N° d'ordre : 15/2012-M/EL

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

MEMOIRE
Présenté pour l'obtention du diplôme de MAGISTER
En : ELECTRONIQUE
Spécialité : **Communication Parlée**

Par
FEDILA Meriem

THÈME

RECONNAISSANCE AUTOMATIQUE DE LOCUTEURS ROBUSTE POUR LES COMMUNICATIONS MOBILES

Soutenu publiquement le 15 janvier 2012, devant le jury composé de :

Mr	**FERGANI Belkacem**	**Maître de Conférences/A, à USTHB**	**Président**
Mr	**AMROUCHE Abderrahmane**	**Maître de Conférences/A, à USTHB**	**Directeur de mémoire**
Mme	**FALEK Lila**	**Maître de Conférences/A, à USTHB**	**Examinatrice**
Mr	**TEFFAHI Hocine**	**Maître de Conférences/A, à USTHB**	**Examinateur**

Dédicaces

Je dédie ce modeste travail :

A la lumière de ma vie, la cause de ma présence, les deux cœurs pleins d'amour de tendresse et de bonté, mes ***chers parents*** *qu'ils méritent tous mon respect et ma grande gratitude. Je vous remercie pour votre soutien, vos encouragement, vos conseilles, votre aide à la réalisation de ce travail. Que dieu vous protège et prolonge votre vie.*

A ma chère sœur ***AICHA*** *et mes abordables frères* ***MOUAD*** *et* ***ABDALLAH.***

A mes chères ***MIMI*** *et* ***Fatima Zohra.***

A mes collègues de **laboratoire 70** *pour leurs aides durant tous le parcours de mon travail.*

MERCI

Remerciement

Après avoir remercie mon ***DIEU***

Un très grand Merci a mon directeur de thèse, monsieur ***AMROUCHE Abderrahmane****, Maitre de conférences classeA à la faculté d'électronique et d'informatique (FEI), USTHB, pour tout l'effort qu'il a consacré au suivi de mon travail. Ce sont ses précieux conseils, remarques, contributions, sa grande attention et son écoute qui m'ont permis d'effectuer ce travail.*

Je tiens à remercier Mr ***FERGANI Belkaceme****, Maître de Conférences à la Faculté d'Electronique et d'Informatique, USTHB, pour l'honneur qu'il me fait de présider ce Jury.*

Je voudrai également exprimer mes remerciements à Mme ***FALEK Lila****, Maître de Conférences à la Faculté d'Electronique et d'Informatique, USTHB, et Mr* ***TEFFAHI Houcine****, Maître de Conférences à la Faculté d'Electronique et d'Informatique, USTHB pour avoir bien voulu accepter de faire partie ce Jury.*

Finalement, je tiens à remercier tous ceux qui ont contribué de près ou de loin à la concrétisation de ce travail.

Résumé

L'émergence de la reconnaissance automatique de locuteur (RAL) pour les applications téléphoniques est née du besoin d'éviter les imposteurs dans certains domaines sensibles. En effet, les transactions boursières, l'accès sécurisé restrictif, etc., nécessitent l'authentification ou la vérification de donneur d'ordre distant.
Pour cela, le développement d'applications qui permettrait une meilleure extraction des paramètres caractéristiques de la voix et leur modélisation, tout en réduisant les influences des systèmes de codage et de transmission des réseaux téléphoniques, constitue un chalenge intéressant pour la recherche dans ce domaine.
Les applications envisagées dans ce travail concernent les systèmes de communications mobiles, et en particulier l'utilisation du codec adaptatif multidébits à large bande AMR_WB (Adaptive Multi-Rate Wideband), destiné à la bande élargie et qui utilise la technologie ACELP (Algebraic Code Excited Linear Prediction). Ce codec a été reconnu comme norme par l'ITU-T sous le nom de G.722.2**.**

L'objectif principal du sujet abordé dans ce présent mémoire est d'étudier et évaluer deux types de modélisation (Modèle de Mélange du Gaussienne - GMM, machine à vecteur de support - SVM), et les méthodes d'extraction des paramètres à savoir les paramètres issus du codeur AMR_WB, et ceux issus du signal resynthétisé (transcodé) et leur impact sur la RAL robuste.

A travers les expériences que nous avons effectuées sur la base de données ARADIGIT pour l'identification automatique de locuteur IAL en mode indépendant du texte, nous avons montré que le modèle GMM peut représenter des distributions aléatoires d'une manière fidèle en environnement calme.
Nous avons également mis en évidence que les SVM sont plus robuste que les GMM à travers l'effet du canal de transmission (simulé par le bruit blanc gaussien additif ou AWGN :Additive White Gaussian Noise, par le bruit de Rayleigh et par un canal binaire symétrique BSC-Binary Symetric Channel), et du bruit de l'environnement (simulé par des bruits issus de la base NOISEX'92 NATO, Varga).

Nous avons montré aussi que l'utilisation des paramètres issus du signal codé permet d'approcher les résultats de la référence (signal original). La reconnaissance de locuteur basée sur l'utilisation des paramètres internes de codeur est une solution robuste. Ceci permet notamment d'éviter les distorsions dues à la resynthèse (reconstruction) du signal et la consommation de ressources supplémentaires, mais aussi de minimiser le temps d'exécution.

Mot clés : IAL, GMM, SVM, AMR_WB, Parole transcodée, RAL robuste.

Table des matières

Chapitre II La modélisation des vecteurs acoustiques

Chapitre III Les systèmes de communications mobiles

Chapitre IV Reconnaissance Automatique de Locuteurs robuste pour les communications mobiles

Liste des figures

Liste des tableaux

Liste des Abréviations

AMR_WB	Adaptatif Multi Rate WideBand
ANN	Artificial Neural Network
AWGN	Additive White Gaussian Noise
BSC	Binary Symetric Channel
CELP	code excited linear prediction
DET	Detection Error Tradeoff
DSR	Distributed Speech/Speaker Recognition
DTW	Dynamic Time Warping
EER	Error Equal Rate
EFR	Enhanced Full Rate
EM	Expectation-Maximisation
GMM	Gaussian Mixture Models
GSM	global System for Mobile communication
HMM	Hidden Markov Models
IAL	Identification Automatique du Locuteur
ISF	Immitance Spectral Frequency
LFCC	Linear Frequency Cepstrum Coefficient
LPC	Linear Prediction Codage
LPCC	Linear Prediction Cepstral Coefficients
LSF	Line Spectral Frequency
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
NSR	Network Speech/Speaker Recognition
RAL	Reconnaissance Automatique du Locuteur
RBF	Radial Basis Function
ROC	Receiver Operating Characteristics
RSB	Rapport Signal sur Bruit

SVM	Support Vector Machines
TSR	Terminal Speech/Speaker Recognition
VAD	Voice Activity Detection
VAL	Vérification Automatique du Locuteur
VQ	Vector Quantization
UBM	Universal Background Model
UMTS	Universal Mobile Telecommunications System

Introduction générale

La parole est un signal aléatoire complexe qui contient plusieurs types d'informations. En premier lieu, le signal de parole véhicule un message linguistique qui sert à la communication entre individus. Mais la parole transporte également des informations sur l'identité de l'individu ayant prononcé le message. Les humains se servent de ces informations pour identifier les personnes qu'ils connaissent, en particulier lorsqu'ils ne peuvent pas voir leur interlocuteur, au téléphone par exemple.

Les systèmes de reconnaissance automatique du locuteur RAL s'intéressent précisément à ces caractéristiques particulières du signal de parole pouvant permettre de déterminer automatiquement l'identité de celui qui parle.

Les applications des systèmes de reconnaissance automatique de locuteur se distinguent par leur contexte applicatif et leur niveau de sécurité. En effet, L'émergence de la RAL pour les applications téléphoniques est née du besoin d'éviter les imposteurs dans certains domaines sensibles. Les personnalités occupant des responsabilités stratégiques, les transactions boursières, l'accès sécurisé restrictif, etc., nécessitent l'authentification ou la vérification de donneur d'ordre distant.

Les applications envisagées dans ce projet concernent les systèmes de communications mobiles, en particulier avec l'utilisation du codec AMR_WB (Adaptive Multi-Rate Wide-Band)**.** La technologie AMR (Adaptive Multi-Rate) a été développée du fait sa robustesse aux erreurs de transmission. Il s'agit d'un codeur multi débits, c'est-à-dire pouvant fonctionner à plusieurs débits. Le rôle des ces codecs AMR 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.

Le codeur AMR-WB destiné à la bande élargie a quant à lui été normalisé en 2000 au 3GPP (3d Generation Partnership Projet). Il a été également reconnu comme norme par l'ITU-T (Telecommunication Standarisation Sector of ITU) en Juillet 2003 sous le nom de G.722.2 **[1].** Ce codeur utilise la technologie ACELP (Algebraic Code Excited Linear Prediction), technologie qui est utilisée également dans les codecs AMR_NB et EFR.

Ce travail de mémoire, dédié à la reconnaissance automatique de locuteurs robuste pour les communications mobiles, s'intéresse en particulier à l'influence du transcodage AMR_WB sur la performance d'un système de reconnaissance automatique de locuteur. Il s'agit aussi d'étudier et de mettre en œuvre d'une méthode d'extraction des paramètres dans le domaine compressé, c.-à.-d à partir des paramètres internes du codeur AMR_WB.

Notre système de RAL est bâti sur l'utilisation de différentes méthodes de classification. En effet, les systèmes RAL nécessitent une étape de modélisation statistique des paramètres acoustiques, extraits du locuteur. La technique de modélisation prédominante est basée sur des modèles de mélange de gaussiennes (GMM) **[2]**. Citons cependant certaines méthodes alternatives, comme par exemple les réseaux de neurones probabilistes **[3].** Plus récemment, les approches par "Support Vector Machine : machine à vecteur de support" (SVM) **[4]** ont fait une percée parmi les méthodes les plus performantes en RAL.

Ce mémoire se compose de quatre chapitres, organisés comme suit :

Le chapitre 1 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 et la vérification de locuteur) et enfin les méthodes de reconnaissance les plus utilisées actuellement.

Dans le chapitre 2, nous décrirons les méthodes de modélisation statistique que nous avons utilisées dans ce travail. Nous commençons d'abord par détailler le modèle statistique le plus utilisé dans les systèmes RAL en mode indépendant du texte, à savoir, le modèle GMM (Gaussian Mixture Model). Puis nous abordons la théorie des machines à vecteurs de support ou SVM utilisé comme classifieur discriminatif.

Dans le chapitre 3, nous donnons un aperçu général sur les systèmes de communications mobiles. Après cela, nous allons décrire le codeur et le décodeur d'adaptation de haute qualité à des débits multiples en large bande (AMR_WB).

Le chapitre 4 contient l'ensemble des tests effectués et les résultats que nous avons obtenus.

Le dernier chapitre conclue ce travail et met l'accent sur quelques problèmes qui peuvent être traités par des futurs travaux.

Généralités sur la Reconnaissance Automatique du Locuteur

I. 1 Introduction

La reconnaissance automatique du locuteur RAL est le processus qui identifie automatiquement celui qui parle en se basant sur des informations individuelles incluses dans le signal de parole.

La reconnaissance automatique du locuteur par leur voix est une tâche très complexe et nécessite la prise en compte de plusieurs paramètres. Le type d'application pour lequel cette reconnaissance est faite, impose l'utilisation des méthodes bien définies pour avoir des résultats significatifs. L'objectif principal de ce chapitre est de présenter les différents types des systèmes de reconnaissance automatique du locuteur**.**

I.2 Domaine d'application de la RAL

Les applications directes de la RAL concernent les problèmes de confidentialité et d'authentification. Nous distinguerons :

1. les applications "sur site" : serrures vocales pour contrôle d'accès, cabines bancaires en libre service **[5]**.
2. 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.
3. les applications judiciaires: recherche de suspects, orientations d'enquêtes, preuves lors d'un jugement.

La difficulté de la tâche de reconnaissance n'est pas la même d'une application à l'autre. Dans le cas des applications 'sur site', l'environnement de prononciation de la phrase ou du mot de passe est plus facilement contrôlé que dans le cas des applications via le réseau téléphonique (distorsions dues au canal, différences entre les combinés téléphoniques, bande passante limitée).

Les systèmes RAL ont pour mission de décoller l'information portée par le signal vocal. On classe également les systèmes RAL comme suit :

- On sépare reconnaissance du locuteur dépendante du texte et reconnaissance du locuteur indépendante du texte. Dans le premier cas, la phrase à prononcer est fixée dès la conception du système; et n'est pas précisée dans le deuxième cas.
- En reconnaissance du locuteur, on fait la différence entre la vérification et l'identification du locuteur, selon que le problème est de vérifier que la voix analysée correspond bien à la personne qui est censée de la produire, ou qu'il s'agit de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé.

I.3 L'identification du locuteur

C'est le processus de détermination de celui susceptible de produire un enregistrement vocal. D'un point de vue schématique (Figure.1.1), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL.

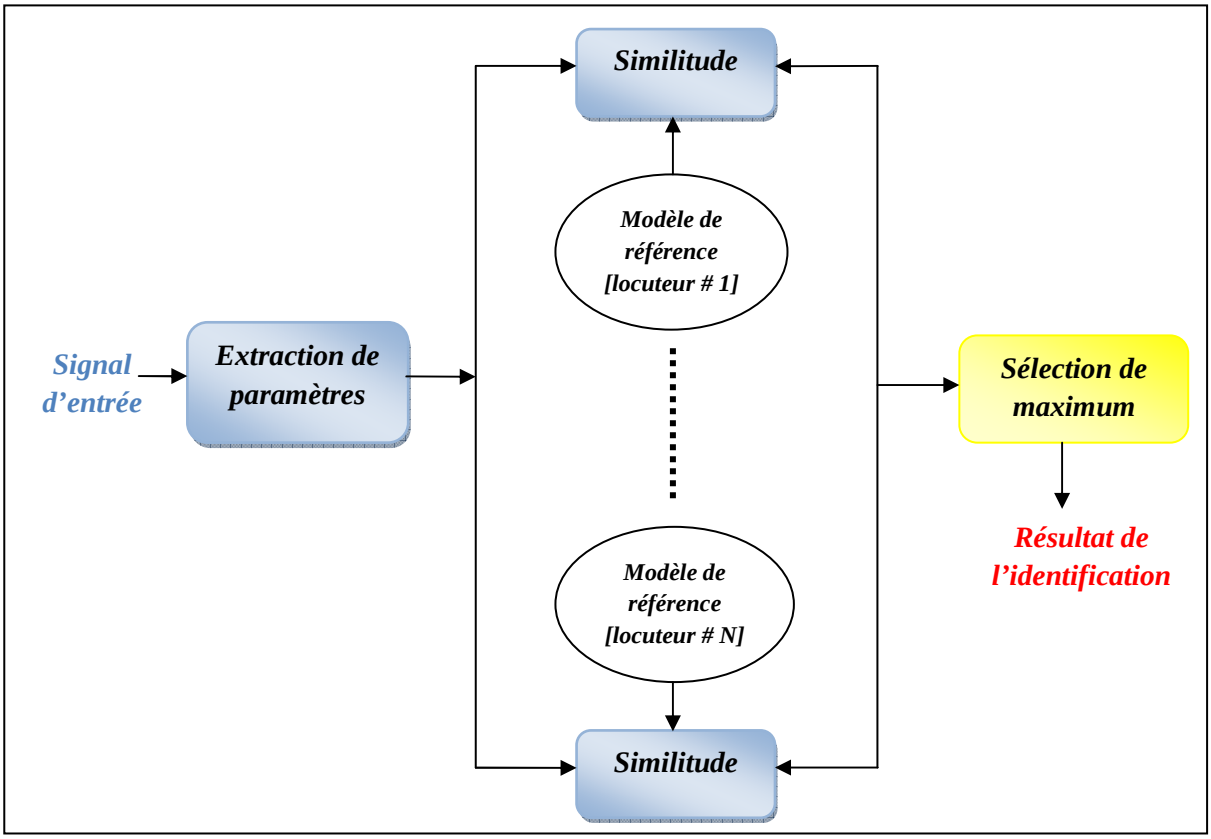


Figure.1.1 *Schéma d'un système d'identification du locuteur.*

I.4 La vérification du locuteur

Il s'agit de vérifier que la voix analysée correspond bien à la personne qui est censée de la produire. C'est le processus qui prend la décision d'accepter ou de rejeter l'identité d'un locuteur susceptible d'être la source d'un enregistrement vocal. Le schéma de principe de la vérification du locuteur est donné par la Figure.1.2.

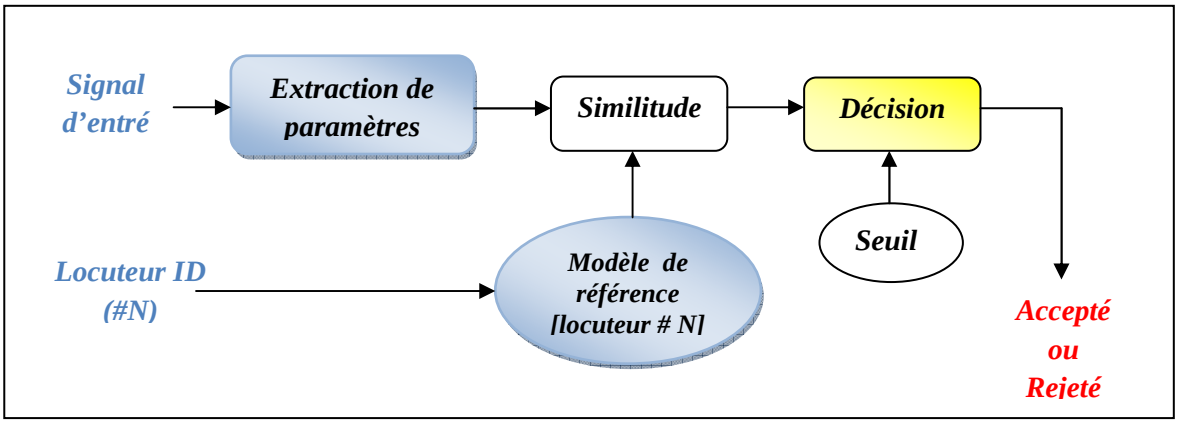


Fig.1.2 *Schéma d'un système de vérification du locuteur.*

I.5 Une comparaison entre l'identification et la vérification automatique du locuteur

Il est clair que les deux systèmes de reconnaissance automatique du locuteur, à savoir, l'identification et la vérification, ont une tâche commune qui est la comparaison d'une voix connue avec une autre non connue pour répondre à la question suivante : *est ce que les deux voix sont produites par le même locuteur?* Malgré que le but des deux systèmes est le même (trouver la source d'une voix), ces derniers présentent des grandes différences. La distinction la plus importante concerne les propriétés de l'ensemble de référence des locuteurs, elle est faite selon que, premièrement, l'ensemble est *fermé* ou *ouvert*, deuxièmement, l'ensemble est *connu* ou *inconnu* **[6]**.

I.5.1 L'identification à ensemble fermé ou ouvert

Dans les systèmes d'identification du locuteur, l'ensemble de référence des locuteurs peut être de deux types : *fermé* ou *ouvert*. Dans un ensemble fermé, on suppose que la voix traitée est produite par l'un des locuteurs de l'ensemble de référence. Par contre, dans un ensemble ouvert, on ne sait pas si le propriétaire de la voix traitée appartient à l'ensemble de référence ou non.

Il est très important de faire cette distinction. L'identification à ensemble fermé est plus simple que celle à ensemble ouvert.
Comme on sait que la voix provient de l'un des locuteurs de référence, l'identification se fait comme suit :
• Calculer la distance entre la voix traitée et chaque locuteur de l'ensemble de référence.
• Choisir le locuteur qui a la distance la plus petite.
Dans l'identification à ensemble ouvert, on ne peut pas dire que le locuteur qui a la distance la plus petite est la source de la voix traitée. On doit avoir un seuil prédéfini d'une façon que le locuteur qui a une distance inférieur au seuil sera considéré comme étant la source de la voix traitée.

I.5.2 L'ensemble connu et l'ensemble inconnu

La distinction entre l'ensemble fermé et l'ensemble ouvert n'a aucun sens dans un système de vérification du locuteur. Dans ce dernier, on suppose que les identités à proclamer appartiennent à l'ensemble de référence des locuteurs. Cependant, il y a une autre propriété de l'ensemble de référence qui est : *connu* ou *inconnu*. Dans la vérification traditionnelle l'ensemble de référence est connu, il peut être les employés d'une entreprise, les clients d'une banque ou en général l'identité de toutes personnes qui doivent être vérifiées de temps en temps.

I.5.3 Nombre de comparaisons

Le nombre de comparaisons faites lors d'une tâche d'identification ou de vérification diffère selon cette tâche. La vérification du locuteur nécessite une seule comparaison entre l'identité proclamée et la voix traitée. Dans l'identification du locuteur, la comparaison apparaît itérativement entre la voix traitée et chaque locuteur de l'ensemble de référence

I.6 Structure d'un système de RAL

Un système de reconnaissance automatique du locuteur (Figure.1.3) est composé de trois modules importants : *Paramétrisation, classification et décision.*

- ✓ Un premier module de traitement du signal réalise l'extraction des vecteurs acoustiques, qui représentent les données pertinentes à l'apprentissage des systèmes RAL.
- ✓ Dans le deuxième module, on fait une modélisation des vecteurs acoustiques obtenus dans l'étape précédente le signal est représenté par des vecteurs de coefficients. Dans la phase de classification, les vecteurs du signal de test (ou leur modèle) sont comparés aux vecteurs des locuteurs de référence (ou à leurs modèles).
- ✓ La phase de décision désigne le locuteur finalement reconnu.

