégradation dans la qualité du signal codé. Le codeur G.726 utilisé en VoIP est basé sur la technique ADPCM [29].

Le codage différentiel adaptatif comme le CVSD (Continuously Variable Slope Delta Modulation) effectue un ajustement en continu de la taille du pas de quantification. La technique CVSD est utilisée pour le codage de parole dans les communications sur Bluetooth [29].

Les algorithmes de codage de formes d'onde décrit ci-dessus codent tous la parole entièrement dans le domaine temporel. Le codage de forme d'onde dans le domaine fréquentiel est pourtant possible et il a certains avantages. Par exemple, lors du codage en sous-bandes SBC (Sub-Band Coding), la parole à l'entrée est séparée en un nombre de bandes

de fréquence ou sous-bandes, puis chacune est codée indépendamment en utilisant par exemple le codage ADPCM. À la réception, les signaux des sous-bandes sont décodés et recombinés ensemble pour donner le signal de parole reconstitué. L'avantage provient du fait que le bruit dans chaque sous-bande dépend uniquement du type de codage dans cette sous-bande. Donc, on peut allouer plus de bits pour les sous-bandes perceptuellement importantes de façon à maintenir un faible niveau de bruit [29].

Les codeurs de forme d'onde permettent de coder tout signal qui se situe dans la même bande fréquentielle que la parole. Il est possible de transmettre des données à travers un canal équipé d'un codeur de forme d'onde, tant que le spectre du signal modulé est compris entre 0et 4 KHz. Les réseaux téléphoniques filaires sont communément utilisés pour la transmission de données. Il existe plusieurs systèmes de modulation/démodulation des données destinés à ces canaux, notamment les standards de l'ITU connues sous le nom de modems « série V » [ITU-T, V] [29].

3.3.2. Codage paramétrique:

Les codeurs paramétriques modélisent le signal vocal à l'aide d'un modèle paramétrique propre aux signaux de parole. Le codeur extrait les paramètres du modèle et les code sur un nombre limité de bits. Le décodeur utilise ces paramètres pour alimenter le modèle et resynthétiser un signal de parole. Ces codeurs permettent d'obtenir de faibles débits de codage, typiquement inférieurs à 4 Kb/s, tout en conservant une bonne qualité de parole (MOS proche de 3,5) même si elle est généralement inférieure à celle des codeurs hybrides. Les algorithmes de codage paramétriques sont destinés à des applications de sécurité et ils sont souvent développés et standardisés par des organismes militaires, comme le DoD ou l'OTAN [29].

Les codeurs de parole utilisant le codage paramétrique diffèrent par le type de modèle utilisé et par la façon de coder les paramètres. Deux types de modèles sont couramment utilisés : le modèle « source-filtre » décrit précédemment et les modèles sinusoïdaux STC (Sinusoidal Transform Coder). Ces codeurs calculent différents types de paramètres, en particulier des paramètres spectraux qui représentent l'enveloppe spectrale du signal. Les paramètres spectraux sont généralement obtenus par prédiction linéaire mais on utilise aussi parfois une transformée de Fourier à court terme. D'autres paramètres comme le voisement ou

les niveaux de voisement, la fréquence fondamentale, l'énergie (souvent représentée par un gain), les valeurs de quelques harmoniques sont aussi utilisés [29].

Les codeurs source-filtre utilisent différentes approches pour construire le signal d'excitation. On peut distinguer les codeurs à deux états d'excitation (LPC classique), les codeurs à excitation multi-bande MBE (Multi Band Excited), les codeurs à excitation mixte MELP (Mixed Excitation Linear Prédiction), les codeurs HSX (Harmonic Stochastic excitation) et les codeurs à interpolation de forme d'onde prototypes WI (Waveform Interpolation) [29].

3.3.3. Codage hybrides :

La fréquence d'échantillonnage étant fixe, la réduction de débit des codeurs d'onde fait chuter rapidement la qualité d'écoute pour des débits inférieurs à 16 kbit/s. Les codeurs hybrides utilisent les deux méthodes temporelle et paramétrique de façon complémentaire, ce qui permet un codage de parole de bonne qualité à des débits relativement faibles. Ces codeurs sont basés sur des techniques de codage temporel auxquelles des modèles de production de parole sont associés pour améliorer leur efficacité. Cependant, ce type de codage nécessite des coûts de calculs plus importants. Tous les codeurs hybrides s'appuient sur une analyse LPC pour obtenir les modèles de synthèse de parole. Les deux techniques paramétrique et temporelle modélisent respectivement le conduit vocal et le signal résiduel ou erreur de prédiction (l'excitation) [30].

Différentes techniques efficaces ont été proposées pour représenter le signal d'excitation comme, par exemple, le codeur à prédiction linéaire excité par des impulsions MPE-LPC (Multi Pulse Excitation - LPC) qui utilise comme signal d'excitation une suite d'impulsions qui peuvent prendre des amplitudes et être situées à des positions arbitraires. La forme d'onde d'excitation est obtenue en optimisant les positions et les amplitudes d'un nombre fixe d'impulsions par trame. Le codeur excité par des impulsions régulièrement espacées RPE (Regular Pulse Excitation) représente le signal d'excitation par un ensemble d'impulsions régulièrement espacées. Ces codeurs fournissent une bonne qualité de la parole avec une complexité raisonnable pour un débit autour de 10 kbit/s [30].

Le codeur à prédiction linéaire excité par code CELP (Code Excited Linear Prédiction), détermine une forme d'onde optimale du signal résiduel de prédiction en utilisant la technique d'analyse par synthèse. Le codage CELP réalise une modélisation paramétrique du signal de parole sous la forme d'un signal d'excitation, généralement issu d'un dictionnaire (codebook) de formes d'ondes prédéterminées, passant au travers d'un ou plusieurs filtres. La technique CELP est parmi une des idées les plus influentes dans le codage de la parole et ses principes constituent la base de beaucoup de codeurs normalisés. Par exemple, un codeur LD-CELP (Low Delay-CELP) ciblant un faible délai de codage/décodage a été normalisé par la recommandation G.728 de l'UIT. Ensuite, le codeur G.729 (CS-ACELP), basé sur un codage de parole ACELP (Adaptive Code Excited Linear Predictive) à 8 Kbits/s [30].

3.4. Les différents codeurs de la parole:

Les codeurs de parole ont une influence majeure sur le signal transmis. Le développement des codeurs de parole est guidé par le double objectif de compresser le débit binaire nécessaire pour transmettre le signal de parole et d'optimiser la qualité perçue du signal restitué par le décodeur [29].

3.4.1. ACELP:

L'algorithme ACELP est une implémentation de l'algorithme CELP destinée à diminuer la complexité de calcul. La structure standard de l'algorithme CELP utilise une technique D'analyse par synthèse avec prédiction à long terme et à court terme Le principe est de rechercher une séquence d'innovation dans un dictionnaire stochastique selon un critère qui sera précisé plus loin. La différence entre le signal de parole originale et le signal synthétisé, représentant l'erreur objective, subit un filtrage linéaire à pondération perceptuelle [29].

La séquence élue est celle qui minimise l'erreur quadratique moyenne pondérée. Cette recherche de la séquence d'innovation optimale présente une grande complexité de calcul, notamment les filtrages de prédiction à long et court terme sur toutes les séquences [29].

La simplification proposée par l'algorithme ACELP consiste à utiliser des séquences algébriques pour construire le dictionnaire stochastique. Les séquences considérées sont

constituées par les seules valeurs +1, 0 et -1. Une séquence ne contient qu'un nombre limité de valeurs non nulles (+1 ou -1). L'avantage de cette structure est qu'elle permet l'utilisation d'algorithmes de calcul rapide lors de la recherche [29].

