République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Electronique et d'Informatique

Mémoire
Présenté pour l'obtention du diplôme de MAGISTER
En : Electronique
Spécialité **: Communication Parlée**

Par
KROBBA Ahmed

THEME

Reconnaissance Automatique du Locuteur Via une Ligne de Communication GSM

Soutenu le : 07 /02 /2010.

Devant le jury composé de :

Mme	R .TOUHAMI	Professeur	USTHB	Présidente
Mr	M. DEBYECHE	Maître de conférences	USTHB	Directeur de Thèse
Mr	A. AMROUCHE	Maître de conférences	USTHB	Examinateur
Mr	H. TEFFAHI	Maître de conférences	USTHB	Examinateur

Remerciements

Je remercie avant tout, Dieu le tout puissant, pour m'avoir permt de présenter ce modeste travail.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse Monsieur Mohamed DEBYECHE maître de conférence et Directeur du laboratoire LCPTS à la faculté d'électronique et d'informatique, pour m'avoir fait l'honneur de diriger ce mémoire. Qu'il trouve ici mes respectueux remerciements pour l'intérêt et le soutien sans relâche qu'il m'a toujours accordé.

Mes remerciements s'adressent aussi à Madame Rachida Touhami, Professeur à l'USTHB, pour avoir accepté de présider le jury.

Mes remerciements vont à A. AMROUCHE, Maître de conférences et chercheur à la faculté d'électronique et d'informatique l'USTHB et H. TEFFAHI, Maître de conférences et chercheur à faculté d'électronique et d'informatique l'USTHB, qui ont bien voulu me faire l'honneur d'être membres du jury.

Je tiens à remercier aussi l'ensemble de l'équipe du laboratoire Communication Parlée et traitement de signal (LCPTS).

Enfin, je n'oublierais pas d'associer à mon travail mes parents, ma famille, mes amis qui se reconnaîtront et tous ceux que j'ai pu oublier.

Résumé

Le GSM (Global System Mobile) s'est imposé comme le moyen de communication téléphonique le plus répandu actuellement. La reconnaissance du locuteur à travers des enregistrements de parole de type GSM devient alors un sujet de recherche très important. Dans ce travail, nous avons mis en œuvre un système de reconnaissance du locuteur via une ligne de communication GSM. La ligne de communication GSM simulée et mise en œuvre est basée sur le codeur GSMEFR (Enhanced Full Rate). Le codeur GSMEFR correspond au protocole GSM 6.60 actuellement utilisé par les opérateurs de téléphonie mobile, il est basé sur l'algorithme ACELP (Algebraic Code Excited Linear Prediction). La simulation a été effectuée sous la boite outils de **MATLAB version 7.4**. Le système de Reconnaissance Automatique du locuteur (RAL) développé est lui basé sur l'approche de modélisation par mélange de fonctions de densité de probabilité gaussiennes GMM (Gaussian Mixture Models). Une étude sur l'influence de la parole transcodée GSM sur les performances du système de RAL mis en œuvre à été réalisée avec différentes bases de données. Nous avons utilisé dans un premier temps la base de données ARADIGIT16K version original échantillonnée à 16KHz pour avoir la structure optimale de note système de RAL. Ensuite, cette base de données à été sous-échantillonnée à 8KHz avec le logiciel **Praat** pour aboutir à la base de données bande téléphonique ARADIGIT8K. Cette base de données ARADIGIT8K a été passée à travers le codec GSM pour obtenir la base de données transcodée GSM (source uniquement) à savoir la base ARADIGIT_GSM. Le canal de transmission a été simulé par deux types de bruits, le bruit blanc additif AWGN (Additive White Gaussian Noise) et le bruit de Rayleigh à différents niveaux RSB (Rapport Signal Bruit). La base de données d'évaluation ARADIGIT16K est composée d'un vocabulaire de 10 chiffres arabes (de zéro à neuf) prononcés par soixante (60) locuteurs (31 locuteurs masculins et 29 locutrices), chaque chiffre est répété trois fois par chaque locuteur. Les résultats obtenus par les différentes bases de données ont montré une dégradation du taux d'identification du locuteur de l'ordre 6% pour la base transcodée GSM (source uniquement) et l'ordre de 16% pour la base de données transcodée GSM (source uniquement + canal simulé par le bruit AWGN) et de l'ordre de 24 % pour la base de données transcodée GSM (source uniquement + canal simulé par le bruit Rayleigh).

Abstract

The GSM (Global System Mobile) has become most means of widely used over the telephone channel. The speaker recognition through recordings of GSM speech type becomes a very important research topic. In this work, we have implemented a system for speaker recognition via a communication line phone. The communication line GSM is simulated and implemented using the encoder GSMEFR (Enhanced Full Rate). The encoder GSMEFR corresponding to GSM 6.60 protocol is currently used by mobile operators. It is based on the ACELP algorithm (Algebraic Code Excited Linear Prediction). The simulation was conducted under a tool box of MATLAB version 7.4. The automatic speaker recognition system (ASR) is developed based on the modeling approach of mixing functions of Gaussian probability density GMM (Gaussian Mixture Models). A study on the influence of GSM speech coding on ASR system performance was conducted with different databases, initially, we used the original ARADIGIT16K database and its 16KHz sampled version, for the optimal structure of rating system ASR. Second, this database was down-sampled to 8KHz with the Praat software, leading to the ARADIGIT8K database. This ARADIGIT8K database was passed through the GSM codec to get the GSM transcoded database (only source) to know the ARADIGIT_GSM database. The transmission channel has been simulated by two types of noise, AWGN (Additive White Gaussian Noise) and Rayleigh noise levels at different SNR (Signal-to-Noise Ratio). The ARADIGIT16K database evaluation consists of 10 Arabic digits (zero to nine) spoken by sixty (60) speakers (31 men and 29 women). Each utterance is repeated three times by each speaker. The results obtained by using the different databases have shown a decrease in the rate identification of the speaker by 6% for the basic GSM transcoded database (only source) and by approximately 16% for GSM transcoded database (only source + channel simulated by AWGN noise), and about 24% for GSM transcoded database GSM (only source+channel simulated noise Rayleigh).

Abstract

The GSM (Global System Mobile) has become most means of widely used over the telephone channel. The speaker recognition through recordings of GSM speech type becomes a very important research topic. In this work, we have implemented a system for speaker recognition via a communication line phone. The communication line GSM is simulated and implemented using the encoder GSMEFR (Enhanced Full Rate). The encoder GSMEFR corresponding to GSM 6.60 protocol is currently used by mobile operators. It is based on the ACELP algorithm (Algebraic Code Excited Linear Prediction). The simulation was conducted under a tool box of MATLAB version 7.4. The automatic speaker recognition system (ASR) is developed based on the modeling approach of mixing functions of Gaussian probability density GMM (Gaussian Mixture Models). A study on the influence of GSM speech coding on ASR system performance was conducted with different databases, initially, we used the original ARADIGIT16K database and its 16KHz sampled version, for the optimal structure of rating system ASR. Second, this database was down-sampled to 8KHz with the Praat software, leading to the ARADIGIT8K database. This ARADIGIT8K database was passed through the GSM codec to get the GSM transcoded database (only source) to know the ARADIGIT_GSM database. The transmission channel has been simulated by two types of noise, AWGN (Additive White Gaussian Noise) and Rayleigh noise levels at different SNR (Signal-to-Noise Ratio). The ARADIGIT16K database evaluation consists of 10 Arabic digits (zero to nine) spoken by sixty (60) speakers (31 men and 29 women). Each utterance is repeated three times by each speaker. The results obtained by using the different databases have shown a decrease in the rate identification of the speaker by 6% for the basic GSM transcoded database (only source) and by approximately 16% for GSM transcoded database (only source + channel simulated by AWGN noise), and about 24% for GSM transcoded database GSM (only source+channel simulated noise Rayleigh).

Table des matières

Liste des Figures

Liste des tableaux

Liste des abréviations

GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
DTW	Dynamic Time Warping
VQ	Vector Quantization
MSSO	Méthodes Statistiques du Second Ordre
LPC	Linear Predictive Coding
LPCC	Linear Predictive Cepstrum Coefficient
EMV	Estimation du Maximum de Vraisemblance
EM	Expectation-Maximization
MFCC	Mel Frequency Cepstral Coefficient
TCDI	Transformée en Cosinus Discrète Inverse.
TFD	Transformée de Fourier Discrète
RAL	Reconnaissance Automatique du locuteur
RAR	Reconnaissance Automatique de la parole
VAL	Vérification Automatique du locuteur
IAL	Identification Automatique du locuteur
RSB	Rapport Signal sur Bruit
GSM	Global System for Mobile communication
FR	Full Rate
HR	Half Rate
EFR	Enhanced Full Rate
ACELP	Algebraic Code Excited Linear Prediction
CELP	Code Excited Linear Predictor
LTP	Long Term Prediction
AWGN	Additive White Gaussian Noise

Introduction générale

Introduction Générale

La caractérisation automatique du locuteur est un vaste domaine dans lequel la ‘’machine’’ a pour tâche d'extraire du signal de parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc.

La Reconnaissance Automatique du Locuteur (RAL) est un sous-problème de la caractérisation automatique du locuteur, elle consiste à reconnaître l'identité d'un individu à partir de sa voix **[1] [2].** Les applications des systèmes de RAL se distinguent par leur contexte applicatif et leur niveau de sécurité. Ces contraintes peuvent être prises en compte pour la définition d'une tâche spécifique de la RAL. Il est communément admis de regrouper ces tâches dans des grandes catégories : identification, vérification et indexation. L'application téléphonique (téléphone, Internet), pour des services à distance, est intéressante car la reconnaissance automatique du locuteur est l'unique accès biométrique offrant cette alternative. De plus, elle propose des services dans divers domaines entre autres la validation des transactions bancaires par téléphone (pour venir en complément d'un code d'accès confidentiel et améliorer ainsi le service bancaire) **[3].**

Le GSM (Global System for Mobile communications) est le standard pour la téléphonie mobile dans toute l'Europe. La norme propose trois types de codage de la parole désignés par GSMFR (Full Rate), GSMHR (Half Rate) et GSMEFR (Enhanced Full Rate) **[4].** Le rôle des codecs GSM est de compresser le signal de parole avant sa transmission, de manière à réduire le nombre de bits nécessaires à sa représentation, tout en maintenant une qualité de signal acceptable en sortie. Il s'agit dans tous les cas d'une compression avec pertes, affectant les propriétés spectrales du signal. Dans ce travail, on s'intéresse à l'influence du transcodage GSM sur la performance d'un système de reconnaissance du locuteur. Pour cela, on doit disposer de bases de données significatives de signaux réels GSM utilisés dans les télécommunications. Malheureusement, nous ne disposons pas pour l'heure de telles bases, c'est ainsi que nous avons opté pour une simulation du codec GSM sous le logiciel Matlab. Cette simulation est basée sur le codeur GSMEFR qui correspond au protocole GSM 6.60 actuellement utilisé par les opérateurs de téléphonie mobile. Le système de reconnaissance du locuteur mis en œuvre est lui basé sur l'approche de modélisation par mélange de fonctions de densité de probabilité gaussiennes GMM (Gaussian Mixture Models) pour la modélisation du locuteur. En effet la modélisation par un mélange de gaussiennes (GMM) est très indiquée pour la reconnaissance du locuteur en mode indépendante du texte.

Ce mémoire est organisé comme suit :

Le premier chapitre présente une introduction à la reconnaissance automatique du locuteur. Sont également présentées, les différentes étapes du système de RAL, les paramètres les plus efficaces pour représenter le signal de parole, les tâches de la RAL (l'identification, la vérification et l'indexation en locuteurs) et enfin les méthodes de reconnaissance les plus utilisées actuellement.

Dans le deuxième chapitre nous présentons le formalisme des modèles de Mélange des Gaussiennes (GMM), leur théorie, leurs principes pour la modélisation du locuteur ainsi que les algorithmes d'apprentissage des modèles et d'identification y afférents.

Dans le troisième chapitre, nous présentons les différentes normes GSM utilisées pour le codage de la parole et notre implémentation du codec GSMEFR sous les boites outils MATALB version 7.4

Dans le dernier chapitre 4, nous présentons le contexte expérimental, la structure optimale du système d'identification du locuteur et les résultats obtenus avec les différentes bases de données mises au point ainsi que l'évaluation objective du système.

Nous terminerons notre mémoire par une conclusion générale dans laquelle nous donnerons quelques perspectives de recherche dans ce domaine.

Reconnaissance Automatique du Locuteur

Chapitre 1

Reconnaissance Automatique du Locuteur

1.1 Introduction

La Reconnaissance Automatique du Locuteur (RAL), contrairement à la Reconnaissance Automatique de la Parole (RAP), s'intéresse tout particulièrement aux informations extra-linguistiques véhiculées par le signal vocal (signal de parole). La RAL a très souvent bénéficié des avancées de la RAP adaptées au domaine de la RAL. Les applications de la RAL sont principalement liées aux problèmes d'authentification et de confidentialité.
Ce chapitre est une introduction au domaine de la RAL, nous présentons des notions concernant l'authentification biométrique et les différentes tâches liées à la RAL telles que l'Identification, la vérification automatique du locuteur, l'indexation du locuteur et le suivi du locuteur.

1.2 L'authentification biométrique

Biométrie : bios=vivant , metron=mesure

Pour identifier une personne, trois approches (éventuellement complémentaires) sont possibles:

- Utiliser un identifiant : ce que l'on possède (carte, badge, document) ;
- Utiliser une connaissance : ce que l'on sait (un mot de passe) ;
- Utiliser une biométrie : ce que l'on est.

Les deux premiers points diffèrent de la biométrie au sens où ils sont concernés par la possession d'un élément (mots de passe, ou clés) qui peut être utilisé par des fraudeurs pour usurper l'identité d'un tiers, tandis que la biométrie correspond à des critères morphologiques de l'être humain, qui doivent être uniques pour chaque personne.

1.2.1 La biométrie

Le terme biométrie est défini dans le Petit Robert par la Science qui étudie à l'aide des mathématiques (statistiques, probabilités) les variations biologiques à l'intérieur d'un groupe déterminé. Le terme « biométrie » se réfère étymologiquement à la mesure du vivant.

Les modalités biométriques sont subdivisées selon un critère de variation dans le temps. Les biométries invariantes au cours du temps pour un individu, tenant de la morphologie humaine, sont souvent nommées « morphologiques » ou « statiques ». Celles qui présentent une grande variation temporelle sont nommées « comportementales » ou « dynamiques ».

Quelques exemples sont répertoriés ci dessous :

- morphologiques : l'iris, les empreintes digitales, les empreintes palmaires, le visage **[5][6].**

- comportementales : signature (image et dynamique), frappe au clavier, et la voix **[7][8].**

1.3 La reconnaissance du locuteur et applications

Comme on peut le constater sur la (figure 1.1), la reconnaissance automatique du locuteur s'inscrit dans le domaine plus général du traitement de la parole. Elle exploite la variation inter locuteurs et s'intéresse aux informations extra linguistiques du signal vocal.

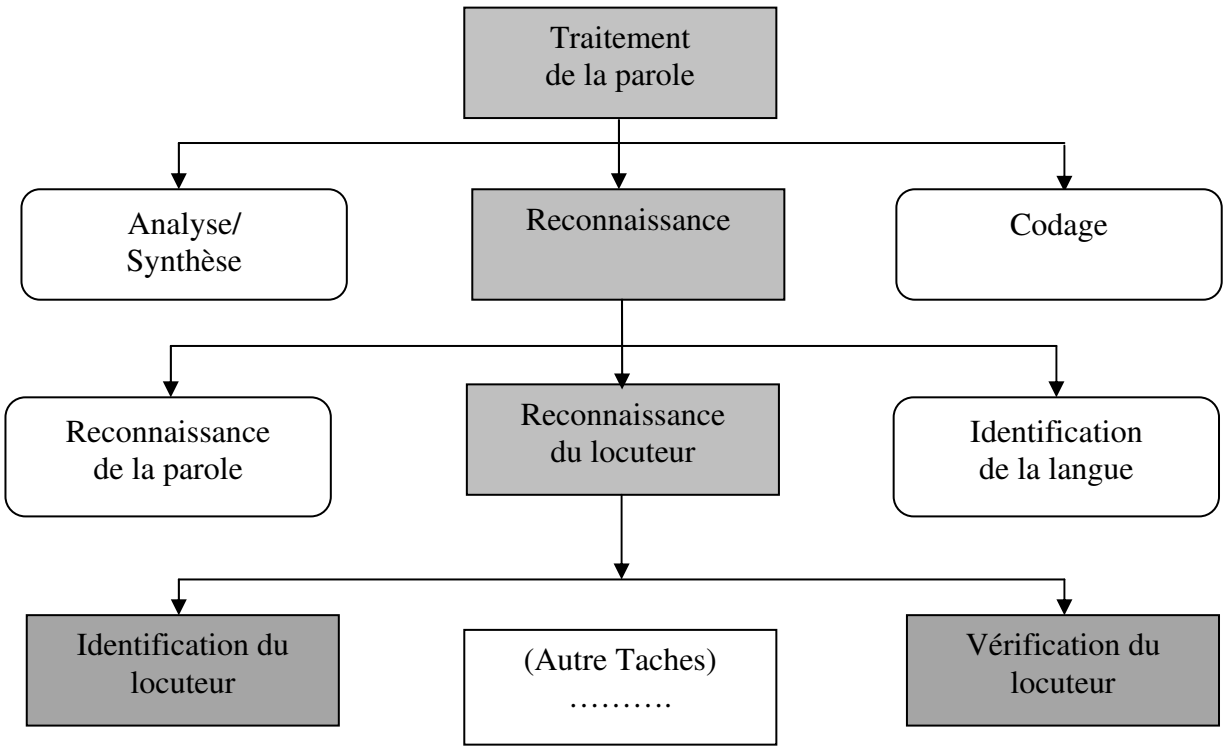

Figure 1.1 : Les différentes tâches du Traitement de la parole

Un message parlé ne véhicule pas seulement le sens de ce que veut exprimer l'individu qui l'émet, il porte également des informations sur l'individu lui-même et en premier lieu sur son identité. Par conséquent, il est important de ne pas confondre reconnaissance du locuteur et reconnaissance de la parole, dans un cas on cherche à déterminer l'identité d'un individu

grâce à sa voix et dans l'autre cas on cherche le contenu linguistique du signal émis. Les applications directes de la RAL concernent les problèmes de confidentialité et d'authentification.

Nous distinguerons :

- les applications "sur site" : serrures vocales pour contrôle d'accès, cabines bancaires en libre service, etc.

- les applications liées aux télécommunications : ces applications concernent l'identification du locuteur à travers le réseau téléphonique pour accéder à un service de transactions bancaires à distance ou pour interroger des bases de données en accès privé.

- les applications judiciaires (*forensic applications*) : recherche des suspects, d'éléments de preuve, de preuves, Tests auditifs par des experts, Lecture de spectrogrammes par des experts, etc, [9] [10].

1.4. Les différents tâches de la RAL

On distingue quatre tâches différentes en reconnaissance du locuteur : les tâches d'*identification du locuteur et de vérification du locuteur* et les tâches d'*indexation en locuteurs et de suivi de locuteur.*

1.4.1. Identification automatique du locuteur

L'Identification Automatique du Locuteur (IAL) consiste à reconnaître le vrai locuteur parmi une population (ou base) composée de N locuteurs connus. L'entrée du système est l'échantillon de parole d'un locuteur inconnu. La sortie du système correspond à l'identité du locuteur de la base de référence qui est le plus "proche" du signal de parole inconnu. Dans cette tâche, on fait l'hypothèse que le signal de parole à identifier est prononcé par un des locuteurs de la base de référence **[11] [12].**

D'un point de vue schématique (Figure 1.2), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est ''comparée'' 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.

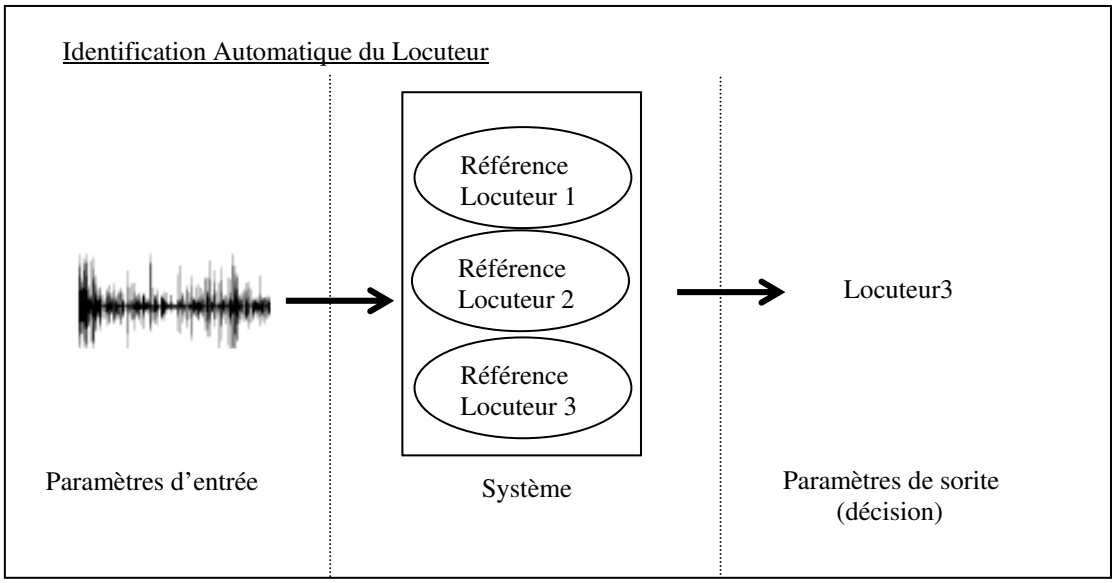

Figure. 1.2: La tâche d'identification Automatique du Locuteur (IAL).