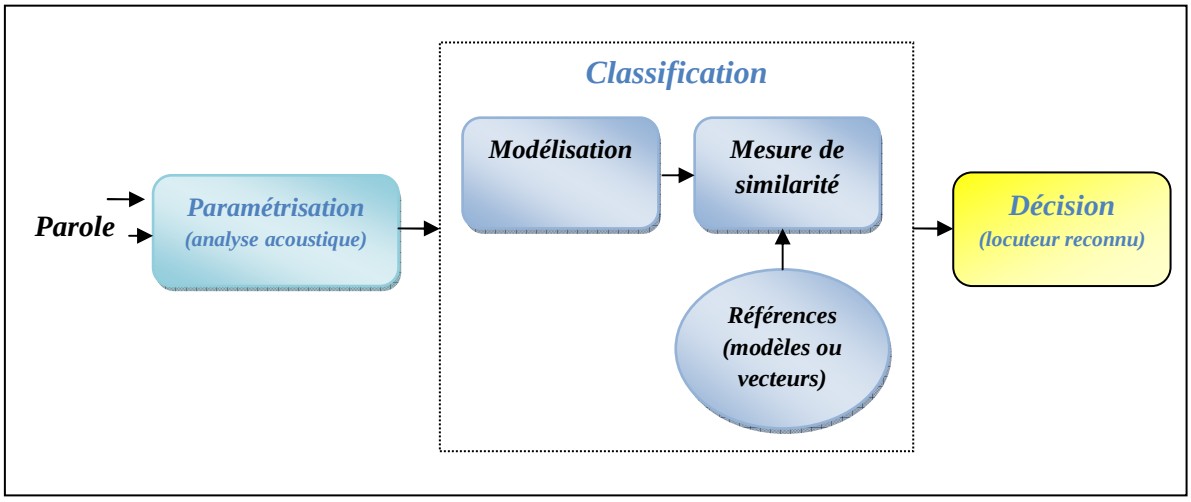


Figure.1.3 *Structure générale d'un système de RAL.*

I.7 Paramétrisation

L'extraction des paramètres acoustiques est une étape très importante dans un système de reconnaissance automatique du locuteur. Le type de paramètres utilisés a un grand effet sur la performance du système. Dans le cas idéal, les paramètres acoustiques utilisés dans un système de reconnaissance automatique du locuteur doivent satisfaire les contraintes suivantes **[7]** :

- être fréquents (ne pas correspondre à des événements ne survenant que très rarement dans le signal de parole d'un locuteur).
- être facilement mesurables.
- ne pas être trop sensibles à la variabilité intra-locuteur.
- ne pas être affectés par le bruit ambiant ou par les variations dues au canal de transmission.
- être robustes à la transmission et au bruit.

I.7.1 Paramètre de l'analyse spectrale

Les principaux paramètres de l'analyse spectrale utilisés en RAL sont :

- 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 **[8]** : 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) et les coefficients LSF (Line Spectral Frequency).

Les paramètres de l'analyse spectrale sont généralement répartis suivant une échelle linéaire ou une échelle Mel.

❖ **Paramètres dynamique**

La prise en compte d'une information de type dynamique est un facteur d'amélioration des performances d'identification du locuteur. Les dérivées du premier et du second ordre, appelées aussi coefficients de delta et delta-delta, permettent d'introduire une information concernant le contexte temporel d'une trame courante **[9]**.
D'après Harb **[10]** l'utilisation de ces paramètres reste l'approche la plus populaire actuellement en raison de la simplicité pour sa mise en œuvre et l'amélioration des performances que nous pouvons observer.

I.7.2 Paramètre prosodique

Le terme "paramètres prosodiques" réunit l'énergie, la durée et la fréquence fondamentale (ou pitch) **[11]**. Ces paramètres s'avèrent cependant fragiles en pratique et ne permettent pas, à eux seuls, de discriminer les locuteurs. En conséquence, ils sont souvent associés aux paramètres de l'analyse spectrale (surtout l'énergie).

I.7.3 Extraction des paramètres

Presque la plus part des informations qui peuvent être extraites d'un signal de parole se trouvent dans la bande fréquentielle 200Hz-8KHz. Les étapes principales pour extraire les vecteurs acoustiques sont illustrées dans la Figure.1.4.

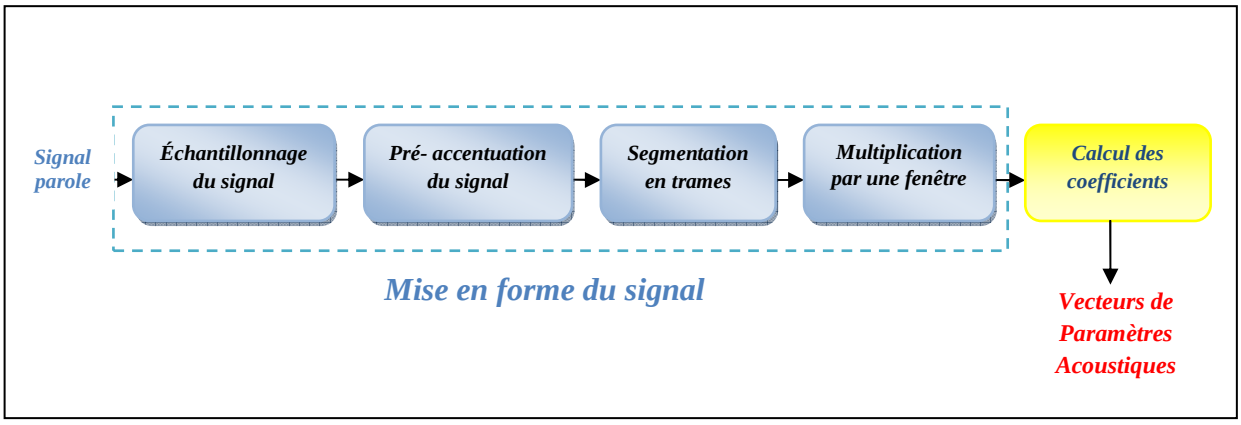


Figure.1.4 *Schémas synoptique de l'extraction de paramètres acoustiques.*

a. Echantillonnage

L'échantillonnage consiste à transformer un signal analogique x(t) (continu) en un signal numérique x(n) (discret), en capturant des valeurs à intervalle de temps régulier.
Le signal de parole est continu, ce qui rend son traitement par la machine difficile, il s'avère nécessaire de le numériser avant tout traitement. La fréquence d'échantillonnage doit vérifier la condition de Shannon donnée comme suite :

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

f_e: Fréquence d'échantillonnage.
f_{max}: Fréquence maximale du signal a traiter.

b. La préaccentuation

Le spectre d'un signal de parole a une décroissance globale de l'énergie. Pour compenser cette décroissance, on effectue une préaccentuation en utilisant un filtre passe-haut. Le filtre le plus utilisé est le filtre à réponse impulsionnelle finie décrit ci-dessous:

$$H(z) = 1 - 0.95z^{-1}. \tag{1.2}$$

La réponse de ce filtre peut être vue dans la Figure.1.5.

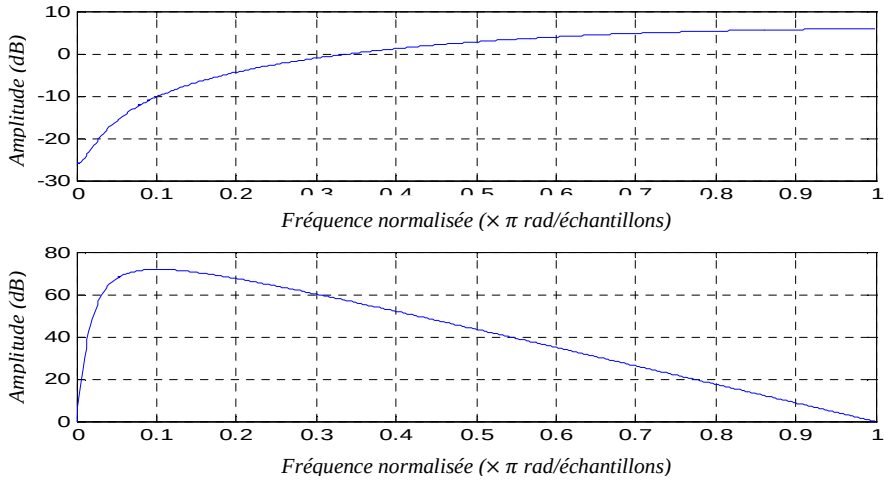


Figure.1.5 *Le filtre de la préaccentuation.*

c. L'élimination du silence

Dans la production d'un énoncé de parole, toute la séquence n'est pas utilisable pour identifier un locuteur **[12]**. Il est essentiel de retirer les trames de silence ou de bruits qui ne reflètent pas les caractéristiques du locuteur et qui diminuent les performances de RAL.
L'élimination des zones de silence est une tâche très importante. Cette tâche semble relativement triviale, mais elle présente quelques difficultés dans la pratique. Les mesures les plus utilisées

pour trouver et éliminer le silence sont : l'énergie du signal, la puissance du signal et le rapport de passage par zéro. Pour un signal de parole s(n) ces mesures sont calculées comme suit :

$$E_s(m) = \sum_{n=1}^{m} s^2(n) \tag{1.3}$$

$$P_s(m) = \frac{1}{L}\sum_{n=1}^{m} s^2(n) \tag{1.4}$$

$$Z_s(m) = \frac{1}{L}\sum_{n=1}^{m} \left| sgn\big(s(n)\big) - sgn(s(n-1)) \right| \tag{1.5}$$

$$sgn\big(s(n)\big) = \begin{cases} +1, s(n) \geq 0 \\ -1, s(n) < 0 \end{cases} \tag{1.6}$$

Il est à noter que l'index pour ces fonctions est *m* et pas *n*, car ces mesures ne sont pas calculées pour chaque échantillon. L'énergie s'accroît quand le signal s(n) contient de la parole et c'est le cas aussi pour la puissance. Le rapport de passage par zéro donne une mesure du nombre de fois où le signal s(n) change de signe. Ce rapport est en général plus grand dans les régions non voisées.

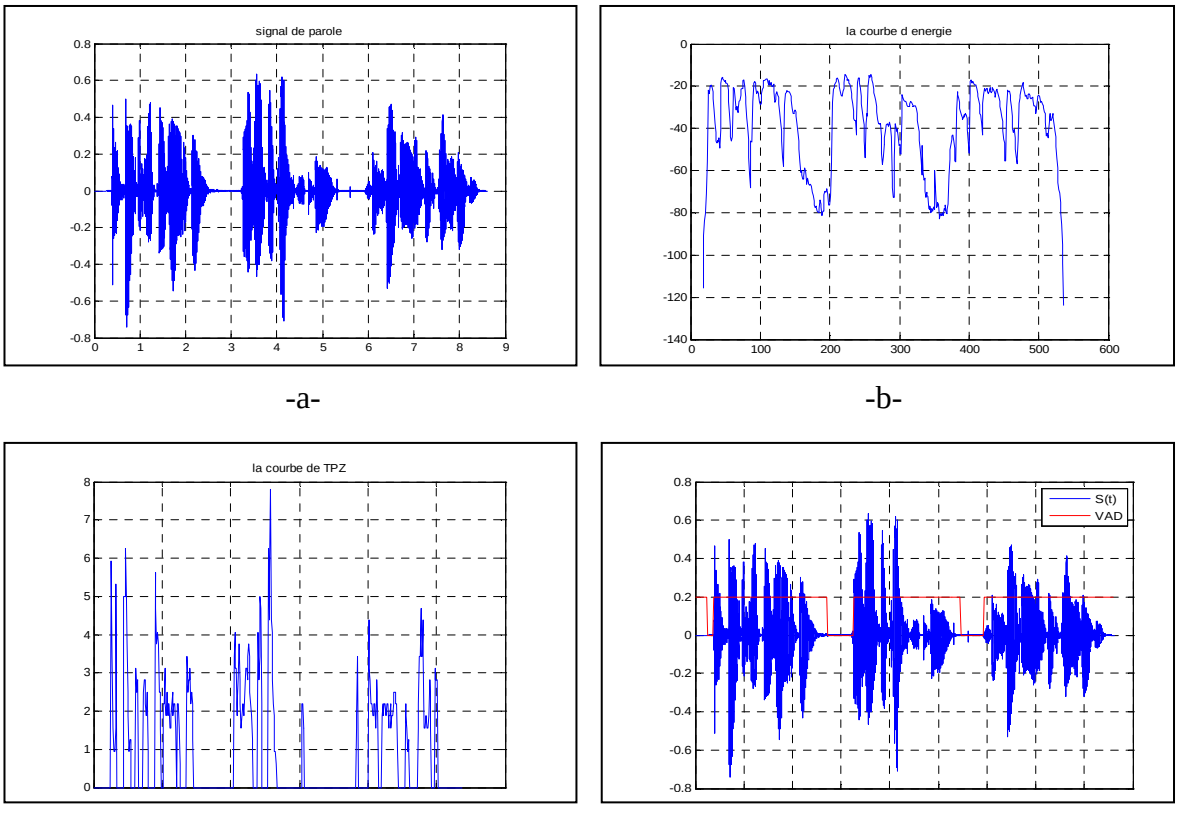


Figure.1.6 *Les différentes mesures utilisées pour éliminer le silence.*

a : le signal de parole.

b : la courbe d'énergie.

c : la courbe du taux de passage par zéros.

d : l'utilisation de l'algorithme VAD pour éliminer les zones du silence.

- A**lgorithme VAD (Voice Activity Detection)**

La classification effectuée par le VAD utilise une grande variété de méthodes qui sont : la comparaison de l'énergie à un seuil, la détection de voisement et l'estimation du pitch, l'analyse spectrale, le rythme de passage par zéro ou de changement de signe, la mesure de périodicité, l'étude statistique du résidu de la prédiction linéaire...etc. Les VAD incluent souvent une combinaison de plusieurs de ces méthodes.
La fonction d'élimination du silence, VAD (m) peut être définie comme suit :

$$\mathrm{VAD(m)} = \begin{cases} 1 & \text{si } s(n) \geq \text{seuil} \\ 0 & \text{si } s(n) \leq \text{seuil} \end{cases} \quad (1.7)$$

d. Le fenêtrage

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 10 à 30 ms, et le décalage entre deux trames consécutives est d'environ 10 à 15 ms. Ensuite, on applique une fenêtre à chaque trame 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 par un autre signal $w(n)$ possédant N échantillons. Le signal de la trame résultante à N échantillons est donc :

$$s(n) = \sum_{n=1}^{N} x(n)w(n) \quad (1.8)$$

Avec $s(n)$: Signal de la trame.
$x(n)$: Signal à fragmenter.
$w(n)$: Fenêtre de longueur N. n=*1,…..…..N.*

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

$$w(n) = 0.54 - 046\ cos\ (\frac{2\pi n}{n-1}) \quad (1.9)$$

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

e. Les paramètres acoustiques

Les performances d'un système d'identification du locuteur dépendent essentiellement de la qualité des paramètres choisis. La plus part des systèmes RAL utilisent les *mel frequency cepstral coefficients (MFCC)* et *Linear Prediction Cepstrum Coefficients (LPCC)* **[13]** et ceci pour les raisons suivantes :

- Ces mesures fournissent un bon modèle de signal de paroles, cela est particulièrement vrai dans des régions quasi stationnaires du signal de paroles.
- Des expériences ont montré que ces mesures donnent de bons résultats dans les applications de reconnaissance automatique du locuteur.

I.7.4 L'extraction des paramètres MFCC

Toutes les étapes d'extraction des coefficients mel-Cepstral à partir d'une trame de parole seront décrites dans cette section (Figure.1.7).

✓ La première étape consiste à calculer la transformée de Fourrier rapide (TFR), permet le passage du domaine temporel au domaine fréquentiel. Elle est définie comme suit :

$$S[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \qquad (1.10)$$

✓ Ensuite, on calcul l'amplitude, on obtient donc le spectre.

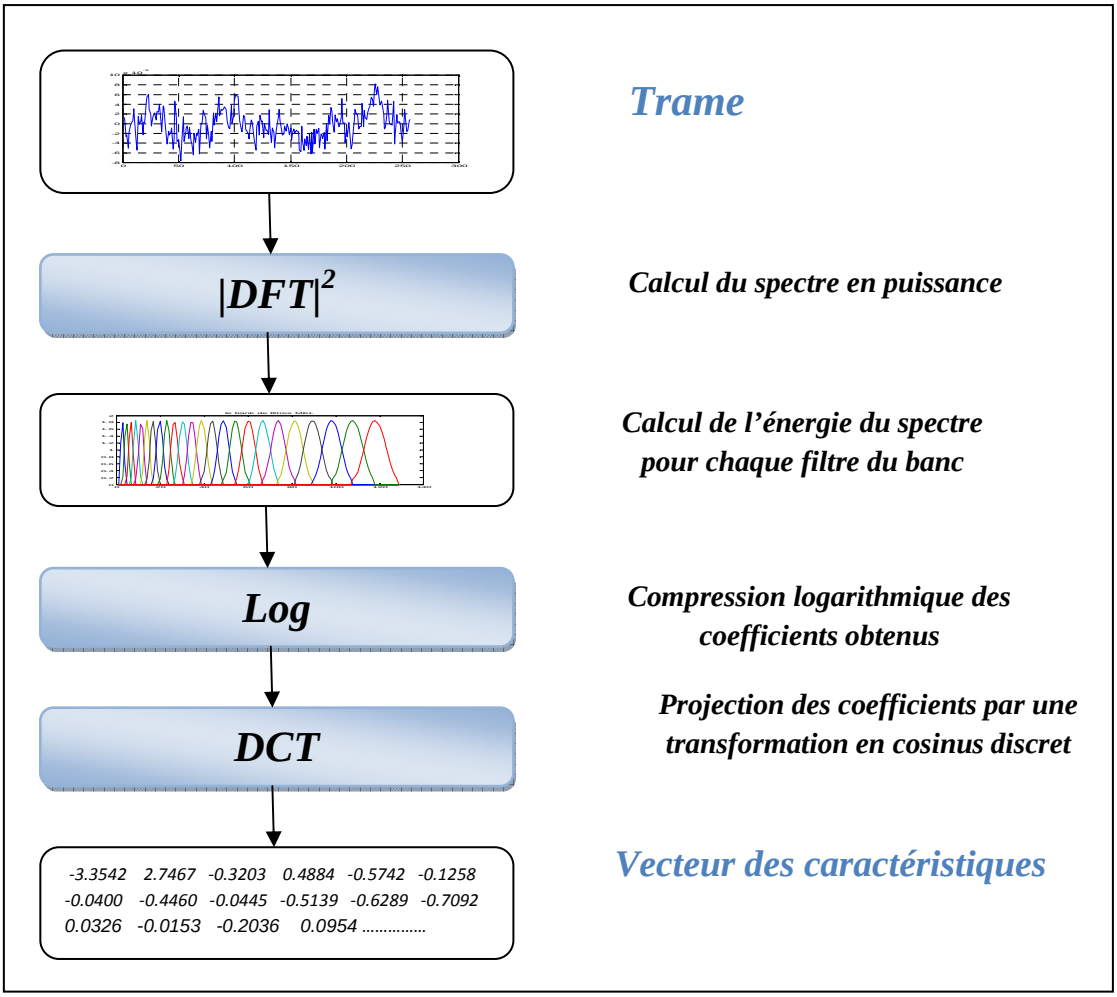


Figure.1.7 *Principe du codage MFCC.*

✓ 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 :

$$B_m = B_b + m\,\frac{B_h - B_b}{M+1}\ ,\quad 0 \le m \le M+1 \tag{1.11}$$

M : Le nombre de filtres.
Bh : La fréquence la plus haute du signal.
Bb : La fréquence la plus basse du signal.

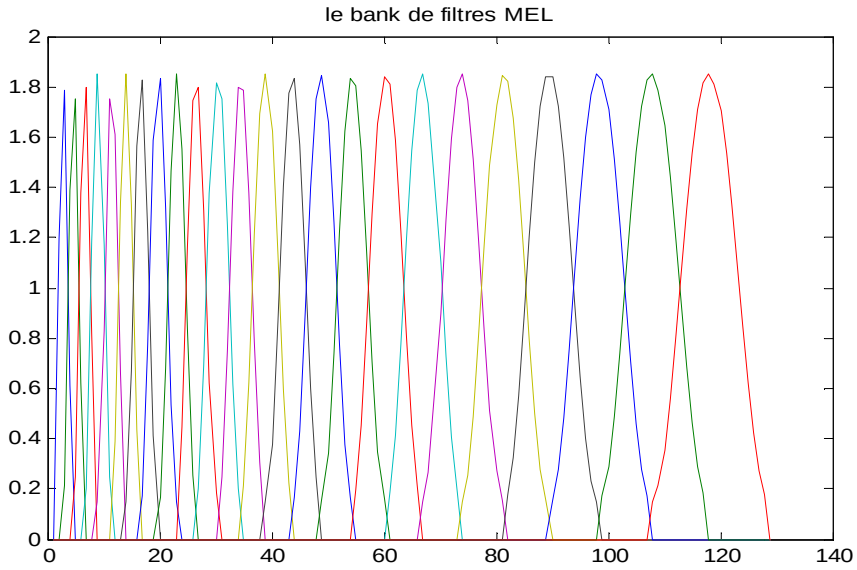


Figure.1.8 *La répartition des filtres triangulaires sur les échelles Mel.*

❖ Echelle de Mel

En effet, après des études psychophysiques, il a été montré que la perception humaine des fréquences contenues dans un signal de parole ne suit pas une échelle linéaire. Par conséquent, on fait correspondre à chaque fréquence F, mesurée en Hz, une valeur F_{mel} mesurée en Mel, tel que :

$$B(f) = 2595 * log_{10}(1 + \frac{f_{Hz}}{700}) \tag{1.12}$$

Où :
f_{Hz} est la fréquence en Hz,
B(f) est la fréquence Mel-échelle de f_{Hz}.