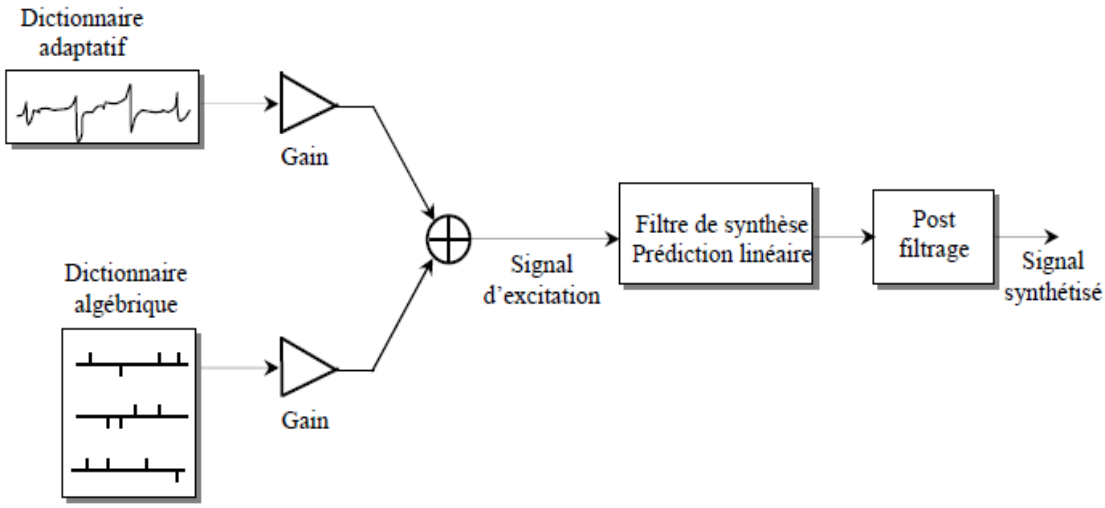


Fig 3.3: Schéma simplifié du modèle de synthèse de l'algorithme ACELP [29].

3.4.2. Le codeur FR (Full Rate):

Ce codeur appartient à RPE-LTP (Regular pulse excitation-long term prediction-linear predictive). Dans la partie de codeur une trame de 160 échantillons de parole est codée sous la forme d'un bloc de 260 bits conduisant à un débit de 13 kb/s, le décodeur cartographie les blocs codés de 260 bits à des blocs de sortie de 160 échantillons de parole reconstruits. Le canal à pleine fréquence GSM supporte 22.8 kb/s, ainsi, les 9.8 kb/s restants sont utilisés pour la protection contre les erreurs. Le codeur FR est décrit dans le GSM 6.10 jusqu'au niveau de bit, ce qui permet sa vérification au moyen d'un ensemble de séquences de test numérique qui sont également données dans le GSM 6.10 [31].

3.4.3. Le codeur HR (Half Rate):

Le codeur HR a été établi pour faire face au nombre croissant d'abonnés. Ce codeur est un codeur VSELP (Vector Sum Excited Lineary Prédiction) de Motorola. Afin de doubler la capacité du système cellulaire GSM, le canal à demi-débit supporte 11.4kb/s. Par conséquent

5.8kb/s sont utilisés pour la protection contre les erreurs. La qualité de la parole de sortie mesurée pour le codeur HR est comparable à la qualité du codeur FR dans toutes les conditions testées, sauf pour les conditions de tandem et de bruit de fond. Le standard normatif GSM 06.06 donne le code ANSI-C exact de bit pour cet algorithme tandis que le GSM 06.07 donne un ensemble de séquences de test numérique pour la vérification de conformité [31].

3.4.4. Le codeur EFR:

Le codeur EFR (Enhanced Full Rate), basé sur un algorithme ACELP, est le standard de codage de parole le plus récent pour les réseaux mobiles GSM, il remplace les deux standards précédents HR (Half Rate) et FR (Full Rate). Il fournit une qualité de parole équivalente à celle de la téléphonie filaire utilisant le codeur ADPCM à 32 Kb/s. Les paramètres fournis par ce codeur sont les coefficients du filtre de prédiction linéaire, les indices et les gains des dictionnaires fixe et adaptatif. Chaque trame de 20 ms est codée sur 244 bits offrant ainsi un débit de codage de 12.2 Kb/s. Un codage canal est ajouté pour la protection contre les erreurs, conduisant à un débit de 22.8 Kb/s [29].

L'EFR opère sur des trames de 20 ms, soit 160 échantillons à la fréquence d'échantillonnage de 8 KHz. Chaque trame est divisée en 4 sous-trames de 5 ms, soit 40 échantillons. Le codeur analyse le signal et extrait les paramètres du modèle ACELP. Deux séries de coefficients de prédiction linéaire à court terme sont calculées pour chaque trame. Les indices des dictionnaires adaptatif et algébrique fixe sont calculés avec leurs gains respectifs pour chaque sous-trame [29].

3.4.5. Le codeur AMR :

Le codeur AMR est un autre standard ITU destiné aux réseaux GSM et UMTS. Il possède plusieurs modes de codage pour produire l'un des débits suivants : 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 et 4.7 Kb/s. L'allocation binaire dans une trame codée change en fonction du débit, notamment le nombre de bits alloués pour les vecteurs LSP et le dictionnaire algébrique.

L'AMR est basé sur un algorithme ACELP et il effectue des fonctions de codage similaires à celles du codeur EFR présentées précédemment. De fait ce codeur se confond

avec l'EFR lorsque le débit de codage est à son maximum 12.2 Kb/s. Pour les autres modes, quelques modifications ont lieu au cours du processus de codage [29].

L'analyse spectrale n'est conduite qu'une seule fois par trame, avec une fenêtre centrée sur la quatrième sous-trame, résultant en un seul vecteur LSP. La quantification utilisée est alors une SVQ (Split Vector Quantization). Les vecteurs correspondants aux trois premières sous-trames sont obtenus par interpolation linéaire du seul vecteur disponible [29].

La recherche en boucle ouverte ne se déroule qu'une seule fois par trame pour les 2 modes aux plus bas débits 5.15 et 4.75 Kb/s [29].

3.4.6. AMR-WB :

Le codeur AMR-WB, standardise sous la recommandation de l'ITU-T G.722.2, utilise le codage ACELP décrit précédemment. II fonctionne pour 9 débits différents allant de 6.6 kbit/sa 23.85 kbit/s. Les modes allant de 12.65 kbit/s jusqu'au débit maximal offrent une excellente qualité de codage de la parole large bande tandis que les deux modes inferieurs (6.6 et 8.85 kbit/s) sont prévus pour être utilises temporairement, en général dans des cas de saturation du réseau. Le codec opère à la fréquence d'échantillonnage de 16 kHz et sur des blocs de 20 ms. Le bloc d'entrée est d'abord sous-échantillonne à 12.8 kHz de cette manière tous les calculs de nombre d'échantillons seront des multiples de 2. Chaque trame de 20 ms est découpée en 4 sous-trames de 5 ms chacune. Au niveau du codeur, le calcul des coefficients du filtre LPC (Linear Predictive Coding) est effectué pour chaque trame de 20 ms. En revanche, les paramètres nécessaires aux dictionnaires adaptatif et fixe sont calcules pour chaque sous-trame. En plus de ces différents paramètres, une valeur de VAD (Voice Activity Detection) est transmise. Comme la bande de fréquence basse contient presque toute 1'information, la grande majorité des bits est allouée à son traitement [32].

3.4.7. AMR-NB :

Le codeur AMR-NB, standardisé par le 3GPP, Ce codeur sera à la base pour toutes les simulations effectuées dans cette thèse. L'AMR-NB encode des signaux échantillonnés à 8 kHz et est capable de produire un débit variable aussi appelé mode, en fonction des ressources du canal. Les débits sont les suivants : 4.75 kbps, 5.15 kbps, 5.90 kbps, 6.70 kbps (identique à

PDC-Japon), 7.4 kbps (identique à DAMPS-IS136), 7.95 kbps, 10.2 kbps and 12.2 kbps (identique au GSM-EFR). Le nombre total de bits attribué dépend du mode sous lequel fonctionne le codeur, mais c'est la quantification du mot de code fixe qui requiert le plus débits. Ce codeur comprend d'autres fonctionnalités, entre autre deux options de détection d'activité vocale [33].

3.5. Les différents Architecture du système de reconnaissance du locuteur dans la communication mobile :

Dans cette section, nous citerons les caractéristiques et les limites que rencontrent les applications de reconnaissance automatique de la parole sous les différentes structures dans lesquelles elles peuvent être déployées en utilisant des liaisons de communication numériques sans fil. La reconnaissance vocale de mobile peut être caractérisée en fonction de l'endroit où la reconnaissance a lieu. Ceci définit trois structures principales: la reconnaissance déporté, la reconnaissance embarquée et la reconnaissance répartie. Chacune de ces structures peut avoir des effets très différents et des caractéristiques sur la performance des systèmes de reconnaissance de la parole. Nous décrivons ces structures plus en détail:

3.5.1. La reconnaissance déportée:

Le principe de la RAP déportée consiste à se servir du terminal uniquement comme un système d'acquisition du son. Une fois le signal enregistré, il est transmis via un réseau, généralement sans fil, à un serveur qui effectue le travail de reconnaissance [34].