1.4.2. Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) consiste à vérifier l'identité proclamée par un individu (Figure 1.3) par la comparaison d'un signal vocal et d'un modèle de référence du locuteur présumé, préalablement appris par le système. Un système de VAL a donc deux entrées : une identité et un accès de test. Le résultat de cette comparaison est considéré comme une mesure de similarité avant d'être comparé à un seuil d'acceptation. Lorsque la mesure de similarité est supérieure à ce seuil, l'individu est accepté, il est rejeté dans le cas contraire **[13][14].**

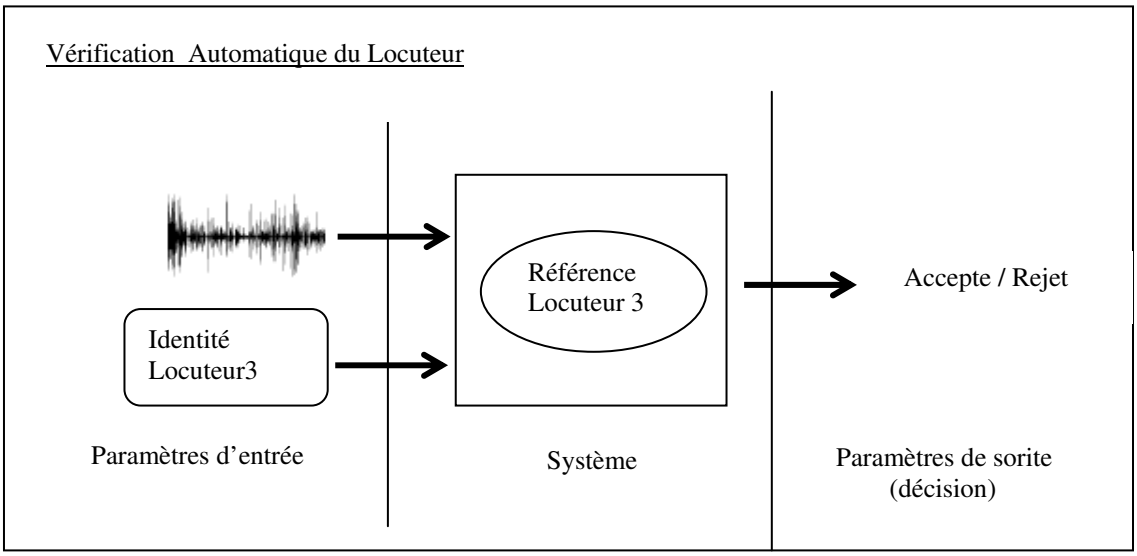

Figure. 1.3: La tâche de vérification Automatique du Locuteur (VAL).

1.4.3 Indexation en Locuteur

L'indexation en locuteur de document audio regroupe plusieurs tâches : segmentation, regroupement et suivi de locuteur. La segmentation en locuteur consiste à découper le signal audio en segments homogènes du point de vue du locuteur. Le regroupement en locuteur consiste à déterminer le nombre des locuteurs présents dans l'enregistrement et les interventions associées à chacun d'entre eux. Ces deux tâches réunies constituent l'indexation en locuteur proprement dite, nommée Speaker diarization. La difficulté de la tâche provient du manque d'information a priori : nombre de locuteurs, identités ou un échantillon de voix pour établir une première référence ne sont pas disponibles. Au niveau applicatif, cette tâche s'inscrit dans le cadre des applications d'indexation en locuteurs de bases de données multimédia (télévisuelles, radiophoniques, réunions**...)[15][16]**.

1.4.4. Suivi de locuteur

La tâche de suivi de locuteur **[17][18]** consiste, à partir d'un ensemble de locuteurs référencés, à délimiter et à identifier les zones du signal où chacun des locuteurs est intervenu. Il est possible d'avoir un seul locuteur à suivre dans un document où de multiples locuteurs interviennent. Les applications correspondant à ce type de tâche intéressent les gouvernements et les services de renseignement dans le cadre d'écoutes téléphoniques mais aussi la recherche d'information dans les documents multimédia.

1.4.5 Dépendance et Indépendance au Texte

Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte. En mode dépendant du texte, la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, phrase...) est connu du système. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne.

1.5 Les variabilités du signal vocal

Les informations transmises par le signal de parole sont multiples [19]. La variabilité du signal de parole entre locuteurs est majoritairement utilisée en reconnaissance du locuteur pour reconnaître les individus. Il s'agit de la variabilité inter locuteurs. La capacité des systèmes de RAL à authentifier une personne repose essentiellement sur la capacité à discriminer les individus grâce à cette variabilité. Mais d'autres facteurs de variations altèrent la parole. Le signal de parole est par exemple considéré comme non reproductible par son locuteur. Il existe une variabilité propre au locuteur dépendante de son état physique mais aussi psychologique. Il s'agit de la variabilité intra-locuteur. De plus les conditions d'environnement influent sur l'onde de parole; les bruits additifs ambiants ou les bruits de convolution engendrés par la prise de son modifient le signal de parole.

1.5.1 Variabilités inter-locuteurs

Le signal de parole permet la communication entre les individus. Il véhicule un message linguistique mais aussi quantités d'informations extra linguistiques et des informations liées au locuteur. La personnalité, au sens large, du locuteur influence la production du signal de parole. Ceci permet notamment de discriminer les locuteurs. Le signal de parole est un signal très complexe où se mêlent différents types d'informations classées par leur « niveau » de représentation. Les informations dites « bas-niveau » sont facilement utilisables à partir de l'analyse numérique du signal de parole. Elles regroupent des informations liées principalement à des traits physiques de l'individu (facteurs morphologiques et physiologiques). Les informations de « haut-niveau », comme la linguistique ou l'état émotionnel du locuteur sont beaucoup plus complexes à caractériser. Ces informations sont relatives aux facteurs socio-culturels de l'individu [**20**] propose sur une hiérarchie composée sur six niveaux d'informations différents :

- le niveau acoustique : les paramètres sont liés à l'analyse de l'enveloppe spectrale du signal ;
- le niveau prosodique qui désigne la « mélodie » de l'énoncé de parole ;
- le niveau phonétique : la distinction des différents sons identifiables d'une langue ;
- le niveau idiolectal qui se rapporte aux particularités langagières propres à un individu ;
- le niveau dialogue qui définit la façon de communiquer d'un individu, comme ses temps de parole dans une conversation ;
- enfin, le niveau sémantique qui caractérise la signification du discours.

Le niveau acoustique est le plus utilisé en RAL où l'influence de l'anatomie du locuteur sur l'émission du signal de parole est retenue. Ces approches se basent sur une représentation numérique de l'enveloppe du signal, définie comme une suite de paramètres, les cepstres. Ces informations, présentes au niveau de l'enveloppe du signal, sont facilement extraites. Les niveaux prosodique et phonétique sont basés sur la représentation du signal de parole à un niveau supérieur. La prosodie caractérise le style d'élocution du locuteur. Les méthodes pour analyser cette information sont plus complexes bien que souvent basées sur l'analyse numérique du signal. Le niveau phonétique implique l'utilisation d'une segmentation en phonèmes, le plus souvent réalisée par un reconnaisseur de parole. Ces approches sont de plus en plus employées en RAL, en combinaison avec les approches acoustiques. Enfin les niveaux idiolectal, dialogal et sémantique ne sont, à notre connaissance, pas utilisés en RAL.

1.5.2 Variabilités intra-locuteur

Les systèmes de RAL se basent sur le principe de la variation du signal de parole entre les individus. Il est cependant clair qu'une variation du signal de parole est aussi présente pour un même individu. Cette variation peut être due à beaucoup de facteurs dont :

- la nature intrinsèquement variable de la communication parlée ;
- des facteurs pathologiques de type fatigue, rhume,... ou émotionnels. Ces facteurs provoquent des altérations momentanées de la voix ;
- à plus long terme, une altération de la voix due à l'âge est présente chez tous les individus.

1.5.3 Facteurs de distorsion

Les facteurs de distorsion dans le signal de parole peuvent être de deux natures, la première provient des facteurs environnementaux, la seconde provient des changements des conditions d'acquisition et de transmission entre deux enregistrements. Lorsque la chaîne de transmission est maîtrisée et stable d'un enregistrement à un autre, les variations de celle-ci ne posent pas de problèmes particuliers et toute normalisation par rapport à ces variations est peu utile. Dans un cadre applicatif, ce n'est souvent pas le cas et cette variation doit être modélisée. Le protocole opératoire peut apporter, de par sa définition, des variabilités entre deux enregistrements. Celles-ci apparaissent par la quantité de données disponible, ou par l'utilisation d'un microphone différent entre les deux enregistrements. Par exemple, pour une application embarquée dans un véhicule, l'enregistrement peut être fait chez l'utilisateur dans un environnement calme à l'aide d'un microphone de qualité alors que l'accès aux données se fait dans la voiture, la fenêtre potentiellement ouverte (environnement) avec un téléphone de qualité réduite (acquisition et transmission).

1.6 Structure d'un système de RAL

Un système de RAL, quelle que soit la tâche (IAL, VAL, suivi de locuteur, indexation en locuteur) possède trois modules qui sont l'analyse acoustique du signal de parole, la modélisation du locuteur et la décision.

- Le premier module de traitement du signal réalise l'analyse acoustique du signal de parole. A l'issue de cette étape, le signal est représenté par des vecteurs de coefficients, ce qui permet de réduire l'information en quantité et en redondance. Ces vecteurs sont éventuellement représentés par un modèle mathématique, on parle alors de méthodes paramétriques**.**

- Dans le module modélisation, on fait une modélisation des caractéristiques propres de chaque locuteur connu du système (modèle de locuteur). Cette modélisation est réalisée à partir des données d'apprentissage. Une mesure de similarité est ensuit calculé entre un modèle locuteur et un signal parole, puis transmise au module de décision.

- Le module de décision désigne le locuteur finalement reconnu.

La structure générale d'un système de reconnaissance automatique du locuteur est représentée sur la (Figure 1.4).

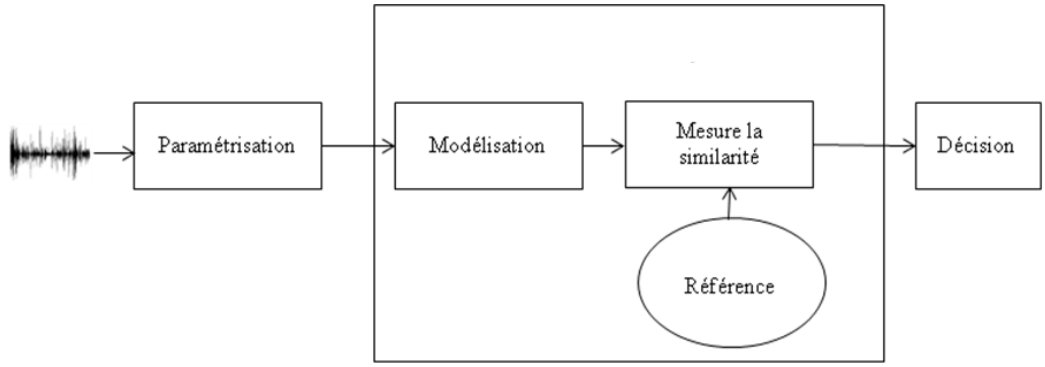

Figure. 1.4 : Structure générale d'un système RAL.

1.7 Paramétrisation (Analyse acoustique)

Le processus de paramétrisation consiste à extraire du signal de parole les informations pertinentes en vue de la reconnaissance. Le signal de parole, de par sacomplexité (multitudes d'informations et redondance), ne peut être exploité directement.Une représentation simplifiée du signal de parole est par conséquent nécessaire. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques calculéspériodiquement sur le signal de parole.

1.7.1 Paramètres de l'analyse spectrale

L'analyse acoustique du signal de parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. Généralement, on calcule un jeu de coefficients acoustiques à des intervalles de temps réguliers, sur des blocs de signal de longueur fixe. Ce jeu de coefficients constitue un vecteur acoustique. Les techniques de paramétrisation acoustique sont nombreuses. Néanmoins, on peut les regrouper en trois grandes familles :

- Les analyses par bancs de filtres : il s'agit d'une analyse assez simple du système auditif de l'homme qui consiste à calculer l'énergie du signal vocal dans différentes bandes de fréquences.

- Les analyses à base de transformée de Fourier : les coefficients issus de la transformée de Fourier peuvent être utilisés pour une analyse en bancs de filtres ainsi que pour les calculs de coefficients cepstraux. Parmi les plus connus se trouvent les coefficients LFCC (Linear Frequency Cepstrum Coefficient) et les coefficients MFCC (Mel Frequency Cepstral Coefficients).

- Les analyses par prédiction linéaire [**21**]: les coefficients issus d'un modèle autorégressif (de l'analyse LPC) peuvent être utilisés comme paramètres caractéristiques. Parmi les plus connus se trouvent les coefficients LPCC (Linear Predictive Cepstrum Coefficient). Le lecteur pourra se reporter à **[22]** pour une présentation des diverses méthodes de codages de la parole.

1.7.2 Extraction des paramètres

Le signal de parole présente de la redondance et il est porteur de plusieurs types d'informations, ce qui justifie la recherche d'une représentation spécifiquement pertinente pour la reconnaissance. L'extraction des paramètres consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques. Cette extraction débute par une mise en forme du signal de parole suivie par le calcul de coefficients acoustiques comme indiqué sur la figure 1.5 suivante :

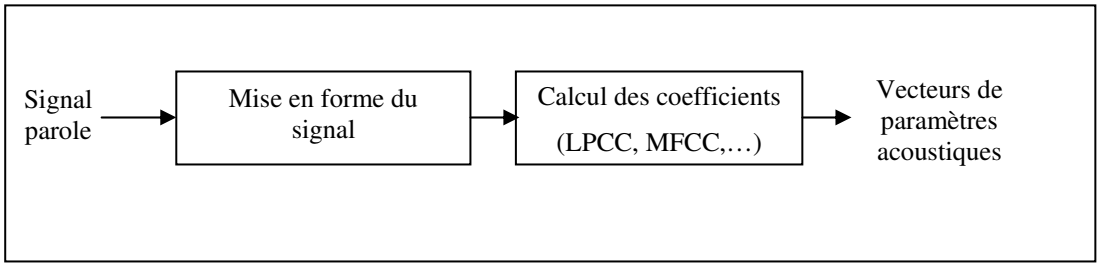

Figure 1. 5 : Schémas synoptique de l'extraction de paramètres acoustiques.

1.7.2.1 Mise en forme du signal de parole

L'étape de mise en forme de signal de parole consiste à lui appliquer une série d'opérations de traitement du signal dont le but est de l'adapter au calcul de coefficients acoustiques. La figure 1.6 illustre ces différentes opérations.

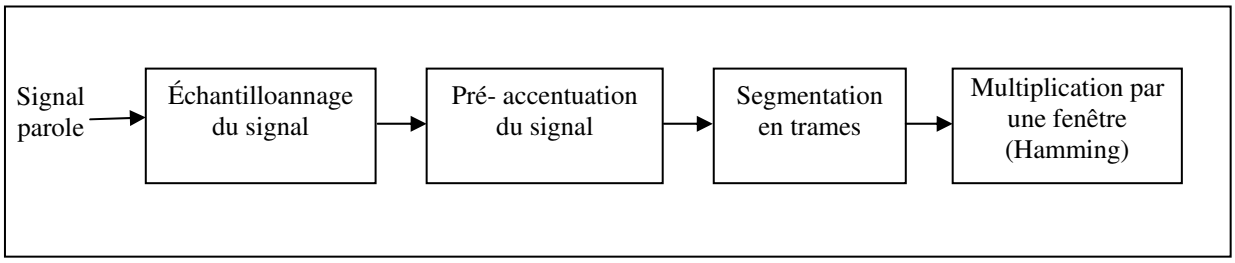

Figure 1.6 : Etapes de mise en forme du signal de parole.

a- Echantillonnage

Le signal de parole capté par le microphone étant analogique, il s'avère nécessaire de le numériser avant tout traitement. La numération du signal consiste à échantillonner ce signal puis à quantifier chaque échantillon. La fréquence d'échantillonnage doit verifier la condition de Shannon donnée comme suite :

$$f_e \geq 2 f_{\max} \tag{1.1}$$

f_e **:** Fréquence d'échantillonnage.

$f_{\max}$ **:** Fréquence maximale du signal a traiter**.**

b- Pré- accentuation

L'onde acoustique sortante des lèvres subit, à cause de la désadaptation entre les deux milieux intérieur et extérieur, une distorsion assimilable à une désaccentuation de 6 *dB* par octave sur tout le spectre**.** Pour pouvoir compenser cette distorsion, et accentuer les hautes fréquences, on applique un filtre de pré-accentuation simulée par un filtre passe haut de transmittance :

$$H(z) = 1 - az^{-1} \tag{1.2}$$

a **:** étant un coefficient de pondération généralement compris entre 0.95 et 1 **(** $0.95 \leq a \leq 1$ **).**

c- Segmentation en trames

Cette opération consiste à découper le flot de parole continue en trames pendant lesquelles le signal est supposé quasi-stationnaire. Chaque trame a habituellement une durée identique d'environ 20 à 30 ms, et le décalage entre deux trames consécutives est d'environ 15 à 10 ms.

d- Multiplication par une fenêtre

L'étape de fenêtrage consiste à appliquer au signal vocal une fenêtre glissante de durée limitée, et ce afin de limiter le nombre d'échantillons et de réduire les effets de bords (phénomène de Gibbs).
Parmi les différentes fenêtres de pondération, on retrouve : la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée.
Donc, pour limiter la durée d'un signal $x(n)$, on le multiple pour un autre signal $w(n)$ possédant *N* échantillons. Le signal de la trame résultante à *N* échantillons est donc :

$$s(n) = \sum_{k=1}^{N} x(n) w(n) \tag{1.3}$$

Avec $s(n)$: Signal de la trame

$x(n)$: Signal à fragmenter, n=1,….. , ∞

$w(n)$: Fenêtre de longueur N, k=*1,*……..*N*.

L'équation de la fenêtre de Hamming est donnée par l'expression suivante :

$$w(n) = 0.54 + 0.46\cos(\frac{2\pi n}{N-1}) \tag{1.4}$$

Avec $0 \leq n \leq N-1$

1.7.2.2 Les paramètres acoustiques

a- L'énergie du signal

L'énergie du signal est un indice qui peut contribuer à augmenter les performances d'un système de reconnaissance, cette énergie correspond à la puissance du signal. Elle est souvent évaluée sur plusieurs trames de signal successives pour pouvoir mettre en évidence ses variations. La formule de calcul de ce paramètre est :

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{1.5}$$

Comme autre variante de cette énergie, on peut aussi utiliser l'énergie logarithmique qui est définie comme suit :

$$E = \ln\left(\sum_{n=0}^{N-1} s^2(n)\right) \tag{1.6}$$

Où N est le nombre d'échantillons du signal et les $s(n)$ sont les échantillons du signal. L'énergie ainsi obtenue est sensible au niveau d'enregistrement; généralement elle est normalisée, exprimée en décibels (par rapport à un niveau de référence).

b- Les coefficients MFCC (Mel Frequency Cepstral Coefficients)

Les coefficients cepstraux issus d'une analyse par Transformée de Fourier, caractérisent bien la forme du spectre et permettent de séparer l'influence de la source glottique de celle du conduit vocal. Le cepstre du signal de parole est défini comme étant la Transformée de Fourier Inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal **[25].**

$$s(t) = e(t) * h(t) \tag{1.7}$$

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

L'application du logarithme sur le module de la Transformée de Fourier de *s(t)* dans l'équation (1.7) donne :

$$\log|S(f)| = \log|E(f)| + \log|H(f)| \tag{1.8}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{1.9}$$

Ce domaine est intéressant pour faire la séparation des contributions du conduit vocal et de la source d'excitation dans le signal de parole. En effet, si les contributions relevant du conduit vocal et les contributions de la source d'excitation évoluent avec des vitesses différentes dans le temps, alors il est possible de les séparer par l'application d'un simple fenêtrage dans le domaine fréquentiel (filtrage passe-bas) pour le conduit vocal **[26]**.
Les coefficients cepstraux les plus répandus sont les MFCC (Mel Frequency Cepstral Coefficients). Ils présentent l'avantage d'être faiblement corrélés entre eux, et donc on peut approximer leur matrice de covariance par une matrice diagonale. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'informations utiles dans le signal de parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark **[27].**

- Echelle Mel

$$Mel(f) = 2595\left(1 + \frac{f}{700}\right) \tag{1.10}$$

- Echelle Bark

$$Bark(f) = 6\,Arc\sinh\left(\frac{f}{1000}\right) \tag{1.11}$$

f représente la fréquence [Hz].