Cette transformation non linéaire peut être vue dans Figure.1.9**.**

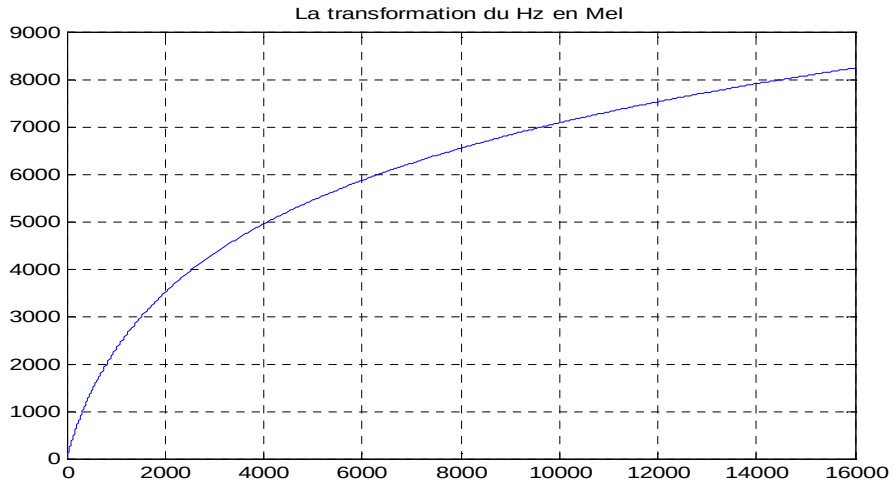


Figure.1.9 *La transformation du Hz en Mel.*

- ✓ Après le calcul des coefficients Mel-Cepstraux, on prend le logarithme et l'inverse de la transformation en cosinus discrète comme illustrée par la formule suivante:

$$c_k = \sum_{i=1}^{M} \log E_i \cos\left[\frac{\pi k}{M}(i - \frac{1}{2})\right]. \qquad 1 \leq k \leq d \tag{1.13}$$

I.8 La modélisation des paramètres acoustiques

Dans un système de reconnaissance automatique du locuteur, les paramètres acoustiques sont utilisés pour estimer un modèle, qui peut être statistique ou basé sur le calcul d'une distance euclidienne. Un modèle idéal doit satisfaire les contraintes suivantes **[14]**:

- Il doit nécessiter le plus faible espace de stockage possible.
- Il doit avoir une méthode d'estimation la moins complexe possible.
- Il doit permettre une décision rapide lors de la phase de test.
- Il doit être le plus robuste possible aux variations intra-locuteur.
- Il doit permettre la meilleure séparation des locuteurs entres eux.
- Il doit avoir la représentation la plus complète possible des paramètres acoustiques des locuteurs.

Il existe plusieurs méthodes de modélisation utilisées dans les systèmes de reconnaissance automatique du locuteur. Nous pouvons distinguer deux grandes familles de modèles. La première famille correspond aux méthodes basées sur le calcul d'une distance euclidienne entre les vecteurs acoustiques extraits d'un signal de parole et d'autres représentant le locuteur.
La seconde famille correspond aux méthodes basées sur une représentation statistique du locuteur dans l'espace de paramètres acoustique.

I.8.1 Reconnaissance de locuteur à base de DTW

La reconnaissance par DTW (Dynamique Time Warping **[15]**) repose sur le principe que chaque mot est représenté par une prononciation de référence (template). Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (time warping). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendent du texte. Cette approche est facile à mettre en œuvre et donne des performances relativement bonnes **[16].**

I.8.2 Quantification vectorielle

La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces (Figure.10). Chaque sous-espace est associé à un vecteur centroide. 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 test, on fait la comparaison des trames extraites d'un signal vocal avec le dictionnaire des vecteurs. Cette comparaison est faite en calculant la distance qui sépare les trames en entrée avec les centres des partitions. La distance mesure le degré de similarité entre la voix de test et le modèle d'un locuteur. La quantification vectorielle s'applique en mode dépendant ou indépendant du texte **[17], [18].** Cependant, cette méthode élimine beaucoup d'information sur les locuteurs et elle nécessite des paroles très longues pour avoir des informations statistiques stables.

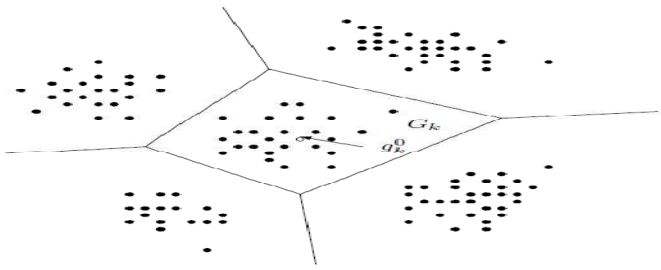


Figure.1.10 *Définition des classes de données en quantification vectorielle.*

I.8.3 Le model de Markov caché (HMM- Hidden Markov Model)

Les modèles de Markov caché 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'un forme acoustique à une référence, mais de la probabilité que la forme ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaine de Markov, une succession d'états avec des probabilités de transition d'un état à l'autre, et de lois de probabilités. Quant à l'utilisation des modèles de Markov cachés en reconnaissance du locuteur, on peut se référer à **[19] et [20]**. La Figure.1.11 représente un exemple modèle de Markov caché à deux états.

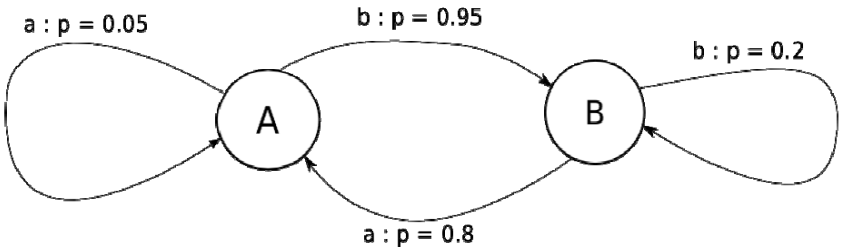


Figure.1.11 *Un modèle de Markov caché.*

I.8.4 Le modèle de mélanges de gaussiennes (Gaussians Mixture Model GMM)

C'est le modèle le plus utilisé dans la reconnaissance automatique du locuteur indépendante du texte. Il a été objet de plusieurs études et a montré de très bonne performance sur plusieurs types de systèmes RAL **[21] [22]**. 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 ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques (Figure. 1.12). 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.

Nous avons consacré tout un chapitre pour étudier en détails ce modèle et présenté toutes les étapes qui le composent.

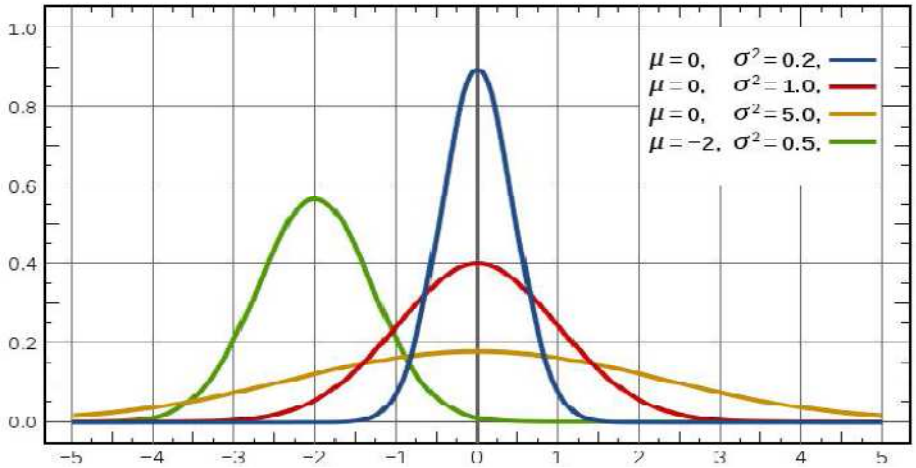


Figure.1.12 *Mélanges des gaussiennes.*

I.8.5 Le modèle du non locuteur ou modèle du monde UBM

Le modèle du monde est utilisé comme a priori pour l'estimation des modèles de locuteur. Il représente la couverture exhaustive des paramètres acoustiques d'un grand nombre de locuteurs enregistrés dans diverses conditions. Ce modèle est appris par maximum de vraisemblance avec une très grande quantité de données. Ainsi les enregistrements collectés pour la création du modèle du monde doivent représenter :

1. les différentes langues de la population de locuteurs à reconnaître,
2. les genres considérés (homme, femme, les deux),
3. les conditions d'enregistrements (type de microphone, type de canal de transmission).
4. le type de parole : lue, spontanée, conversationnelle.

Ce modèle constitue une représentation moyenne d'un ensemble de locuteurs et des conditions d'enregistrement. Nous pouvons observer que le modèle du monde intervient à différents niveaux dans un système de reconnaissance et surtout pour la vérification automatique de locuteur VAL, il représente l'hypothèse inverse dans le test de vérification, et sert d'initialisation pour l'estimation des modèles de locuteur. Le lecteur pourra se référer aux travaux de **[23]** pour une description plus complète de son importance dans un système de VAL.

I.8.6 Les réseaux de neurones

Un réseau de neurones ANN (Artificial Neural Network) **[24] [25],** est un modèle de calcul dont la conception est très schématiquement inspirée du fonctionnement de vrais neurones (humains ou non). Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage de type statistique. Ils sont classés d'une part dans la famille des applications statistiques, et d'autre part dans la famille des méthodes de l'intelligence artificielle.

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. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée. La Figure. 1. 13 représente un réseau de neurones de type MLP (Multilayer Perceptron).

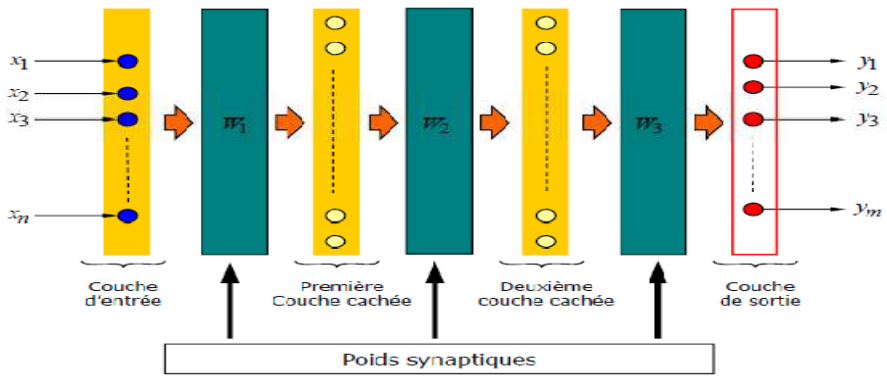


Figure.1.13 *Exemple d'un réseau de neurone à quatre couches.*

I.8.7 Les Machines à Vecteurs de Support SVM

Les SVM ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la Théorie de Vapnik-Chervonenkis. L'idée d'une classification par SVM **[26]** est de calculer l'hyperplan permettant de séparer au mieux deux nuages de points. Il s'agit de reconsidérer le problème dans un espace de dimension supérieure. Dans ce nouvel espace, il existe un séparateur linéaire permettant de classer au mieux les deux nuages de points.

I.9 Décision et mesure des performances

Pour l'identification du locuteur 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. 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 terme de taux d'identification correcte (TIC). Par exemple, un système d'identification du locuteur peut avoir un pourcentage d'erreur de 2% sur un corpus d'évaluation.

$$TIC = \frac{nombre\ de\ test\ correctement\ identifié}{nombre\ totale\ de\ test} \times 100$$

Pour la vérification du locuteur, le processus est plus compliqué. Pour évaluer les performances d'un système de vérification, deux types d'erreurs peuvent être commis. Premièrement, le client est rejeté alors que l'identité proclamée est la sienne. Dans ce cas, on parle de *« Faux Rejet»* (FR). La deuxième erreur survient lorsqu'un imposteur est accepté alors que l'identité annoncée n'est pas la sienne. Ce cas est appelé *« Fausse Acceptation »* (FA).

Un système biométrique est évalué par les taux des FR et FA (en anglais, False Rejection Rate, (*FRR*) et False Acceptation Rate (*FAR*)). Un bon système de vérification minimise les deux taux d'erreurs FA et FR. Le plus souvent, les performances du système RAL sont exprimées en taux d'égale erreur (Equal Error Rate, *EER*), qui correspond à la valeur où le FRR est égale au FAR.

Plus la valeur de l'EER est faible, plus le système est meilleur. L'EER est utilisé pour comparer les systèmes RAL entre eux.

La relation entre FRR et FAR est représentée par une courbe appelée ROC (Receiver Operating Characteristics) **[27]**. Cette courbe est obtenue en faisant varier le seuil de décision et en calculant à chaque fois les deux valeurs FRR et FAR. La figure.1.14 illustre un exemple de la courbe ROC. L'EER correspond au point d'intersection des deux courbes FRR et FAR.

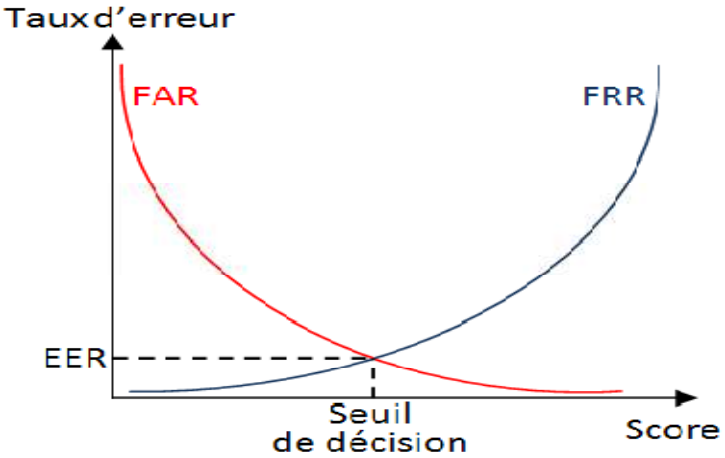


Figure.1.14 *Exemple de courbe ROC.*

La relation entre FRR et FAR est peut être aussi représentée par la courbe DET (Detection Error Tradeoff). La courbe DET diffère principalement de la courbe ROC par l'échelle basée sur une distribution normale qui se substitue à l'échelle linéaire **[28].**

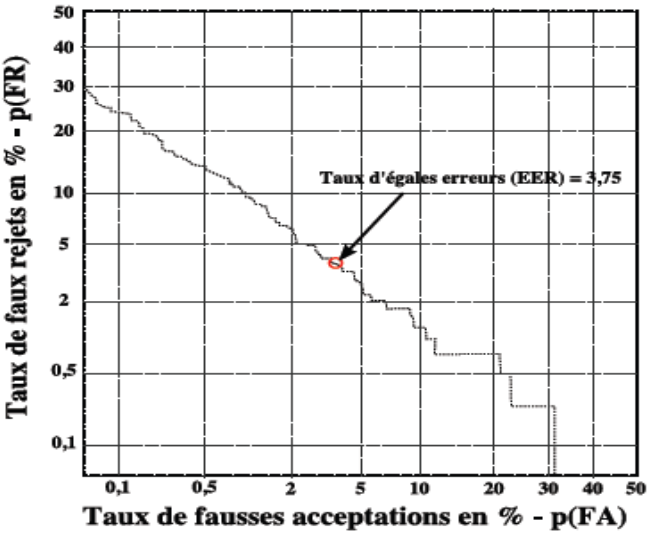


Figure.1.15 *Exemple de représentation des performances d'un système de vérification de locuteur par une courbe DET.*

I.10 Conclusion

Dans ce chapitre, nous avons présenté les différents types des systèmes de reconnaissance automatique du locuteur, à savoir, l'identification et la vérification automatique du locuteur.
Un système de reconnaissance de locuteur procède généralement en trois étapes : l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision. En analyse acoustique, les coefficients MFCC sont les coefficients les plus répandus. Quant à la modélisation, l'approche GMM constitue l'état de l'art en RAL, en mode indépendant du texte. La décision d'un système de reconnaissance automatique du locuteur est basée sur les deux processus d'identification et/ou de vérification de locuteur.

2

La modélisation des vecteurs acoustiques

II.1 Introduction

Dans ce chapitre, nous allons présenter un module important de la reconnaissance automatique du locuteur. Comme nous l'avons déjà mentionné dans le chapitre précédent, les systèmes RAL nécessitent une étape de modélisation statistiques des paramètres (ou vecteurs) acoustiques, extraits du signal de parole du locuteur. En effet, plusieurs méthodes de modélisation statistique ont été utilisées et chacune d'elle présente des avantages et des inconvénients. Nous commençons d'abord par détailler le modèle statistique le plus utilisé dans les systèmes RAL en mode indépendant du texte, à savoir, le modèle **GMM** (Gaussian Mixture Model). Puis nous allons détailler la théorie des machines à vecteurs de support ou SVM utilisé comme classifieur discriminatif.

II.2 Pourquoi les GMMs ?

Le modèle de mélange de distributions gaussiennes (Gaussian mixture model (GMM)) consiste à supposer que la distribution des données peut être décrite comme une somme pondérée de densités gaussiennes multidimensionnelles **[21]**. Ce modèle de mélange est classique dans le domaine de la reconnaissance de forme car il correspond à une situation où les données appartiennent à un ensemble de classes distinctes, avec une probabilité d'appartenance propre à chaque classe. Le cas particulier considéré ici est celui où dans chaque classe les données suivent une loi gaussienne. Ce choix tient essentiellement au fait que la loi gaussienne appartient à une famille de distributions dite exponentielles pour lesquelles le problème de l'identification des composantes du mélange se trouve simplifié. Pour le signal de parole, ce modèle ne paraît donc pas déraisonnable et il est d'autre part assez proche de la caractérisation

fournie par la quantification vectorielle. La différence étant qu'avec la quantification vectorielle, on se contente de mettre en évidence un certain nombre de "points d'accumulation" des paramètres mesurés, alors qu'avec le modèle de mélange de distributions gaussiennes, on cherche en plus à décrire la distribution des paramètres mesurés autours de ces points d'accumulation.

D'après plusieurs études faites par plusieurs chercheurs, il apparaît que le modèle de mélange de gaussiennes est le plus adapté à la reconnaissance automatique du locuteur indépendante de texte. Ces bons résultats peuvent s'expliquer par certaines de ses propriétés, telles que **[14]** :

- ✓ La capacité à modéliser des formes complexes de distributions sans être théoriquement limité par sa complexité en augmentant le nombre de composantes du mélange.
- ✓ La segmentation implicite des vecteurs acoustiques en *K* classes de son dans l'espace des paramètres. Chacune d'elle possède sa probabilité d'occurrence a priori mais la modélisation ne donne aucune information sur leur dynamique.
- ✓ L'existence d'un outil très puissant pour l'estimation des paramètres qui leur sont associés : l'algorithme *Expectation-Maximisation (EM).*

II.3 Modèle à mélanges du gaussiennes

Les mélanges de gaussiennes sont utilisés pour modéliser un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle GMM à un HMM (Hidden Marcov Model) à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance de locuteur en mode indépendant du texte. Les travaux de D.A.Reynolds **[2]** consistent l'état de l'art en la matière. Le modèle GMM est une segmentation en plusieurs classes des paramètres acoustiques représentants l'identité d'un locuteur (Figure. 2.1).

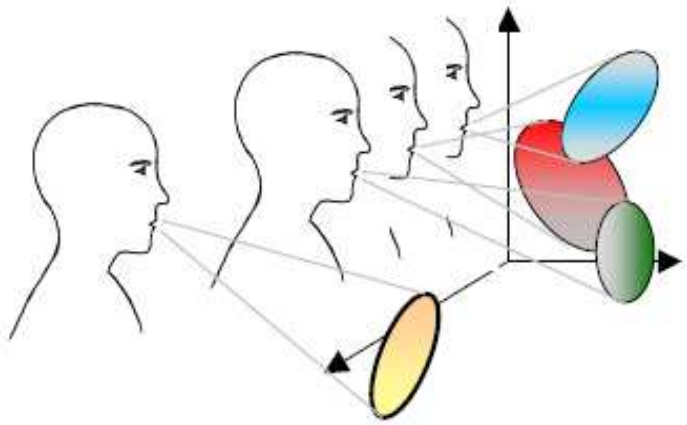


Figure.2.1 *Illustration de nuages acoustiques représentants l'identité d'un locuteur.*

Un mélange de gaussiennes (Figure.2.2) est une somme pondérée de M distributions gaussiennes [**21**].

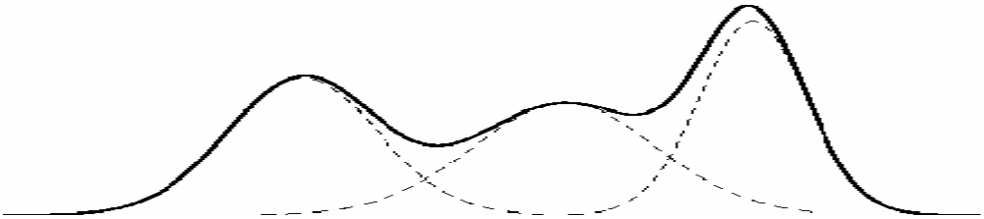


Figure.2.2 *Approximation de la distribution d'un paramètre acoustique par une combinaison de gaussiennes.*

Cela nous amène à considérer que les vecteurs acoustiques (étudiés dans le chapitre précèdent) sont des réalisations de variables aléatoires mutuellement indépendantes de densité de probabilité $b_m(\bar{x})$ gaussienne.