Cette approche permet d'envisager l'utilisation de serveurs beaucoup plus puissants et donc de fournir des services plus divers et généralement de meilleure qualité. La présence d'un serveur performant permet d'adjoindre d'autres modules à la reconnaissance vocale telle qu'un système de dialogue ou encore des vocabulaires dédiés à chaque application. Pour résumer, l'avantage majeur de cette architecture est la levée des contraintes liées aux ressources limitées des systèmes embarqués [34].

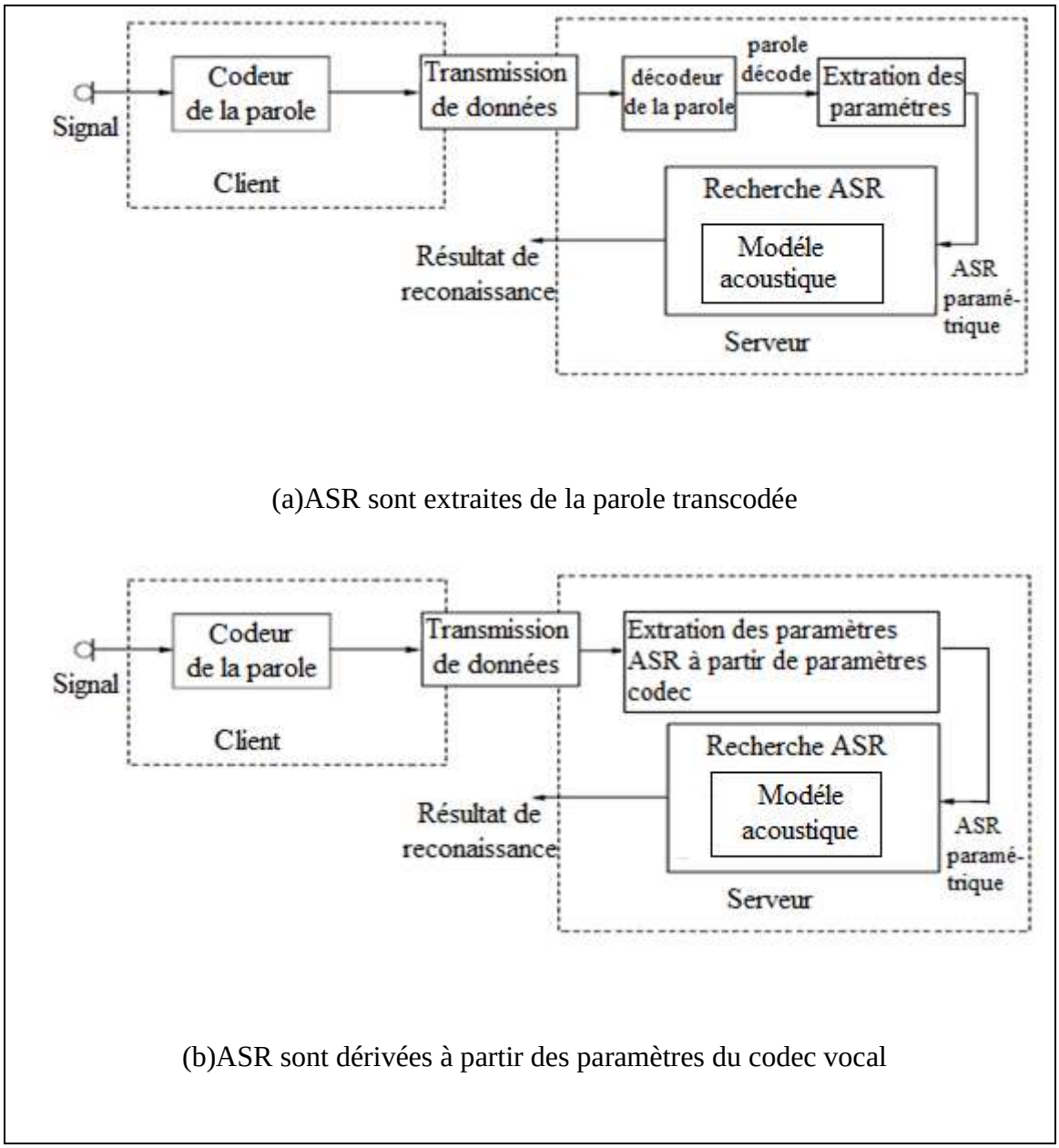


Fig. 3.4: L'architecture du système de la reconnaissance déportée [35].

Un des inconvénients de ce type d'architecture est la perte d'informations due à la transmission du signal. En effet, avant d'être envoyé au(x) serveur(s) de reconnaissance, le signal doit être encodé suivant un des codecs de l'opérateur (HR, FR, EFR, AMR . . .). Ces codages (avec pertes) sont nécessaires car l'émission du signal sur le réseau a un coût non négligeable. De plus, au cours d'une transmission sur un réseau (sans fil ou non), différents phénomènes, comme la perte de paquets d'informations ou l'ajout de bruits viennent perturber le signal. Parmi les inconvénients, nous pouvons encore citer le coût (tant sur le plan financier

que sur l'aspect du trafic réseau) de transmission du signal au serveur et de la réponse du serveur au terminal [34].

Enfin, l'absence d'autonomie représente certainement l'inconvénient majeur de cette architecture. En effet, elle suppose que le téléphone soit connecté au réseau pour pouvoir fournir les services de reconnaissance, interdisant l'emploi de ceux-ci en local pour des applications comme les jeux, l'agenda. . . [34].

3.5.2. La reconnaissance embarquée:

Avec cette approche, toutes les étapes de la reconnaissance sont effectuées dans le système embarqué. Aucun serveur à distance est nécessaire, le système est donc entièrement autonome : il réalise lui-même l'acquisition du son puis le décodage [34].

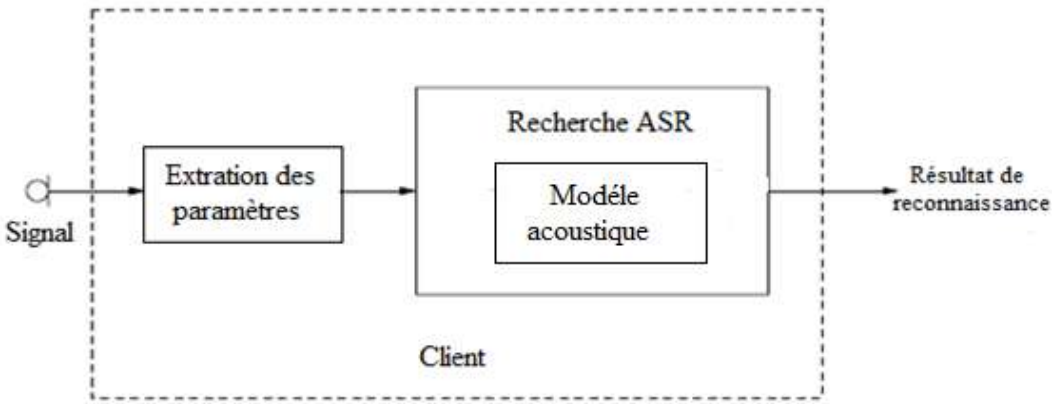


Fig. 3.5: L'architecture du système de la reconnaissance embarquée [35].

L'intérêt principal de cette approche est l'absence de contrainte : en effet, il n'est pas nécessaire de se limiter à l'utilisation des codecs de parole définis par l'ETSI1, car aucune compatibilité avec un serveur n'est requise. De plus, les inconvénients de la reconnaissance déportée sont résolus par l'absence de transmission [34].

Cette solution présente toutefois un inconvénient majeur : les ressources nécessaires. En effet, toutes les phases (paramétrisation du signal et reconnaissance) sont effectuées au sein

du téléphone. Les différentes phases de la reconnaissance nécessitent beaucoup de ressources comparées aux capacités d'un téléphone portable. De plus, l'augmentation des ressources d'un téléphone (mémoire et puissance de calcul), bien que possible, est très couteuse [34].