La procédure de calcul des coefficients MFCC est illustrée dans la (figure 1.7)

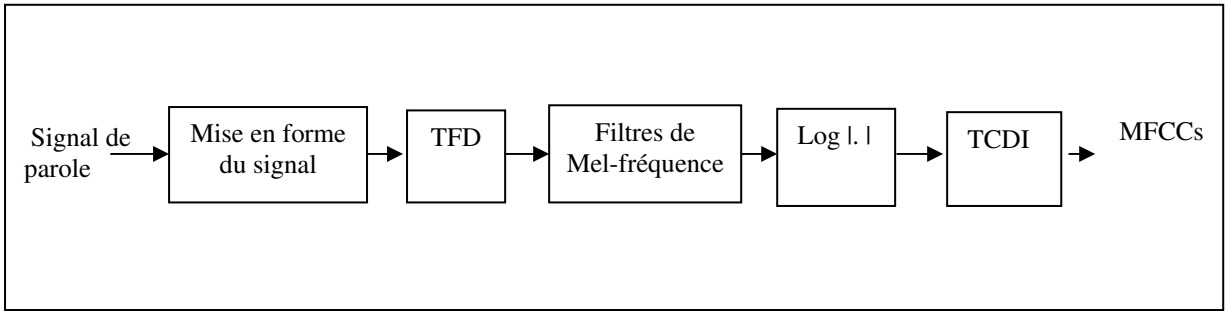

Figure. 1.7 Calcul des coefficients $MFCC_s$

Après la mise en forme du signal, une TFD (Transformée de Fourier Discrète) est calculée pour faire passer le signal de parole dans le domaine spectral :
Pour un signal discret $\{s[n]\}$ avec $0 < n < N$, *où* N est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète $S[k]$ est obtenue par :

$$S(k) = \sum_{n=à}^{N-1} s(n) \exp\left(\frac{-2j\pi nk}{N}\right), \ 0 \le k \le N-1 \tag{1.12}$$

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

$$B_m = B_b + m\frac{B_n - B_b}{M+1}, \ 0 \le m \le M+1 \tag{1.13}$$

M : Le nombre de filtres.
B_h : La fréquence la plus haute du signal.
B_b : La fréquence la plus basse du signal.

Dans le domaine fréquentiel, les points f_m discrets correspondants sont calculés par équation :

$$f_m = B^{-1}\left(B_b + m\frac{B_h - B_b}{M+1}\right) \tag{1.14}$$

Où $B^{-1}(x)$ désigne la fréquence correspondante à la fréquence x sur l'échelle Mel,

$$B^{-1}(x) = 700\left(10^{\frac{x}{2595}} - 1\right) \tag{1.15}$$

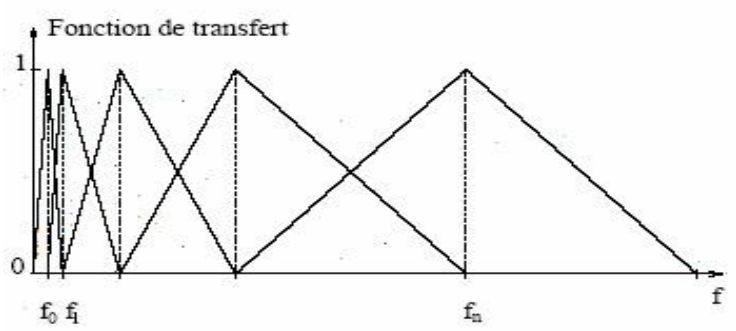

Figure. 1.8 Banc de filtres sur l'échelle linéaire

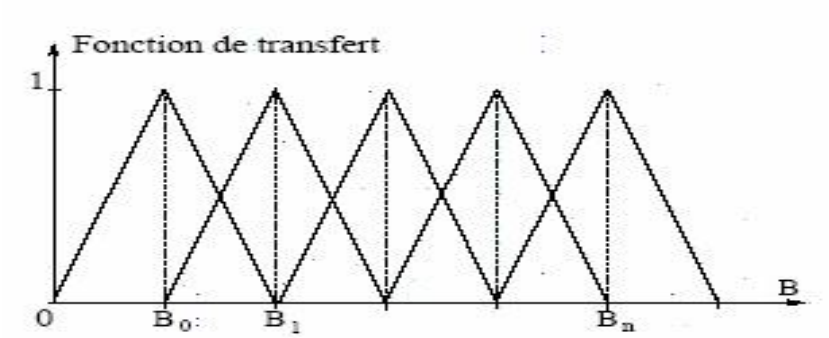

Figure. 1.9 Banc de filtres sur l'échelle Mel.

Les coefficients cepstraux de fréquence en échelle Mel (MFCC) peuvent être obtenus par une Transformée de Fourier Inverse à partir des énergies d'un banc de filtres. Les *d* premiers coefficients cepstraux peuvent être calculés directement à partir du logarithme des énergies *Ei* issues d'un banc de *M* filtres par la transformée en cosinus discrète définie par :

$$c_k = \sum_{i=1}^{M} \log E_i \cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2}\right)\right], \quad 1 \le k \le d \tag{1.16}$$

Le coefficient c_0 qui est la somme des énergies n'est pas utilisé, il est éventuellement remplacé par le logarithme de l'énergie totale *E* normalisée calculée dans le domaine temporel.

c- Les coefficients différentiels

Pour prendre en compte la dynamique temporelle du signal de parole, on utilise en plus des paramètres cités précédemment, des coefficients différentiels du premier ordre et du second ordre issus des coefficients cepstraux, de l'énergie ou de tout autre paramètre. Soit $c_k(t)$ le coefficient cepstral d'indice *k* de la trame t, alors le coefficient différentiel $\Delta c_k(t)$ correspondant est calculé sur $2n_\Delta + 1$ trames par :

$$\Delta c_k(t) = \frac{\sum_{i=n_\Delta}^{n_\Delta} i c_k(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{1.17}$$

La dérivée première de l'énergie ΔE est calculée de la même façon par :

$$\Delta E(t) = \frac{\sum_{i=n_\Delta}^{n_\Delta} iE(t+i)}{\sum_{i=n_\Delta}^{n_\Delta} i^2} \tag{1.18}$$

Les coefficients différentiels du second ordre peuvent aussi contribuer à l'amélioration des systèmes de reconnaissance. Les coefficients $\Delta\Delta ck$ et $\Delta\Delta E$ sont calculés par régression linéaire des coefficients Δck et ΔE respectivement, et sur $n\Delta\Delta$ (typiquement $n\Delta = n\Delta\Delta = 2$).

1.7.2.3 Paramètres prosodiques

Les paramètres prosodiques illustrent en grande partie le style d'élocution d'un locuteur : vitesse d'élocution (débit), durée et fréquence des pauses, ... ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement,...).
Néanmoins, ces paramètres caractéristiques du locuteur, notamment la fréquence fondamentale et ses variations, ne sont pas suffisamment discriminants pour être utilisés seuls dans un système de RAL. Ils sont généralement associes aux paramètres de l'analyse spectrale pour améliorer les performances des systèmes de RAL.

1.8 Modélisation

La modélisation des caractéristiques de chaque locuteur connu du système (modèles de locuteurs ou modèles clients) est réalisée à partir données d'apprentissages collectées au cours des sessions d'enrôlement.

On peut distinguer quatre grandes approches pour la construction des modèles clients :

- L'approche vectorielle : le locuteur est représenté par un ensemble de vecteur de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW (Dynamic Time Warping) et par quantification vectorielle.

- L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. 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.

- L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux neurones.

- L'approche relative : il s'agit de modéliser un locuteur relativement par rapport à d'autre locuteur de référence dont les modèles sont bien appris.

1.8.1 L'approche vectorielle

Dans l'approche vectorielle, un modèle de locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique construit lors de la phase de parametrisation des signaux d'apprentissage. Lors de la reconnaissance, une distance entre cet ensemble de vecteurs et les vecteurs de paramètres issus des signaux de test est calculée.
L'approche vectorielle compte deux grandes techniques : la programmation dynamique et la quantication vectorielle

1.8.1.1 Programmation dynamique

La programmation dynamique (Dynamic Time Warping : DTW) consiste a aligner temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage, le modèle de locuteur est tout simplement l'ensemble des vecteurs de paramètres obtenus après parametrisation des signaux d'apprentissage. Une distance est calculée entre vecteurs d'apprentissage et de test et moyennée sur l'ensemble de la séquence.

De par son principe, la programmation dynamique est utilisée exclusivement en mode dépendant du texte **[28], [29], [30].** Très rapide et montrant des performances relativement bonnes, la programmation dynamique est toutefois très sensible à la qualité d'alignement et notamment au choix du point de départ.

1.8.1.2 Quantification vectorielle

La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces. Chaque sous-espace est associé à un vecteur centroide i.e. à un vecteur de paramètres représentant l'ensemble des vecteurs composant le sous-espace. Dans ces conditions, un modèle de locuteur est composé d'un ensemble de vecteurs centroides, appelé dictionnaire de quantification (codebook).

Lors de la phase de reconnaissance, une distance est calculée entre un vecteur de test et chaque vecteur centroide du dictionnaire. La distance minimale est assignée au vecteur de test. La distance d'une séquence de vecteurs de test est obtenue par moyenne des distances minimales attribuées à chacun des vecteurs de test.

La quantification vectorielle s'applique en mode dépendant ou indépendant du texte **[31], [32], [33].** La rapidité et les performances de cette technique dépendent fortement de la taille du

dictionnaire : plus la taille du dictionnaire augmente, meilleures sont les performances ; néanmoins, le processus devient d'autant plus lent.

1.8.2 L'approche statistique

L'approche statistique consiste à représenter une séquence de vecteurs acoustiques issus de la parametrisation par des statistiques à long terme. Les premiers travaux suggèrent d'utiliser les paramètres du spectre moyen à long terme comme seul modèle des locuteurs **[34].** Lors de la reconnaissance, le spectre moyen estime sur les vecteurs de test est compare, à l'aide d'une distance spectrale, au spectre moyen issu de l'apprentissage.

Par la suite, l'approche statistique à été enrichie par l'introduction de statistiques d'ordre supérieur (statistiques d'ordre 2) qui permettent notamment de caractériser la variation des paramètres acoustiques (matrice de covariance).

1.8.2.1 Méthodes Statistiques du Second Ordre

Cette partie présente une famille de mesures de similarité entre locuteurs. Ces mesures reposent sur les caractéristiques du second ordre d'une séquence de vecteur. C'est-à-dire sur le vecteur moyen et la matrice de covariance de cette séquence. Plusieurs mesures de distance ont été utilisées, on peut citer : Le rapport de vraisemblance, la distance de *kulbak-leibler*, maximum de vraisemblance, test de sphéricité, déviation absolue des valeurs propres. Ces mesures donnent des résultats très encourageant sur la parole propre, et naturellement, voient leurs performances se dégrader sur la parole téléphonique. De part leur relative simplicité, ces mesures peuvent également servir de référence pour évaluer la qualité d'une base de données **[35] [36].**

1.8.2.2 Les mélanges de gaussiennes

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour *Gaussian Mixture Modes*) consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes **[37]** ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble être bien prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs. Et il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes sont considérés comme un cas particulier des HMM et une extension de la quantification vectorielle.

1.8.2.3 Modèles de Markov cachés

Les modèles de Markov (ou HMM pour *Hidden Markov Models*) ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Dans cette approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme

acoustique ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur. Le modèle est constitué de l'association d'une chaine de Markov, une succession d'états avec des probabilités de transition d'un état à autre, et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état) **[38] [39].**

1.8.3 L'approche connexionniste

L'approche connexionniste, telle que nous l'entendons ici, repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs clients afin que ce dernier apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle client se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée. Différents types de modèles de réseaux sont proposes dans la littérature **[40] [41]**. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée.

Le principal inconvénient de l'approche connexionniste est la complexité d'apprentissage. En outre, elle pose le problème de l'ajout d'un nouveau client qui nécessite dans la majorité des mises en œuvre le ré-apprentissage de tous les modèles. En effet, une nouvelle phase de classification est nécessaire afin de prendre en compte le nouveau client au sein du processus de discrimination entre locuteurs.

1.8.4 L'approche prédictive

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes. De par ce concept, cette approche est considérée dans la littérature comme une approche dynamique i.e. une approche tenant compte des informations dynamiques véhiculées par le signal de parole. Elle s'appuie principalement sur l'estimation d'une fonction de prédiction, propre à chaque locuteur et apprise sur les signaux d'apprentissage. Lors de la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test. L'erreur de prédiction moyenne constitue alors la mesure de similarité entre le signal de test et le modèle de locuteur (fonction de prédiction). Une autre solution envisagée est d'estimer une fonction de prédiction sur la séquence de test et de la comparer, à l'aide d'une distance, à la fonction de prédiction estimée lors de l'apprentissage**.**

1.9 Décision

La structure générale commune entre les différentes tâches d'un système de RAL, Le module de décision est différente selon la tâche choisie. Nous allons présenter le module de décision ci-dessous pour les tâches de vérification et d'identification.

1.9.1 Identification Automatique du Locuteur

En identification, un signal de test est comparé à toutes les références de locuteurs connus du système, résultant en un ensemble de mesures de similarité en entrée du processus de décision détaillé au chapitre 2. Aussi, ce processus a pour tâche de rechercher la mesure de similarité maximale (ou minimale dans le cas de mesures de distance) et de designer l'identité du locuteur correspondant.
Dans ce contexte, la mesure des performances d'un système d'IAL est généralement donnée en termes de taux d'identification correcte.

1.9.2 Vérification Automatique du Locuteur

En vérification, le processus de décision consiste à comparer la mesure de similarité entre le signal de test et le modèle client à un seuil de décision. Si la mesure est supérieure au seuil, le locuteur est considéré comme un client et il est accepté. Dans le cas contraire, le locuteur est considéré comme un imposteur et il est rejeté.

Plusieurs solutions sont proposées dans la littérature pour mesurer les performances d'un système de VAL. Toutes s'appuient sur deux erreurs caractéristiques des systèmes de VAL que sont :

- l'erreur de fausse acceptation : le système reconnaît à tort l'identité revendiquée (acceptation d'un imposteur).

- l'erreur de faux rejet : le système rejette à tort l'identité revendiquée (rejet d'un client).

1.10 Conclusion

Dans ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes d'un système de RAL. Le système de reconnaissance de locuteur procède généralement en trois étapes : l'analyse acoustique du signal parole, la modélisation du locuteur et l'étape de décision. Dans notre travail nous avons utilisé pour l'étape analyse acoustique les coefficients MFCC. Pour la modélisation de locuteur, l'approche statistique GMM, utilisée en mode indépendant de texte et enfin l'étape de décision est basée sur le processus d'identification de locuteur.

Modèle de mélange de gaussiennes

Chapitre 2

Modèle de mélange de gaussiennes (GMM)

Les modèles de Mélange de lois Gaussiennes (GMM : Gaussian Mixture Models, en anglais) ont étés utilisés dans de nombreux domaines, par exemple pour le traitement et la reconnaissance des images ou de la parole. Dans le cadre de la reconnaissance du locuteur, un GMM modélise un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle de GMM à un modèle de Markov cachés (HMM : Hidden Markove Model, en anglais) à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est plus utilisée en ce qui concerne la reconnaissance du locuteur en mode indépendant du texte.

Dans ce chapitre, nous présenterons le modèle du mélange de gaussiennes. Nous aborderons ensuite le problème d'estimation de l'ensemble de ces paramètres et enfin nous détaillerons la phase de décision d'un système de reconnaissance de locuteur par GMM.

2.1 Introduction

L'utilisation des GMM_s pour la modélisation des locuteurs a été initiée par les travaux de thèse de Douglas Reynolds [**42**], cette approche a donné, depuis plus de 10 ans maintenant, les meilleures performances pour les systèmes de reconnaissance du locuteur en mode indépendant du texte basé sur l'approche probabiliste. La plupart des systèmes actuels utilisent une modélisation de locuteurs par GMM. Ceci n'est pas un hasard, deux raisons principales font des densités de mélanges de gaussiennes une modélisation incontournable pour la représentation de l'identité du locuteur.

La première raison est de nature intuitive, étant donné que les densités multimodales telles que les GMM permettent de modéliser les ensembles fondamentaux des paramètres acoustiques puisque rappelons que ces vecteurs acoustiques issus de la phase de paramétrisation du signal de parole en une distribution complexe. Les modèles simples monogaussiennes sont insuffisants pour prendre en compte tous les détails de cette distribution. Aussi, l'espace acoustique qui défini la voix du locuteur peut être caractérisé par un groupe de classe acoustique (nuages) représentant de larges événements phonétiques, ces classes acoustique sont le reflet des configurations générales du conduit vocal (déformations physique) qui caractérisent l'identité du locuteur. Le positionnement du 1^{er} nuage acoustique (zone de concentration de vecteurs acoustique) peut être représenté par un vecteur moyen, tandis que sa forme (dispersion : sphérique, ovale,..) et son orientation (horizontales, verticale, oblique gauche,…) sont définis par la matrice de covariance.

La deuxième raison, les densités de mélange de gaussiennes est qu'une combinaison linéaire de monogaussiennes est capable de représenter une large gamme de distribution. La puissance des GMM est son habilité à approximer fidèlement des distributions aléatoires que celles des paramètres acoustiques.

2.2 Définitions

En statistiques, on appelle densité mélange, ou loi mélange une fonction de densité qui est issue d'une combinaison linéaire de plusieurs fonctions de densité.

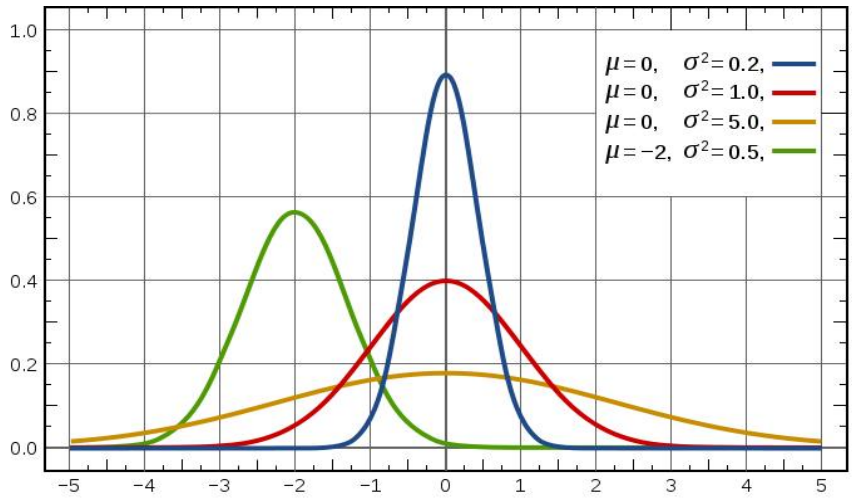

Figure 2.1 Mélanges des gaussiennes

Soit un individu *x*, (ou objet, ou événement, image, signal parole), une fonction $f(x/\mu,\Sigma)$, densité d'une variable aléatoire x représentée sur l'espace vectoriel R^d et paramétrée par (μ,Σ). Par exemple, si *f* est la densité d'un tel individu selon la loi normale (ou Gaussienne) de moyenne μ et matrice de covariance Σ , cette densité est donnée par :

$$f(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \qquad (2.1)$$

En superposant et pondérant *K* Gaussiennes, on définit un mélange de Gaussiennes.

On note habituellement π_k les poids avec $\sum_{k=1}^{K}\pi_k=1$, respectivement la moyenne μ_k et la matrice de covariance Σ_k de la *k*-ème composantes. On note également $\lambda_k(\pi_k,\mu_k,\Sigma_k)$, ainsi que $\lambda=\{\lambda_k\}$. La densité d'un individu *x* selon la distribution de probabilité paramétrée par λ est donnée par :

$$p(x|\lambda) = \sum_{k=1}^{K} \pi_k , f\left(x|\mu, \Sigma\right) \tag{2.2}$$

Un jeu de données (ou échantillon, ou *data* set en anglais) est une matrice dont chaque ligne caractérise un individu x_n. On note un tel échantillon $X = \{x_1, x_2, ..., x_N\}^T$; si on suppose que celui-ci est indépendamment et identiquement distribué, la probabilité jointe de cet échantillon est :

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda_n) \tag{2.3}$$

$\varphi = (\pi_1, \pi_2, ..., \pi_K, \lambda_1, \lambda_2, ...\lambda_K)$ le paramètre global du mélange.

La principale difficulté de cette approche consiste à déterminer le meilleur paramètre φ**.** Pour cela, on cherche habituellement le paramètre qui maximise la vraisemblance,

2.3 Approche statistique à base de GMM pour la modélisation

La problématique de la reconnaissance du locuteur peut être formulée dans un cadre statistique simple dont le principe est le suivant :

Soient $\lambda_1 \lambda_{2....} \lambda_S$ les S modèles probabilistes correspondants aux *S* locuteurs.

Soit $X = x_1 x_{2....} x_N$ la suite d'observations acoustiques (vecteurs acoustiques, par exemple de type MFCC) du locuteur à reconnaître.