Le mélange de gaussiennes $p(\bar{x}|\theta)$ est donné par la formule (2.1) qui est une somme pondérée par les coefficients π_m des distributions gaussiennes $b_m(\bar{x})$.

$$p(\bar{x}|\theta) = \sum_{m=1}^{M} \pi_m b_m(\bar{x}) \quad (2.1)$$

Où :

- $\bar{x}$ est un vecteur aléatoire de dimension d qui représente un vecteur acoustique.
- $b_m(\bar{x})$, m=1,2,…, M sont les densités de probabilités gaussiennes qui composent le modèle.
- π_m , m =1,2,…, M, sont les poids des densités $b_m(\bar{x})$.

Chaque densité de probabilité $b_m(\bar{x})$ est définie par :

$$b_m(\bar{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} exp\left\{-\frac{1}{2}(\bar{x} - \overline{\mu_m})' \Sigma_m^{-1} (\bar{x} - \overline{\mu_m})^T\right\} \quad (2.2)$$

Tel que :

$\overline{\mu_m}$ est le vecteur moyen et Σ_m est la matrice de covariance. Les poids π_m vérifient la contrainte donnée par la formule suivante :

$$\sum_{m=1}^{M} \pi_m = 1 \quad (2.3)$$

Ce modèle statistique est composé par des vecteurs moyens $\overline{\mu_m}$, des matrices de covariances Σ_m et des coefficients π_m de toutes les densités de probabilités $b_m(\bar{x})$. Ces paramètres sont représentés par la notation suivante :

$$\theta = \{\pi_m, \overline{\mu_m}, \Sigma_m\}, m = 1, \dots, M \quad (2.4)$$

La modélisation par GMM comprend deux phases principales [**21**] [**29**]:

- Une phase d'apprentissage sur un ensemble de fichiers audio qui représentent les différents locuteurs du système.
- Une phase d'identification qu'un son quelconque est produit par un locuteur bien déterminé.

II.3.1 La phase d'apprentissage

Dans cette phase, on estime les paramètres des gaussiennes qui composent un modèle GMM en se basant sur les vecteurs acoustiques déterminés dans l'étape d'extraction de paramètres. 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 étape est l'optimisation des paramètres obtenus dans la première étape en utilisant l'algorithme **EM** *(Expectation Maximisation)* [**30**] [**31**].

II.3.1.1 L'initialisation des paramètres

a. Algorithme des *K*-moyennes

L'algorithme des k-moyennes est un algorithme de classification non supervisée ayant pour objectif de trouver k vecteurs $x_1 \dots x_k$ permettant de représenter au mieux un ensemble de n vecteurs $y_1 \dots y_n$. Un vecteur représentant une classe est choisi comme étant le représentant de cette classe au sens d'une distance. L'algorithme des K-moyenne n'est que localement optimal, par conséquence, il est influencé par ses conditions initiales **[32]**.

Étape 1 La première étape est l'initialisation du dictionnaire.

Étape 2 : La deuxième étape consiste à appliquer deux règles, définies ci-dessous

- **La règle de centroide :** Cette règle exige que tous les centroides soient les moyennes des vecteurs acoustiques des régions représentées par ces centroide. Cela peut être formulé comme suit:

$$c_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \bar{x}_k \quad (2.5)$$

- **La règle de plus proche voisin :** Le vecteur $\overline{x_k}$ est affecté à la région **i** si la distance euclidienne entre ce vecteur et le centroide de cette région est minimale. La formule suivante décrit explicitement cette règle.

$$region(\bar{x}_k) = i, \quad c_i = \min\left(d_{j=1}^{K}(\overline{x_k}, c_j)\right) \tag{2.6}$$

K est le nombre de régions.
$\mathbf{C}_i$ est le centroide de la *i* région.
$d(x_k, c_i)$ est la distance euclidienne entre les vecteurs x_k et c_i .

Étape 3 : Si il y'a une amélioration importante de la distorsion moyenne donnée par la formule ci-dessous, alors aller à l'étape 2 et si la différence de la distorsion est inférieur à un seuil prédéfini alors Fin.

$$D_m = \frac{1}{N}\sum_{k=1}^{N} d(\overline{x_k}, cent(\overline{x_k})) \tag{2.7}$$

b. **Algorithme LBG**

L'algorithme LBG **[32]** [**33**], C'est une version étendue d'algorithme k-moyennes, c'est un l'algorithme itératif utilisé pour la création d'un dictionnaire optimal basé sur des vecteurs d'apprentissage.
1. On initialise le dictionnaire, par exemple, par le vecteur moyen de toute la base d'apprentissage.
2. Connaissant ce dictionnaire, on étiquette chaque vecteur de base d'apprentissage par le numéro de son plus proche voisin. On détermine la partition optimale (la règle de plus proche voisin).
3. A partir de tous les vecteurs étiquetés par le même numéro, on en déduit un nouveau représentant par un calcul de la moyenne (la règle de centroide).
4. Si le nombre de centroides désirés n'est pas atteint, on applique une technique de « Spliting », celle-ci consiste à découper chaque centroide C_i en deux nouveaux vecteurs $C_i + \varepsilon$ et $C_i - \varepsilon$ (ε étant un vecteur de perturbation), avant d'appliquer au nouveau dictionnaire obtenu les itérations 2 et 3.
On arrête cet algorithme si on atteint le nombre de centroides désirés pour représenter les vecteurs acoustiques.

II.3.1.2 Optimisation des paramètres par l'algorithme EM

L'idée principale de l'algorithme EM est, en commençant par les paramètres initiaux θ du modèle, on estime les nouveaux paramètres θ, telle que la vraisemblance du nouveau modèle soit supérieur ou égale à la vraisemblance du modèle initial. L'algorithme EM fait intervenir à la fois des observations *X* et des variables manquantes (l'indice de la gaussienne *m = 1,...,* M). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation

n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ 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, à savoir :

$$Q(\theta, \theta^{(t)}) = \sum\sum p(m|x_n, \theta^{(t)}) \times \log p(x_n, m|\theta) \quad (2.8)$$

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

$$Q(\theta, \theta^{(t)}) = \sum\sum \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m|\right] - \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2}(x_n - \mu_m)\right] \quad (2.9)$$

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

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)}\, p(x_n|\mu_m^{(t)}.\Sigma_m^{(t)})}{\sum_{m=1}^{M} \pi_k^{(t)}\, p(x_n|\mu_m^{(t)}.\Sigma_m^{(t)})} \quad (2.10)$$

En supposant que p (x_m/θ) sont des densités gaussiennes à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[Cste + \log \sigma_m^2 + \frac{(x_n - u_m)^2}{\sigma_m^2}\right] \quad (2.11)$$

Où σ_m^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_m est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte qui existe sur ces paramètres. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\pi_m^{\{t+1\}} = \frac{1}{N}\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}} \quad (2.12)$$

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

$$\mu_m^{\{t+1\}} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}}.x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}}} \quad (2.13)$$

Et pour les variances :

$$\sigma_m^{2\{t+1\}} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}} (x_n - \mu_m^{\{t\}})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{\{t\}}} \quad (2.14)$$

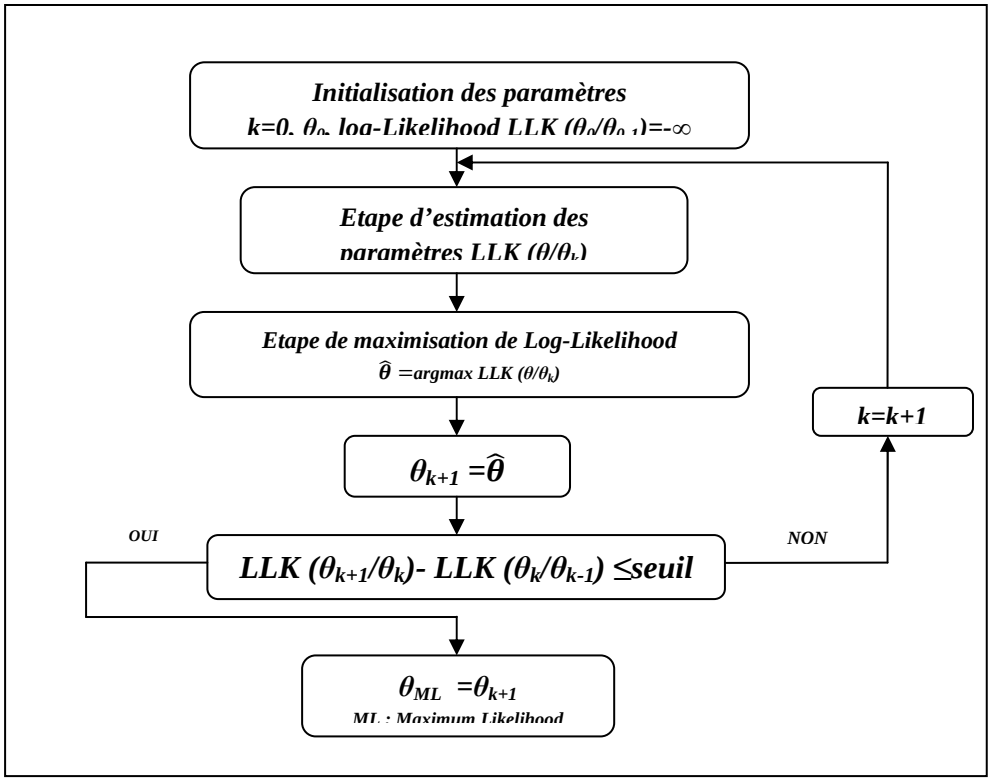


Figure.2.3 *Schéma de fonctionnement de l'algorithme EM.*

II.3.2 La phase de classification ou de décision

Dans La phase de classification (la sélection du locuteur le plus probable), on a un groupe de L locuteurs *G = {1,2,..., L}*, qui sont représentés, respectivement, par les modèles GMM : $\theta_1, \theta_2,..., \theta_L$. Dans une identification du locuteur, l'objective est de trouver le modèle du locuteur qui à le maximum de vraisemblance pour une séquence d'observations données. Formellement, cela peut être écrit comme suit :

$$\hat{l} = arg\max_{1\le l\le L} Pr(\theta_l|x) = \arg\max_{1\le l\le L} \frac{p(x|\theta_l)Pr(\theta_l)}{p(x)} \quad (2.15)$$

Supposons que tous les locuteurs ont la même probabilité $p(x) = \frac{1}{L}$, la formule de classification se simplifie comme suit :

$$\hat{l} = arg\max_{1\le l\le L} p(x|\theta_l) \quad (2.16)$$

En utilisant le logarithme et l'indépendance entre les observations, la formule précédente peut être écrite comme suit :

$$\hat{l} = arg\max_{1\le l\le L} \sum_{t=1}^{T} \log p(\overline{x_t}|\theta_l) \quad (2.17)$$

Où $p(\overline{x_t}|\theta_l)$ est donnée par la formule (2.1).

II.4 Les Machines à Vecteurs de Support

Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais : *Support Vector Machines,* SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires. Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance, etc.). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un modèle de mixture gaussienne (GMM) **[34].** Les SVM sont des classifieurs qui reposent sur deux idées clés, permettant de traiter des problèmes de discrimination non-linéaire, et de reformuler le problème de classification comme un problème d'optimisation quadratique.
La première idée clé est la notion de *marge maximale*. La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Ces derniers sont appelés *vecteurs supports*. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge.
Afin de pouvoir traiter des cas où les données ne sont pas linéairement séparables, la deuxième idée clé des SVM est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension, dans lequel il est probable qu'il existe une séparatrice linéaire. Ceci est réalisé grâce à une fonction noyau, qui doit respecter certaines conditions, et qui a l'avantage de ne pas nécessiter la connaissance explicite de la transformation à appliquer pour le changement d'espace. Les fonctions noyau permettent de transformer un produit scalaire dans un espace de grande dimension, ce qui est coûteux, en une simple évaluation ponctuelle d'une fonction. Cette technique est connue sous le nom de "*kernel trick*" **[4] [34]**.

II.4.1 Principe Général des SVM

Les SVM peuvent être utilisés pour résoudre des problèmes de discrimination, c'est-à-dire décider à quelle classe appartient un échantillon, ou des problèmes de régression, c'est-à-dire prédire la valeur numérique d'une variable. La résolution de ces deux problèmes passe par la construction d'une fonction *f* qui à un vecteur d'entrée *x* fait correspondre une sortie *y* :

$$y = f(x) \tag{2.18}$$

II.4.2 Discrimination Linéaire et Hyperplan Séparateur

A titre de rappel, le cas simple est le cas d'une fonction discriminante linéaire, obtenue par combinaison linéaire du vecteur d'entrée $x = (x_1,\ldots, x_N)^T$:

$$h(x) = w^T x + w_0 \tag{2.19}$$

Avec :
w^T : vecteur de pondération
w_0 : biais

Il est alors décidé que *x* est de classe 1 si $h(x) \geq 0$, et de classe -1 dans le cas contraire.

La frontière de décision $h(x) = 0$ est un hyperplan, appelé *hyperplan séparateur*, ou *séparatrice.* Rappelons que le but des algorithmes à apprentissage supervisé est d'apprendre la fonction $h(x)$ par le biais d'un ensemble d'apprentissage :

$$\{(x_1, l_1), (x_2, l_2), ..., (x_p, l_p)\} \in IR^N \times \{-1, 1\} \qquad (2.20)$$

Où les l_k sont les étiquettes, *p* est la taille de l'ensemble d'apprentissage, *N* la dimension des vecteurs d'entrée. Si le problème est linéairement séparable, on doit alors avoir :

$$l_k h(x_k) \geq 0 \quad 1 \leq k \leq p \text{ , autrement dit } l_k (w^T x_k + w_0) \geq 0 \quad 1 \leq k \leq p \qquad (2.21)$$

Prenons un exemple pour bien comprendre le concept de séparation. Imaginons un plan dans lequel sont répartis deux groupes de points. Ces derniers sont associés à un groupe : les points $(+)$ pour y > x et les points $(-)$ pour y < x.

On peut trouver un séparateur linéaire évident dans cet exemple, la droite d'équation y = x. Dans ce cas le problème est dit *linéairement séparable* (figure.2.4.a).

Pour des problèmes plus compliqués, en général, il n'existe pas de séparateur linéaire. Imaginons par exemple un plan dans lequel les points $(-)$ sont regroupés en un cercle, avec des points $(+)$ tout autour (figure.2.4.b) : aucun séparateur linéaire ne peut correctement séparer les groupes : le problème n'est pas *linéairement séparable* et il n'existe pas d'hyperplan séparateur.

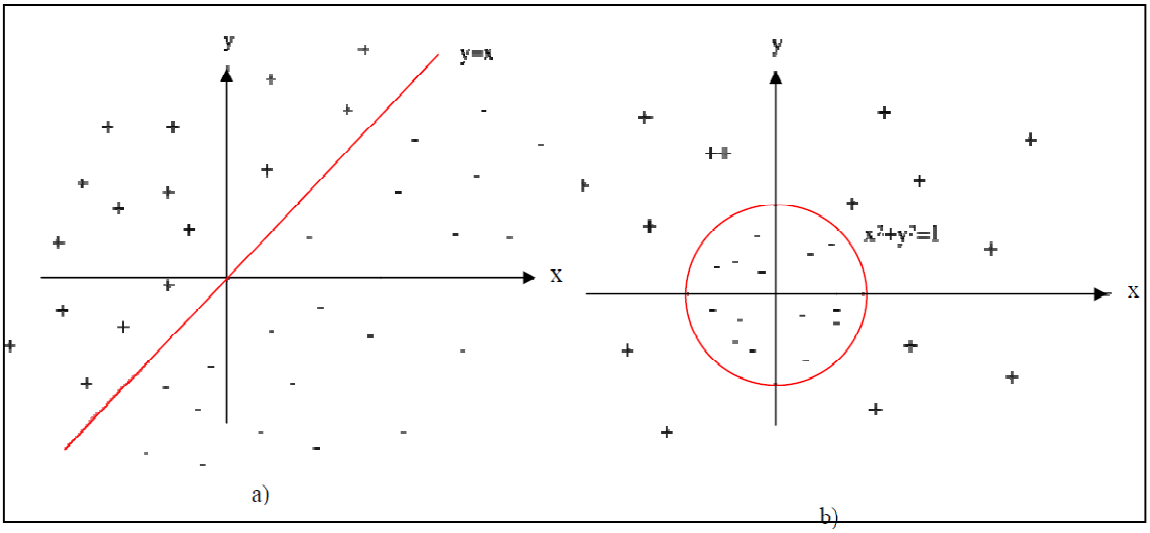


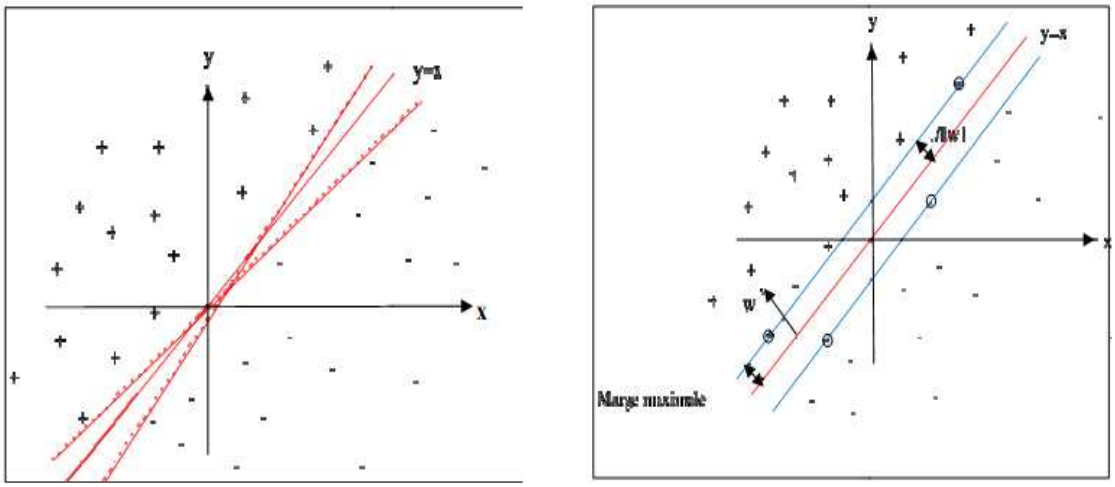
Figure.2.4 *Exemples d'un problème de discrimination a deux classes, avec une séparatrice :*

a) linéaire *b) non- linéaire.*

II.4.3 Marge Maximale

On se place désormais dans le cas où le problème est linéairement séparable. Même dans ce cas simple, le choix de l'hyperplan séparateur n'est pas évident. Il existe en effet une infinité d'hyperplans séparateurs, dont les performances en apprentissage sont identiques (le risque empirique est le même), mais dont les performances en généralisation peuvent être très différentes (figure.2.5.a).

Pour résoudre ce problème, il a été montré qu'il existe un unique hyperplan optimal, défini comme l'hyperplan qui maximise la marge entre les échantillons et l'hyperplan séparateur (figure.2.5.b) **[34].**



a) Ensemble de points linéairement séparables.

b) Hyperplan optimal (en rouge) avec la marge maximale.

Figure.2.5 *choix d'un hyperplan séparateur.*

La marge est la distance entre l'hyperplan et les échantillons les plus proches. Ces derniers sont appelés *vecteurs supports*. L'hyperplan qui maximise la marge est donné par :

$$\arg\max_{w,w_0}\min_{k}\left\{\left\|x-x_k\right\| : x\in IR^N,\ w^T x+w_0=0\right\} \tag{2.22}$$

Il s'agit donc de trouver w et w_0 remplissant ces conditions, afin de déterminer l'équation de l'hyperplan séparateur :

$$h(x)=w^T x+w_0=0 \tag{2.23}$$

II.4.4 Recherche de l'Hyperplan Optimal

La marge est la plus petite distance entre les échantillons d'apprentissage et l'hyperplan séparateur qui satisfasse la condition de séparabilité (à savoir $l_k\left(w^T x_k + w_0\right) \geq 0$ comme expliqué précédemment). La distance d'un échantillon x_k à l'hyperplan est donnée par sa projection orthogonale sur l'hyperplan $\frac{l_k\left(w^T x_k + w_0\right)}{\|w\|}$

L'hyperplan séparateur (w, w_0) de marge maximale est donc donné par :

$$\arg\max_{w,w_0}\left\{\frac{1}{\|w\|}\min_k\left[l_k\left(w^T x_k + w_0\right)\right]\right\} \tag{2.24}$$

Afin de faciliter l'optimisation, on choisit de normaliser w et w_0, de telle manière que les échantillons à la marge (x^+_{marge} pour les vecteurs supports sur la frontière positive, et x^-_{marge} pour ceux situés sur la frontière opposée) satisfassent :

$$\begin{cases} w^T x^+_{\text{marge}} + w_0 = 1 \\ w^T x^-_{\text{marge}} + w_0 = -1 \end{cases}$$

d'où pour tous les échantillons, $k = 1,..., p$:

$$l_k\left(w^T x_k + w_0\right) \geq 1 \tag{2.25}$$

Cette normalisation est parfois appelée la forme canonique de l'hyperplan, ou *hyperplan canonique*. Avec cette mise à l'échelle, la marge vaut désormais $\frac{1}{\|w\|}$, il s'agit donc de maximiser $\|w\|^{-1}$.