Dans la littérature plusieurs implémentations dans des systèmes embarqués sont proposées. Les ressources disponibles dans ces systèmes sont variables suivant les projets. Trois grandes classes peuvent être distinguées [34]:

- **Très peu de ressources** : DSPFactory propose un système nécessitant un DSP[2]/CPU à 4MHz. Un système présenté par Wang permet débarquer un système de reconnaissance de mots isolés sur un MCU[3] à 8MHz.
- **Peu de ressources** : le système d'IBM recommande un DSP/CPU à 50 MHz et 1Mo de mémoire.
- **Système de demain**: NEC propose un système ayant besoin d'un processeur triple cœur (3 systèmes à 200MHz) et 6Mo de mémoire.

Ce dernier système est beaucoup plus gourmand que les autres. Il semble plutôt destiné aux futures applications et paraît difficilement intégrable aujourd'hui au regard des ressources des systèmes embarqués disponibles actuellement [34].

3.5.3. La reconnaissance répartie:

Ce mode de reconnaissance se situe entre la RAP déportée et la RAP embarquée. En effet, une partie du travail est effectuée sur le système embarqué et une autre partie sur un serveur distant. Généralement, la phase d'extraction des paramètres est réalisée sur le terminal et la partie reconnaissance - à proprement parler- se trouve sur un serveur [34].

Cette approche permet de contourner les problèmes liés à la perte d'informations due au transport. Avec cette architecture, une seule phase "d'extraction de l'information" est nécessaire : les deux premières étapes (encodage et décodage) de l'architecture deviennent inutiles. Les pertes engendrées par l'encodage effectué sur le téléphone sont alors évitées (HR, FR, EFR et AMR sont tous des encodeurs avec perte) [34].

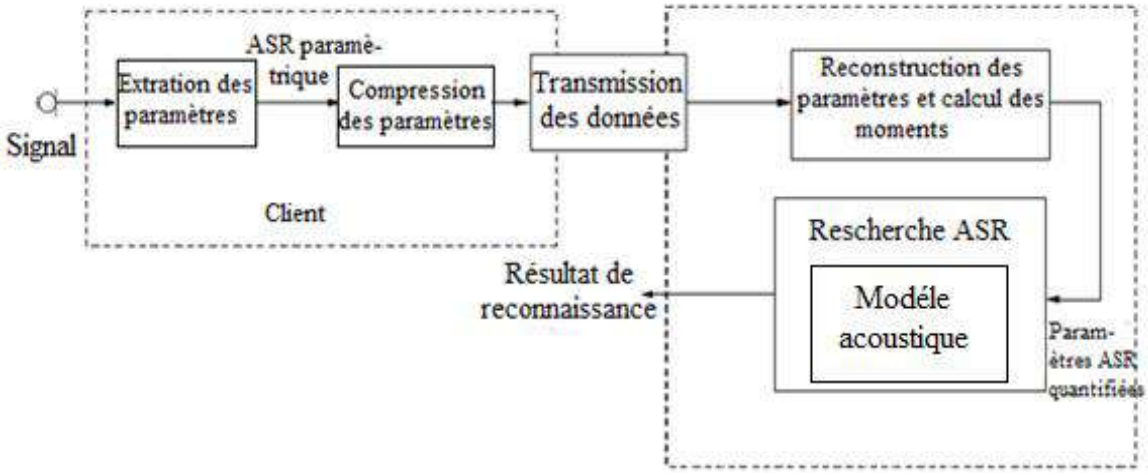


Fig.3.6 : L'architecture du système de la reconnaissance répartie [35].

Cette approche implique, par ailleurs, que le téléphone (et donc son fabricant) s'accorde avec l'opérateur désirant fournir le service de reconnaissance vocal, étant donné que l'extraction des paramètres est effectuée directement dans le téléphone. Ces paramètres sont ensuite utilisés par le système de reconnaissance durant la phase de décodage. Les codecs classiques, définis par l'ETSI, sont plutôt basés sur les coefficients LPC alors que les systèmes actuels de RAP sont essentiellement basés sur les coefficients MFCC ou PLP ou encore sur des dérivés de l'une de ces deux paramétrisation, il est donc nécessaire d'ajouter des fonctions spécifiques de paramétrisation au sein des téléphones [34].

Enfin, cette approche (comme la RAP déportée) impose une connexion constante au réseau GSM pour bénéficier des services de RAP [34].

3.6. Conclusion :

Ce chapitre a présenté une vue d'ensemble sur la reconnaissance automatique du locuteur dans la communication mobile. Nous avons donné un aperçu sur le signal parole dans les réseaux mobiles, dont les performances d'un système de reconnaissance automatique du locuteur dans les réseaux mobiles se dégradent à cause de la distorsion du canal et la présence de bruits. Pour remédier au maximum à ce problème de dégradation, il est nécessaire d'utiliser

des codeurs qui ont une influence majeure sur le signal transmis, et qui sont guidés par le double objectif de compresser le débit binaire nécessaire pour transmettre le signal de parole et d'optimiser la qualité perçue du signal restitué.

Bien que la technique de "codage du signal parole" soit un facteur important, nous avons présenté les différentes techniques de codages de la parole qui ont pour but de représenter le signal par le moins de bits possible, mais maintenir un certain niveau de qualité du signal parole décodé. En général, les techniques de codage sont basées sur la réduction de la redondance du signal, afin que le signal résiduel (après réduction de redondance) puisse être encodé avec moins de bits que l'originale.

En fin nous avons cité les trois architectures envisageables (reconnaissance déportée, reconnaissance embarquée et reconnaissance répartie), dans les systèmes embarqués, le choix d'une reconnaissance embarquée a été fait car cette architecture est la seule qui permette une utilisation entièrement autonome des fonctions de reconnaissance vocale. En effet, la reconnaissance répartie et la reconnaissance déportée font toutes deux appel à des serveurs de reconnaissance externes.

Chapitre IV

Expériences et Résultats

4.1. Introduction :

Dans ce chapitre, nous allons présenter les expériences et les résultats obtenus après l'évaluation de notre système de reconnaissance du locuteur. Notre objectif étant de développer un système de reconnaissance du locuteur basé sur l'approche i-vecteur dans un environnement mobile.

Les protocoles de développement et d'évaluation de cette technique de modélisation sont décrits dans cette section. Ils mettent en jeu des modules d'extraction des paramètres acoustiques à savoir MFCCs et PLPs. Les expériences ont été menées en mode indépendant du texte. Nous citerons aussi dans ce chapitre les bases de données et logicielles utilisées.

4.2. Logiciels et langages utilisés dans l'application:

a) Logiciel Matlab:

Nous avons utilisé un logiciel MATLAB (matrix laboratory) version 12.02 sous Windows 7, Matlab est un langage de programmation émulé par un environnement de développement, il est utilisé à des fins de calcul numérique et commercialisé par la société américaine the Math Works. Ce langage est basé essentiellement sur le calcul matriciel à syntaxe simple, avec des fonctionnalités mathématiques spécialisées et graphiques étendues.

b) logiciel HTK:

HTK est une boîte à outils de modèles de Markov cachés HMM. Conçue pour la construction des systèmes de reconnaissance de parole. Cette boîte est constituée d'un ensemble de modules bibliothèque et d'outils disponibles en codes sources C. Dans notre travail nous avons utilisé cette boîte pour extraire les paramètres acoustiques MFCCs et PLPs.

4.3 Description des bases de données utilisées:

Les expériences ont été effectuées en utilisant les bases de données NIST-2000 et la base de bruit standard.

a) La base de données NIST:

L'évaluation de système de reconnaissance du locuteur fait une partie des séries continue d'évaluations annuelles menées par le NIST (National Institute of Standards and Technology). En effet, depuis 1996 une série d'évaluations SRE (Speaker Recognition Evaluation), dans le but de promouvoir la recherche en reconnaissance automatique du locuteur. NIST-SRE 2000, qui était d´éditée à la vérification du locuteur. D'autres travaux se sont attaques à l'identification et l'indexation des locuteurs. La base de données contient des fichiers audio (hommes et femmes) sous format (WAV.file), avec une fréquence d'échantillonnage égale à 8 KHz. La durée des fichiers varie de 10 s À trois minutes. Le découpage est le suivant: 3 enregistrements servent à l'apprentissage des modèles et 1 à la reconnaissance (tests). Le répertoire « WAV » est divisé en 2 sous répertoire « Apprentissages » pour modèles et « test» pour la reconnaissance (tests).