La tâche de la RAL consiste à déterminer, parmi tous les modèles possibles ($\lambda_1 \lambda_{2....} \lambda_S$), le modèle λ, le plus probable connaissant les observations acoustiques X :

$$\lambda = \arg\max_{\lambda} \quad p(\lambda|X) \tag{2.4}$$

Grâce à la règle de Bayes, il est possible d'écrire :

$$p(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{p(X)} \tag{2.5}$$

Avec :

$p(X|\lambda)$: La probabilité d'observer la séquence *X* sachant le modèle λ.

$p(\lambda)$: la probabilité du modèle λ.

$p(X)$: représente la probabilité de la suite d'observations acoustiques X.

Comme *P(X)* ne dépend pas de λ et si l'on suppose que tous les modèles ont la même probabilité ($\forall \lambda \in \{\lambda_1, ...\lambda_S\}, p(\lambda) = \frac{1}{S}$), les deux équations précédentes s'écrirons de la façon suivante :

$$p(\lambda | X) = p(X | \lambda) \tag{2.6}$$

$$\lambda = \arg\max_{\lambda} \quad p(X | \lambda) \tag{2.7}$$

La probabilité $p(X|\lambda)$ est classiquement représentée par des modèles à mélange du gaussiennes.

2.3.1 Modèle à mélanges du gaussiennes

Un mélange de gaussiennes est une somme pondérée de *K* densité gaussiennes. Soit un locuteur *s* et un vecteur acoustique *x* de dimension *D*, le mélange de gaussiennes est défini comme suit :

$$p(x | \lambda_s) = \sum_{k=1}^{K} \pi_k^s b_k^s(x) \tag{2.8}$$

Où les $b_k^s(x)$ représentent les densités gaussiennes paramétrées par le vecteur de moyenne μ_k^S et matrice de covariance $\sum_k^S$:

$$b_k^S(x) = \frac{1}{(2\pi)^{D/2}.\left[\sum_k^S\right]^{1/2}} \exp\left[-\frac{1}{2}\left(x - \mu_k^S\right)'.\left(\sum_k^S\right)^{-1}.\left(x - \mu_k^S\right)\right] \tag{2.9}$$

et les π_k^S représentent les poids du mélange avec $\sum_k^S .\pi_k^S = 1$.

Un locuteur est donc modélisé par un ensemble de paramètres noté λ_S :

$$\lambda_s = \left\{\pi_k^s, \mu_k^s, \sum_k^s\right\}_{k=1,.....K} \tag{2.10}$$

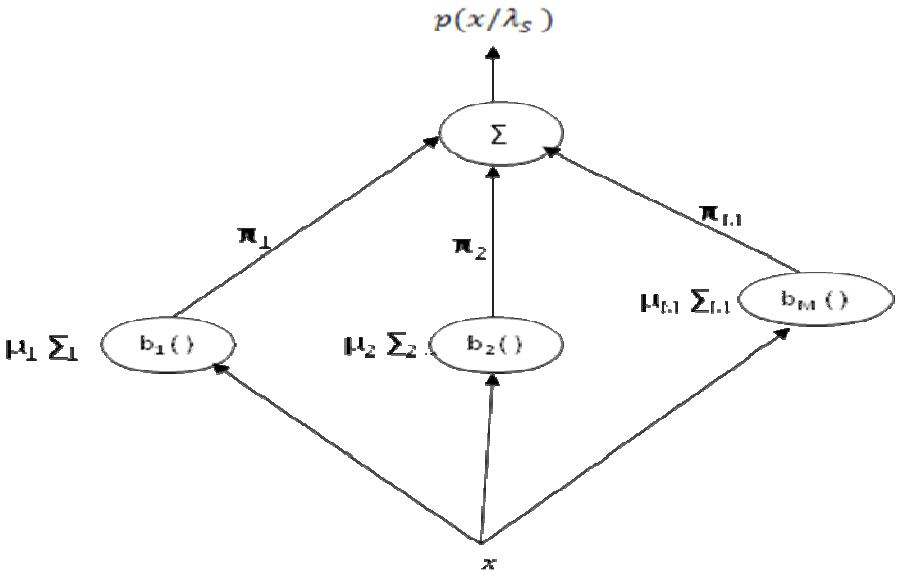

Figure 2.2 Modèle de Mélange de Gaussiennes

Les paramètres d'un modèle GMM, à savoir, le poids de chacune des *K* gaussiennes, le vecteur moyen et la matrice de covariance, sont entrainés dans un premier lieu sur une base d'apprentissage avant d'être utilisés pour identifier le locuteur.

La figure 2.3 illustre le principe de modélisation des vecteurs acoustiques par GMM

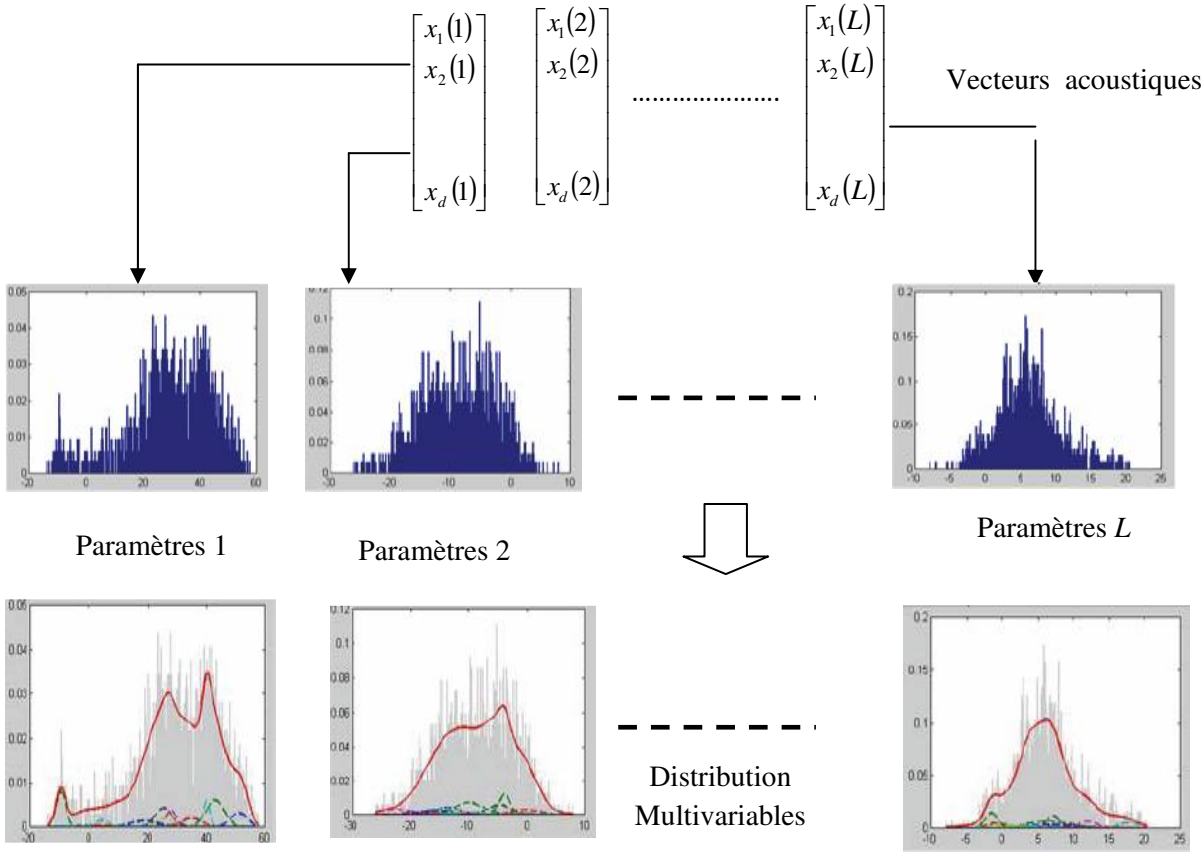

Figure 2.3 GMM pour modéliser des distributions

2.4 Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM de locuteur. La méthode conventionnelle est celle du Maximum de vraisemblance (MV)**,** dont le but est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage. Pour une séquence de N vecteurs d'apprentissages $X = (x_1, x_2, x_3 ....., x_N)$(suffisamment indépendants), la vraisemblance du modèle GMM est :

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(x_n|\pi_k, \mu_k, \Sigma_k) \qquad (2.11)$$

L'apprentissage se fait en deux étapes : la première étape est l'initialisation des paramètres du modèle en utilisant l'algorithme K-moyennes (k-means) ou l'algorithme LBG, la deuxième étapes est l'optimisation des paramètres obtenus dans la première étape en utilisant l'algorithme EM (*Expectation Maximisation*) [**43**][**44**].

2.5 L'initialisation des paramètres

2.5.1 Algorithme des K-moyennes

L'algorithme des K-moyennes consiste à faire la répartition des vecteurs acoustiques *x* d'une classe (locuteur) en *N* sous-ensemble disjoints caractérisées par un centroide. Le résultat de cette répartition est appelé dictionnaire. L'algorithme des K-moyenne n'est que localement optimal, par conséquence, il est influencé par ses conditions initiales.

Étape 1 : choisir un ensemble de vecteurs initiaux y_i, $1 \leq i \leq N_s$.

Étape 2 : Attribuer chaque élément de l'ensemble d'apprentissage (x^k, $1 \leq i \leq N_s$) à une des régions R_i en choisissant le centre de gravite y_i le plus proche (loi de sélection du plus proche voisin).

Étape 3 : Mettre à jour les centres de gravité y_i de chaque région R_i en calculant sa nouvelle position à l'aide des échantillons d'apprentissage attribué à cette région.

Étape 4 : Si la différence de distorsion globale *D* entre cette itération et la précédente est inférieur à un seuil prédéfini ε, alors Fin ; si non aller à l'étape 2

Cette procédure divise la tache de minimisation de la distorsion globale en deux étapes. La première suppose que les centres de gravité soient connus, elle réparti les échantillons d'apprentissage parmi les L régions possibles en fonction de la mesure de distorsion. La seconde suppose que les régions ou classes sont connus et calcule la nouvelle position du centre de gravité qui minimise la distorsion. La seconde suppose que les régions ou classe sont connues la nouvelle position du centre de gravité qui minimise la distorsion intra-classe. Il a été démontre que cet algorithme ne conduit qu'à un minimum local de la fonction de distorsion [**45**].

2.5.2 Algorithme LBG

C'est une version étendue d'algorithme K-moyennes. Il permet de résoudre l'un des problèmes associés à l'algorithme des K-moyennes, celui de l'initialisation des centres de gravité. En fait, l'algorithme LBG divise itérativement l'ensemble des échantillons d'apprentissage en 2,4 ,……. ,2^P régions, et calcule le centre de gravité de chacune des régions. La structure de cet algorithme est la suivante :

Étape 1 : on attribue la valeur 1 à *N* (nombre de classes ou régions) et on calcule le centre de gravité y_1 associé à l'ensemble des échantillons d'apprentissage.

Étape 2 : On perturbe légèrement le vecteur centre de gravité y_i en deux vecteurs $y_1^{(0)} = y_1+\mu$ et $y_2^{(0)} = y_1-\mu$ qui forment le dictionnaire initial de l'algorithme des K-moyennes pur construire un dictionnaire (CodeBook) de taille N=2.

Étape 3 : On perturbe à nouveau les deux Mettre à jour les centres de gravité y_i de chaque région R_i en calculant sa nouvelle position à l'aide des échantillons d'apprentissage attribuées à cette région.

Étape 4 : Si la différence de distorsion globale *D* entre cette itération et la précédente est inférieure à un seuil prédéfini ε, alors Fin ; si non aller à l'étape 2

2.6 Optimisation des paramètres

Après la phase d'initialisation des paramètres du modèle GMM, en utilisant l'algorithme K-means ou LBG, ces derniers doivent être optimisés au moyen d'un algorithme de type *EM* (Expectation Maximisation en anglais). L'idée principale de l'algorithme *EM* est, en commençant par les paramètres initiaux λ du modèle, on estime les nouveau paramètres $\overline{\lambda}$, telle que la vraisemblance du nouveau modèle soit supérieur ou égale à la vraisemblance du modèle initial. En d'autre terme, $p(X|\lambda) \geq p(X|\overline{\lambda})$ ou X est la séquence des vecteurs d'apprentissage.

2.6.1 Apprentissage par Maximum de Vraisemblance (MV)

2.6.1.1 L'algorithme Expectation Maximisation (EM)

L'algorithme EM fait intervenir à la fois des observations *X* et des variables manquantes (l'indice de la gaussienne $k = 1,\ldots, K$). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(B, B^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée à la base des paramètres courants [**46**] à savoir :

$$Q(\lambda,\lambda^{(t)}) = \sum\sum p(m|x_n,\theta^{(t)}).\log p(x_n,m|\theta) \tag{2.12}$$

Où θ désigne l'ensemble des paramètres à estimer $(\pi_k,\mu_k,\sum_k)$ et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération t. Ce qui donne, après calcul :

$$Q(\theta,\theta^{(t)}) = \sum\sum \gamma_{n,k}^{(t)}\left[\log\pi_k - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left|\sum{}_k\right|\right] - \sum_{k=1}^{K}\sum_{N=1}^{N}\gamma_{n,k}^{(t)}\left[\frac{1}{2}(x_n-\mu_n)\right] \tag{2.13}$$

Où $\gamma_{n,k}^{(t)}$ est une probabilité a posteriori estimée à l'itération *t* :

$$\gamma_{n,k}^{(t)} = \frac{\pi_k^{(t)}\,p(x_n|\mu_k^{(t)},\sum_k^{(t)})}{\sum_{k=1}^{K}\pi_k^{(t)}\,p(x_n|\mu_k^{(t)},\sum_k^{(t)})} \tag{2.14}$$

En supposant que $p(x_n|\theta)$ sont des densités gaussiennes à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient :

$$Q(\theta,\theta^{(t)}) = \sum_{k=1}^{K}\sum_{n=1}^{n}\gamma_{n,k}^{(t)}\log\pi_k - \frac{1}{2}\sum_{k=1}^{K}\sum_{n=1}^{n}\gamma_{n,k}^{(t)}\left[Cste+\log\sigma_k^2 + \frac{(x_n-\mu_k)^2}{\sigma_k^2}\right] \tag{2.15}$$

Où σ_k^2 est un élément diagonal de la matrice de covariance.

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire *Q* par rapport à chacun de ceux-ci. Le cas des poids des composantes de mélange π_k est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte qui existe sur ces paramètres $\left(\sum_{k=1}^{K}\pi_K = 1\right)$. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\pi_k^{(t+1)} = \frac{1}{N}\sum_{n=1}^{n}\gamma_{n,k}^{(t+1)} \tag{2.16}$$

En ce qui concerne les vecteurs des moyennes, on montre que les formules de réestimations sont données par :

$$\pi_k^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,k}^{(t+1)}}{\sum_{n=1}^{N} \gamma_{n,k}^{(t+1)}} \tag{2.17}$$

et pour les variances :

$$\sigma_k^2 = \frac{\sum_{n=1}^{N} \gamma_{n,k}^{(t)} \left(x_n - \mu_k^{(t)}\right)^2}{\sum_{n=1}^{N} \gamma_{n,k}^{(t)}} \tag{2.18}$$

2.6.2 Apprentissage par Maximum A Posteriori (MAP)

L'algorithme EM est un des algorithmes les plus importants et les plus puissants en estimation statistique. De plus, il bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum de la fonction de vraisemblance [**46**][**47**]. Cependant, ses limites apparaissent lorsqu'on dispose de peu de données. Donc, il est important d'introduire de l'information a priori. Par conséquent, on ne cherche plus à maximiser la vraisemblance des données mais plutôt la probabilité a posteriori.

Les formules de ré-estimation, pour une gaussienne *k*, sont les suivantes [**48**] :

- Les poids des gaussiennes :

$$\pi_k = \frac{n_k^0 + n_m}{\sum_{k=1}^{K} (n_k^0 + n_k)} \tag{2.19}$$

- Les vecteurs des moyennes :

$$\mu_k = \frac{n_k^0 \overline{X} + n_k \overline{X}}{n_k^0 + n_k} \tag{2.20}$$

- Les variances :

$$\sigma_k^2 = \frac{n_k^0 \overline{X_K^0 X_k^{0'}} + n_k \overline{X_k X_k'}}{n_k^0 + n_k} - \mu_k \mu_k' \tag{2.21}$$

Où n (respectivement n^0) représente le poids, $\overline{X}$ (respectivement $\overline{X^0}$) le moment l'ordre 1 et $\overline{XX'}$ (respectivement $\overline{X^0 X^{0'}}$) le moment l'ordre 2 des données à adapter X (respectivement des données initiales X^0)

L'apprentissage incrémental consiste à effectuer quelques itérations d'apprentissage sur les données d'adaptation en conservant l'information apportée par les données initiales X^0. Dans le cas où de nombreuses données sont disponibles, l'apprentissage incrémental (ou plus généralement l'estimateur MAP) converge vers les estimateurs du maximum de vraisemblance. Il permet d'obtenir de nouveaux modèles avec peu de données. Ces estimées seront plus fiables que celle obtenue par MV étant donné qu'elles intègrent des connaissances a priori. Cette approche est la plus utilisée en reconnaissance du locuteur en mode indépendant du texte [**49**].

2.7 Identification du locuteur

L'identification du locuteur consiste à associer un des modèles d'apprentissage au signal à identifier, ceci se réalise par le calcul de la vraisemblance entre la séquence d'observations acoustiques du signal à reconnaître et tous les modèles GMMs estimés dans l'étape d'apprentissage. Le locuteur reconnu est celui qui correspond au modèle GMM qui engendre la vraisemblance maximale.

Soit un groupe de S locuteurs, représentés par les modèles GMM : λ_1 λ_2,……., λ_S. L'objectif de la phase d'identification est de trouver, à partir d'une séquence observée X, le modèle qui a la probabilité a posteriori maximale, c'est-à-dire :

$$\hat{S} = \arg \max_{1 \le s \le S} \quad p(\lambda_s | X) \tag{2.22}$$

Ce qui donne, d'après la loi de Bayes :

$$\hat{S} = \arg \max_{1 \le s \le S} \quad \frac{p(X | \lambda_s)}{p(X)} p(\lambda_s) \tag{2.23}$$

En supposant l'équiprobabilité d'apparition des locuteurs $p(\lambda_s) = \frac{1}{S}$, la loi de classification devient :

$$\hat{S} = \arg \max_{1 \leq s \leq S} p\left(X \mid \lambda_s\right) \tag{2.24}$$

2.8 Conclusion

Les mélanges de gaussiennes constituent l'état de l'art en reconnaissance automatique du locuteur, en mode indépendant du texte. Il existe plusieurs techniques pour faire apprendre un modèle GMM. En premier lieu, l'initialisation des paramètres par l'algorithme K-moyennes (k-means) ou l'algorithme LBG. En deuxième étape, l'optimisation des paramètres par l'algorithme EM permet d'estimer les paramètres du modèle tout en offrant un formalisme théorique et une preuve de convergence. Malheureusement, ses limites apparaissent lorsqu'on dispose de peu de données. Dans ce cas, il est plus judicieux d'utiliser un apprentissage par adaptation MAP. Cette estimation est plus fiable que celle obtenue par MV sachant qu'elle intègre des connaissances a priori.

La norme GSM et implémentation du codec GSMEFR

Chapitre 3

La norme GSM et implémentation du codec GSMEFR

3.1 Introduction

Le système GSM (Global System for Mobile communication) a été défini à la fin d'années 80 par des opérateurs et industriels réunis sous l'égide de l'ETSI (Européen Télécommunications Standard Institute). L'objectif était de créer une norme paneuropéenne de téléphonie cellulaire numérique palliant les lacunes des systèmes de téléphonie cellulaire analogique (taille importante des mobiles, incompatibilité entre les pays, non confidentialité des communications et efficacité spectrale faible). La norme GSM utilisée pour la codage de la parole est de trois types le GSMFR (Full Rate), le GSMHR (Half Rate) et le GSMEFR (Enhanced Full Rate). Dans ce chapitre, nous présentons le fonctionnent du système GSM et la simulation de codec GSMEFR.

3.2 Description de la norme GSM

3.2.1 La cellule

Dans un réseau GSM, le territoire est découpé en petites zones appelées cellules, chaque cellule est équipée d'une station de base fixe munie de ses antennes installées sur un point haut (château d'eau, immeuble…).

Les cellules sont dessinées hexagonales (Figure. 3.1) mais la portée réelle des stations dépend de la configuration du territoire arrosé et du diagramme de rayonnement des antennes d'émission dans la pratique, les cellules se recouvrent donc partiellement entre elles.

Deux cellules adjacentes ne peuvent utiliser la même bande de fréquence afin d'éviter les interférences la distance antre des cellules ayant la même bande doit être de 2 à 3 fois le diamètre d'une cellule la taille des cellules peut atteindre jusqu'à 35km et dépend de la densité d'utilisateur et de la topographie.