La formulation dite *primale* des SVM s'exprime alors sous la forme suivante :

Minimiser $\frac{1}{2}\|w\|^2$ sous les contraintes $l_k\left(w^T x_k + w_0\right) \geq 1$.

Une manière de résoudre ce type de problème d'optimisation sous contraintes est d'introduire les multiplicateurs de Lagrange, où le lagrangien est donné par :

$$L(w, w_0, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{k=1}^{p}\alpha_k\left\{l_k\left(w^T x_k + w_0\right) - 1\right\} \tag{2.26}$$

Le lagrangien doit être minimisé par rapport à w et w_0, et maximisé par rapport à α. En annulant les dérivées partielles du lagrangien, selon les conditions de Kuhn-Tucker, on obtient :

$$\begin{cases} \sum_{k=1}^{p} \alpha_k l_k x_k = w^* \\ \sum_{k=1}^{p} \alpha_k l_k = 0 \end{cases} \tag{2.27}$$

*w** représente la solution.
En réinjectant ces valeurs dans l'équation (2.26), on obtient la formulation duale :

$$\text{Maximiser } \tilde{L}(\alpha) = \sum_{k=1}^{p} \alpha_k - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j l_i l_j x_i^T x_j \tag{2.28}$$

$$\text{sous les contraintes } \alpha_k \geq 0, \text{ et } \sum_{k=1}^{p} \alpha_k l_k = 0 .$$

Ce qui donne les multiplicateurs de Lagrange optimaux α_k^*.
Afin d'obtenir l'hyperplan solution, on remplace *w* par sa valeur optimale w^*, dans l'équation de l'hyperplan *h*(*x*), ce qui donne :

$$h(x) = \sum_{k=1}^{p} \alpha_k^* l_k (x . x_k) + w_0 \tag{2.29}$$

Il y a trois remarques intéressantes à faire à propos de ce résultat. La première découle de l'une des conditions de Kuhn-Tucker, qui donne :

$$\alpha_k [l_k h(x_k) - 1] = 0 \quad 1 \leq k \leq p . \tag{2.30}$$

$$\text{d'où } \begin{cases} \alpha_k = 0 \\ l_k h(x_k) = 1 \end{cases} \tag{2.31}$$

- les seuls points pour lesquels les contraintes du lagrangien sont actives sont donc les points tels que $l_k h(x_k) = 1$, qui sont les points situés sur les hyperplans de marges maximales. En d'autres termes, seuls les *vecteurs supports* participent à la définition de l'hyperplan optimal ;
- la deuxième remarque découle de la première. Seul un sous-ensemble restreint de points est nécessaire pour le calcul de la solution, les autres échantillons ne participant pas du tout à sa définition. Ceci est donc efficace au niveau de la complexité. D'autre part, le changement ou l'agrandissement de l'ensemble d'apprentissage a moins d'influence que dans un modèle de mélanges gaussiens par exemple, où tous les points participent à la solution. En particulier, le fait d'ajouter des échantillons à l'ensemble d'apprentissage qui ne sont pas des vecteurs supports n'a aucune influence sur la solution finale ;

- la dernière remarque est que l'hyperplan solution ne dépend que du produit scalaire entre le vecteur d'entrée et les vecteurs supports. Cette remarque est l'origine de la deuxième innovation majeure des SVM: le passage par un espace de redescription grâce à une fonction noyau.

Cependant, nous avons toujours considéré que les données d'apprentissage sont linéairement séparables. Si ce n'est pas le cas, la méthode décrite précédemment ne fonctionne pas. De plus, comme nous pouvons le constater à la figure.2.6, même si les données d'apprentissage sont linéairement séparables, en présence d'une donnée aberrante, la maximisation de la marge de séparation peut conduire à l'obtention d'un hyperplan qui n'est plus du tout optimal.

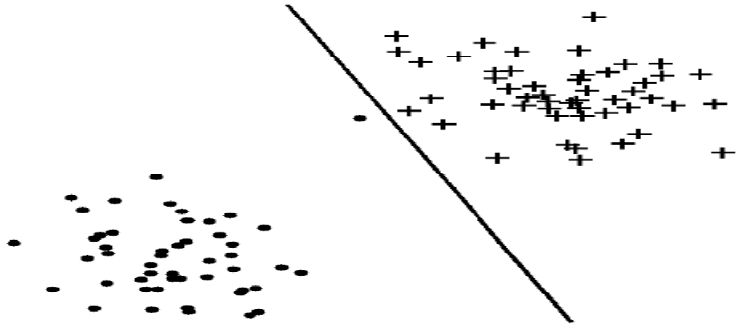


Figure.2.6 *Hyperplan sous-optimal lié à la présence d'une donnée aberrante.*

Pour contourner ce problème, Cortes et Vapnik **[34]** introduisent la notion de classifieurs à marge molle (« soft margin » en anglais) qui consiste à tolérer que certaines données d'apprentissage violent les contraintes de séparation (2.25). Ils proposent ainsi d'introduire des variables d'écart (« slack variables » en anglais) :

$$\varepsilon_k \geq 0 \tag{2.32}$$

et de nouvelles contraintes :

$$l_k(w.x_k + w_0) \geq 1 - \varepsilon_k \tag{2.33}$$

La formulation primale du problème d'optimisation revient alors à minimiser sous les contraintes (2.32) et (2.33) la fonction objective suivante :

$$\frac{1}{2}\|w\|^2 + c\sum_{k=1}^{n}\varepsilon_k \tag{2.34}$$

Où *C* est une constante positive, dite de régularisation. Ainsi, plus la valeur de *C* est grande, moins l'algorithme tolère que les données violent les contraintes de séparation.
Finalement, dans le cas de la formulation duale, on maximise toujours l'expression (2.28) sous les contraintes $\alpha_k \geq 0$, et $\sum_{k=1}^{p}\alpha_k l_k = 0$, aux quelles on ajoute les contraintes : $\alpha_k \leq c$

II.4.5 Technique du '' Kernel trick'' (Cas non séparable)

La technique du *"Kernel trick"* est utilisée pour résoudre les cas où il n'existe pas d'hyperplan séparateur dans l'espace où sont répartis les deux groupes de points (exemple de la figure 2.4.b). La notion de marge maximale et la procédure de recherche de l'hyperplan séparateur telles que présentées précédemment, ne permettent de résoudre que des problèmes de discrimination linéairement séparables. C'est une limitation sévère qui condamne à ne pouvoir résoudre que des problèmes très particuliers. Afin de remédier au problème de l'absence de séparateur linéaire, l'idée des SVM est de reconsidérer le problème dans un espace de dimension supérieure, éventuellement de dimension infinie. Dans ce nouvel espace, il est alors probable qu'il existe un séparateur linéaire. Plus formellement, on applique aux vecteurs d'entrée *x* une transformation non-linéaire ϕ et l'espace d'arrivée $\phi(x)$ est appelé *espace de redescription*. Dans cet espace, on cherchera l'hyperplan $h(x)=w^T\phi(x)+w_0$ qui vérifie la condition : $l_k h(x_k)>0$, pour tous les points x_k de l'ensemble d'apprentissage, c'est-à-dire l'hyperplan séparateur dans l'espace de redescription. En utilisant la même procédure que dans le cas sans transformation, on aboutit au problème d'optimisation suivant :

Maximiser $\tilde{L}(\alpha)=\sum_{k=1}^{p}\alpha_k-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j l_i l_j \phi(x_i)^T\phi(x_j)$ (2.35)

sous les contraintes $\alpha_i\geq 0$, et $\sum_{k=1}^{p}\alpha_k l_k=0$.

Le problème de cette formulation est qu'elle implique un produit scalaire entre vecteurs dans l'espace de redescription, de dimension élevée, ce qui est couteux en termes de calculs. Pour résoudre ce problème, on utilise une astuce connue sous le nom de *"Kernel trick"*, qui consiste à utiliser une fonction noyau, qui vérifie la condition suivante:

$$K(x_i,x_j)=\phi(x_i)^T.\phi(x_j) \quad (2.36)$$

D'où l'expression de l'hyperplan séparateur en fonction de la fonction noyau:

$$h(x)=\sum_{k=1}^{p}\alpha_k^* l_k K(x_k,x)+w_0 \quad (2.37)$$

L'intérêt de la fonction noyau est, alors, double :

- Le calcul se fait dans l'espace d'origine, ceci est beaucoup moins coûteux qu'un produit scalaire en grande dimension.

- La transformation ϕ n'a pas besoin d'être connue explicitement : seule la fonction noyau intervient dans les calculs. On peut donc envisager des transformations complexes, et même des espaces de redescription de dimension infinie **[34].**

II.4.6 Choix de la Fonction Noyau

En pratique, on construit directement une fonction noyau qui doit respecter certaines conditions, elle doit correspondre à un produit scalaire dans un espace de grande dimension.

Les noyaux usuels employés pour les SVM sont le noyau:

- polynomial : $K(x_i, x_j) = (x_i^T . x_j + 1)^d$ (2.38)
- gaussien ou RBF (*Radial Basis Function*) : $K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$ (2.39)

II.5 Conclusion

Dans ce chapitre, nous avons présenté différentes morphologies qui sont utilisées en RAL pour réaliser les références de locuteur. L'approche dominante pour représenter la référence du locuteur, en RAL, est le modèle de mélange de gaussiennes (GMM, *Gaussian Mixture Model*). Elle constitue l'état de l'art des systèmes de RAL et a montré de très bonnes performances sur plusieurs types du RAL. L'autre approche discriminante employée est le modèle SVM. A l'origine, il a été conçu comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à deux classes.

3

Les systèmes de communications mobiles

Dans le présent chapitre, nous donnons un aperçu général sur la reconnaissance automatique parole et locuteur dans les systèmes de communications mobiles. Après cela, nous allons décrire le codeur et le décodeur d'adaptation à multidébits en large bande (AMR-WB), qui sont principalement destinés à traiter des signaux de parole d'une largeur de bande de 7 kHz. L'adaptation AMR-WB fonctionne à une multitude de débits binaires compris entre 6,6 kbit/s et 23,85 kbit/s. Le débit peut être modifié à toute limite de trame de 20 ms.

III.1 Bref historique sur les systèmes de communications mobiles

Depuis le début des années 90, la téléphonie fixe est fortement concurrencée par la téléphonie mobile. Le mobile a connu une croissance fulgurante, avec l'introduction de la modulation numérique il est devenu possible d'avoir des signaux robustes aux interférences, qui améliore la capacité des systèmes cellulaires. De tel système de seconde génération (2G) sont le global System for Mobile communication (GSM) standardisée par l'ETSI (European Telecommunications Standards Institute) **[35]** en 1990. Les principales dégradations identifiées pour la téléphonie fixe (écho, bruit, dégradation due au codage de la parole, bande étroite) sont également présentes avec la téléphonie mobile. Cependant, la mobilité de l'utilisateur ajoute une variabilité des dégradations dans le temps, tels que des bruits non stationnaires (e.g. voiture qui passe dans la rue), et un délai plus long dû aux traitements numériques. La téléphonie mobile est actuellement en pleine convergence avec Internet : c'est la troisième génération de téléphones mobiles (3G), notamment avec la technologie UMTS (Universal Mobile Telecommunications System). Elle offre une bande passante plus large et un débit plus élevé pour l'échange de données et pour de nouvelles applications telles que la visiophonie ou la télévision.

III.2 La reconnaissance automatique dans les systèmes mobiles

III.2.1 Différentes architectures pour la reconnaissance de locuteur/parole dans les systèmes de communication mobiles

Nous pouvons dénombrer trois grandes architectures pour la reconnaissance de la parole/locuteur dans les systèmes de communications mobiles **[36]**:

- La reconnaissance déportée ou reconnaissance dans le réseau NSR (Network Speech/Speaker Recognition).
- La reconnaissance embarquée ou reconnaissance dans le terminal TSR (Terminal Speech/Speaker Recognition).
- La reconnaissance répartie ou reconnaissance distribuée DSR (Distributed Speech/Speaker Recognition).

III.2.1.1 La reconnaissance déportée NSR

Le principe de la reconnaissance déportée (Figure.3.1) consiste à se servir du terminal uniquement comme un système d'acquisition du son. Une fois le signal enregistré, il est transmis via un réseau, à un serveur qui effectue le travail de reconnaissance. Cette approche permet d'envisager l'utilisation de serveurs beaucoup plus puissants et donc de fournir des services plus divers et généralement de meilleure qualité. La présence d'un serveur performant permet d'adjoindre d'autres modules à la reconnaissance vocale, tels qu'un système de dialogue ou encore des vocabulaires dédiés à chaque application. Pour résumer, l'avantage majeur de cette architecture est la levée des contraintes liées aux ressources limitées des systèmes mobiles.

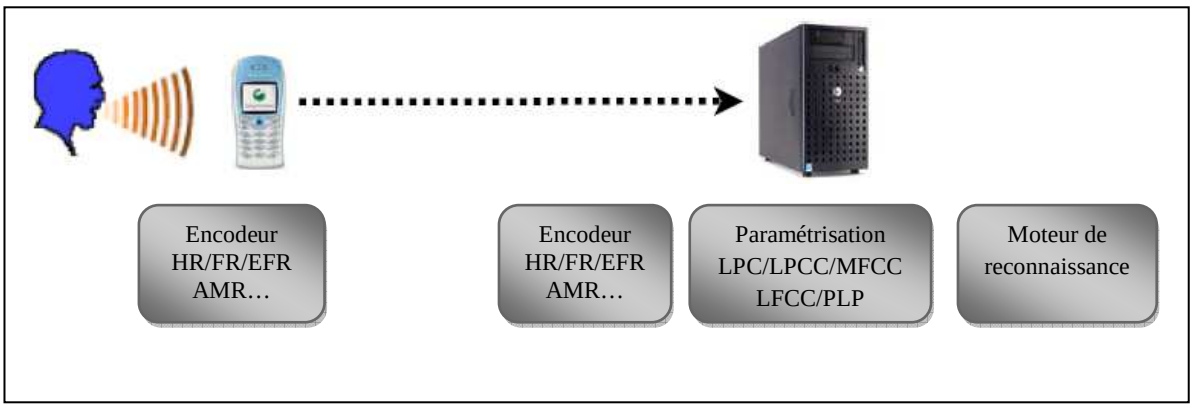


Figure.3.1 *Architecture déportée.*

Un des inconvénients de ce type d'architecture est la perte d'informations due à la transmission du signal. En effet, avant d'être envoyé au(x) serveur(s) de reconnaissance, le signal doit être encodé suivant un des codecs de l'opérateur (HR **[37]**, FR **[38]**, EFR **[39]**, AMR **[40].** . .). Ces codages (avec pertes) sont nécessaires car l'émission du signal sur le réseau a un coût non négligeable. De plus, au cours d'une transmission sur un réseau (sans fil ou non), différents phénomènes, comme la perte de paquets d'informations ou l'ajout de bruits viennent perturber le signal.
Parmi les inconvénients, nous pouvons encore citer le coût (tant sur le plan financier que sur l'aspect du trafic réseau) de transmission du signal au serveur et de la réponse du serveur au terminal.
Dans la littérature, les problèmes généraux de la reconnaissance déportée (pas seulement via des réseaux téléphoniques mobiles) sont abordés notamment dans **[41]** ou **[42]**. Des exemples plus spécifiques aux réseaux GSM sont présentés dans **[43], [44]** et **[45].**

III.2.1.2 La reconnaissance embarquée TSR

Avec cette approche, toutes les étapes de la reconnaissance sont effectuées dans le système embarqué (Figure.3.2). Aucun serveur à distance est nécessaire, le système est donc entièrement autonome : il réalise lui-même l'acquisition du son puis le décodage.
L'intérêt principal de cette approche est l'absence de contrainte : en effet, il n'est pas nécessaire de se limiter à l'utilisation des codecs de parole définis par l'ETSI, car aucune compatibilité avec un serveur n'est requise. De plus, les inconvénients de la reconnaissance déportée (perte d'informations due à la transmission ou à la compression avant émission) sont résolus par l'absence de transmission. Cette solution présente toutefois un inconvénient majeur : les ressources nécessaires. En effet, toutes les phases (paramétrisation du signal et reconnaissance) sont effectuées au sein du téléphone. Les différentes phases de la reconnaissance nécessitent beaucoup de ressources comparées aux capacités d'un téléphone portable. De plus, l'augmentation des ressources d'un téléphone (mémoire et puissance de calcul), bien que possible, est très coûteuse.

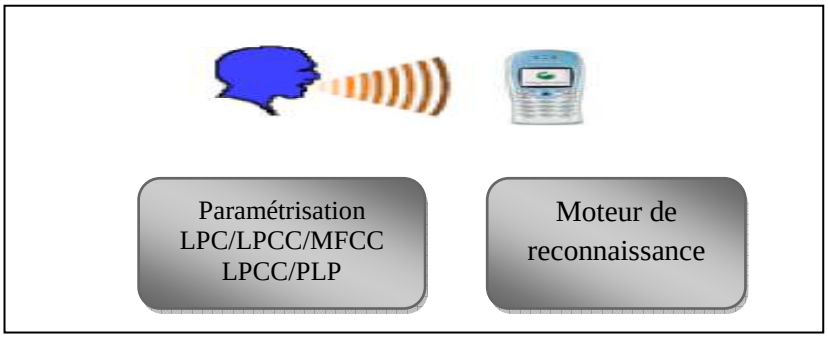


Figure.3.2 *Architecture embarquée.*

III.2.1.3 La reconnaissance distribuée (répartie) DSR

Ce mode de reconnaissance se situe entre la reconnaissance déportée et la reconnaissance embarquée. En effet, une partie du travail est effectuée sur le système embarqué et une autre partie sur un serveur distant (Figure.3.3). Généralement, la phase d'extraction des paramètres est réalisée sur le terminal ([**46], [47], [48].** . .) et la partie reconnaissance -à proprement parler- se trouve sur un serveur.

Cette approche permet de contourner les problèmes liés à la perte d'informations due au transport.

Figure.3.3 *Architecture répartie.*

L'extraction des paramètres est effectuée directement dans le téléphone. Ces paramètres sont ensuite utilisés par le système de reconnaissance durant la phase de décodage. Les codecs classiques, définis par l'ETSI, sont plutôt basés sur les coefficients LPC alors que les systèmes actuels de RAL sont essentiellement basés sur les coefficients MFCC ou PLP ou encore sur des dérivés de l'une de ces deux paramétrisation, il est donc nécessaire d'ajouter des fonctions spécifiques de paramétrisation au sein des téléphones.

La DSR présente les inconvénients suivants :

- un traitement supplémentaire doit être implanté dans le terminal,
- l'interopérabilité de tous les terminaux mobiles n'est assurée que si un standard est défini,
- une charge réseau supplémentaire est induite pour la transmission des paramètres.

Une architecture distribuée de paramétrisation, Aurora, a été standardisée par l'ETSI dans ce sens. Ce standard est brièvement expliqué dans la section suivante.

III.2.2 Le standard ETSI Aurora pour une architecture distribuée

Le standard ETSI Aurora **[49]** a été créé à l'origine pour effectuer la reconnaissance de la parole sur des architectures distribuées.

III.2.2.1 Définition du standard

Les performances de reconnaissance sont altérées par le codage bas débit de la parole mis en place sur les réseaux de communication, mais aussi par les erreurs de transmission. Sur ce type d'architecture, le terminal a alors pour charge d'extraire les paramètres cepstraux et de les transmettre sur un canal de données protégé, après compression.

Le flux compressé est ensuite reçu par le serveur distant pour effectuer la reconnaissance. Les dégradations dues au codage bas débit de la parole et aux pertes de transmission sont ainsi évitées. L'architecture distribuée, entre le terminal mobile et un serveur d'applications distant, permet notamment de s'affranchir des problèmes dus au codage bas-débit de la parole **[50].**
Le standard Aurora est issu d'un groupe de travail de l'ETSI et a été publié pour la première fois en février 2000, dénommé Aurora-1. Le standard a régulièrement bénéficié d'améliorations. Le standard Aurora-2 intègre ainsi un étage de réduction de bruit, basée sur le filtrage de Wiener, avant le calcul des paramètres cepstraux et une méthode d'égalisation aveugle des paramètres cepstraux **[51], [52].** Elle permet de compenser les variations, à court terme, du canal d'acquisition acoustique. L'évolution Aurora-3 permet la reconstruction du signal au niveau du serveur d'application distant (figure.3.4). Pour reconstruire le signal, à partir des paramètres transmis, et améliorer la reconnaissance sur les langues tonales, le standard utilise les paramètres de fréquence fondamentale (pitch) et de classes de voisement.