Nous avons pris 200 locuteurs à partir de cette base. La base de données est passée ensuite à travers le codec standard AMR-WB pour aboutir à la base de données transcodée mobile.

b) La base de données NOISEX-92

NOISEX-92 est une base de données de bruits standards. Elle fournit différents bruits enregistrées militaire dans des environnements réels avec de taux d'échantillonnage 19.9KHz.

4.4. Étapes de la réalisation des systèmes de reconnaissance du locuteur:

Différentes étapes sont nécessaires à la construction et à la validation du système:

- Créer l'ensemble d'apprentissage: la base de données que nous avons utilisée est constituée de 600 fichiers, prononcés par 200 locuteurs de différents sexes qui répètent la même phrase 3 fois.
- Effectuer l'analyse acoustique: l'ensemble des fichiers de la base de données ont été convertis en une série de vecteurs.
- Faire l'apprentissage des modèles de mélanges de lois gaussiennes: il s'agira de construire le modèle du monde « GMM-UBM » et le modèle du locuteur cible.

- L'apprentissage d'un sous-espace de variabilité totale à partir des statistiques de base « GMM-UBM».
- Créer un modèle PLDA Gaussian avec des i-vecteurs de développement.
- Calculer le score entre les modèles d'i-vecteurs et le test.
- Faire la reconnaissance et évaluer la performance des systèmes de reconnaissance du locuteur « MFCCs et PLPs ».

4.5. La structure générale de Système de reconnaissance du locuteur:

La structure générale d'un système de reconnaissance automatique du locuteur est représentée sur la (Figure 4 .1).

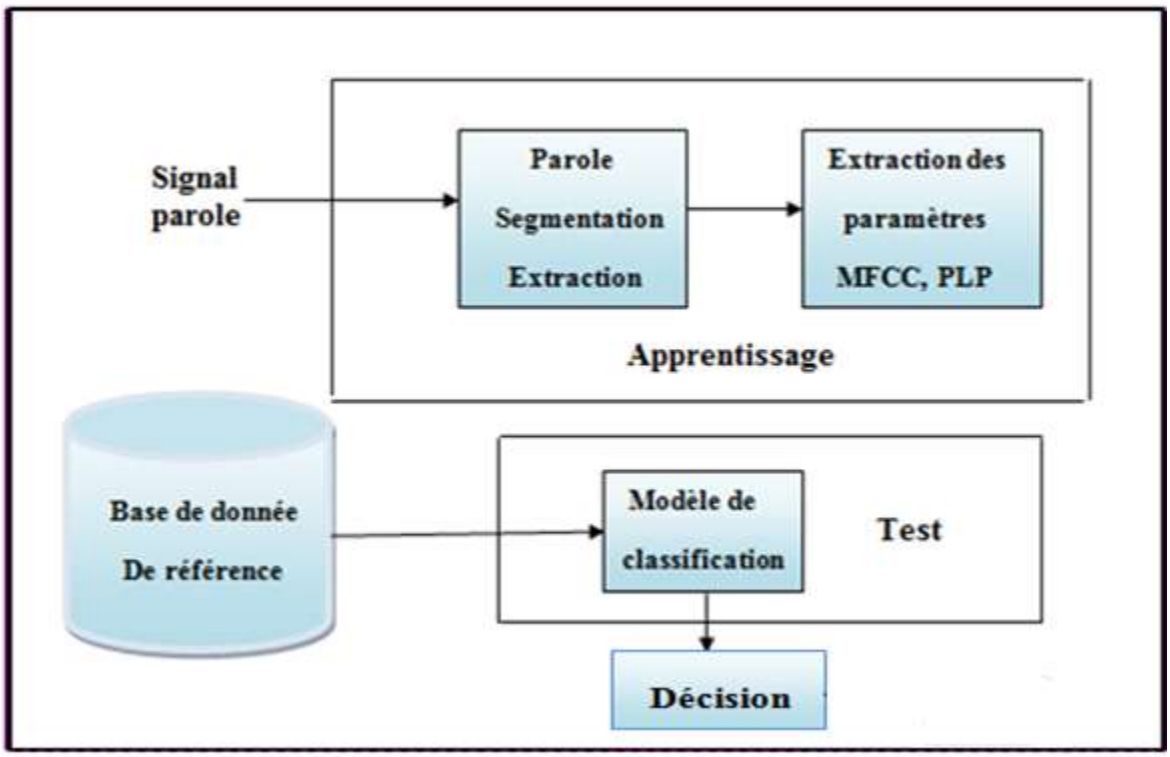


Figure 4.1: Structure générale d'un système RAL.

4.6. Extraction des paramètres:

4.6.1. Paramétrisation des coefficients MFCCs:

Le signal de parole est paramétré en utilisant les coefficients MFCCs. Le signale de parole passe à la préaccentuation après, les paramètres MFCCs sont extraits toutes les 10ms, sur des fenêtres de 25 ms. La fenêtre de pondération utilisée est la fenêtre de Hamming et puis appliquer Douze (12) coefficients MFCCs, ils sont calculés pour chaque trame dans l'échelle fréquentielle Mel. Les coefficients différentiels du premier et du second ordre (ΔΔ), sont ensuite calculés pour former un vecteur de dimension 36 (12MFCC + 12 Δ MFCC+ 12 Δ Δ MFCC).

4.6.2. Paramétrisation des coefficients PLPs:

Pour le calcul des coefficients PLPs, on a procédé comme suit : Le signale de parole passe à la préaccentuation et le fenêtrage puis par la méthode d'auto corrélation en récupère les paramètres PLPs comme suit : On prend un vecteur de dimension égal à 12, qui représente les coefficients PLPs statique. On rassemble les PLPs et les (Δ) PLPs et (Δ Δ) PLPs dans un seul vecteur qui aura une dimension égale à 36**.**

4.7 Protocole d'évaluation des résultats:

Pour évaluer notre système de reconnaissance, nous avons structuré les bases de données comme suivant:

- **La base d'apprentissage:** elle est composée de 600 fichiers vocaux prononcés par 200 locuteurs comprenant les deux sexes. Chaque locuteur répète la phrase 3fois.

- **La base de test:** elle est composée de 200 fichiers vocaux prononcés par 200 locuteurs, comprenant les deux sexes Chaque locuteur répète la phrase 1 fois.

4.8. Expériences et résultats:

Pour évaluer la robustesse des systèmes de reconnaissance du locuteur à base de coefficients « MFCCs-I vecteur » et « PLPs-I vecteur ». Nous avons réalisé des expériences

de reconnaissance du locuteur pour deux types de bruits (Canal et environnement) qui sont respectivement: bruit de (AWGN, Rayleigh) et bruit (babble, Car) Pour l'entrainement des modèles acoustiques a été réalisé avec la base de données NIST-2000 non bruitée, le bruit a été seulement ajouté pour la base de test : (AWGN, Rayleigh, Babble : le bruit de la foule, Car : le bruit de la voiture) issus de la base de données de bruits standards NOISE, en utilisant les niveaux SNR suivants : 0dB, 5dB, 10dB, 15dB.

Pour les bruits d'environnement la qualité d'un signal est souvent représentée par le rapport de l'énergie du signal divisé par l'énergie du bruit, appelée "Rapport signal/bruit" (SNR en anglais).

Pour :

$$s(t) = x(t) + n(t) \tag{4.1}$$

Le rapport signal sur bruit est défini par :

$$SNR = \frac{E_S}{E_n} \tag{4.2}$$

E_S : représente l'énergie moyenne du signal *x(t).*
E_n *:* représente l'énergie moyenne du bruit *n(t).*

Pour le canal de transmission les bruits AWGN and Rayleigh fading, on définit la capacité C du canal par :

$$C = W \log(1 + \frac{S}{N}) \tag{4.3}$$

Avec: $N = W.N_0$ et $S = \frac{l.E_b}{T} = D.E_b$.