La taille limitée des cellules permet de limiter la puissance d'émission nécessaire pour la liaison donc augmenter l'autonomie des mobile [**50**]

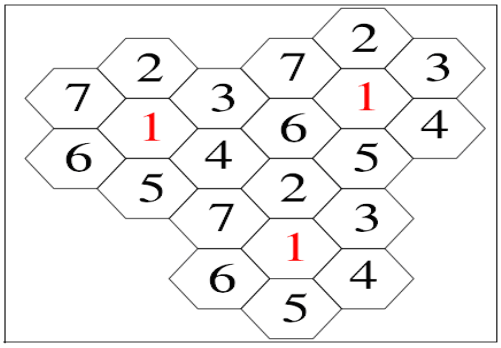

Figure 3.1 Structure cellulaire du réseau GSM

3.2.2 La communication dans le réseau GSM

Le mobile transmet par radio la communication vers la station de base de sa cellule. La conversation est ensuite acheminée de façon plus classique (câble, fibre optique ...) vers le correspondant s'il est raccordé au réseau téléphonique filaire, ou à sa station de base s'il est équipé d'un mobile. Cette station de base transmet finalement la conversation par radio au correspondant même si deux personnes se trouvent dans la même cellule et se téléphonent, la conversation ne passe jamais directement d'un **GSM** à l'autre.

Au cours d'un déplacement, il est possible qu'on sorte d'une cellule. Il est nécessaire alors de changer la station de base tout en maintenant la communication : c'est le transfert intercellulaire ou (handover) ce processus oblige tous les mobiles **GSM** à écouter les stations de base des cellules voisines en plus de la station de base de la cellule dans laquelle il se trouve.

3.2.3 Les équipements du GSM :

Le téléphone **GSM** ou station mobile (**M**obile **S**tation) est caractérisée par deux identités :

- Le numéro d'équipement, **IMEI** (**I**nternational **M**obile **E**quipment **I**dentity) mis dans la mémoire du mobile lors de sa fabrication.
- Le numéro d'abonner **IMSI** (**I**nternational **M**obile **S**ubscriber **I**dentity) se trouvant dans la carte **SIM** (**S**ubscriber **I**dentity **M**odule) de l'abonné.

Le système de communication radio et l'équipement qui assure la couverture de la cellule sont illustrés dans la figure 3.2

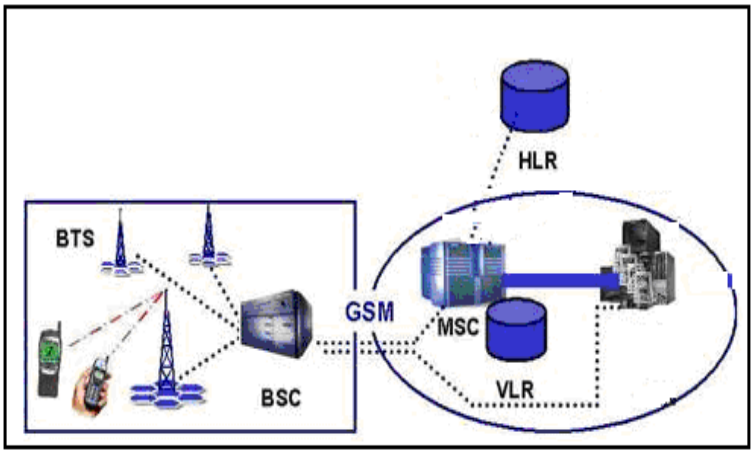

Figure 3.2 Les équipements du réseau GSM

a- La station de base (BTS) :

La station de base est l'élément central, que l'on pourrait définir comme un ensemble émetteur/récepteur pilotant une ou plusieurs cellules dans le réseau **GSM**, chaque cellule principale au centre de laquelle se situe une station base.

b- Le contrôleur de station de base (BSC) :

Le contrôleur de station de base gère plusieurs stations de base et il agit comme un concentrateur puisqu'il transfère les communications provenant des différentes stations de base vers une sortie unique, le **BSC** possède son registre d'abonnés visiteurs **VLR** stockant les informations de l'abonné liées à sa mobilité.

c- Le centre de commutation mobile(MSC) :

Le **MSC** est relié au sous-système radio via des interfaces, son rôle principal est d'assurer la commutation entre les abonnés du réseau mobile et ceux des autres réseaux

De plus, il participe à la fourniture des différents services aux abonnés tels que les services supplémentaires et les services de messagerie.

d- Le registre des abonnés nominaux (HLR) :

Il s'agit d'une base de données avec des informations essentielles et il contient les informations suivantes :

- Toutes les informations relatives aux abonnés: le type d'abonnement, le numéro de l'abonné (**IMSI**), etc.
- Ainsi qu'un certain nombre de données dynamiques telles que la position de l'abonné dans le réseau en fait, et l'état de son terminal (allumé, éteint, en communication, libre,...).

e- Enregistreur de localisation des visiteurs (VLR) :

L'enregistreur de localisation des visiteurs VLR, "Visitor Location Register", est une base de données qui mémorise les données d'abonnement des abonnés présent dans sa zone géographique. Chaque VLR est associé à un commutateur " MSC ".

Les données mémorisées par le VLR sont similaires aux données du HLR, vient se rajouter l'identité temporaire TMSI. Le VLR peut avoir une information de localisation plus précise que le HLR.Un ensemble MSC/VLR peut gérer de l'ordre d'une centaine de milliers d'abonnés pour un trafic moyen par abonné de 0,025 Erlang

3.2.4 Canaux de transmission

Le GSM utilise plusieurs canaux pour réaliser les communications : des canaux de transport d'information (voix et données), naturellement, mais également des canaux de contrôle d'appel, et un canal pour informer les terminaux d'une cellule des paramètres de connexion à l'antenne (fréquences utilisées, etc.).

Ces canaux sont :

- **TCH :** Transport Channel, le canal de transport de communication. Deux sont nécessaire par communication pour assurer le full-duplex.
- **BCCH** : Braodcast Control Channel, canal de diffusion des informations sur la cellule. Un par cellule.
- **FCH** : Frequency Channel, même voix que le BCCH, il permet au BTS de diffuser la fréquence du signal auquel il fonctionne.
- **SCH :** Signal Channel, même voix que le BCCH, signalisation.

- **SDCCH:** Stand-alone Dedicated Control Channel. Canal de control dédié à une communication. A chaque Canal de Transport est associe 1/8 de canal pour le SDCCH.

3.3 Codage de la parole

Les codeurs de parole compressent le signal de parole avant de le transmettre. Ils réduisent ainsi le nombre de bits requis pour sa représentation numérique, tout en gardant un niveau de qualité acceptable à la sortie du décodeur. Ils permettent ainsi une meilleure exploitation des ressources disponibles, tant au niveau utilisateur qu'au niveau réseau.

3.4 La norme GSM

De nombreux algorithmes de codage ont été normalisés depuis quelques années pour les systèmes de communications. On retient notamment le GSM plein débit (*Full Rate* GSM), demi-débit (*Half Rate* GSM) et le GSM plein débit amélioré (GSMEFR : Global System Mobile *Enhanced Full Rate*). La numérisation du signal de parole permet une meilleure protection contre les distorsions et les bruits introduits par les canaux radio mobiles [**51**].

3.4.1 Codeur de parole Plein Débit (FR)

Le codeur plein débit à été standardisé en 1988. Il appartient à la classe des codeurs de prédiction linéaire RPE-LTP (Regular Pulse Excitation with Long Term Prediction). A l'entrée du codeur, une trame de 160 échantillons est encodée en bloc de 260 bits à un débit de 13 kb/s. Le décodeur ordonne les séquences de 260 bits en blocs de 160 échantillons. Le canal GSM plein débit peut supporter un trafic de 28,8 kb/s [**52**].

3.4.2 Codeur de parole demi-débit (HR)

Le codeur demi-débit a été établi afin de contrer l'augmentation du nombre d'utilisateurs. C'est un codeur VSELP (Vector Sum Excited Linear Prediction) standardisé en 1995 par Motorola et ayant un débit de 5,6 kb/s. Le canal supporte un trafic de 11,4 kb/s conformément au fait de doubler la capacité du système. Ainsi, 5,8 kb/s sont utilisés en gestion d'erreurs. Le niveau de qualité du signal de parole en sortie est comparable au codeur plein débit et offre aussi des performances moyennes dans les communications mobile à mobile [**53**].

3.4.3 Le codec de parole plein débit amélioré GSMEFR

Le codec GSM plein débit amélioré fut standardisé en 1996 par Nokia dans l'intention de proposer une alternative au codeur plein débit (FR : Full Rate). Il apporte une certaine amélioration de la qualité par rapport aux codeurs FR et HR, notamment dans les communications mobile à mobile. Le codeur EFR utilise 12,2 kb/s pour le codage de la parole et 10,6 kb/s pour la gestion des erreurs. Le codage de la parole suit un modèle ACELP (Algebraic Code Excited Linear Prediction) [**54**].

3.5 Le codec GSM avec transmission et réception

Si les principales caractéristiques techniques de GSM sont bien connues du grand public, la méthode permettant de coder la voix humaine et d'obtenir une qualité honorable est très ténébreuse pour beaucoup.

Un système de codage/décodage de la parole comprend 2 parties : le codeur et le décodeur. Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage ou transmission. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique.
La parole est tout d'abord filtrée entre 300 Hz et 3400 Hz (comme pour le téléphone fixe) puis échantillonnée à 8 kHz. Les échantillons sont ensuite codés linéairement à 13 bits ce qui donne un débit résultat de 104 kbps. Ceci est bien sûr beaucoup trop élevé et les étapes suivantes vont permettre de réduire ce débit.
Toutes les 20 ms, 160 échantillons codés sont stockés et analysés en vue de calculer les paramètres d'un filtre et d'un signal d'excitation qui imiterait la séquence d'échantillons stockée. Ce calcul permet un codage linéaire prédictif de l'échantillon de parole considéré, tenant compte du fait que la parole possède de nombreuses corrélations dont la prédiction peut tirer parti.

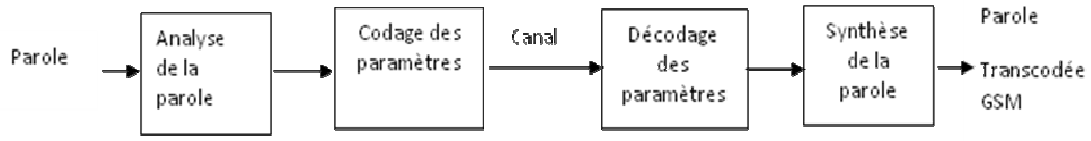

Figure 3.3 Schéma de parole transcodée GSM

3.6 Simulation du codec GSMEFR sous MATLAB.

L'utilisation des signaux transcodés GSM réels dans le domaine de la reconnaissance est très important mais comment obtenir ces signaux transcodés GSM. On peut les obtenir, soit par des opérateurs de télécommunications qui disposent de base de donnés prêtes pour l'emploi, soit par simulation du codec de la parole dans le réseau GSM. Dans notre travail, nous avons simulé le fonctionnement de codec GSMEFR version GSM 6.60 sous les boites outils de MATLAB. Cette simulation est réalisée en se basant sur l'algorithme de codage de la parole de type ACELP.

3.6.1 Codeur GSMEFR

Le codeur GSMEFR utilise donc l'algorithme ACELP (Algebraic Code Excited Linear Prediction) basé sur le principe des codeurs CELP (Code Excited Linear Predictor). Comme tout codeur CELP, l'EFR calcule, quantifie et transmet les paramètres de la prédiction linéaire à court-terme LPC et de celle à long terme LTP : dictionnaire d'excitation adaptatif, dictionnaire d'excitation fixe et les gains associés. Le codeur opère sur des trames de 20 ms (160 l'échantillon à 8 kHz), divisées en quatre sous-trames de 5 ms (40 échantillons) pour la détermination de l'excitation.

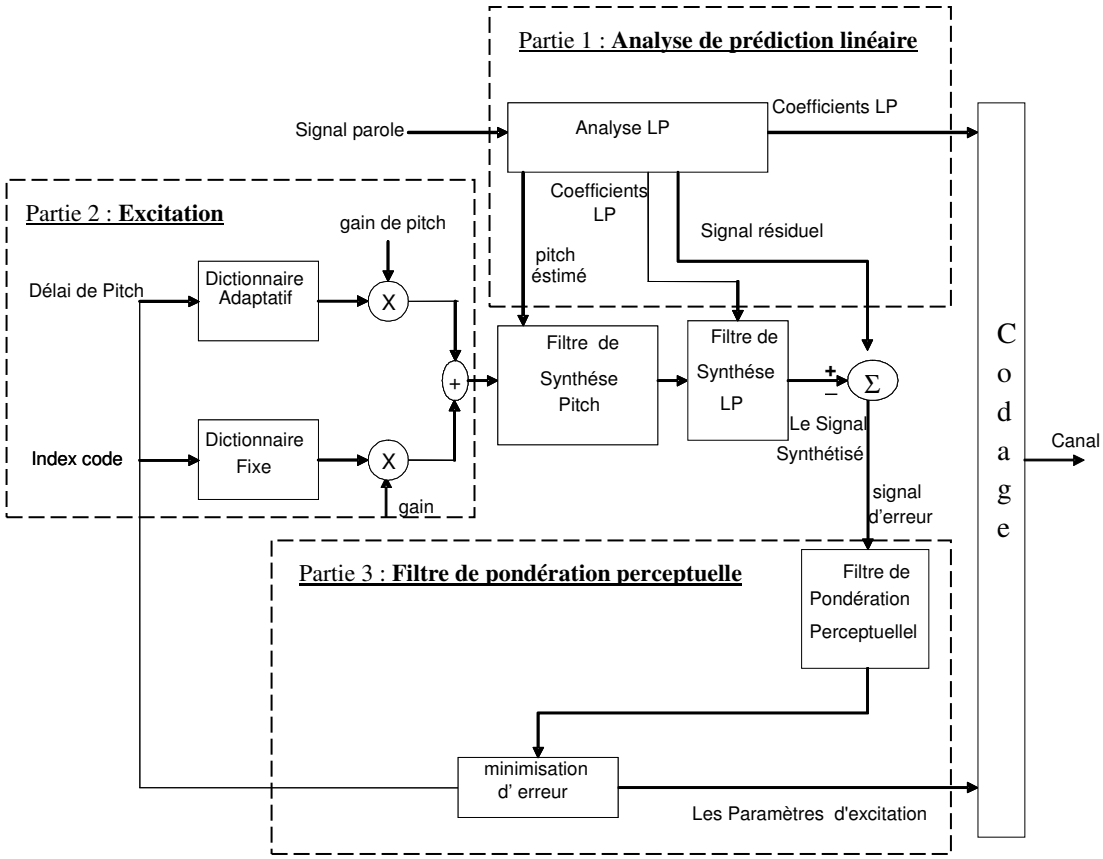

Figure 3.4 Le schéma du codeur GSMEFR (Analysé par synthèse) basé sur l'algorithme de codage ACELP

Le schéma blocs du codeur GSMEFR (Analyse par synthèse) est montré dans la figure 3.4. Le codeur de la parole est basé sur le codage prédictif et l'analyse par synthèse (*LPAS : Linear Prediction Analysis by Synthesis*). Le signal d'entrée est analysé et un signal d'excitation est déterminé. La fonction d'analyse basée sur le codage prédictif exploite les redondances du signal de parole en le modélisant par un nombre restreint de paramètres sous la forme d'un filtre linéaire définit par les coefficients du filtre de prédiction LP tandis que le

signal d'excitation est modélisé par l'analyse par synthèse. Dans ce schéma bloc, on distingue trois parties.

Partie 1 : Analyse de prédiction linéaire

Le signal de parole original est analysé par le filtre d'analyse LP, à la sortie du filtre nous avons :

- un signal de sortie de faible énergie correspondant à l'erreur de prédiction appelée signal résiduel.
- les coefficients de prédiction LP qui peuvent être utilisés pour supprimer les redondances à court terme du signal.
- Le pitch qui permet d'estimer par la périodicité du signal représentée par la prédiction à long terme (LTP).

Partie 2 : Excitation

Pour déterminer les gains, paramètre très important pour la synthèse de la parole, deux dictionnaires sont utilisés :

- Le dictionnaire adaptatif pour modéliser les composantes périodiques de l'excitation.
- Le dictionnaire fixe pour modéliser les composantes non périodiques de l'excitation.

L'excitation est calculée par sous-trames de 5ms. Elle est le résultat de l'addition de deux sous excitations. La sous excitation vecteur-code extrait du dictionnaire adaptatif utilisant l'indice (délai de pitch) et la pondération par le gain de pitch. La sous excitation des impulsions-codes extrait du dictionnaire fixe à l'indice (index code) et pondérée par un (gain).

Partie 3 : Le filtre de pondération perceptuelle

L'analyse du codeur ACELP consiste à trouver les paramètres d'excitation (les indices et les gains) de manière à minimiser l'erreur perceptuelle (sortie du filtre perceptuel). La différence entre le signal synthétisé et le signal résiduel génère un signal d'erreur. Ce signal est pondère avec le filtre de pondération, incorporant dans sa structure une fonction qui prend en compte la perception de l'appareil auditif humain. Le signal d'erreur est utilisé pour sélectionner le meilleur signal d'excitation. La somme des sorties pondérées des deux dictionnaires, le pitch estimé et les coefficients de prédiction linéaire LP excitent les filtres de synthèse de pitch et de synthèse LP pour générer un signal de parole synthétisé.
Les blocs importants de chacune de ces parties sont détaillés dans ce qui suit.

3.6.1 .1 Analyse par prédiction linéaire

a- Analyse LP à court-terme

Le principe fondamental de la prédiction linéaire à court terme est qu'un échantillon donné peut être prédit à partir d'une combinaison linéaire des échantillons finis qui le précèdent. Un

seul jeu de coefficients du prédicteur est déterminé en minimisant les différences entre les échantillons actuels et ceux prédits. La technique de prédiction linéaire est basée sur le modèle de la production de la parole, le conduit vocal est simplement assimilé par un filtre linéaire tout pole dont la fonction de transfert peut être représentée par l'équation suivante :

$$H(Z)=\frac{G}{A(Z)}=\frac{1}{1+a_1z^{-1}+a_2z^{-2}+.....+a_pz^{-p}}=\frac{G}{\sum a_iz^{-p}} \qquad 3.1$$

Les a_i sont appelées coefficients de prédictions.

G représente le gain du modèle

Dans le domaine temporel, la sortie x(n) associée à H(z) satisfait l'équation aux différence suivante :

$$s(n)=-\sum_{k=1}^{p}a_k s(n-k)+Gu(n) \qquad 3.2$$

U(n) étant l'entre du filtre prédicateur suivant :

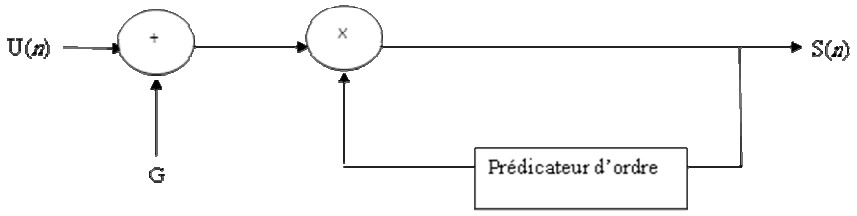

Figure 3.5 Modèle du conduit vocal

En supposant que cette entré U(n) soit totalement inconnue, le signal de parole à un instant n peut être prédit par une combinaison linéaire des p échantillons précédents, soit :

$$\hat{s}(n)=-\sum_{k=1}^{p}a_k s(n-k) \qquad 3.3$$

L'erreur de la prédiction *e*(*n*) est donnée par :

$$e(n)=s(n)-\hat{s}(n)=s(n)+\sum_{k=1}^{p}a_k s(n-k) \qquad 3.4$$

L'énergie résiduelle de prédiction est définie par le somme :

$$E_p = \sum_n e^2(n) = \sum_{n_1}^{n_2} \left[s(n) - \sum_{k=1}^{P} a_k s(n-k) \right]^2 \qquad 3.5$$

Si les coefficients sont choisis tels qu'ils minimisent l'énergie résiduelle de prédiction, il suffit pour les obtenir de poser

$$\partial E / \partial a_i = 0 \qquad \text{pour } i = 1,....., p. \qquad 3.6$$

Le calcul de cette relation (équation 3.6) conduit aux équations suivantes :

$$\sum_{k=1}^{p} a_k \phi(i,k) = \phi(i,0), \quad i = 1,....., p. \qquad 3.7$$

$$\phi(i,k) = \sum_n s(n-i)s(n-k) \qquad 3.8$$

Ces équations normales dites de Yule Walker, constituent un système linéaire de *p* équation à *p* inconnues. La résolution de ce système permettre d'obtenir les coefficients du filtre. Parmi les méthodes de minimisation de l'énergie de l'erreur donc de résolution du système, on trouve la méthode de covariance (basée sur la matrice de covariance), la métode d'autocorrélation et la méthode harmonique. La méthode d'autocorrélation est la plus utilisée parce qu'elle conduit à un système facile à résoudre et elle la stabilité du modèle auto régressif trouvé.

b- Méthode d'autocorrélation

L'énergie résiduelle de prédiction définie dans l'équation (3.5) est minimisée sur une durée infinie :

$$E_p = \sum_n e^2(n) \qquad 3.9$$

La fonction d'autocorrélation est définie par :

$$R(i) = \sum_{-\infty}^{+\infty} s(n)s(n-i) \qquad 3.10$$

avec $R(i) = R(-i)$ 3.11

Le signal vocal est défini par toutes les valeurs du temps ; il est identiquement nul en dehors d'une séquence de N échantillons ceci équivaut à multiplier le signal vocal par une fenêtre de

longueur finie correspondant à N échantillons. La sommation infinie de l'équation se ramène donc à une somme finie, soit :

$$E_p = \sum_{n=0}^{n-1+p} e^2(n) \sum_{n} e^2(n) \qquad 3.12$$

L'équation devient alors :

$$\phi(i,k) = \sum_{-\infty}^{+\infty} s(n-i)s(n-k) = \sum_{-\infty}^{+\infty} s(n)s(n+i+k) \qquad 3.13$$