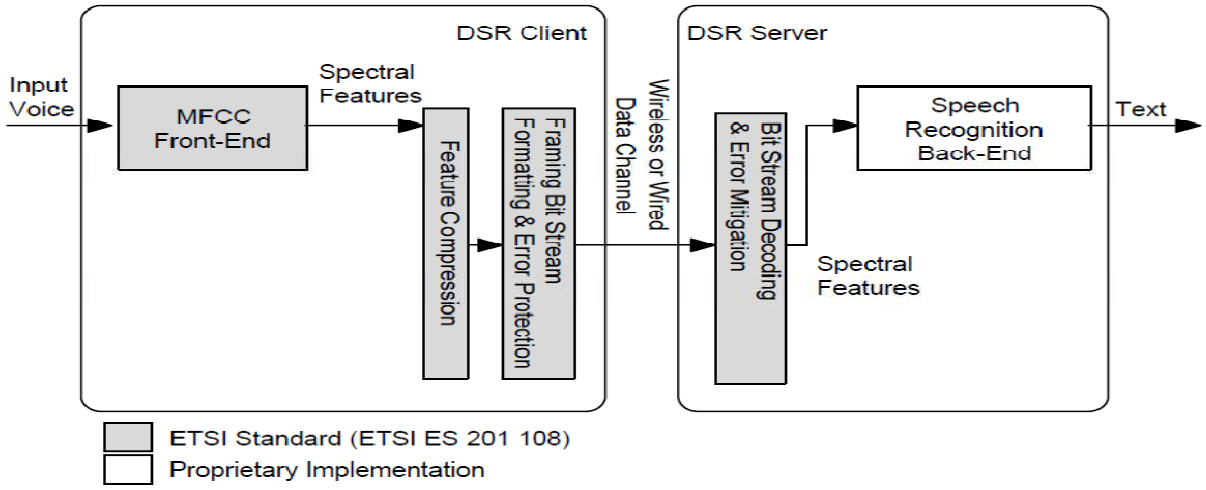


Figure.3.4 *Reconnaissance de parole distribuée et le standard ETSI Aurora.*

Destiné aux réseaux de communication téléphonique, le standard est compatible avec des signaux d'une fréquence d'échantillonnage de 8kHz, mais aussi de 11kHz et de 16kHz. L'étage d'extraction des paramètres cepstraux s'effectue comme suit :

- le signal d'entrée est débruité par un filtrage de Wiener.
- les paramètres cepstraux sont calculés avec 23 filtres, de fréquence centrale fixée selon l'échelle de MEL.
- les paramètres cepstraux sont égalisés par un algorithme du gradient (Least Mean Square, LMS), à partir d'un cepstre de référence estimé sur un bruit.
- Le standard Aurora utilise 12 paramètres cepstraux statiques et un paramètre d'énergie, calculés toutes les 10 ms sur une fenêtre d'analyse de 25 ms (Hamming).
- L'étage de quantification est ensuite appliqué aux paramètres cepstraux avant transmission. Le débit total utilisé pour la transmission est alors de 5.6 kbit/s.

A la réception, le serveur déquantifie les paramètres et extrait les dérivées, premières et secondes, des cepstres et de l'énergie. Un horizon de 9 trames (90 ms) est utilisé pour le calcul des dérivées. Au total un vecteur de 39 paramètres est disponible au niveau du serveur.
Le standard Aurora ne se contente pas seulement d'effectuer une analyse cepstrale, il propose sa propre détection d'activité vocale. Elle utilise la mesure de l'accélération de l'énergie, calculée sur la trame, sur une sous-région du spectre qui contient la fréquence fondamentale et sur la bande basse du spectre **[49].** L'accélération de l'énergie est basée sur un calcul de variance spectrale. L'état de voisement de la trame est aussi disponible.
Au total quatre classes, dont trois sur l'état de voisement et une sur la présence de parole, sont définies :

- ✓ bruit/silence.
- ✓ non voisée.
- ✓ moyennement voisée.
- ✓ voisée.

L'information de voisement est obtenue par différents indicateurs :

- ✓ la mesure du taux de passage par zéro du signal (ZCM, zero Crossing Measure).
- ✓ un seuil sur l'énergie.
- ✓ l'indicateur d'activité vocale.

Si le traitement au niveau du serveur peut être facilement implémenté, l'utilisation de ce standard implique une compatibilité avec les terminaux mobiles, i.e., les routines algorithmiques du standard doivent être embarquées dans les terminaux. C'est aujourd'hui de plus en plus le cas, mais il est encore difficile de proposer une solution basée sur le standard Aurora, sur les réseaux de communication existants.
L'utilisation du standard Aurora présente de multiples avantages. Nous proposons une synthèse des avantages et des inconvénients du standard.

a. Avantages

– utilisation du signal originel acquis sur le terminal pour l'extraction des paramètres,
– Détection d'activité vocale VAD incluse,
– débruitage et égalisation,
– architecture standardisée,
– paramètres de voisement et pitch disponibles pour chaque trame.

b. Inconvénients :

– débit supplémentaire nécessaire pour la transmission,
– 12 premiers cepstres disponibles seulement.

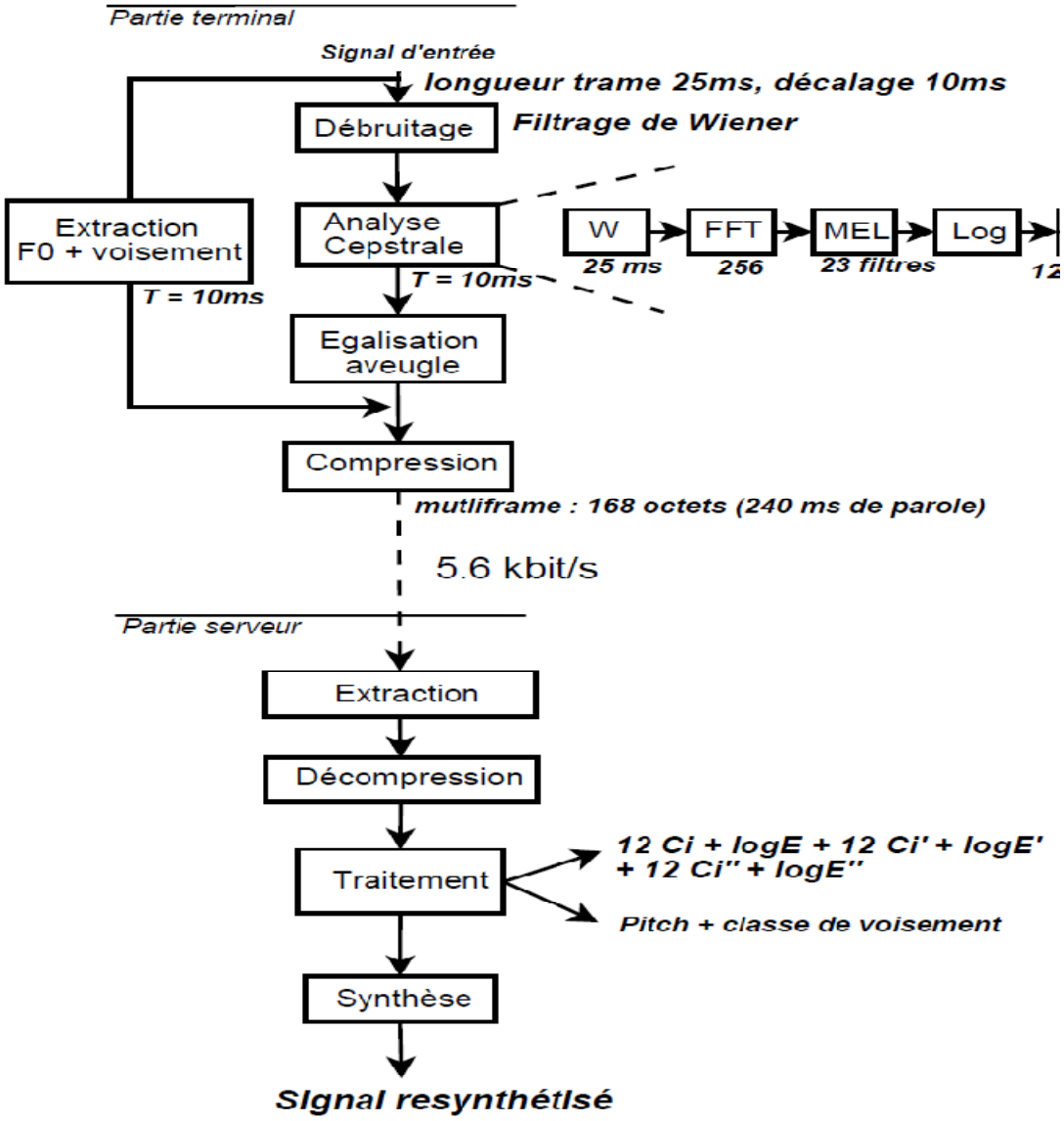


Figure.3.5 *DSR front-end.*

III.2.3 La reconnaissance dans le domaine compressé

Pour exploiter le flux audio qui transite sur les réseaux, dans le but d'effectuer la RAL, la méthode la plus simple est de calculer les paramètres cepstraux à partir du signal resynthétisé. Ceci introduit alors des distorsions dues à la resynthèse (décodage) du signal. Des travaux ont été menés pour utiliser des paramètres utiles à la RAL, à partir des paramètres internes des codeurs **[53]**. Ceci permet notamment d'éviter les distorsions dues à la resynthèse du signal et la consommation de ressources supplémentaires.
Dans notre travail et dans le but d'évaluer un autre schéma d'extraction des paramètres, compatible avec les caractéristiques des systèmes de communications mobiles, une solution utilisant l'extraction des paramètres cepstraux dans le domaine compressé, c.-à-d., à partir des paramètres internes des codeurs de parole **[54]**.

III.3 Le codage de la parole

Le codage de la parole est très utilisé sur les réseaux de communication. Il permet de diminuer la bande passante nécessaire au transfert de la parole. Les distorsions apportées par les codeurs de parole sont d'autant plus importantes que le débit alloué est faible. Les codeurs utilisés dans les réseaux mobiles exploitent les caractéristiques des signaux de la parole pour réduire le débit à des valeurs typiquement comprises entre 8 et 24 Kb/s tout en conservant une bonne qualité, mais inférieure à celle des réseaux téléphoniques filaires. Outre les codeurs de parole, on rencontre souvent dans les réseaux mobiles des mécanismes de transmission discontinue notés DTX *Discontinuous Transmission*. Ils sont destinés à réduire la charge du réseau en limitant la transmission sur le canal de tout signal qui ne contient pas de la parole. La distinction entre un signal qui contient de la parole et un autre sans parole est effectuée par des détecteurs d'activité vocale VAD *Voice Activity Detector*, qui s'appuient sur des paramètres caractéristiques du signal de parole.
Le codeur de parole, utilisé dans le cadre de cette thèse, est le codeur AMR_WB **[1],** et qui sera détaillé par la suite dans la deuxième partie de ce chapitre.

Le codeur ou codec est le dispositif qui dans la chaîne de communication adapte le signal à transmettre – ici la parole - au réseau de transport. La fonction première du codeur est la compression de débit. A l'émission, le codeur produit un flux codé c'est-à-dire une représentation des échantillons de parole sur un nombre de bits compatible avec les capacités du réseau de transmission. C'est ce format codé qui circule sur le réseau. A la réception, le décodeur régénère un signal qui d'un point de vue perceptuel doit être le plus proche possible du signal de parole original. La réduction de débit doit se faire en respectant plusieurs contraintes:

- une dégradation la plus faible possible de la qualité de la parole,
- un retard introduit dans la chaîne de communication le plus faible possible
- une robustesse aux erreurs de transmission survenant sur la chaîne de communication.

Pour parvenir à ces fins, le codage de parole fait appel à des techniques algorithmiques qui tirent parti des propriétés de la parole et de celles de l'audition humaine.
Il existe plusieurs familles de codeurs de parole (référence). Certains sont normalisés c'est-à-dire développés ou sélectionnés dans le cadre d'une action de normalisation lancée par un organisme international tel le 3GPP (3d Generation Partnership Projet) **[55]** ou l'ITU-T **[56]**. La solution retenue résulte généralement d'une compétition entre plusieurs candidats selon des critères connus et partagés par les acteurs.

La technologie AMR (Adaptive Multi-Rate) a été développée du fait sa robustesse aux erreurs de transmission. Il s'agit d'un codeur multi débits c'est-à-dire pouvant fonctionner à plusieurs débits (ou modes). Le choix du mode répond à un compromis entre la qualité de la parole (i.e. le débit du codeur), le niveau de qualité du lien radio (C/I, niveau de champ rapporté aux interférences) et la capacité de la cellule. Le codec AMR a été normalisé au 3GPP (instance de normalisation pour les réseaux mobiles cellulaires) en 1999. S'appliquant à des signaux bande étroite, il comporte 8 modes: 12,2 kbit/s, 10,2 kbit/s, 7,95 kbit/s, 7,40 kbit/s, 6,70 kbit/s, 5,90 kbit/s, 5,15 kbit/s and 4,75 kbit/s. Il est largement employé sur les réseaux mobiles 2G et 3G. Bâti sur le même modèle, le codeur AMR_WB destiné à la bande élargie a quant à lui été normalisé en 2000 au 3GPP. Il a été également reconnu comme norme par l'ITU-T en Juillet 2003 sous le nom de G.722.2 **[1].**

III.4 Adaptative Multi Rate Wideband AMR_WB

Le codeur AMR_WB (Adaptatif multidébit à large bande) comporte 9 modes dont seuls 5 sont obligatoire dans les terminaux: 6,6 kbit/s, 8,85 kbit/s, 12,65 kbit/s, 15,85 kbit/s, 23,85 kbit/s **[1].** Sur les réseaux 2G, seuls les 3 débits inférieurs peuvent être employés. Sur les réseaux 3G, les 5 débits sont utilisables. Toutefois, en termes de qualité, pour la parole, les 2 débits supérieurs se distinguent peu du mode 12,65 kbit/s tout en consommant notablement plus de ressource radio. Le codec AMR_WB est fondé sur le modèle de codage prédictif linéaire avec excitation par code (CELP, *code excited linear prediction*).

Le signal d'entrée est préaccentué au moyen du filtre de préaccentuation :

$$H_{pre-emp}(z) = 1 - \mu\, z^{-1} \tag{3.1}$$

Le modèle CELP est appliqué ensuite au signal préaccentué. Un filtre de synthèse à prédiction linéaire (LP) du 16 ordre (ou à court terme) est utilisé, ce qui est donné par:

$$H(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1+\sum_{i=1}^{m} \hat{a}_i z^{-i}} \tag{3.2}$$

Où $\hat{a}_i$, $i = 1,\ldots, m$ sont les paramètres de prédiction linéaire (LP, *linear prediction*) et où m = 16 est l'ordre du prédicteur.

Le modèle de synthèse vocale par prédiction CELP est représenté dans la Figure.3.6. Dans ce modèle, le signal d'excitation injecté à l'entrée du filtre de synthèse LP à court terme est construit par adjonction de deux vecteurs d'excitation issus des répertoires adaptatif et fixe (d'innovation). La parole est synthétisée par injection des deux vecteurs convenablement choisis dans ces répertoires par le filtre de synthèse à court terme. La séquence d'excitation optimale est choisie dans un répertoire au moyen d'une procédure de recherche par analyse et synthèse dans laquelle l'erreur entre le signal vocal original et le signal vocal synthétisé est minimisée en fonction d'une valeur mesurée de distorsion à pondération perceptive.

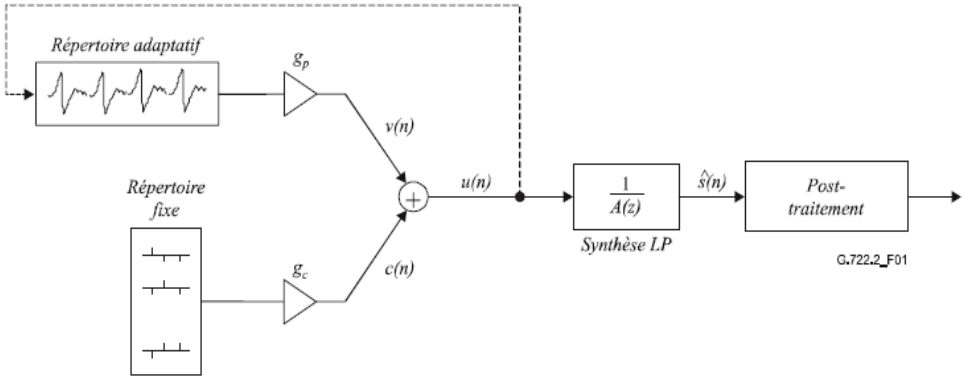


Figure.3.6 *Schéma fonctionnel simplifié du modèle de synthèse CELP***.**

Le filtre à pondération perceptive utilisé dans la technique de recherche par analyse et synthèse est donné par la relation suivante:

$$W(z) = A\,(z/\gamma_1)\, H_{de-emph}(z) \tag{3.3}$$

Où *A(z)* est le filtre de prédiction linéaire non quantifiée, γ_1=0.92, et $H_{de-emp}(z) = \frac{1}{1-0.68z^{-1}}$

III.4.1 Description du codeur

Le codeur effectue l'analyse des paramètres de codage LPC, de prédiction LTP et de répertoire fixe à la fréquence d'échantillonnage de 12,8 kHz. Le codeur travaille sur des trames vocales de 20 ms. A chaque trame, le signal vocal est analysé afin d'extraire les paramètres du modèle de prédiction CELP (coefficients de filtre LP, indices et gains de répertoire adaptatif et de répertoire fixe). En plus de ces paramètres, des indices de gain dans la bande des hautes fréquences sont calculés en mode 23,85 kbit/s. Le flux des signaux dans le codeur est décrit par la Figure 3.5.

III.4.1.1 Prétraitement

Le codeur effectue l'analyse des paramètres de codage LPC, de prédiction LTP et du répertoire fixe à la fréquence d'échantillonnage de 12,8 kHz. Le signal d'entrée doit donc subir une décimation de 16 kHz à 12,8 kHz. Après la décimation, deux fonctions de prétraitement sont appliquées au signal: un filtrage passe-haut et une préaccentuation. Le filtre passe-haut sert de protection contre des composants parasites à basse fréquence. Lors de la préaccentuation, un filtre passe-haut du premier ordre est utilisé afin d'accentuer les très hautes fréquences. Ce filtre est donné par : $H_{pre\text{-}emp}(z) = 1 - 0.68z^{-1}$

III.4.1.2 Analyse par prédiction linéaire et quantification

Une analyse par prédiction à court terme (LP) est effectuée à chaque trame vocale au moyen de la méthode d'autocorrélation dans une fenêtre asymétrique de 30 ms. Une redondance de 5 ms est utilisée lors du calcul d'autocorrélation.
Les autocorrélations de signaux vocaux fenêtrés sont converties en coefficients de prédiction linéaire au moyen de l'algorithme de Levinson-Durbin. Les coefficients LP sont ensuite transformés en paires du domaine ISP aux fins de la quantification et de l'interpolation. Les filtres interpolés, quantifiés et non quantifiés, sont reconvertis en coefficients de filtre LP (afin de construire les filtres de synthèse et de pondération à chaque sous-trame).

a. Conversion de la prédiction linéaire (LP) en paires spectrales d'immitance (ISP)

Les coefficients du filtre LP a_k, k = 1,…, 16, sont convertis en représentation de paires ISP aux fins de la quantification et de l'interpolation. Pour un filtre LP du 16^{e} ordre, les paires ISP sont définies comme étant les racines des polynômes de somme et de différence suivants:

$$f_1^{'}(z) = A(z) + z^{-16} A(z^{-1}) \quad (3.4)$$

$$f_2^{'}(z) = A(z) - z^{-16} A(z^{-1}) \quad (3.5)$$

Les polynômes $f_1^{'}(z)$ et $f_2^{'}(z)$ sont respectivement symétrique et antisymétrique. L'on peut démontrer que toutes les racines de ces polynômes sont sur le cercle des unités et qu'elles alternent l'une avec l'autre. Le polynôme $f_2^{'}(z)$ a deux racines, à z = 1 (ω= 0) et à z = –1 (ω= π).

Pour éliminer ces deux racines, l'on définit les nouveaux polynômes suivants:

$$f_1\ (z) = f_1^{'}(z) \tag{3.6}$$

$$f_2\ (z) = f_2^{'}(z)/(1 - z^{-2}) \tag{3.7}$$

Les polynômes *f*1(*z*) et *f*2(*z*) ont respectivement 8 et 7 racines conjuguées sur le cercle des unités $e^{\mp jw_i}$. Donc, ces polynômes peuvent être écrits comme suit:

$$F_1\ (z) = (1 + a[16]) \prod_{i=0,2,\ldots,14}(1 - 2q_i z^{-1} + z^{-2}) \tag{3.8}$$

$$F_2\ (z) = (1 + a[16]) \prod_{i=1,3,\ldots,13}(1 - 2q_i z^{-1} + z^{-2}) \tag{3.9}$$

où q_i = *cos* (ω_i) avec ω_i qui représente les fréquences spectrales d'immitance (ISF) et où $a[16]$ est le dernier coefficient du prédicteur. Les paires ISF répondent à la relation d'ordonnancement $0 < \omega_1 < \omega_2 < \ldots < \omega_{16} < \pi$. L'on considère q_i comme étant les paires ISP dans le domaine cosinusoïdal.
Étant donné que les deux polynômes *f*1*(z)* et *f*2*(z)* sont symétriques, seuls les 8 et 7 premiers coefficients de chaque polynôme, respectivement, et le dernier coefficient du prédicteur, ont besoin d'être calculés. Les coefficients de ces polynômes sont trouvés par les relations récursives

pour i = 0 à 7

$$f_1(i) = a_i + a_{m-i}$$

$$f_2(i) = a_i - a_{m-i} + f_2(i-2) \tag{3.10}$$

$$f_1(8) = 2a_8$$

où *m* = 16 est l'ordre du prédicteur, et $f_2\ (-2) = f_2\ (-1) = 0$.
Les paires ISP sont trouvées par évaluation des polynômes *F*1(*z*) et *F*2(*z*) à 100 points équidistants entre 0 et π et par vérification des changements de signe. Un changement de signe signifie l'existence d'une racine et l'intervalle du changement de signe est ensuite divisé 4 fois afin de mieux localiser la racine.