Où *W :* est la largeur de bande du canal en Hertz.
S : est la puissance moyenne du signal
N_0 : est la densité spectrale de puissance du bruit

l : est la longueur du signal codé

T : est la période du signal

D : est le débit (taux) de transmission

E_b : L'énergie par bit.

L'efficacité spectrale est le nombre moyen de bits d'informations transmis dans un intervalle de signalisation de durée *T*. Si *T* est normalisée et reliée à l'inverse de la longueur de bande.

4.8.1. Expériences:

Nous avons validé notre système de reconnaissance du locuteur en utilisant comme entrée les coefficients « MFCCs-I vecteur » et « PLPs-I vecteur », cette étude est faite en fonction de bruit du canal et l'environnement mobile.

a) Effet du bruit du canal de transmission:

Dans cette expérience à chaque fois quand nous augmentons le niveau du bruit du canal (AWGN), le taux d'erreurs dans les coefficients « MFCCs-I vecteur » et les coefficients « PLPs-I vecteur » augmente, avec les coefficients « PLPs-I vecteur » sont plus élevés que les coefficients « MFCCs-I vecteur », ce que signifie que les coefficients « MFCCs-I vecteur » donnent de meilleurs résultat de taux d'erreurs par rapport aux coefficients « PLPs-I vecteur ».

AWGN (E_b/N_0)	**Taux d'erreurs EER (%)**	
	PLPs-I vecteur	**MFCCs-I vecteur**
0	30.15	25.62
5	22.61	20.10
10	19.02	17.08
15	10.62	9.44

Tab 4.1: Comparaison des taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec les différents niveaux bruit du canal (AWGN).

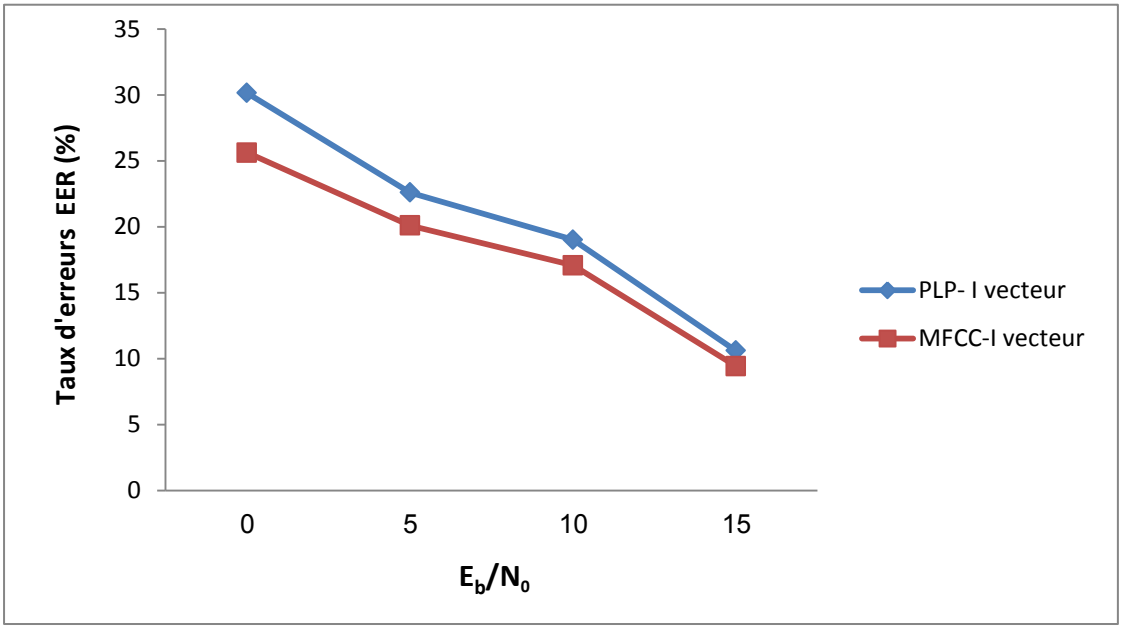


Figure 4.2: Taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit du canal AWGN

Rayleigh (SNR)	Taux d'erreurs EER (%)	
	PLPs-I vecteurs	MFCCs-I vecteur
0	50	40.00
5	40.81	32.16
10	26.31	20.35
15	18.09	12.56

Tab 4.2: Comparaison des taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit du canal (Rayleigh).

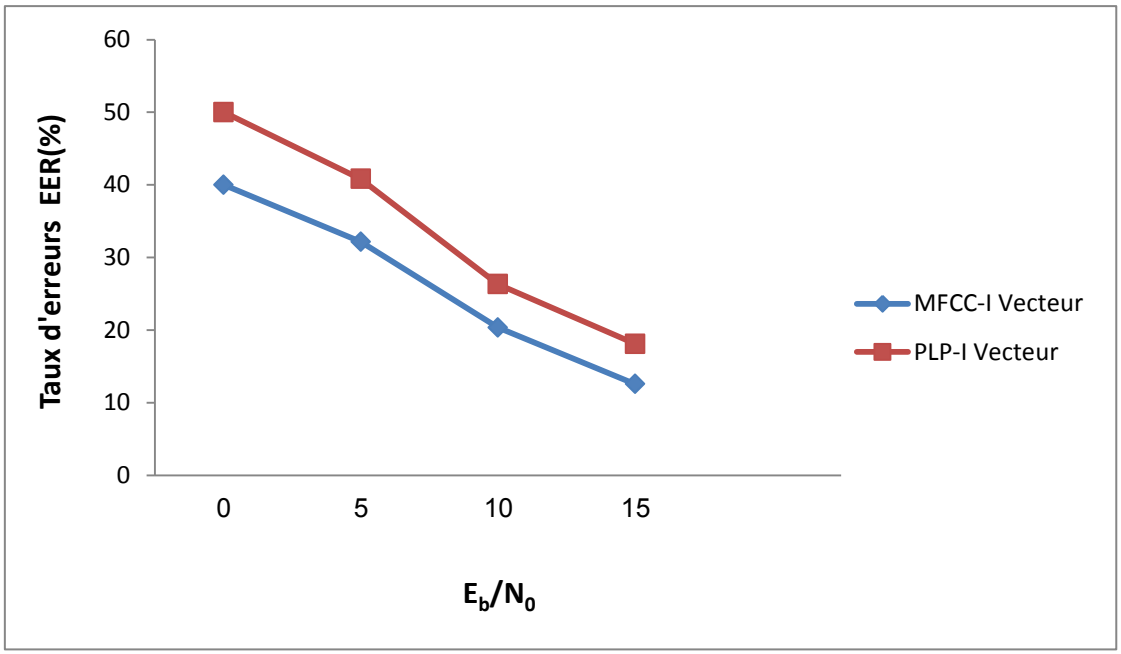


Figure 4.3: Taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit du canal Rayleigh.

À partir de cette expérience nous constatons que les résultats des taux d'erreurs obtenus en bruit du canal (Rayleigh) sont plus élevés que ceux obtenus en bruit du canal (AWGN).

Nous constatons aussi que les coefficients « MFCCs-I vecteur » donnent de meilleurs résultats des taux d'erreurs par rapport aux coefficients « PLPs-I vecteur ».

b) Effet des bruits d'environnement mobile:

Cette expérience concerne l'environnement mobile bruité avec deux type de bruit telle que le bruit (bruit de la foule, voiture), nous avons vari le niveau SNR : 5dB, 10dB, 15dB, l'entrainement des modèles acoustiques a été réalisé avec la base de données NIST non bruitée, le bruit a été ajouté pour le corpus de test.

Bruit de la foule (SNR)	Taux d'erreurs EER (%)	
	PLPs-I vecteur	MFCCs-I vecteur
0	51.02	49.05
5	47.02	46.16
10	38.25	32.75
15	25.85	18.45

Tab 4.3: Comparaison des taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit d'environnement (la foule).