Dans ce cas $\phi(i,k)$ n'est autre que la fonction d'autocrrélation évaluée pour (*i*-*k*), soit

$$\phi(i,k) = R(i-k) \qquad 3.14$$

Le système donné par l'équation, s'écrira alors sous la forme suivante :

$$\sum_{K=1}^{p} a_k R(i-k) = R(i) \qquad i = 1,....., p. \qquad 3.15$$

La matrice, de dimension $p \times p$, des valeurs d'autocorrelations est une matrice de Toeplitz symétrique, tous les éléments d'une diagonale donnée sont égaux. Cette propriété peut être exploitée pour obtenir un algorithme efficace de résolution du système d'équations. La solution la plus efficace est une méthode itérative connue sous le nom de l'algorithme de Wiener Livinson Durbin. Dans notre implémentation nous disposons de 10 coefficients estimés dans chaque trame de 20ms avec de (160 échantillons à 8khz), les coefficients LP résultants sont stockés dans un vecteur **ar**. Les étapes de la prédiction linéaire peuvent être résumées dans la figure (3.6) suivant :

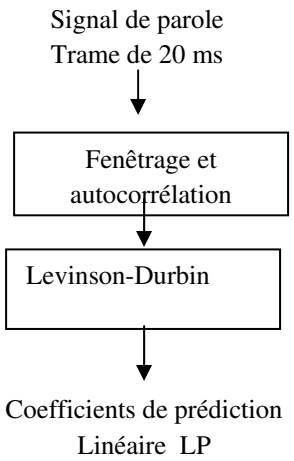

Figure 3.6 Les étapes de calcul des coefficients LP

3.6.1.2 Analyse à longe-terme (estimation du pitch)

La mise en œuvre d'une prédiction à longe-terme (LTP : *Long Term Prediction*), dans le codage par prédiction linéaire, est un moyen efficace de représenter la périodicité du signal de parole. Cette analyse n'a pas d'effet sur les trames de parole non voisées, qui n'ont pas de structure harmoniques. En effet, elle a pour but de modéliser les vibrations des cordes vocales produisant les sons voisés qui possèdent des corrélations à long terme et se traduisent au niveau du signal de parole par une fréquence fondamentale F0, appelée pitch

L'estimation du pitch est un des problèmes les plus difficiles de l'analyse de la parole. Car l'oreille est plus sensible aux changements de fréquence F0. La redondance à long terme peut être modélisée en utilisant un filtre linéaire $B(z)$ du premier ordre. Le filtre d'analyse de pitch est donné par l'équation suivante :

$$B(z) = 1 - \beta z^{-D} \qquad 3.16$$

où β est le coefficient prédiction, correspondant au degré de périodicité associé avec $0 \leq \beta \leq 1.4$ et D le délai prédicteur, Le délai D correspond au pitch estimé. Ils s'obtiennent en minimisant l'erreur de prédiction :

$$E_{LTP} = \sum_{-\infty}^{\infty} \left[e_w(n) - \beta e_w(n-D) \right]^2 \qquad 3.17$$

où $e_w(n)$ est le résidu de prédiction à court-terme $e(n)$ pondérée par une fenêtre d'analyse.
En dérivant E_{LTP} par rapport à β, on obtient :

$$\beta(D) = \frac{\sum_{-\infty}^{\infty} e_w(n) - e_w(n-D)}{\sum_{-\infty}^{\infty} e_w(n-D)} \qquad 3.18$$

Puis on remplace β par sa nouvelle valeur pour calculer D **:**

$$E_{LTP}(D) = \sum_{-\infty}^{\infty} e_w^2(n) - \frac{[\sum_{-\infty}^{\infty} e_w(n) - e_w(n-D)]^2}{\sum_{-\infty}^{\infty} e_w^2(n-D)} \qquad 3.19$$

Le gain β du prédicteur à long-terme peut être très élevé, ce qui rend instable le schéma en boucle ouverte. Aussi, le prédicteur à long-terme $B(z)$ est toujours implanté selon le schéma en boucle fermée illustré par la Figure (3.7). Le critère d'erreur doit donc être modifié pour intégrer le signal d'excitation reconstruit $\hat{e}(n)$:

$$E_{LTP} = \sum_{-\infty}^{\infty} \left[e_w(n) - \beta \hat{e}_w(n-D) \right]^2 \qquad 3.20$$

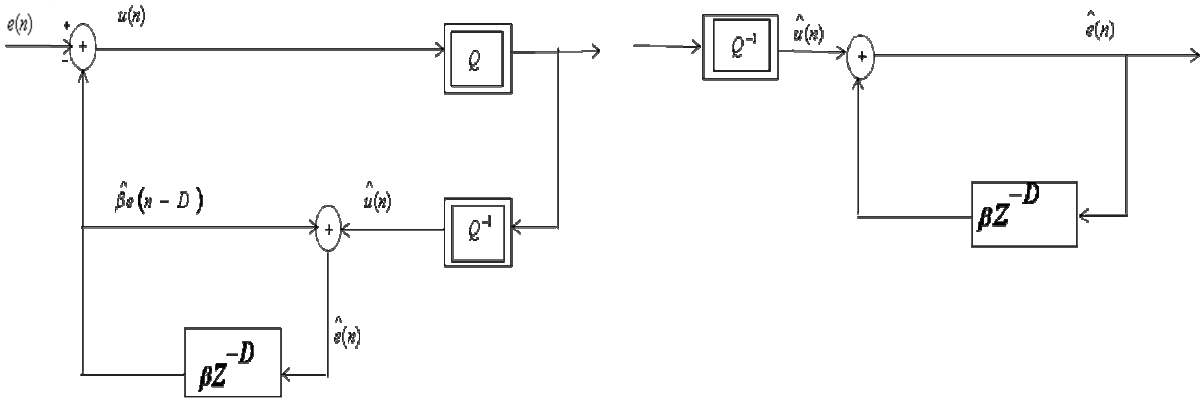

Figure 3.7 Prédiction à long terme en boucle fermée

3.6.1.3 Excitation

Concernant la prédiction à long terme : pour modéliser les composantes périodiques de l'excitation, le codec GSMEFR utilise un dictionnaire adaptatif dont l'indice (délai de pitch) et le gain associé (gain de pitch) sont déterminés en minimisant une erreur perceptuelle sur la parole. De manière analogue, pour modéliser les composantes non périodiques de l'excitation, le codec GSMEFR utilise un dictionnaire fixe dont l'indice (index code) est associé à un gain. Les index et gains des dictionnaires adaptatifs et fixes sont calculés et transmis pour chaque sous trame de 5ms.

a- Dictionnaires adaptatifs

Par un son voisé, le signal d'excitation présente une forte périodicité à long-terme. Le codeur GSMEFR à analyse par synthèse modélise le plus souvent cette périodicité à l'aide d'un dictionnaire adaptatif.

Le dictionnaire adaptatif modélise la périodicité du signal à long terme due au Pitch. Il est constitué d'une mémoire tampon composée des vecteurs-codes. Dans la figure (3.8) le signal d'excitation $\hat{e}$, et compose avec l'indice dans le dictionnaire est associé à la valeur du pitch D et le gain de pitch λ joue un rôle équivalent au coefficient β du filtre prédicteur $B(z)$, le signal $\hat{u}$ modélise les composantes non-périodiques de l'excitation $\hat{e}$.

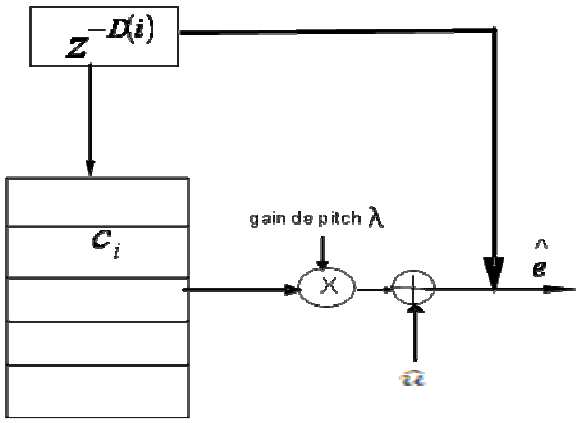

Figure 3.8 Synthèse d'une excitation voisée par dictionnaire adaptatif

Le dictionnaire adaptatif apparaît ainsi comme un buffer à décalage dont chaque vecteur d'indice i est formé par :

$$c_i = \left[\hat{e}(n - D(i)), \hat{e}(n - D(i) + 1)....., \hat{e}(n - D(i) + L - 1) \right] \qquad 3.21$$

Où L correspond à la longueur d'une sous-trame et $D(i)$ au délai, associé à l'indice i.

Le délai de pitch D (indice dans le dictionnaire) et le gain de pitch sont cherchés en minimisant une distance perceptuelle entre le signal parole résiduelle et le signal parole synthétisé, selon le schéma d'analyse par synthèse illustré par la Figure (3.9).

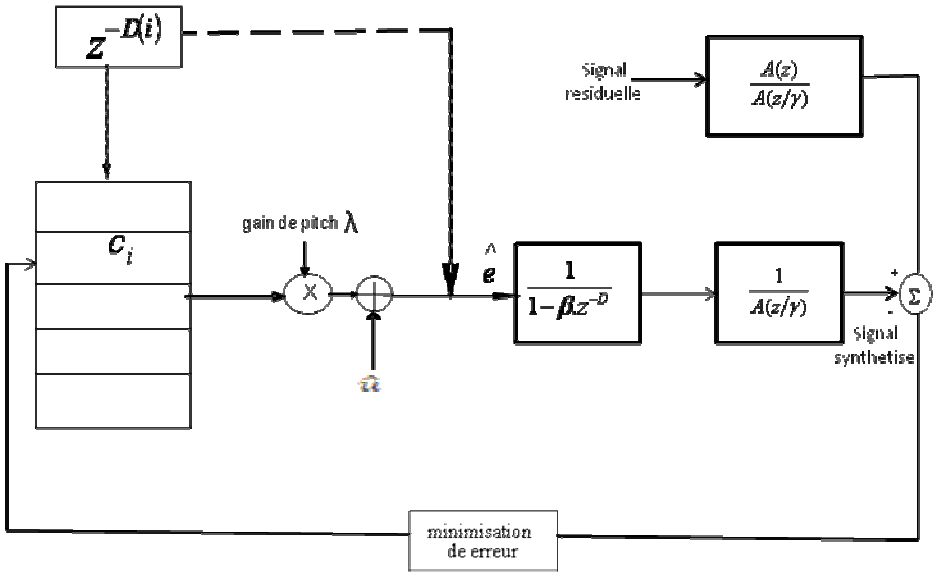

Figure 3.9 Détermination de la composante périodique de l'excitation

Pour quantifier le vecteur ***u*** de composantes non-périodiques de l'excitation, on utilise un dictionnaire fixe associé à un gain.

b- Dictionnaire fixe

La quantification de l'excitation résiduelle est très importante pour restituer le naturel de la parole. Pour la Modélisation de l'excitation résiduelle, Les codeurs à analyse par synthèse comme le ACELP utilisent des dictionnaires fixes associés à un gain. Dans les principaux types de dictionnaires utilisés par les codeurs de type ACELP, les éléments sont quantifiée à 3 niveaux (-1, 0,1) et distribués selon une gaussienne de moyenne nulle et de variance unitaire**.** On détermine la composante aléatoire en sélectionnant chaque 5 ms une excitation j_r (n) du dictionnaire fixe ou *r* présente l'indice de cette excitation.

Le schéma suivant résume la détermination de l'excitation :

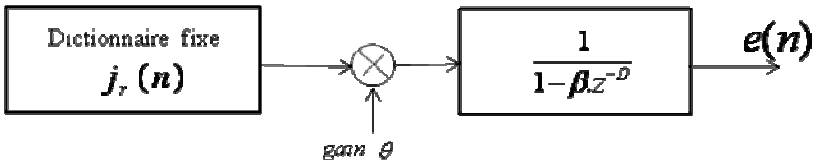

Figure 3.10 Détermination de la composante non-périodique de l'excitation

n : nombre d'échantillonnes

θ : gain,

$e(n)$ c'est l'excitation à déterminer. Elle peut être calculée par l'équation suivante :

$$e(n) = \theta j_r(n) + \beta e(n-j) \qquad 3.22$$

3.6.1.4 Filtre de pondération perceptuelle

Le principe de l'analyse par synthèse utilisée pour déterminer le signal d'excitation est de minimiser un critère d'erreur E entre une séquence du signal résiduel $s(n)$ et du signal synthétisé $\hat{s}(n)$ (figure 3.4). L'excitation optimale sera établie par minimisation de l'erreur quadratique moyenne, définie comme suit :

$$E = \left[s(n) - \hat{s}(n) \right] \qquad 3.23$$

Le critère d'erreur est modifié par une pondération perceptive E_w dont le but est de contrôler la répartition fréquentielle du bruit.

Le caractère physiologique de l'oreille possède la propriété intéressante de percevoir les sons d'une manière fréquentielle. De cette propriété découle la pondération perceptive qui a pour fonction de répartir la puissance du bruit d'une fréquence à une autre. Cette répartition est efficace lorsque la distribution de la puissance du bruit est augmentée au niveau des régions formantiques. Dans ce cas, le bruit est masqué par la puissance du signal et l'écoute sera plus agréable. Le nouveau critère, appelé critère d'erreur subjectif, est défini comme la minimisation de l'erreur quadratique

$$E_W = \left[\left(s(z) - \hat{s}(z) \right) w(z) \right]^2 \qquad 3.24$$

Avec le filtre de pondération perceptive $w(z)$ donné par :

$$w(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 + \sum_{k=1}^{p} a_k \, z^{-k}}{1 + \sum_{k=1}^{P} a_k \gamma^k z^{-k}} \qquad 3.25$$

Le filtre de pondération est construit à partir du filtre de prédiction à court terme et du coefficient γ permettant de contrôler au mieux la répartition de la puissance du bruit. La valeur de γ est généralement comprise entre 0 et 1, ($0 < \gamma \leq 1$). Dans notre travail nous avons choisi la valeur optimale γ égale à 0,75 pour une fréquence d'échantillonnage de 8Khz.

3.6.1.5 Filtre de synthèse

Dans notre implémentation deux filtres de synthèses sont utilisés. Le filtre de synthèse LP qui modélise la périodicité à court terme du signal et le filtre de synthèse de pitch qui modélise la périodicité à long terme. Pour le filtre de synthèse LP à court-terme, la fonction de transfert est définie comme suit :

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{10} a_k z^{-k}} \qquad 3.26$$

avec a_k coefficients de l'analyse LP.

Pour le filtre de synthèse de pitch à longe-terme, la fonction transfert est définie par :

$$B(z) = \frac{1}{1 - \beta . z^{-D}} \qquad 3.27$$

3.6.1.6 Codage des paramètres

Le GSMEFR a un débit de codage de parole égale à 12,2Kbit/s. Pour attendre ce débit, on utilise une quantification scalaire sur les paramètres λ (gain de pitch), ***D***(délai de pitch), *j* (index de dictionnaire fixe), θ (gain) et (**ar** : LPC). Ensuite ces paramètres sont envoyés dans le canal de transmission. Le tableau 3.1 suivant illustre les allocations des bits par trame de 20 ms.

Les paramètres	sous trame	Total par trame
Les coefficients LPC ar	8	80
λ (gain de pitch)	4	16
D (délai de pitch)	7	28
j (index de dictionnaire fixe)	40	80
θ (gain)	10	40
Total		244

Tableau 3.1 Allocation des bits par trame de 20ms.

3.6.1.7 Simulation du canal de transmission

a- Canal de type AWGN

Le bruit introduit par le canal peut prendre différentes formes. Le modèle de bruit le plus commun est celui du Bruit Additif Blanc Gaussien (AWGN : Additive White Gaussian Noise). Le bruit est dans ce cas modélisé comme une variable aléatoire *n* qui s'additionne au signal codé. La variable *n* est supposée gaussienne de moyenne nulle et de variance σ^2, sa densité de probabilité est définie par :

$$P(n) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-n^2/2\sigma^2} \qquad 3.28$$

Le terme 'Blanc' indique que la densité spectrale de puissance du bruit est considérée unilatérale (c'est-à-dire de la fréquence 0 à ∞) et constante de valeur *N0*.
Si x_k est la variable qui représente le signal codé correspondant à un bit d'information à l'instant *k*, alors la sortie y_k bruitée est donnée par :

$$y_k = x_k + n \qquad 3.29$$

Figure 3.11 Simulation du canal avec bruit gaussien

Pour le cas de ce type de canal AWGN, on définit la capacité C du canal par :

$$C = W \log_2\left(1 + \frac{S}{N}\right) \qquad 3.30$$

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,

k s'exprime en terme de bits par seconde par Hz (b/s/Hz). La valeur E_b/N_0 en fonction de k satisfait la condition :

$$\frac{E_b}{N_0} > \frac{2^k - 1}{n} \qquad 3.31$$

b- Canal à évanouissement

Dans les communications radiomobiles, l'atténuation du signal varie avec la vitesse du véhicule (qui est le récepteur) et la nature des obstacles. Ce type de canal est appelé canal à évanouissement. L'amplitude a_k du signal est souvent modélisée suivant la loi de Rayleigh, le signal est alors sous la forme :

$$y_k = a_k x_k + n \qquad 3.32$$

Figure 3.12 Simulation du canal avec canal Rayleigh

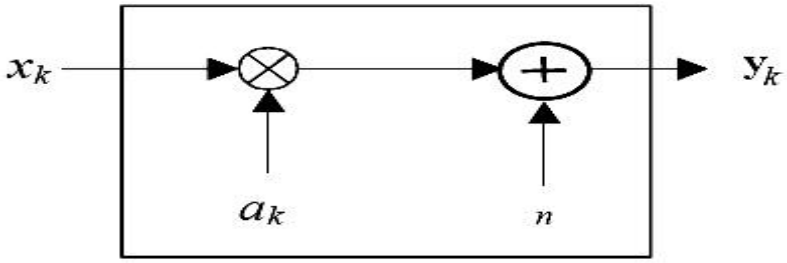

Figure 3.13 Modèle d'un canal de Rayleigh

La fonction de densité de probabilité est dans ce cas donnée par :

$$p(a_k) = \begin{cases} \frac{a_k}{\sigma^2} \exp\left(-\frac{a_k^2}{2\sigma^2}\right) & a_{k\geq} 0 \\ 0 & ailleurs \end{cases} \qquad 3.33$$

Où a_k est appelé aussi l'enveloppe du signal.

3.6.2 Décodeur GSMEFR

Au niveau du décodeur, les paramètres de quantification sont reconstruits tels que les coefficients LP, le gain de pitch, le délai de pitch, l'index de dictionnaire fixe et le gain. Ils sont ensuite stockées dans un buffer de contrôle dans le but de :

- sélectionner la forme d'onde et son gain
- recréer les filtres courts et long terme

Le signal de synthèse s'obtient alors par filtrage long et court terme de la forme d'onde.

Dans le premier temps, l'excitation est produite en additionnant les contributions d'un dictionnaire adaptatif et du dictionnaire fixe :

$$e(n) = e_a(n) + e_f(n)$$

Où $e_a(n)$ est la contribution du dictionnaire adaptatif (pitch) et $e_f(n)$ est la contribution du dictionnaire stochastique (innovation ou fixe). Cette excitation $e(n)$ excite le filtre de synthèse de pitch.

Le signal résultant est ensuite injecté dans le filtre de synthèse LP avec des coefficients LP stockés dans le buffer de contrôle. On obtient alors le signal de parole transcodée GSM à la sortie.

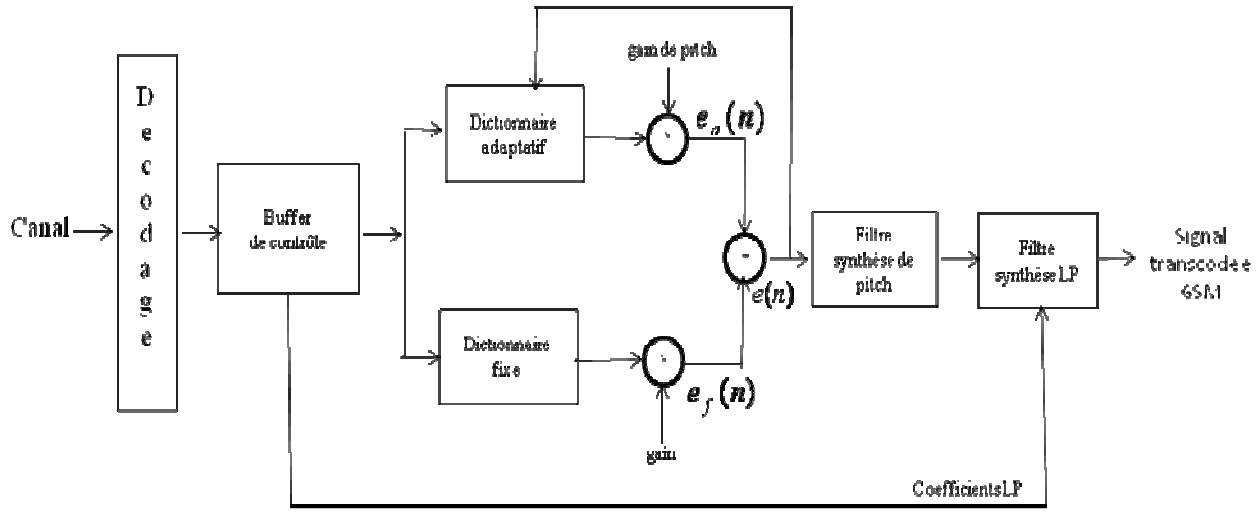

Figure 3.14 Schéma du décodeur GSMEFR.