III.4.1.3 Pondération perceptive

Le filtre de pondération perceptive traditionnel *W*(*z*) = *A*(*z*/γ1)*A*(*z*/γ2) possède des limitations inhérentes dans la modélisation simultanée de la structure du formant et de l'écart requis du niveau spectral à court terme. L'écart de niveau spectral à court terme est plus prononcé dans les signaux en bande large en raison de la large étendue dynamique entre basses et hautes fréquences. Une solution à ce problème consiste à introduire le filtre de préaccentuation à l'entrée, à calculer le filtre LP *A*(*z*) sur la base du signal vocal préaccentué ***s***(*n*), et à utiliser un

filtre $W(z)$ modifié par réglage de son dénominateur. Cette structure découple nettement la pondération par le formant par rapport à l'écart de niveau à court terme. Un filtre de pondération de la forme $W(z)=A(z/\gamma 1)\, H_{de-emph}(z)$ est utilisé, où $H_{de-emph} = \frac{1}{1-\beta_1 z^{-1}}$ et $\beta 1 = 0{,}68$.

III.4.1.4 Analyse en boucle ouverte du délai tonal

Selon le mode, l'analyse du délai tonal en boucle ouverte est effectuée une fois par trame (de 20 ms chacune) ou deux fois par trame (de 10 ms chacune) afin de trouver deux estimations du délai tonal dans chaque trame. Cela est effectué afin de simplifier l'analyse tonale et de limiter la recherche en boucle fermée du délai tonal à un petit nombre de délais autour des délais en boucle ouverte estimés.

III.4.1.5 Calcul de la réponse aux impulsions

La réponse impulsionnelle, *h(n)*, du filtre pondéré de synthèse est calculée à chaque sous-trame. Cette réponse impulsionnelle est requise pour la recherche dans les répertoires adaptatif et fixe.

III.4.1.6 Calcul du signal cible

Le signal cible pour recherche dans le répertoire adaptatif est habituellement calculé par soustraction de la réponse à entrée nulle du filtre pondéré de synthèse $H(z)W(z) = A(z/\gamma 1)H_{de-emph}(z)/\hat{A}(z)$ du signal vocal pondéré $s_w(n)$. Cette soustraction est effectuée sous-trame par sous-trame.

III.4.1.7 Répertoire adaptatif

La recherche dans le répertoire adaptatif est effectuée sous-trame par sous-trame. Elle consiste à effectuer la recherche en boucle fermée du délai tonal, puis à calculer le vecteur de code adaptatif par interpolation de l'excitation antérieure au délai tonal fractionnaire choisi.
Les paramètres du répertoire adaptatif (ou paramètres tonaux) sont le délai et le gain du filtre tonal.

III.4.1.8 Répertoire algébrique

La structure du répertoire est fondée sur un modèle de permutation entrelacée d'impulsion isolée (ISPP, *interleaved single-pulse permutation*). Les 64 positions dans le vecteur de code sont divisées en 4 pistes de positions entrelacées, avec 16 positions dans chaque piste. Les différents répertoires aux différents débits sont construits par insertion d'un certain nombre d'impulsions signées dans les pistes (de 1 à 6 impulsions par piste). L'indice de répertoire, ou mot codé, représente les positions et les signes d'impulsion dans chaque piste.

III.4.1.9 Quantification des gains de répertoire adaptatif et de répertoire fixe

Le gain du répertoire adaptatif (gain tonal) et le gain de répertoire fixe (algébrique) sont quantifiés vectoriellement au moyen d'un répertoire de 6 bits pour les modes 8,85 et 6,60 kbit/s et au moyen d'un répertoire de 7 bits pour tous les autres modes.

III.4.1.10 Mise à jour de la mémoire

Une mise à jour des états du filtre de synthèse et du filtre de pondération est requise afin de calculer le signal cible dans la sous-trame suivante.

III.4.1.11 Production du gain dans la bande des hautes fréquences

Afin de calculer le gain dans la bande des hautes fréquences pour le mode 23,85 kbit/s mode, le signal vocal d'entrée de 16 kHz est filtré dans un filtre FIR passe-bande que possède une bande passante de 6,4 à 7 kHz.

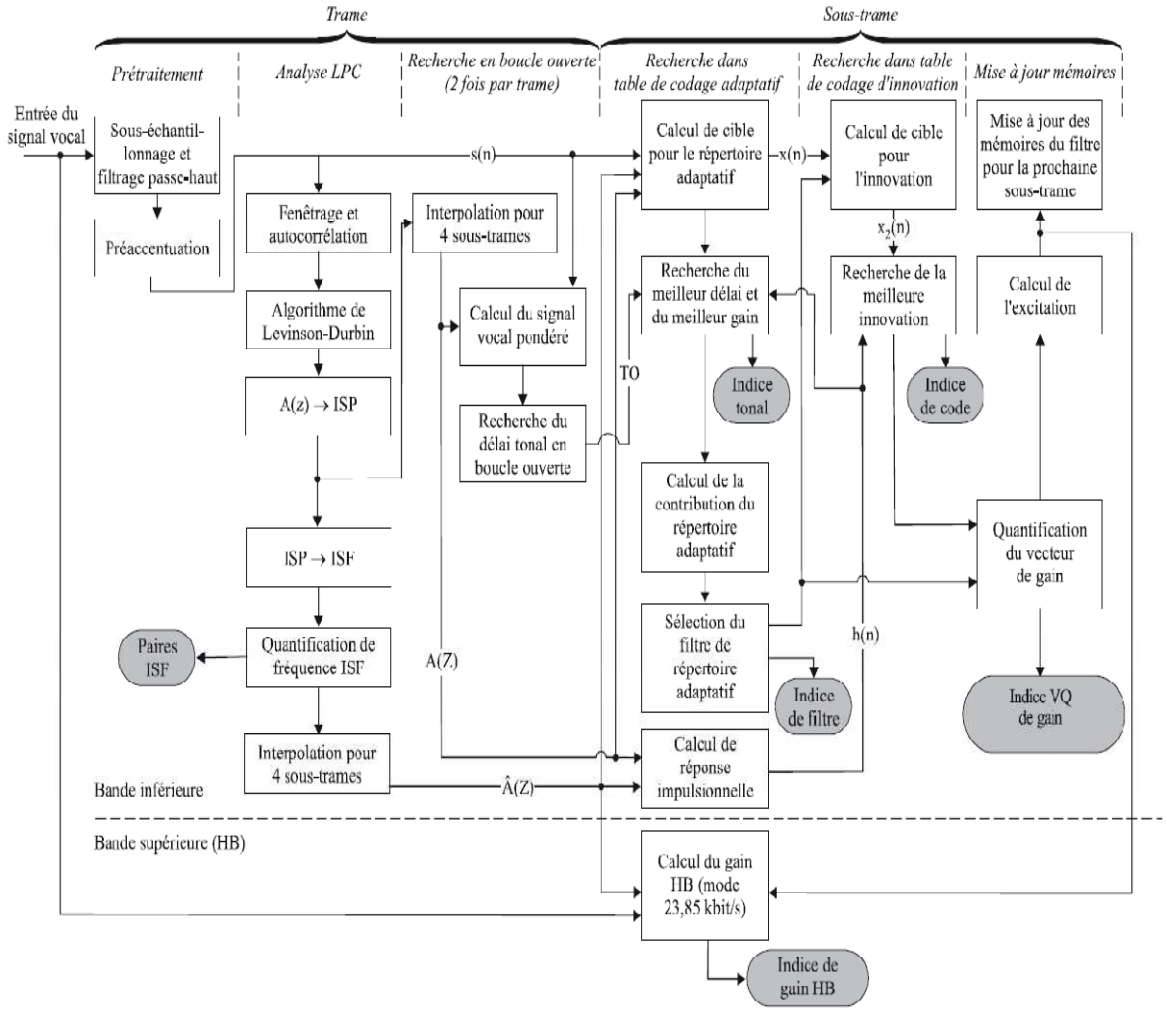


Figure.3.7 *Schéma fonctionnel de codeur AMR_WB.*

III.4.2 Description du décodeur

La fonction du décodeur se compose du décodage des paramètres transmis (paramètres LP, vecteur du répertoire adaptatif, gain du répertoire adaptatif, vecteur du répertoire fixe, gain du répertoire fixe et gain en bande haute) et de l'exécution d'une synthèse afin d'obtenir le signal vocal reconstruit, lequel subit ensuite un post-traitement et un suréchantillonnage. Finalement le signal en bande haute est produit dans la bande de fréquence de 6 à 7 kHz. Le flux de signaux dans le décodeur est présenté dans la Figure.3.7.

III.4.2.1 Décodage et synthèse de parole

Le processus de décodage est effectué dans l'ordre suivant:

- **Décodage des paramètres de filtre LP**: les indices reçus de quantification de paires ISP sont utilisés pour reconstruire le vecteur de paire ISP quantifiée. L'interpolation est effectuée afin d'obtenir 4 vecteurs ISP interpolés (correspondant à 4 sous-trames). Pour chaque sous-trame, le vecteur ISP interpolé est converti dans le domaine des coefficients de filtre LP, *ak*, qui est utilisé pour synthétiser le signal vocal reconstruit dans la sous-trame.

Les échelons suivants sont répétés pour chaque sous-trame:

1. **décodage du vecteur du répertoire adaptatif :** l'indice tonal reçu (du répertoire adaptatif) est utilisé afin de trouver les parties entière et fractionnaire du délai tonal. Le vecteur du répertoire adaptatif $v(n)$ est trouvé par interpolation de l'excitation antérieure $u(n)$ (au délai tonal) au moyen du filtre FIR. L'indice de filtre adaptatif reçu est utilisé afin de trouver si le répertoire adaptatif filtré est $v1(n) = v(n)$ ou $v2(n) = 0{,}18v(n) + 0{,}64\, v(n-1) + 0{,}18\, v(n-2)$;
2. **décodage du vecteur innovatif**: l'indice reçu du répertoire algébrique sert à extraire les positions et les amplitudes (signes) de l'excitation impulsions et afin de trouver le vecteur de code algébrique $c(n)$.
3. **décodage du gain de répertoire adaptatif et de répertoire d'innovation**: l'indice reçu donne le facteur de correction $\hat{\gamma}$ du gain du répertoire fixe.
4. **calcul du signal vocal reconstruit**: les échelons suivants sont pour $n = 0,..., 63$.
   L'excitation totale est construite par:
   $$u(n) = v(n)\hat{g}_p + \hat{g}_c c(n) \quad (3.11)$$
   Avant la synthèse vocale, un post-traitement des éléments d'excitation est effectué;
5. **renforçateur d'excitation en terme de bruit**: une technique de lissage non linéaire du gain est appliquée au gain du répertoire fixe $\hat{g}_c$ afin de renforcer l'excitation en terme de bruit. Sur la base de la stabilité et de la verbalisation du segment vocal, le gain du répertoire fixe est lissé afin de réduire la fluctuation de l'énergie de l'excitation en cas de signaux stationnaires. Cela améliore la performance en présence de bruit de fond stationnaire.
6. **renforçateur de délai tonal**: une procédure de renforcement du délai tonal modifie l'excitation totale $u(n)$ par filtrage de l'excitation contenue dans le répertoire fixe dans un filtre d'innovation dont la réponse en fréquence accentue les fréquences supérieures et dont les coefficients sont associés à la périodicité acheminée par le signal.

III.4.2.2 Filtrage passe-haut, renormalisation multiplicatrice et interpolation

Le filtre passe-haut sert de précaution contre les composants parasites à basse fréquence. Le signal est filtré dans le filtre passe-haut $H_{h1}(z)$ et dans le filtre désaccentué $H_{\text{de-emph}}(z)$.
Finalement, le signal est suréchantillonné à 16 kHz afin d'obtenir le signal de synthèse en bande inférieure $\hat{s}_{16k}(n)$. $\hat{s}_{16k}(n)$ est produit d'abord par suréchantillonnage à 5 fois du signal de synthèse en bande inférieure $\hat{s}_{12.8k}(n)$ à 12,8 kHz, ensuite par filtrage de la sortie dans $H_{\text{decim}}(z)$, et finalement par sous-échantillonnage à 4 fois du signal. (La normalisation multiplicatrice consiste en une multiplication par 2 de la sortie du filtrage passe-haut afin de compenser la normalisation réductrice dans la phase de prétraitement.)

III.4.2.3 Bande des hautes fréquences

Dans la bande de fréquence supérieure (6,4-7,0 kHz), l'excitation est produite de façon à modéliser les fréquences les plus élevées. Le contenu en hautes fréquences est produit par remplissage de la partie supérieure du spectre avec un bruit blanc convenablement échelonné dans le domaine d'excitation, converti ensuite dans domaine vocal par mise en forme dans un filtre issu du filtre LP de synthèse utilisé pour synthétiser le signal sous-échantillonné.

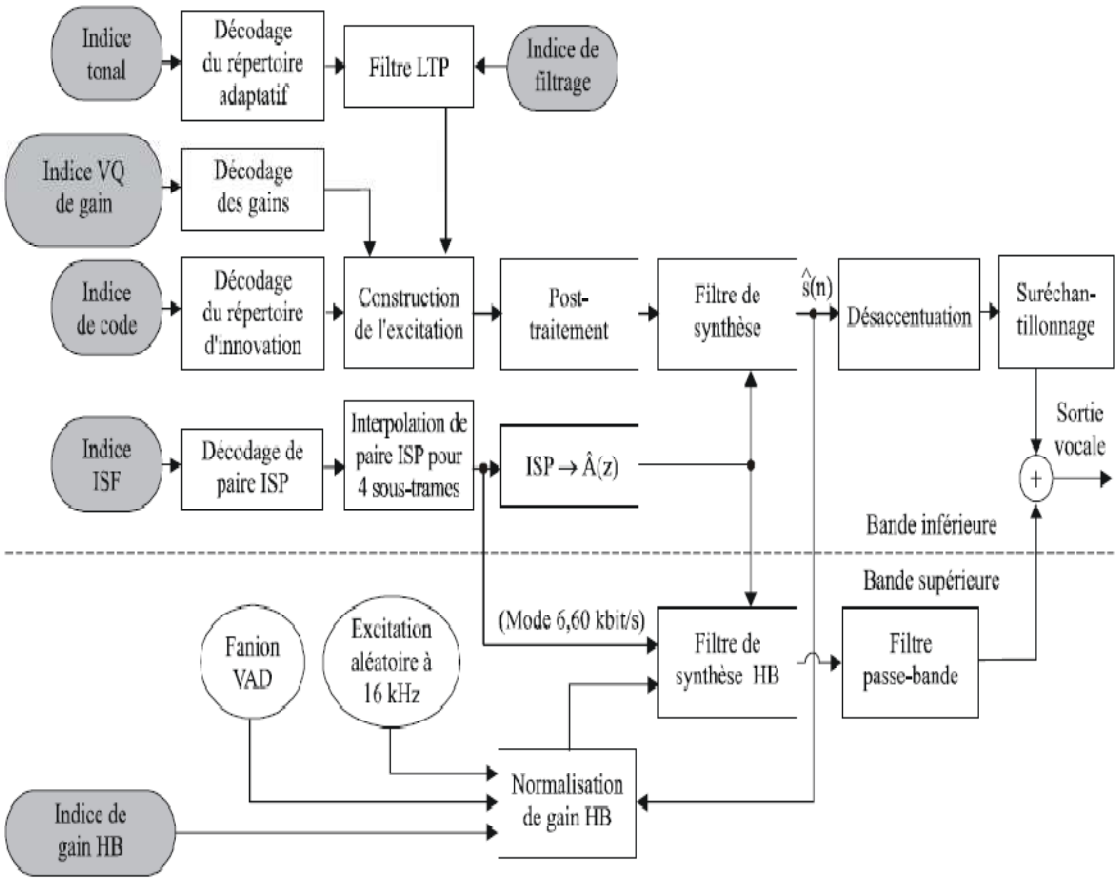


Figure.3.8 *Schéma fonctionnel de décodeur AMR_WB.*

III.4.3 La répartition des bits

La répartition des bits des modes du codec AMR-WB est présentée dans le Tableau 1. Dans chaque trame vocale de 20 ms, 132, 177, 253, 285, 317, 365, 397, 461 et 477 bits sont produits, correspondant à un débit de 6,60, 8,85, 12,65, 14,25, 15,85, 18,25, 19,85, 23,05 ou 23,85 kbit/s.

Tableau 3.1 *Répartition des bits de l'algorithme de codage AMR-WB pour trame de 20 ms.*

Mode	Paramètre	1ère sous-trame	2e sous-trame	3e sous-trame	4e sous-trame	Total par trame
23,85 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	88	88	88	88	352
	Gain de répertoire	7	7	7	7	28
	Energie HB	4	4	4	4	16
	Total					477
23,05 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	88	88	88	88	352
	Gains	7	7	7	7	28
	Total					461
19,85 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	72	72	72	72	288
	Gain de répertoire	7	7	7	7	28
	Total					397
18,25 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	64	64	64	64	256
	Gains	7	7	7	7	28
	Total					365
15,85 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	52	52	52	52	208
	Gains	7	7	7	7	28
	Total					317

Mode	Paramètre	1ère sous-trame	2e sous-trame	3e sous-trame	4e sous-trame	Total par trame
14,25 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	44	44	44	44	176
	Gains	7	7	7	7	28
	Total					285
12,65 kbit/s	Fanion VAD					1
	ISP					46
	Filtrage LTP	1	1	1	1	4
	Délai tonal	9	6	9	6	30
	Code algébrique	36	36	36	36	144
	Gains	7	7	7	7	28
	Total					253
8,85 kbit/s	Fanion VAD					1
	ISP					46
	Délai tonal	8	5	8	5	26
	Code algébrique	20	20	20	20	80
	Gains	6	6	6	6	24
	Total					177
6,60 kbit/s	Fanion VAD					1
	ISP					36
	Délai tonal	8	5	5	5	23
	Code algébrique	12	12	12	12	48
	Gains	6	6	6	6	24
	Total					132

III.4.4 DTX\ VAD\ CNG

Le codeur AMR_WB est également doté de mécanismes de transmission discontinue ("DTX") permettant d'optimiser la consommation de la ressource radio en ne transmettant pas de signal lors des périodes de non-activité vocale. Pour cela, à l'encodeur, un détecteur d'activité vocale (VAD pour "Voice Activity Detection") discrimine les instants de parole de ceux de silence ou de bruit. Au décodeur, un générateur de bruit de confort (CNG pour "Comfort Noise Generator") régénère un signal le plus proche possible du bruit original. Au décodeur, des dispositifs de correction de trames corrompues permettent de réduire l'effet des erreurs survenant sur le canal radio.

III.5 Simulation du canal de transmission

Dans un n'importe quel système de communication, la transmission d'un signal porteur de l'information vers un récepteur est assurée par un support de transmission appelé communément canal. Ce dernier est souvent le siège de plusieurs phénomènes perturbateurs qui peuvent provoquer une perte permanente de l'information utile dans le signal transmis. En effet, ces perturbations ont pour effet de créer une différence entre le message émis et celui reçu. En plus, elles sont de nature aléatoire, c'est-à-dire qu'il n'est pas possibles (ni pour la source, ni pour le destinataire) de prévoir de manière certaine leur effet. Pour une transmission fiable des données à travers un canal, il faut toujours essayer de minimiser la dégradation du signal transmis en développant un système de détection des perturbations éventuelles et à corriger les effets, généralement matérialisés sous forme d'erreurs sur les données reçues. En terme plus générale, le canal inclut le milieu de transmission (lien physique entre l'émetteur et le récepteur), le bruit (perturbation aléatoire issue du milieu, des équipements électroniques,…) et les interférences provenant des autres utilisateurs du milieu de transmission. Il existe plusieurs modèles de canaux bruités. Un intérêt particulier est porté aux modèles suivant : canal a bruit blanc gaussien additif, canal de Rayleigh et canal binaire symétrique.

III.5.1 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 (figure.3.9). La variable *n* est supposée gaussienne de moyenne nulle et de variance σ^2, sa densité de probabilité est définie par :

$$p(n) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(\frac{-n^2}{2\,\sigma^2}\right) \qquad (3.12)$$

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 :

$$y_k = x_k + n \qquad (3.13)$$

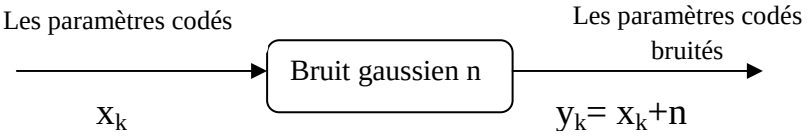


Figure.3.9 *Simulation du canal avec bruit gaussien.*

III.5.2 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 (figure.3.10). L'amplitude a_k du signal est souvent modélisée suivant la loi de Rayleigh, le signal est alors sous la forme :

Figure.3.10 *Simulation du canal avec canal Rayleigh.*

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

$$p(a_k) = \begin{cases} \frac{a_k}{\sigma^2} \exp\left(-\frac{a_k^2}{2\,\sigma^2}\right) & a_k \geq 0 \\ 0 & ailleurs \end{cases} \quad (3.14)$$

III.5.3 Canal binaire symétrique

Le canal binaire symétrique BSC (Binary Symetric Channel) est un modèle de canal simple à erreurs **[57]**. Il possède deux symboles d'entrée {0,1} et deux symboles de sortie {0,1}. Le canal est dit symétrique parce que la probabilité (p) de recevoir 1 lorsque 0 est émis est égale à la probabilité de recevoir 0 lorsque 1 est émis. Ces probabilités de transition qui caractérisent ce canal sont définies par :

$$P(0/1)=p(1/0)=p \quad et \quad p(0/0)=p(1/1)=1-p \quad (3.15)$$

Le fonctionnement du CBS est résumé sous forme de diagramme sur la figure.3.11. Chaque élément binaire à la sortie du canal ne dépendant que de l'élément binaire entrant correspondant, le canal est appelé sans mémoire.