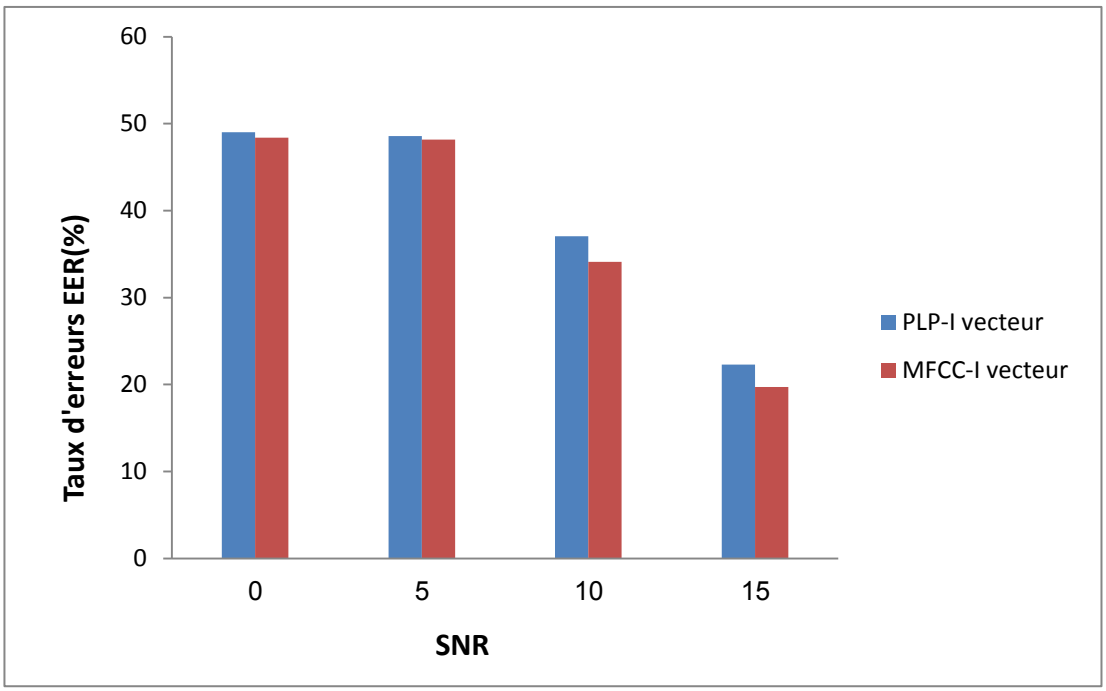


Figure 4.4: Taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit d'environnement (la foule).

La figure 4.4 représente une comparaison des taux d'erreurs des deux systèmes à base des coefficients « MFCCs-I vecteur » et «PLPs-I vecteur » en bruit d'environnement (la foule).

A partir de cette représentation nous constatons que le taux d'erreurs obtenu par le système de RAL à base des «PLPs-I vecteur » est plus élevé que celui obtenu à base de «MFCCs-I vecteur », ce qui signifie que le système de RAL à base de coefficients «MFCCs-I vecteur » donne un meilleur résultat que celui à base des coefficients «PLPs-I vecteur ».

Bruit de la voiture (SNR)	**Taux d'erreurs EER (%)**	
	PLPs-I vecteur	**MFCCs-I vecteur**
0	49.02	48.40
5	48.56	48.16
10	37.05	34.12
15	22.28	19.71

Tab 4.3: Comparaison des taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit d'environnement (la voiture).

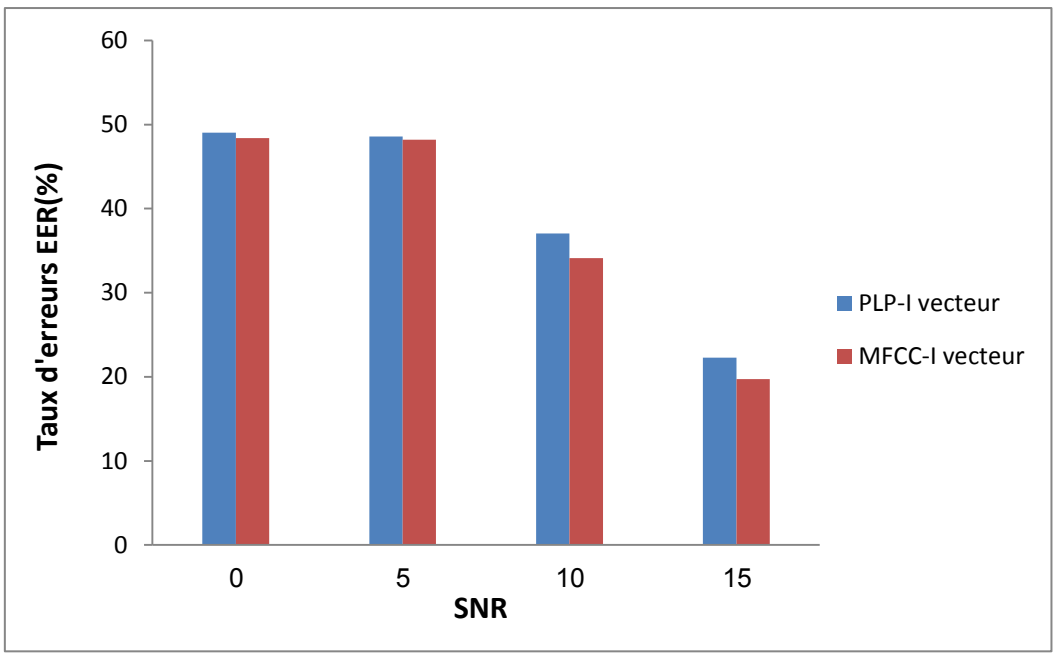


Figure 4.5: Taux d'erreurs (MFCCs-I vecteur et PLPs- I vecteur) avec le bruit d'environnement (la voiture).

Les résultats obtenus à partir de cette expérience montrent que les valeurs des taux d'erreurs obtenus en bruit d'environnement (la voiture) sont sensiblement équivalents que ceux obtenus en bruit d'environnement (la foule).

Nous constatons aussi que les coefficients « MFCCs-I vecteur » donnent de meilleurs résultats des taux d'erreurs par rapport aux coefficients « PLPs-I vecteur ».

D'après ces expériences nous constatons que les résultats des taux d'erreurs obtenus en bruit d'environnement mobile (bruit de la foule, voiture) sont plus élevés que ceux obtenus en bruit du canal (AWGN, Rayleigh), la remarque la plus importante est que le système de reconnaissance à base de coefficients « MFCCs-I vecteur » est plus robuste au bruit contrairement à « PLPs-I vecteur ».

4.9. Conclusion:

Dans ce chapitre, nous avons présenté des expériences et des résultats obtenus pour les différents canaux de transmission et bruit d'environnement mobile, nous avons réalisé deux systèmes de reconnaissance automatique du locuteur basé sur la méthode d'i-vecteurs. Nous avons comparé les deux coefficients « MFCCs-I vecteur » et « PLPs-I vecteur ». D'après les résultats obtenus le taux d'erreurs en utilisant les coefficients « MFCCs-I vecteur » est nettement inférieur aux taux d'erreurs basé sur les coefficients « PLPs-I vecteur » dans tous les cas. Au regard des résultats obtenus, nous avons conclu que le système de RAL à base de « MFCCs-I vecteur » est le plus robuste.

Conclusion générale et perspectives

Conclusion générale et perspectives

L'objectif de ce travail était de construire un système de reconnaissance du locuteur robuste qui résiste aux variabilités acoustiques dans un environnement mobile. La performance du système reconnaissance du locuteur est étroitement liée à la qualité du signal parole. Les signaux transmis par le réseau mobile montrent une dégradation plus importante à cause de l'effet du codage sur le signale parole, canal de transmission et les bruits d'environnement restent encore un grand problème dans une communication mobile, puisque les bruits viennent perturber le signal parole.

Pour aborder le travail dans un environnement mobile, nous avons utilisé une base de données NIST 2000 issue du codeur AMRWB, nous avons simulé le canal de transmission par deux type de canaux de transmission (AWGN et Rayleigh fading). Nous avons étudié l'influence du canal sur la performance du système de reconnaissance du locuteur basé sur l'espace i-vecteurs. Nous avons utilisé comme vecteurs (d'entrée) pour le système de reconnaissance les paramètres : « MFCCs-I vecteur » et « PLPs-I vecteur». Les résultats obtenus ont montré les coefficients « MFCCs-I vecteur » donnent de meilleurs résultats des taux d'erreurs par rapport aux coefficients « PLPs-I vecteur » ce que signifie « MFCCs-I vecteur » est le plus robuste au canal de transmission.