3.7 Conclusion

Le réseau GSM est considéré par ses exploitants comme un investissement qui doit être rentable, offrir une certaine pérennité et pouvoir évoluer. Les qualités de ce réseau sont l'utilisation optimum des ressources (fréquence radio, capacité de transmission, une grande disponibilité, une exploitation simple et efficace, une normalisation réussie et enfin un coût d'infrastructure raisonnable. De nombreux algorithmes de codage de la parole de type GSM sont proposés par l'ETSI pour les systèmes de communications avec les mobiles. Dans notre travail, nous nous sommes intéressés au GSM plein débit amélioré (ou *Enhanced Full Rate* GSM). Le codec GSMEFR a été implémenté en utilisant les outils MATLAB, ce qui nous a permis de disposer de bases de données transcodées GSM nécessaire pour l'évaluation de notre système d'IAL via une ligne de communication GSM.

Évaluation expérimentale

Chapitre 4

Évaluation expérimentale

4.1 Introduction

Dans ce chapitre, nous présentons l'évaluation du système de reconnaissance de locuteur mis en œuvre par l'approche à base de mélange de gaussiennes (GMM) pour le cas des signaux transcodés GSM. Ainsi, nous décrirons les bases de données transcodées GSM réalisées, l'analyse acoustique appliquée aux signaux de parole et les résultats d'Identification Automatique du Locuteur (IAL) obtenus.

4.2 Description des bases de données utilisées

La base de données parole originale utilisée dans ce travail est la base de données ARADIGIT [**55**]. Elle est constituée de prononciations des 10 chiffres de la langue Arabe de zéro jusqu'à neuf prononcés par 120 locuteurs des deux sexes avec trois répétitions pour chaque chiffre. Cette base a été enregistrée par des locuteurs algériens de différentes régions âgés entre 18 et 50 ans dans un environnement calme avec un niveau de bruit ambiant inférieur à 35 dB, sous le format WAV, avec une fréquence d'échantillonnage égale à 16 KHz.

A partir de cette base, nous avons réalisé quatre autres bases de données :

- La base de données ARADIGIT16K sous échantillonnée à 8KHz donnant ainsi la base de données ARADIGIT8K.
- La base de données ARADIGIT8K est passée ensuite à travers le codec GSMEFR que nous avons simulé pour aboutir à la base de données transcodée GSM (source uniquement) à savoir la base nommée ARADIGIT_GSM. La figure 4.1 donne le processus de création de cette base.

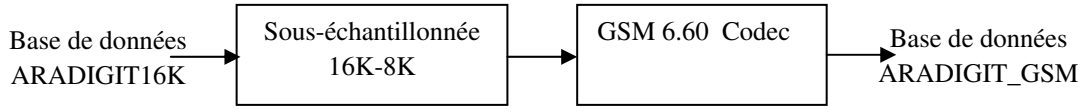

Figure 4.1 Création de la base de données ARADGIT_GSM.

- La base de données transcodée GSM (source + canal), le canal de communication est simulé par le bruit de type AWGN donnant la base de données nommée ARADIGIT_GSM_AW. La figure 4.2 donne le processus de création de cette base.

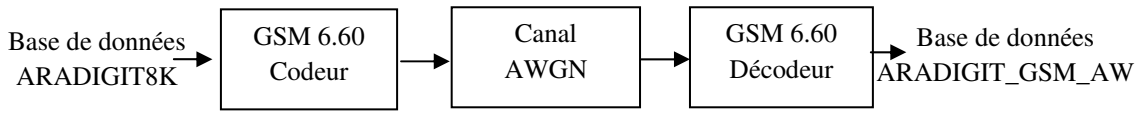

Figure 4.2 Création de la base de données ARADGIT_GSM_AW.

- La base de données transcodée GSM (source + canal), le canal de communication est simulé par le bruit de type Rayleigh donnant la base de données nommée ARADIGIT_GSM_R. La figure 4.3 donne le processus de création de cette base.

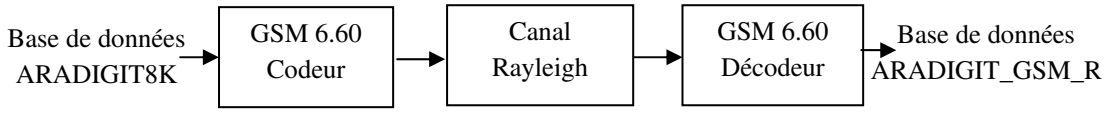

Figure 4.3 Création de la base de données ARADGIT_GSM_R.

4.3 Analyse acoustique et paramétrisation du signal vocal

L'analyse de la parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. On s'intéresse essentiellement à l'information relative à l'identité du locuteur, les systèmes de reconnaissance utilisent souvent les coefficients MFCC (Mel Frequency Cepstral Coefficients) qui permettent une parfaite dé convolution de la contribution du conduit vocal et celle de la source d'excitation [**56][57**].

Dans nos expériences, une analyse est appliquée toutes les 10 ms sur des fenêtres d'analyse de 20 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame, on associe un vecteur de représentation acoustique. Douze (12) coefficients MFCC sont calculés pour chaque trame à partir d'un banc de 24 filtres répartis dans l'échelle fréquentielle Mel. Des coefficients différentiels du premier ordre(△) et du second ordre (△△) sont ensuite calculés pour former un vecteur de dimension 36 (12 MFCC + △12 MFCC+ △△12MFCC).

4.3.1 Extraction des paramètres

La figure 4.4 illustre les étapes suivies afin d'extraire les coefficients MFCC.

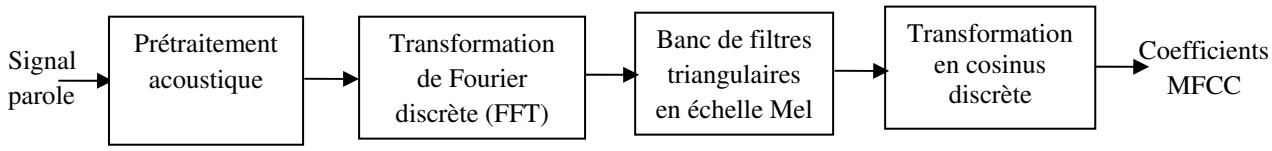

Figure 4.4 Extraction des coefficients MFCC.

La phase de pré-traitement acoustique contient deux étapes :

1. L'étape de pré-accentuation acoustique qui consiste à filtrer le signal vocal par un filtre passe haut de transmittance $H(z) = 1 - 0.95\, z^{-1}$.

2. L'étape de fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération glissante. Dans notre travail, nous avons utilisé une fenêtre de Hamming glissante de durée de 20 *ms* avec un déplacement de 10 *ms.*

La figure 4.5 illustre une fenêtre de pondération de type Hamming sur 512 échantillons, l'équation de cette fenêtre est donnée par :

$$w(n) = 0.54 + 0.46\cos\left[\frac{2m}{N-1}\right] \quad avec \qquad 0 \le n \le N-1 \tag{4.1}$$

N : Nombre d'échantillons dans la fenêtre d'analyse.

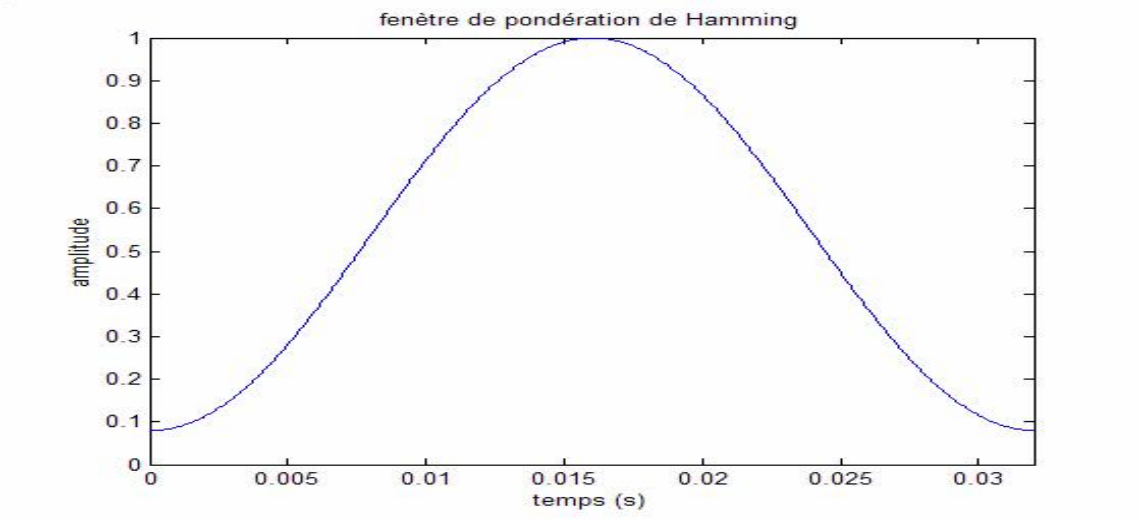

Figure 4.5 Fenêtre de pondération de type Hamming.

Une fois la phase de pré-traitement terminée, on applique aux trames de parole résultantes les traitements suivants :

a. Transformée de Fourier discrète
Elle permet le passage du domaine temporel au domaine fréquentiel. Pour un traitement rapide, on utilise la transformée de Fourier rapide (FFT).

b. Banc de filtres triangulaires dans l'échelle Mel
Le spectre du signal est filtré par un banc de filtres triangulaires, dont les bandes passantes sont de même largeur sur une échelle perceptive de type Mel. Chaque filtre opère sur une bande de fréquence bien déterminée.

c. Transformée en cosinus discrète
Les premiers coefficients cepstraux sont calculés directement à partir du logarithme des énergies *Ei* à la sortie d'un banc de *M* filtres par la transformée en cosinus discrète permettant l'obtention de coefficients c_k (Coefficients MFCC) fortement décorrélés donnés par :

$$c_k = \sum \log E_i \cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2}\right)\right] \tag{4.2}$$

4.4 Protocole d'évaluation

La base de données d'évaluation est divisée en deux bases :

- **La base d'apprentissage** : elle est composée de 1440 fichiers vocaux prononcés par 60 locuteurs comprenant les deux sexes (31 hommes et 29 femmes). Chaque locuteur répète les chiffres (0, 1, 2, 3, 4, 5, 6,7) trois fois.
- **La base de test** : elle est composée de 360 fichiers vocaux prononcés par 60 locuteurs, comprenant les deux sexes (31 hommes, 29 femmes). Chaque locuteur répète les chiffres (8,9) trois fois.

  Ces fichiers de parole sont passés à travers le codec GSMEFR simulé pour donner les différentes bases décrites dans la partie 4.2.

L'évaluation est faite en calculant le Taux d'Identification Correcte (TIC) défini par :

$$TIC(\%) = \frac{Nombre\ de\ test\ correctement\ identifié}{Nombre\ total\ de\ test} \times 100$$

4.5 Les expériences

4.5.1 Expérience 1 : Influence du nombre de gaussiennes sur le *TIC*

Dans cette première expérience l'objectif est d'étudier l'influence du nombre de gaussiennes sur la performance du système d'identification du locuteur mis en œuvre. Nous avons fait varier le nombre de gaussiennes de 1 jusqu' a 64. La base de test est constituée de 60 locuteurs. La figure 4.6 montre les variations des taux d'identification en fonction du nombre de gaussiennes pour la base de données ARADIGIT16K.

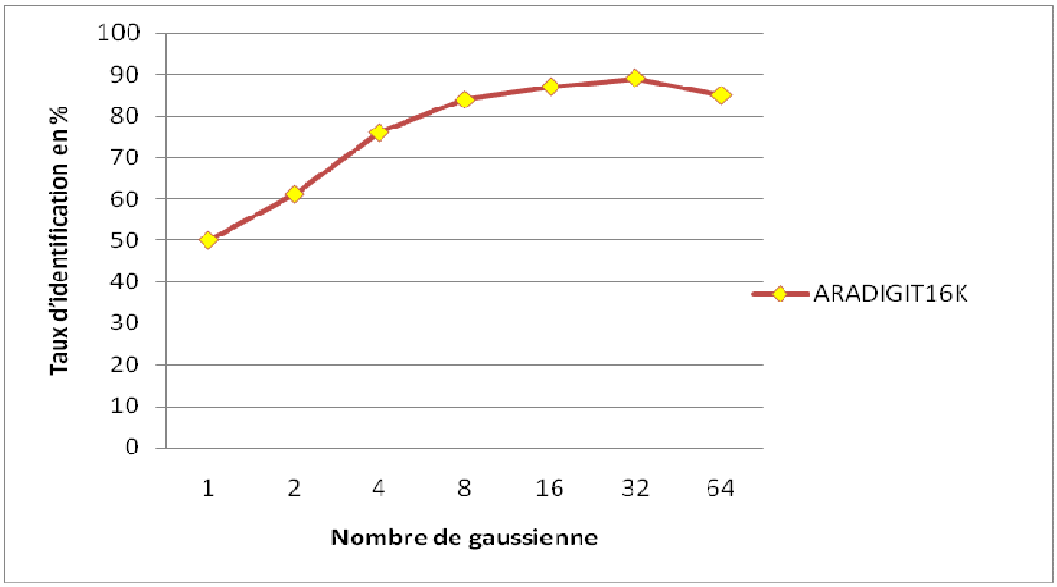

Figure. 4.6 Taux d'identification en fonction du nombre de gaussiennes avec la base de données ARADIGIT16K.

Les résultats montrent que le taux d'identification augmente avec le nombre de gaussiennes, il est de (89,34%) pour un nombre de gaussienne égale à 32, et se dégrade légèrement (85,6 %) au-delà de 32 gaussiennes. L'architecture optimale de notre système correspond donc à un nombre de gaussienne égale à 32.

Le meilleur taux d'identification est obtenu pour un nombre de gaussienne égale à 32 donc un modèle de 32 gaussiennes suffit pour représenter convenablement le locuteur (résultat conforme avec les travaux de Reynolds [**58**]).

4.5.2 Expérience 2

Pour le nombre de gaussiennes optimal 32, des expériences d'identification utilisant la base de données ARADIGIT_GSM ont été réalisées. Les résultats de ces expériences sont comparés à ceux trouvés avec la base de donnéess ARADIGIT8K. Le tableau 4.1 résume ces résultats comparatifs.

Type de base de données	Taux d'identification en (%)
ARADIGIT8K	**84.44**
ARADIGIT_GSM	**78.06**

Tab 4.1 Taux d'identifications comparatifs entre bases ARBDIGIT8K et ARADIGIT_GSM

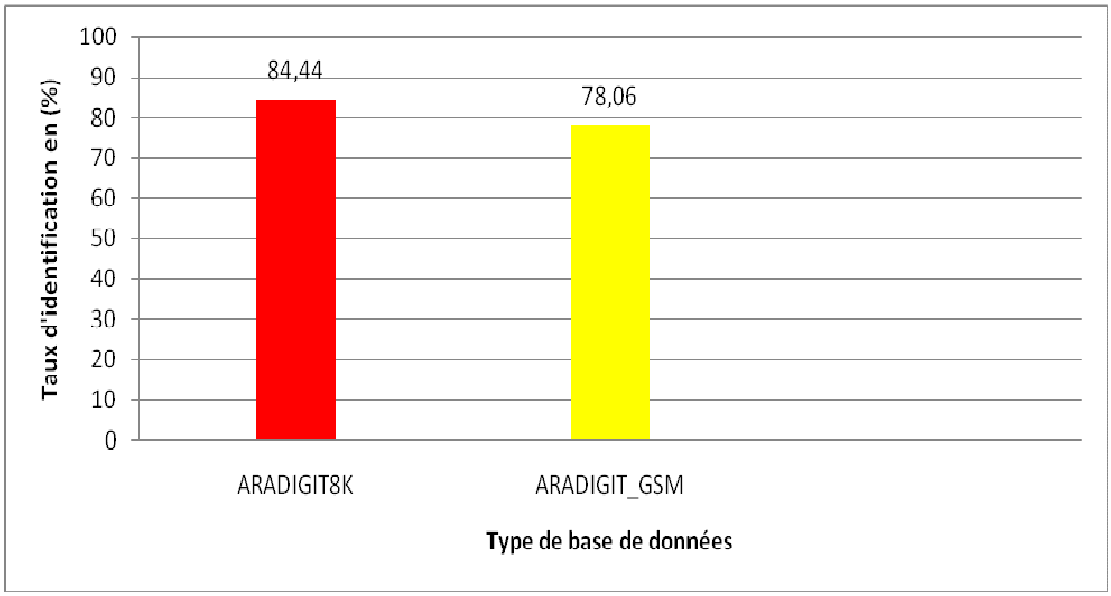

Figure 4.7 Taux d'identifications comparatifs entre bases ARBDIGIT8K et ARADIGIT_GSM.

La figure. 4.7 montre les variations du taux d'identification en fonction du type de base de données utilisée pour le test. Elle montre une dégradation du taux d'identification du locuteur de l'ordre 6% pour une base de données transcodée GSM (source uniquement).

4.5.3 Effet du canal de transmission

4.5.3.1 Expérience 3 : Canal de type gaussien

La simulation du canal est réalisée par l'addition du bruit blanc gaussien. Le canal de type AWGN ajoute du bruit blanc gaussien aux paramètres codés transmis comme décrit dans le chapitre 3, avec un niveau Rapport Signal sur Bruit (RSB : 20dB). Les résultats de cette expérience sont comparés à ceux trouvés avec la base de données ARADGIT8K et la base de données ARADIGIT_GSM. Le tableau 4.2 résume ces résultats comparatifs.

Type de base de données	Taux d'identification en (%)
ARADIGIT8K	**84.44**
ARADIGIT_GSM	**78.06**
ARADIGIT_GSM_AW	**68.73**

Tab 4.2 Les résultats comparatifs.

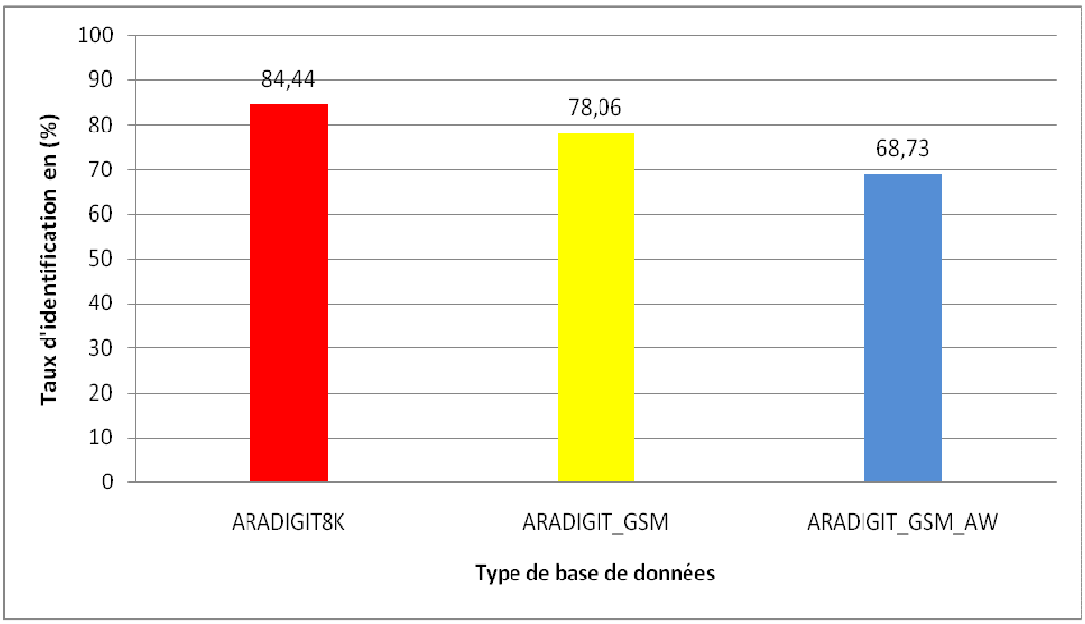

Figure 4.8 Les résultats comparatifs du taux d'identification.

La figure ci-dessus montre une dégradation du taux d'identification du locuteur de l'ordre de 16% pour la base de données transcodée GSM (source+ canal simulé par le bruit AWGN).

4.5.3.2 Expérience 4 : canal de type Rayleigh

Le canal de transmission est simulé par un canal de type Rayleigh décrit dans le chapitre 3, avec un niveau Rapport Signal sur Bruit (RSB : 20dB). Les résultats de cette expérience sont comparés à ceux trouvés avec les bases ARADIGIT16, ARADIGIT8K, ARADIGIT_GSM et ARADIGIT_GSM_AW. Le tableau 4.3 résume les résultats globaux comparatifs.