Nous avons aussi étendu notre étude au cas de l'environnement mobile bruité, cet environnement simulé par une base de données NOISX-92 de bruits standards. Les expériences que nous avons effectuées sur les différents types de bruits (Babel, Car). Les résultats obtenus en utilisant les coefficients « MFCCs-I vecteur » montrent que le système de reconnaissance du locuteur est plus performant que le système de reconnaissance avec des paramètres « PLPs-I vecteur », il offre donc une meilleure robustesse au bruit d'environnement mobile comparé aux paramètres « PLPs-I vecteur ».

Comme perspective de ce travail nous proposons :

• Afin d'améliorer encore plus la performance du système reconnaissance du locuteur dans un environnement mobile, il faut prévoir des méthodes de réduction de bruit ou des méthodes de rehaussement de la parole à l'entrée du système de reconnaissance.

Références bibliographiques

Références bibliographiques

[1] M. Mohamad El Abed, « **Evaluation de systèmes biométriques** ». Thèse de doctorat, Spécialité Informatique et applications, Université de Caen/Basse-Normandie.

[2] Yassine Mami, « **Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence** ».Thèse de doctorat, spécialité d'images et de signal, École nationale supérieure des Télécommunications Paris Tech, 2003.

[3]Alexandre Preti. «**Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur** ». Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des pays de Vaucluse 2008.

[4] Sayoud Halim, « **Reconnaissance Automatique du Locuteur-Approche connexionniste** », Thèse de Doctorat spécialité d'électronique et Electrotechnique, Faculté d'Electronique et d'Informatique USTHB, 2003.

[5] Ajgou Riadh, « **Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux** », Thèse de Doctorat, spécialité d'électronique, Université Mohamed Khider Biskra, 2016.

[6] Vincent Jousse, « **Identification nommée du locuteur: exploitation conjointe du signal sonore et de sa transcription**». Thèse de doctorat, spécialité d'informatique, Université du Maine, 2011. Français.

[7] Teva Merlin. «**Amiral, une plateforme générique pour la reconnaissance automatique du locuteur de l'authentification à l'indexation**". Thèse académie d'Aix-Marseille, spécialité d'informatique, université d'Avignon et des Pays de Vaucluse, 2004.

[8] Mohammed Senoussaoui, « **Amélioration de la robustesse des systèmes de reconnaissance automatique du locuteur dans l'espace des i-vecteurs** ». Thèse de doctorat, spécialité génie électrique, École de technologie supérieure université du Québec 2014.

[9] Ben Mahria Bilal, «**L'approche hybride GMM-SVM pour l'identification des locuteurs en mode indépendant du texte en milieu fermé**», thèse de master, spécialité de système intelligent et réseau, Université Sidi Mohamed Ben Abdallah, 2014-2015.

[10] Fredouille, Corinne, « **Approche statistique pour la reconnaissance automatique du locuteur : informations dynamiques et normalisation bayésienne des vraisemblances**». Thèse de doctorat**,** spécialité d'informatique**,** Université d'Avignon et des Pays de Vaucluse Donneur de grade 2000.

[11] Antonio M. Peinado, José C. Segura, «**Reconnaissance de la parole sur les canaux numériques: Robustesse et normes »,** 2006.
www.onlinelibrary.wiley.com/book/10.1002/0470024720

[12] Nassim Elabed, Zied Sakka, Mounir SAMET et Abdennaceur Kachouri, « **Approche Multi-Bandes pour la Reconnaissance Automatique des Mots Isolés et Phonèmes Arabes ».** Conférence international de Sfax, Tunisie, 2009.

[13] Houda Kadi. « **La reconnaissance automatique du locuteur par la voix IP** » Thèse de master, spécialité de système intelligent et réseau Université sidi Mohamed ben abdallâh, 2014.

[14] Laurent Besacier, «**Un modèle parallèle pour la reconnaissance automatique du locuteur** » thèse de doctorat, Spécialité mathématique et Informatique l'Université d'Avignon et des Pays de Vaucluse, 1998

[15] Claude Barras, « **Reconnaissance automatique des locuteurs pour la structuration en tours de parole** ». Thèse de master, spécialité d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Université Paris-Sud11, 2011.

[16]Mr.Ramou Naim, «**Amélioration des performances de la reconnaissance automatique locuteur par des méthodes de fusion de données »,** thèse de master, spécialité électronique, thèse de master, spécialité électronique, Université des sciences et de la technologie Houari Boumediène (USTHB), 2010.

[17] Imen Trabelsi, Dorra Ben Ayed, « **Evaluation d'une Approche Hybride GMM-SVM pour l'Identification de Locuteurs »,** article volume 8, N°1, pp 61-65, 2011.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, **«Front–end factor Analysis for speaker vérification »,** IEEE Transactions on Audio, Speech, and Language Processing, article volume. 19, no. 4, pages 788–798, 2011.

[19] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, D. Ramos, «**Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques**», speech communication, article volume 59, pages 69-82, 2014.

[20] Pierre-Michel Bousquet, «**Bénéfices et limites des représentations en facteur de variabilité totale pour la reconnaissance du locuteur**», Thèse de doctorat, spécialité d'informatique, Université d'Avignon et des Pays de Vaucluse, 2014.

[21] Hamid Behravan, Ville Hautamaki, Tomi Kinnunen, «**Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish**», Speech Communication, article volume 66, pages 118–129, (2015).

[22] Ahilan Kanagasundaram, David Dean, Sridha Sridharan, Mitchell McLaren Robbie Vogt, «**I-vector based speaker recognition using advanced channel compensation techniques**», Computer Speech and Language, volume 28, pages 121–140, (2014).

[23] Grégor Dupuy, «**Les collections volumineuses de documents audiovisuels: segmentation et regroupement en locuteurs**», thèse de doctorat, spécialité d'informatique, Université Maine, 2015.

[24] Reda Jourani, «**Reconnaissance automatique du locuteur par des GMM à grande marge**». Thèse de doctorat, spécialité Traitement du signal et d'image, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 2012.

[25] Chamek Linda, «**Localisation des mobiles par une strategie de prediction** ». Thèse de master, spécialité du système d'informatique de génie logiciel, Université M'hamed bougara-boumerdes, 2010-2011.

[26] Bouchentouf Hadjer, Boudghene Stambouli Riyad, «**Etude des performances des réseaux 4G (LTE)** ». Thèse de master, spécialité télécommunication, Université Abou Bekr Belkaid-Tlemcen, 2012-2013.

[27] Na TAO, «**Etude des Performances et Optimisation d'un réseau d'accès par Satellite pour les Communications**». Thèse de doctorat, spécialité Aéronautique et Astronautique, Université de Toulouse, 2009.

[28] Benjamin Savouré, «**La téléphonie mobile: technologies, acteurs et usages** ». Thèse de master, spécialité droit des medias, Université de droit et de science politique d'Aix-Marseille, 2005-2006.

[29] Toufic Chmayssani, «**Modulation sur les canaux vocodés**». Thèse de master, spécialité d'informatique, Université Paris-Est, Français 2010.

[30] Mouloud Djamah, «**Codage échelonnable à granularité fine de la parole utilisant la quantification vectorielle arborescente**». Thèse de doctorat, spécialité de télécommunication, Université du Québec2010.

[31] L. Besacier, S.Grassi, A.Dufaux, M.Ansorge, F.Pellandini, «**GSM speech coding and speaker recognition**». Article, 0-7803-6293-4/00/$10.00 02000 IEEE.
www.scribd.com/document/320085636/GSM-SPEECH-CODING-AND-SPEAKER-RECOGNITION

[32] Amelie Marty, «**Debit variable controle par la source pour le codec audio AMR-WB+**». Thèse de master, spécialité génie éléctrique, Université du Sherbrooke (Québec) Canada, 2007.

[33] Juan M. Huerta, «**Speech Recognition in Mobile Environments**». Thèse de doctorat, spécialité génie éléctrique, Université du Pittsburgh, PA 15213, 2000.

[34] Christophe Lévy, «**Modèles acoustiques compacts pour les systèmes embarqués**», Thèse de doctorat, spécialité d'informatique, Université D'Avignon et des Pays de Vaucluse, 2006.

[35] Alexander Schmitt, Dmitry Zaykovskiy, Wolfgang Minker, «**Speech recognition for mobile devices**», Speech Technol, Article volume 1, pages 63–72, 2008.