Type de base de données	Taux d'identification en (%)
ARADIGIT16K	**89.34**
ARADIGIT8K	**84.44**
ARADIGIT_GSM	**78.06**
ARADIGIT_GSM_AW	**68.73**
ARADIGIT_GSM_R	**60.19**

Tab 4.3 Les résultats globaux comparatifs.

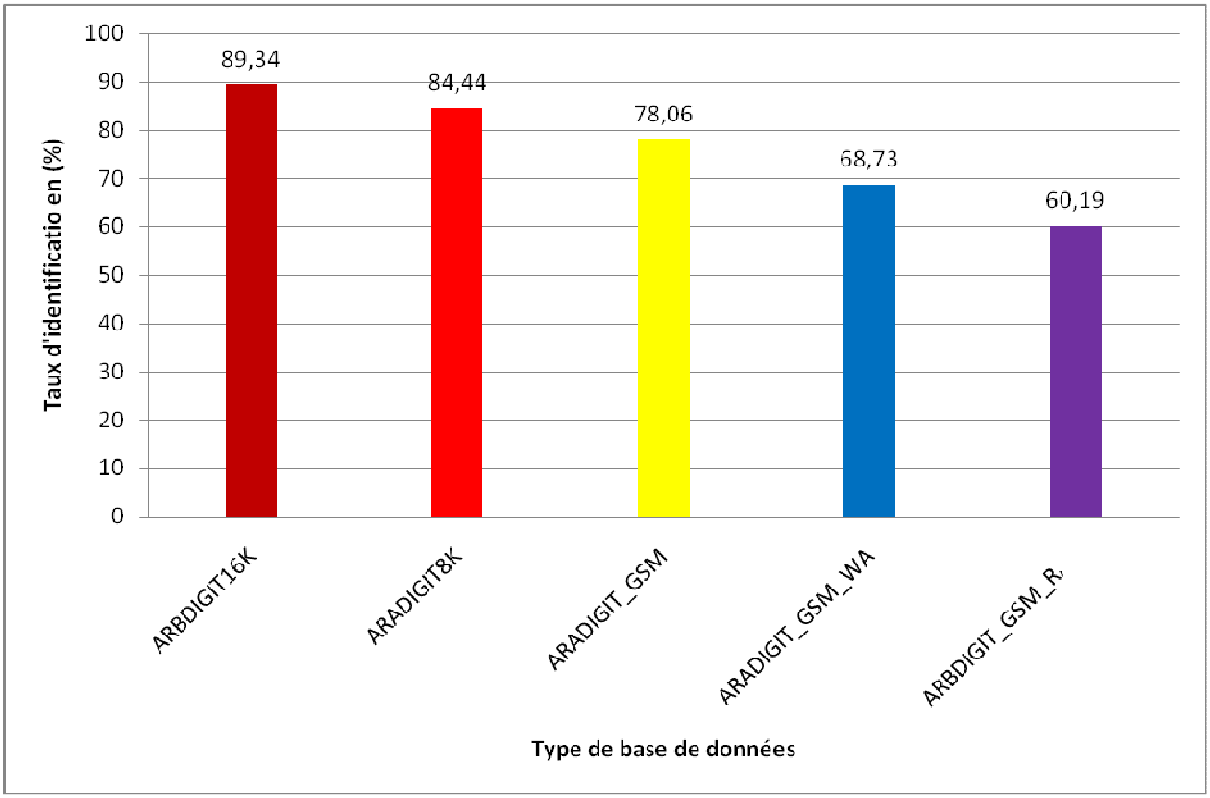

Figure 4.9 Les résultats globaux comparatifs.

Cette figure montre une dégradation du taux d'identification du locuteur de l'ordre de 24 % pour la base de données transcodée GSM (source + canal simulé par le bruit de Rayleigh).

4.5.5 Expérience 5 : Influence du niveau du bruit

Dans cette expérience, la base de test utilise les différentes bases de données simulées à différents niveaux Rapport Signal sur Bruit (RSB : 5dB, 10dB et 20dB). Les résultats expérimentaux obtenus pour le bruit de type AWGN et le bruit de type Rayleigh aux différents niveaux RSB sont résumés dans le tableau 4.4 et le tableau 4.5 respectivement.

Type de base de données	Taux d'identification en (%)
ARADIGIT8K/ ARADIGIT_GSM_AW à 20dB	**32.78**
ARADIGIT8K/ ARADIGIT_GSM_AW à 10dB	**20.24**
ARADIGIT8K/ ARADIGIT_GSM_AW à 5dB	**14.99**

Tab 4.4 Taux d'identification comparatifs entre la base ARADIGIT 8K /ARADIGIT_GSM_AW avec différents RSB.

Type de base de données	Taux d'identification en (%)
ARADIGIT8K/ ARADIGIT_GSM_R à 20dB	**34.79**
ARADIGIT8K/ ARADIGIT_GSM_R à 10dB	**18.14**
ARADIGIT8K/ ARADIGIT_GSM_R à 5dB	**10.88**

Tab 4.5 Taux d'identifications comparatifs entre la base ARADIGIT 8K /ARADIGIT_GSM_R avec différents RSB

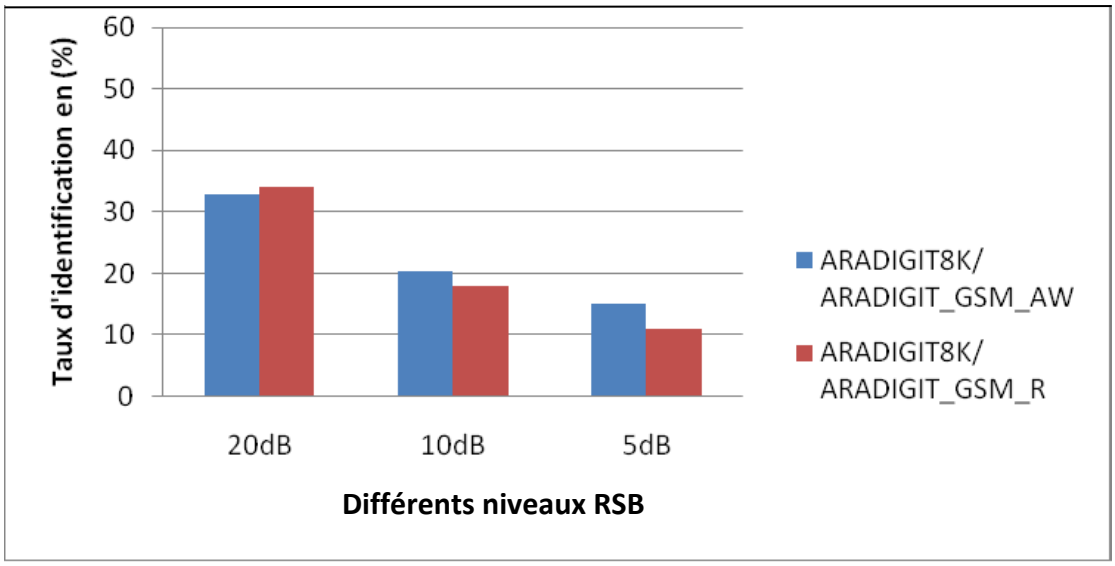

Figure 4.10 Taux d'identifications comparatifs entre la base ARADIGIT 8K /ARADIGIT_GSM_AW et ARADIGIT_GSM_R avec différents RSB.

Les résultats obtenus entre la base d'apprentissage non bruitée et la base de test transcodée bruitée (bruit de gaussien AWGN et bruit de Rayleigh) avec différents niveaux RSB montrent que le taux d'identification diminue avec le niveau du bruit. Cette diminution est plus marquée pour le bruit de type Rayleigh.

4.6 Conclusion

Dans ce chapitre nous avons présenté en premier l'évaluation de notre système d'Identification Automatique du Locuteur (IAL). Ainsi, la base de données ARADIGIT16K version originale échantillonnée à 16KHz a été utilisée, ce qui nous permit d'obtenir la structure optimale de notre système. Les expériences d'évaluation sur la base de données sous-échantillonnée à 8KHz et les bases de données transcodées GSM avec simulation du canal de transmission par des bruits de types gaussien et de type Rayleigh à différents niveaux RSB ont montré que les performances du système d'identification se dégradent avec le niveau du bruit et le type de bruit.

Conclusion et perspectives

Conclusion et Perspectives

Au cours de ce travail, nous avons traité le problème de Reconnaissance Automatique du Locuteur (RAL) en mode indépendant du texte via une ligne de communication GSM. La RAL consiste à reconnaître l'identité d'un individu à l'aide de sa voix. Essentiellement axée sur les problèmes d'authentification et de confidentialité, la RAL est très souvent sollicitée dans un contexte de téléphonie mobile GSM, pour lequel la voix demeure le seul élément disponible pour reconnaître une personne. Les signaux transmis par réseau GSM montrent un niveau de bruit élevé (les appels par téléphones mobiles sont souvent effectués dans des endroits très bruyants voiture, gare, rue que ceux d'un téléphone fixe). Dans notre travail, pour l'utilisation des signaux de type GSM réel nous avons simulé le fonctionnement de codec GSMEFR sous les boites outils de MATLAB version 7.4.

Deux axes ont été abordés dans ce mémoire, la simulation d'une ligne de communication GSM à partir du codec GSMEFR nous permettant d'obtenir les bases de données transcodées GSM et le développement du système de reconnaissance automatique du locuteur basé sur l'approche statistique GMM.
Nous avons ainsi étudié l'influence du codec GSM sur la performance du système de RAL basé sur les GMM. Les tests effectués sur les bases de données transcodées GSM ont montré que le taux d'identification du locuteur se dégrade sous l'effet du transcodage et particulièrement avec l'introduction du canal de transmission simulé dans notre travail par deux types de bruits : le bruit additif de type AWGN et le bruit de type Rayleigh à différents niveaux RSB. La distorsion du signal de parole est très forte par les effets du canal de transmission.

Comme perspectives à ce travail nous proposons :

- Les bruits dans l'environnement mobile restent encore un grand problème d'étude. Pour améliorer la performance du système de Reconnaissance du locuteur dans un environnement mobile GSM, il faut prévoir des méthodes de réduction de bruit ou des méthodes de rehaussement de la parole à l'entrée du système de reconnaissance.

- Les systèmes de RAL actuels reposent majoritairement sur des approches probabilistes. Parmi ces approches, les systèmes sont généralement basés sur une modélisation des locuteurs par des modèles génératifs, comme les modèles à mélange de gaussiennes (GMM), associée à une représentation du signal basée sur

des paramètres cepstraux, l'inconvénient majeur de cette technique est la quantité de signaux d'apprentissage pour une bonne estimation des paramètres des modèles, avec des ressources mémoire limitées. Pour résoudre le problème de classification nous proposons l'utilisation de l'approche discriminante SVM (Support Vector Machine) pour la modélisation du locuteur ou la combinaison des deux méthodes génératives et discriminantes GMM et SVM.

- La reconnaissance effectuée dans les réseaux GSM utilise des signaux codés et décodés ce qui engendre une perturbation de la qualité vocal par la perte des trames dans la synthèse de la parole. Une des solutions est de réaliser la reconnaissance sur les paramètres codés et transmit dans un canal spécial pour une meilleure protection contre les bruits.

Bibliographie

Bibliographie

[1] G. R. Doddington, ''Speaker recognition. Identifying people by their voices''. *IEEE transactions*, Volume 7311, pp. 1651–1664, 1985.

[2] A. E. Rosenberg et F. K. ''Soong, Recent research in automatic speaker recognition''. Advances in *speech signal processing*, 701–737, 1991.

[3] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini, ''Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition'', Proc. Of EUSIPCO, *European Signal Processing Conference*, Tampere, Finland, September 4-8, pp. 437- 440, 2000.

[4] http://www.etsi.fr

[5] Thierry Autret, Robert Bergeron, Muriel Collington, ''Techniques de contrôle d'accès par biométries, commission techniques de sécurité physique'', juin 2006.

[6] Karl Baumgartner, rapport ''Biométrie présentation personnelle'',2002.

[7] Florent Perronnin et Jean-Luc Degelay, ''Introduction à la biométrie Authentification des individus par traitement Audio-Video'', Revue traitement du signal, volume 19, numéro 4, 2002.

[8] J. Bonastre, F. Bimbot, L. Boë, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau, Person authentication by voice : a need for caution. Dans *Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland*, pp. 33–36. 7 2003

[9] HOLLIEN, H., The acoustics of crime. *Applied Psycholinguistics and Communication Disorders*, Plenum Press : New-York & London. 370p, 1990.

[10] KUNZEL, H.J., Current approaches to forensic speaker recognition. *In workshop on automatic Speaker Recogntion and Vérification*, pp 135-141, April Martigny (Switzerland), 1994.

[11] Eagles. Assessment of speaker verification systems. EAGLES Spoken Language Systems, 1995.

[12] FURUI, S., An overview of speaker recognition technology. *In workshop on automatic Speaker Recogntion and Vérification*, pp 1-9, Martigny (Switzerland), April 1994.

[13] NAIK, J.M., Speaker verification : a tutorial. *IEEE Communications Magazine*, pp 42-48, January 1990.

[14] NAIK, J., Speaker verification over the telephone network : databases, algorithms and performance assessment, *In workshop on automatic Speaker Recogntion and Vérification*, pp 31-38, April Martigny (Switzerland), 1994.

[15] S. E. Johnson, Who spoke when - automatic segmentation and clustering for determining speaker turns. Dans *Proceedings of European Conference on Speech Communication and Technology* (Eurospeech 99), 1999.

[16] P. Delacourt, La segmentation et le regroupement par locuteurs pour l'indexation de document audio. Thèse de Doctorat, ENST, 2000.

[17] K. Sonmez, L. P. Heck, et M. Weintraub, Speaker tracking and detection with multiple speakers. Dans *Proceedings of European Conference on Speech Communication and Technology* (Eurospeech 99), 1999.

[18] J.-F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin, et C. J. Wellekens, A speaker tracking system based on speaker turn detection for NIST evaluations. Dans *International Conference on Acoustics, Speech, and Signal Processing* ICASSP, 2000.

[19] M. Rossi. De la quiddité des variables. Actes du séminaire Variabilité et spécificité du locuteur : Etudes et Applications, 11–31, 1989.

[20] G. Ramaswamy, R. Zilca, et O. Alecksandrovich, A programmable policy manager for conversational biometrics. Dans Proceedings of Interspeech, European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 2003.

[21] J.-F. Bonastre, F. Bimbot, L. Boe, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau. Person authentication by voice : a need for caution. Dans les actes de EUROSPEECH, 33–36, 2003.

[22] Y. Grenier. Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique. Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications ENST, 1977.

[23] A. Oppenheim et R. Schafer. Discrete-time signal processing. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, , 1989 .

[24] Bouchefra Khelifa, Thèse de magister à l'ENP : « Contribution à la reconnaissance automatique de la parole continue : Etude et réalisation d'un système de reconnaissance acoustico-phonétique », 1995.

[25] J.Makhoul, « Linear prediction: A tutorial review » *Proc, IEEE,* vol. 63, pp. 561-580. April 1975.

[26] M.Kunt, « Traitement numérique des signaux ». Presses polytechniques romandes, Lausanne, 1980.

[27] D.A. Reynolds, C.Rose, « Robust Text – Independent Speaker Identification Using Gaussian Mixture Speaker Models » IEEE transaction on Speech and Audio Processing, Vol. 3, No. 1. January 1995.

[28] Furui S. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions Acoustics, Speech, and Signal Processing* (ASSP), volume 29(2), pages 254-272, Avril 1981.

[29] Booth I., Barlow M., Watson B. Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, volume 13(3-4), pages 427-433, December 1993.

[30] Yu K., Mason J. S., Oglesby J. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE vision, image and signal processing*, Berlin (Allemagne), 1995.

[31] Soong F. K., Rosenberg A. E., Rabiner L. R., Juang B. H. A vector quantization approach to speaker recognition. *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), pages 387-390, Tampa (USA), 1992.

[32] Mason J. S., Oglesby J., Xu L. Codebooks to optimise speaker recognition. *European Conference on Speech Communication and Technology (Eurospeech),* pages 267-270, Paris (France), 1989.

[33] Matsui T., Furui S. Comparison of text-independent speaker recognition methods using VQ-distorsion and discrete-continuous HMMs. *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), pages 157-160, San Francisco (USA) 1992.

[34] Pruzansky S. Pattern matching procedure of automatic talker recognition. Journal of Acoustical Society of America (JASA), volume 35, pages 354-358, 1963.

[35] Biombot, F,, Magrin-Chagnolleau, I,, et Mathan, I,, Second-ordre statistical measures for text-independent speaker identification. Speech Communication, 17: 177-192, 1995.

[36] Magrin-Chagnolleau, I,, Bouastre, J, F,, et Bimbot, F, Effet of utterance duration and content on speaker identification using second-ordre statistical method. Dans *European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, page 337-340, 1995.

[37] Reynolds, D, Speaker identification and verification using Gaussian mixture speaker models. Speech Communication, 17(1): 91-108, 1995.

[38] Savic, M, et Cupta, S. k Variable parameter speaker verification système based on hidden Markov moding. Dans *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) , volume 1, pages 281-284, 1990.

[39] Rosenberg, E,, Lee, C-H,, et Gokcen, S, Connected word talker verification using whole word hidden Markov modin. Dans *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) , volume 1, pages 381-384,1991.

[40] Oglesby J., Mason J. S. Optimisation of neural models for speaker identification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 261-264, 1990.

[41] Oglesby J., Mason J. S. Radial basis function networks for speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 393-396, Toronto (Canada), 1991.

[42] Reynolds, D. “Speaker identification and verification using gaussian mixture speaker models”, *Speech Communication*, vol. 17, pp. 91–108,1995.

[43] A.P.Dempster, N.M.Laird, and D.B.Rubin. ‘’Maximum-likelihood from incomplete data via the EM algorithm’’. J. Royal Statist. *Soc. Ser. B*, 39, 1997.

[44] J.Bilmes. ‘’A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models’’, Report, University of Berkeley, ICSI-TR-97-021, 1997.

[45] Tapas Kanungo, Senior , David M. Mount, ‘’An Efficient k-Means Clustering Algorithm: Analysis and Implementation’’, IEEE *transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, july 2002

[46] René Boite, Hervé Bourlard, Thierry Dutoit, Joël Hancq et Henri Leich : « Traitement de la parole » – Presses Polytechniques et universitaires Romandes 2000.

[47] N. Ramou et M. Djeddou, '' Détection de genre et technique de normalisation des scores pour la vérification du locuteur'', *1st International Conference on **I**mage and **S**ignal **P**rocessing and their **A**pplications*, Mostaganem, Algeria, octobre, 2009.

[48] Hassan Ezzaidi, Jean Rouat and Douglas O`Shaughnesy « Combining pitch and MFCC for speaker recognition systems » ERMETIS, Université du Québec à Chicoutimi, Québec, Canada, G7H 2B1. INRS-Télécommunications, Université du Québec, 2001.

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

[50] Jean-Philippe Muller "Le réseau GSM et le mobile ''. Version 2002.

[51] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, F. Pellandini, " Influence of GSM Speech Coding on the Performance of Text-Independent Algorithms Speaker Identification", Published in Proceedings of the International COST 254 Workshop on Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communications, 307-311, 1999.

[52] Nemat. S. Abdel Kader, '' effect of GSM system on text-independent speaker recognition performance'', Cairo University, Faculty of Engineering, Communication Dept, *Journal of Theoretical and Applied Information Technology* JATIT. All rights reserved, 2005.

[53] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, F. Pellandini, "GSM Speech Coding and Speaker recognition", ICASSP, International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, June 5-9, Vol.II, pp. 1085-1088, , 2000.

[54] Ahmed.Krobba, Mohmed. Deybeche '' Effet de la Parole Transcodée GSM sur la Performance d'un Système de Reconnaissance Automatique du Locuteur'', *1st International Conference on **I**mage and **S**ignal **P**rocessing and their **A**pplications*, Mostaganem, Algeria, octobre, 2009.

[55] A. Amrouche "Reconnaissance automatique de la parole par les modèles connexionnistes". Thèse de doctorat, faculté d'électronique et d'informatique, USTHB. 2007.

[56] Lukáš Burget, , Pavel Matějka, Petr Schwarz,Ondřej Glembek, and Jan "Honza" Černocký, "Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System", *IEEE Transactions on Audio, Speech, and Language processing* ,Vol. 15, No. 7, September 2007.

[57] Anissa Imen. Amrous, Mohmed.Debyeche, Ahmed. Krobba,'' coopération de connaissances dans les modéles de Markov cachés pour la reconnaissance de mots isolès arabe'', *1st International Conference on **I**mage and **S**ignal **P**rocessing and their **A**pplications*, Mostaganem, Algeria, octobre, 2009.

[58] Douglas A. Reynolds, Thomas F. Quatieri,'' Speaker Verification Using Adapted Gaussian Mixture Models" Digital Signal Processing **10**, 19–41, 2000.

[59] Ahmed Krobba, Mohmed. Deybeche, Abederrahmane. Amrouche. ''Evaluation of Speaker Identification System Using GSMEFR Speech Data'', *International conference on Design & Technology of Integrated Systems* (DTIS), Tunis, March, 2010**.**