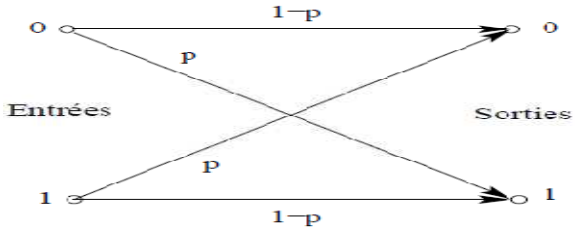


Figure.3.11 *Diagramme du canal binaire symétrique.*

III.6 Conclusion

Dans ce chapitre, nous avons détaillé les approches possibles (reconnaissance déportée, reconnaissance embarquée et reconnaissance répartie) pour faire la reconnaissance dans les systèmes de communications mobiles. L'extraction des paramètres dans le domaine compressé c.à.d. à partir des paramètres internes de codeur de parole a été proposée. Après nous avons décrit le codeur et le décodeur AMR_WB, qui est principalement destiné à traiter des signaux de parole d'une largeur de bande de 7 kHz. L'adaptation AMR-WB fonctionne à une multitude de débits binaires compris entre 6,6 kbit/s et 23,85 kbit/s. Le débit peut être modifié à toute limite de trame de 20 ms.

4

Reconnaissance automatique de locuteurs robuste pour les communications mobiles

IV.1 Introduction

Ce chapitre présente le contexte expérimental des évaluations de reconnaissance de locuteur, en mode indépendant du texte. En premier lieu, nous décrivons les bases de données utilisées. Ensuite, nous présentons le protocole d'évaluation en identification de locuteur, Les résultats sont organisés en trois catégories. La première catégorie regroupe les différents résultats que nous avons obtenus en utilisant différentes techniques de modélisation pour l'identification du locuteur. La deuxième catégorie regroupe les résultats obtenus à travers un ensemble de tests d'évaluation effectués sur les signaux transcodés AMR_WB, et la dernière catégorie regroupe ceux obtenus par simulation du canal et différents milieux bruités.

IV.2 Description des bases de données utilisées

La base de données de parole utilisée dans ce travail est une partie de la base de données **ARADIGIT** [**58**]. Elle est constituée de prononciations des 10 chiffres de la langue Arabe de zéro jusqu'à neuf prononcés par 60 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

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

- La base de données ARADIGIT est passée à travers le codec AMR_WB que nous avons simulé pour aboutir à la base de données transcodée (source uniquement) à savoir la base nommée ARADIGIT_AMR_{wb} (plusieurs bases de données pour les différents modes du codeur AMR_WB). La Figure.4.1 donne le processus de la création de cette base.

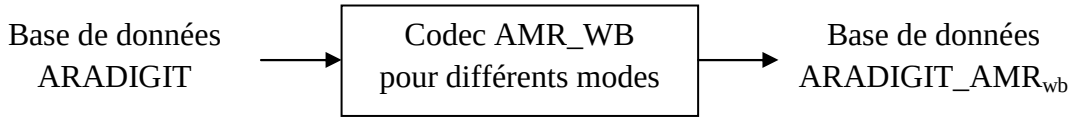


Figure.4.1 *Création de la base de données ARADGIT_AMR$_{wb}$.*

- La base de données transcodée AMR_WB (source + canal), le canal de communication est simulé par le bruit de type AWGN donnant la base de données nommée ARADIGIT_AMR$_{WB}$_AW (Figure.4.2).

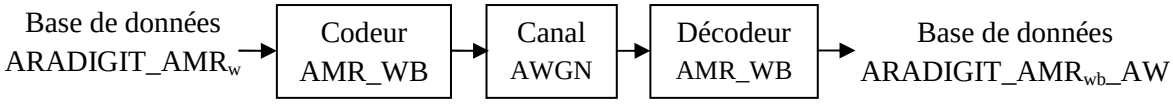


Figure.4.2 *Création de la base de données ARADGIT_AMR$_{wb}$_AW.*

- La base de données transcodée AMR_WB (source + canal), le canal de communication est simulé par le bruit de type Rayleigh donnant la base de données nommée ARADIGIT_AMR$_{WB}$_R (Figure.4.3).

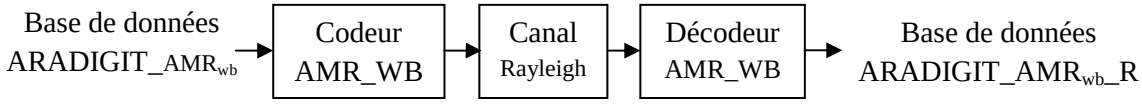


Figure.4.3 *Création de la base de données ARADGIT_AMR$_{wb}$_R.*

- La base de données transcodée AMR_WB (source + canal), le canal de communication est simulé par le canal binaire symétrique donnant la base de données nommée ARADIGIT_AMR$_{WB}$_BSC (Fig.4.4).

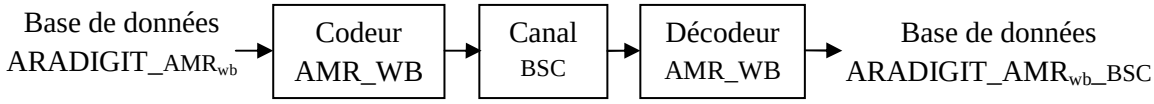


Figure.4.4 *Création de la base de données ARADGIT_AMR$_{wb}$_BSC.*

IV.3 Extraction des paramètres

1. MFCC (*Mel Frequency Cepstrum Coefficient*)

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). Douze (12) MFCC sont calculés pour chaque trame. Des coefficients différentiels du premier ordre (Δ) et du second ordre (ΔΔ) sont ensuite calculés pour former un vecteur de dimension 36 (12 MFCC + 12 Δ MFCC+ 12 ΔΔ MFCC).

2. ISF (*Immittance Spectral Frequency)*

A la réception du flux binaire quantifié au niveau de décodeur AMR_WB, une extraction sur une trame de 20ms de 16 ISF **[1]** est réalisée.

IV.4 Identification en mode indépendant du texte

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

- **La base d'apprentissage** : elle est composée de 180 fichiers vocaux prononcés par 60 locuteurs comprenant les deux sexes (31 hommes et 29 femmes). Chaque locuteur répète une suite de chiffres de 0 à 7 trois fois.
- **La base de test** : elle est composée de 180 fichiers vocaux prononcés par 60 locuteurs, comprenant les deux sexes (31 hommes, 29 femmes). Chaque locuteur répète les chiffres de 8 à 9 trois fois.

L'évaluation est faite en calculant le Taux d'Identification Correcte (TIC) défini par **[59]** :

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

IV.5 Expériences

IV.5.1 Partie 1 les techniques de modélisation pour l'identification du locuteur

1. Expérience 1 : Influence du nombre de GMM sur le taux d'identification

Les modèles GMM sont entraînés par une matrice de vecteurs MFCC de dimension 36, c.-à-d. 36 paramètres MFCC dans chaque vecteur de la matrice d'apprentissage. Les vecteurs MFCC sont extraits en appliquant un fenêtrage de Hamming où la longueur de la fenêtre est de 20ms et avec un facteur de chevauchement de 50%. Au premier lieu, on n'a pas utilisé l'algorithme VAD pour l'élimination du silence.

A travers les tests effectués, nous allons voir l'effet de l'ordre du modèle GMM sur la performance globale du système d'identification du locuteur. Les résultats obtenus sont présentés dans le tableau ci-dessous.

Tableau.4.1 *Taux d'identification en fonction du nombre de gaussiennes sans l'utilisation de VAD.*

Nombre de gaussiennes	Taux d'identification En %
2	*93.33*
4	*95.55*
8	***97.77***
16	***96.66***
32	***95.55***
64	*87.68*
128	*65.75*

2. *Expérience 2 : Détection d'Activité Vocal ou VAD*

Dans cette deuxième expérience nous avons procédé à l'élimination du silence en utilisant l'algorithme VAD (Figure.4.5) et voir son influence sur la performance du système d'identification du locuteur (Tableau 4.2). Enfin la Figure.4.6 fait une étude comparative du taux d'identification de locuteur en fonction du nombre de gaussiennes avec et sans VAD.

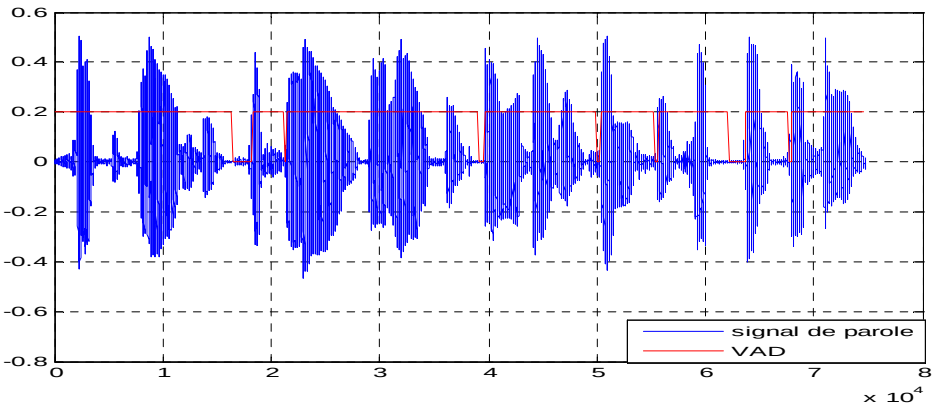


Figure.4.5 *l'utilisation de l'algorithme VAD sur un signal de parole, suite de chiffre [sifer, wahid, ithnani,…, sab'a] prononcé par une locutrice.*

Tableau.4.2 *Taux d'identification en fonction du nombre de gaussiennes avec l'utilisation de VAD.*

Nombre de gaussiennes	Taux d'identification En %
2	94.44
4	95.55
8	**97.77**
16	**98.88**
32	**97.22**
64	89.65

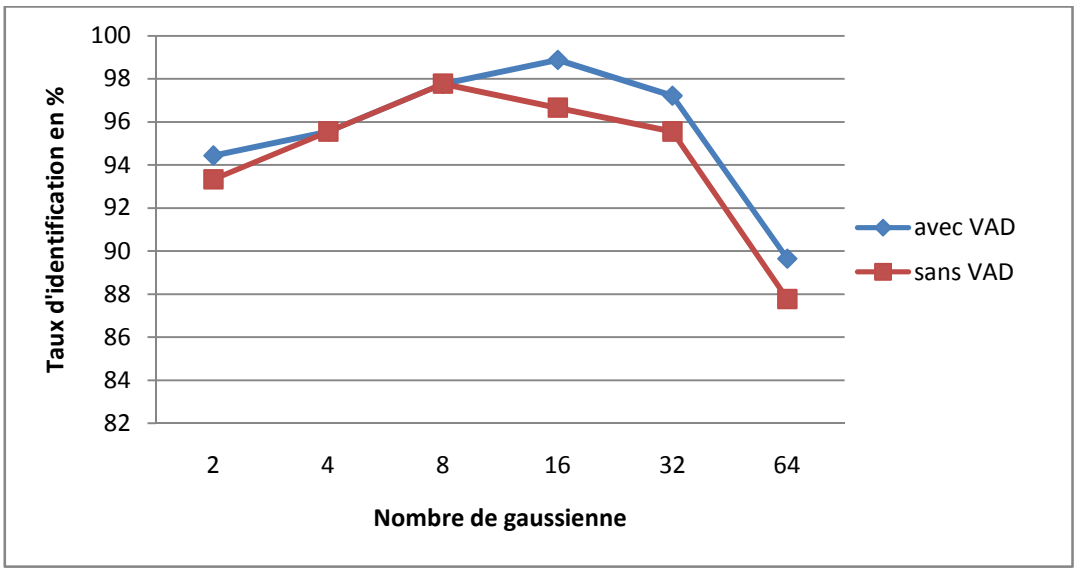


Figure.4.6 *Taux d'identification en fonction du nombre de gaussiennes sans et avec l'utilisation de VAD.*

3. *Expérience 3 : l'influence de la durée de l'enregistrement vocale*

Les résultats expérimentaux obtenus en fonction de la durée de l'enregistrement (la durée de la phrase) sont donnés par le Tableau 4.3. Les courbes de la Figure.4.7 permettent de faire une étude comparative du taux d'identification de locuteurs en fonction du nombre de gaussiennes et de la durée d'enregistrement.

Tableau.4.3 *Taux d'identification en fonction de la durée d'enregistrement vocale et du nombre de gaussiennes.*

modèle / nombre de locuteurs	*N=2*	*N=4*	*N=8*	*N=16*
Apprentissage :0.4s Test : 0.4s	**8.88%**	6.66%	6.11%	5%
Apprentissage : 1s Test : 1s	23.33%	**30%**	10.55%	8.33%
Apprentissage : 4s Test : 1s	94.44%	95.55%	97.77%	**98.88%**

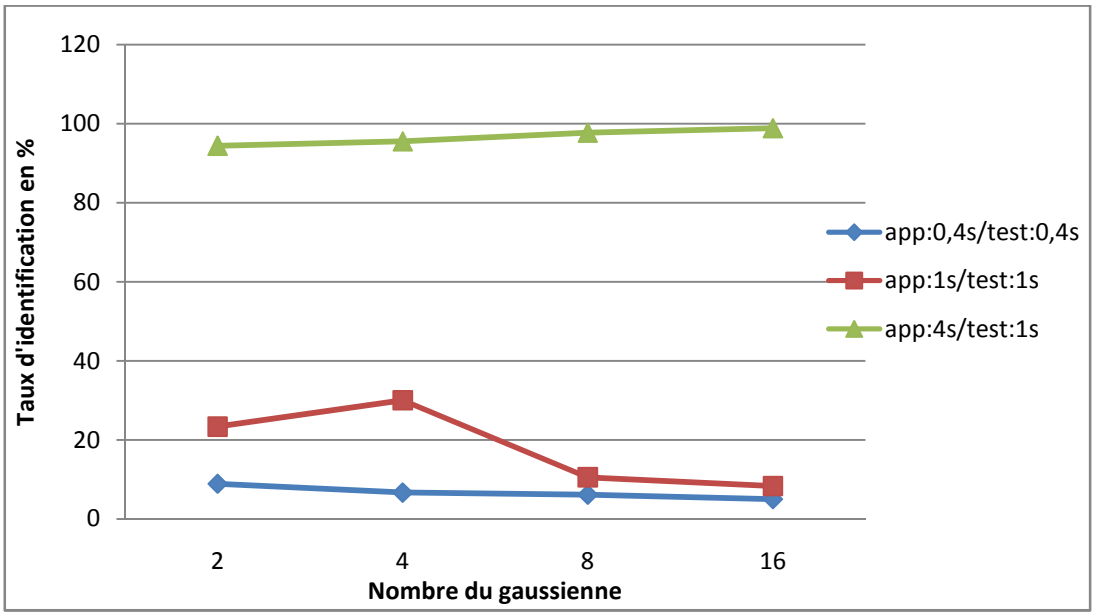


Figure.4.7 *Taux d'identification en fonction du nombre de gaussiennes et de la durée de l'enregistrement vocale.*

- **Interprétation**

Nous remarquons qu'en augmentant l'ordre du modèle GMM nous avons eu une amélioration de la performance du système d'identification du locuteur, il est au dessus de 95% pour un nombre de GMM de 8,16 et 32 (atteint 98.88% pour 16 gaussiennes). Cependant le taux d'identification est peu significatif au-delà de 32 gaussiennes et le temps de calcul augmente considérablement.

En effet, le choix de l'ordre de modèle dépend de la quantité de données et la durée de l'enregistrement vocal utilisée pour l'apprentissage et le test (voir les résultats de la troisième expérience). Choisir un ordre peu élevé va nuire à la précision du modèle. Choisir trop de composante engendra une charge de calcul plus importante.
En plus, les paramètres acoustiques (extraits à partir d'un enregistrement vocal, qui contient l'intervention de plusieurs locuteurs) seront dispersés dans plusieurs régions de l'espace de vecteurs acoustiques ; ici il est nécessaire d'augmenter l'ordre du modèle GMM pour couvrir ces différentes régions.
Dans notre cas, on dispose de peu de donnée (durée de l'enregistrement vocal utilisé lors de la phase d'apprentissage 4 sec et pour la phase de test 1 sec, et le nombre de locuteurs choisit est 60 locuteur), et donc le choix d'un modèle de 16 gaussiennes suffit pour représenter convenablement le locuteur.
Aussi les résultats d'expériences avec et sans VAD, confirment que l'utilisation de l'algorithme VAD augmente les performances du système de reconnaissance de locuteur de l'ordre de 2.22% de 96.66% à 98.88%.

4. Expérience 4 : Influence du nombre de locuteur et le type de classifieur utilisé sur le TIC

Nous allons présenter les résultats d'identification en utilisant différents méthodes de modélisation à savoir : les GMM, les supports vecteurs machines SVM avec l'utilisation de noyaux RBF et le réseau de neurones ANN, appliqués aux trois différents ensembles de locuteurs (10,30 et 60). Les résultats obtenus sont mis sous forme de tableaux et de graphes ci-dessous : Tableau 4.4 et la Figure. 4.8.

Tableau.4.4 *Taux d'identification du locuteur selon le nombre de locuteurs et le classifieur utilisé.*

modèle / nombre de locuteurs	GMM (N=16)	SVM	ANN
10	93.33%	89.12%	26.03%
30	97.77%	90.09%	20.76%
60	**98.88%**	92.79%	19.88%

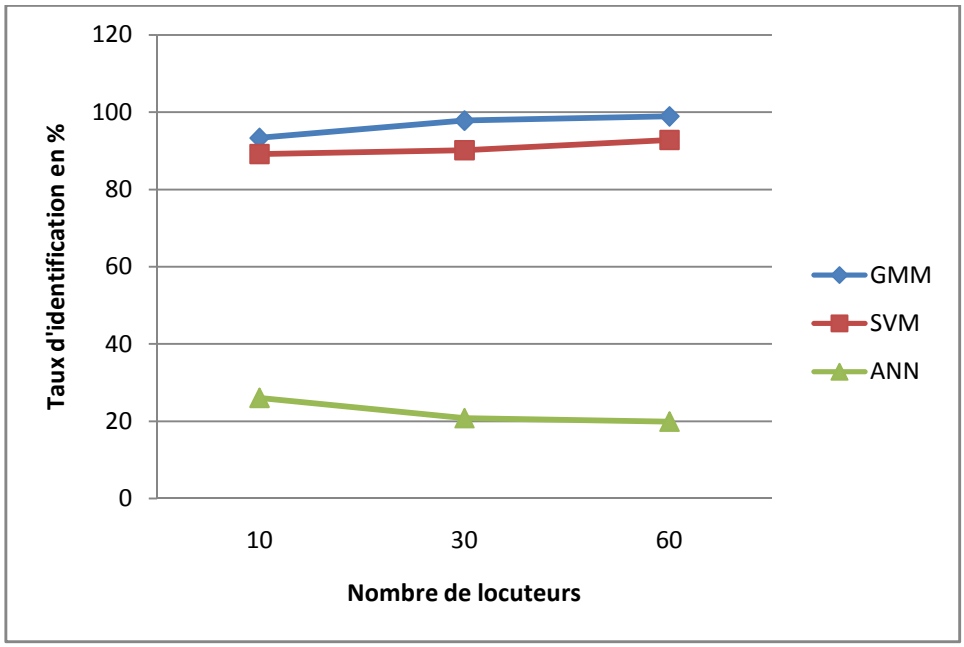


Figure.4.8 *Taux d'identification en (%) en fonction de nombre de locuteur et selon le classifieur utilisé.*

- **Interprétation**

Les performances des systèmes que nous avons mis en œuvre en identification varient d'un système à l'autre, En observant les résultats obtenus, nous pouvons constater que les performances du système sont nettement meilleures lorsqu'on utilise les GMM et qu'au fur et à mesure en augmentant le nombre du locuteur les performances du système s'améliorent (proche de 99%). Pour les SVM le meilleur taux est obtenu avec l'utilisation de 60 locuteurs (environ 93%).

Les réseaux de neurones ne sont pas adaptés dans ce cas, parce qu'ils ne donnent pas de bons résultats, le taux de reconnaissance égal presque à 20%.

IV.5.2 Partie 2 l'influence de la parole transcodée

1. Expérience 5: Comparaison entre les différentes méthodes d'extraction des paramètres

Nous évaluons, dans cette section, les différentes méthodes d'extraction des paramètres possibles pour la reconnaissance de locuteur disponibles aux différents niveaux de l'architecture d'un système de communication mobile, décrites par la Figure. 4.9. Nous détaillons les différentes méthodes utilisées :

- méthode 1 : L'extraction des MFCC est réalisée sur le signal non codé, à partir de la base de données ARADIGIT.
- méthode 2 : réception du flux binaire quantifié au niveau du récepteur. L'extraction des ISF est réalisée après déquantification.
- méthode 3 : décodage de la parole. Le signal est resynthétisé au niveau du récepteur, puis l'extraction des MFCC est réalisée (à partir de la base de données ARADIGIT_AMR$_{wb}$).

La méthode 1 est considérée comme la référence car le signal de parole n'est pas altéré par le codage. La méthode 2 est la solution optimisée que nous proposons, car elle présente de multiples avantages :
1. elle évite l'altération du signal de parole et la consommation supplémentaire de ressources liées à la resynthèse du signal,
2. le flux binaire peut être intercepté n'importe où dans le réseau.

Le codeur considéré dans les expériences présentées est le codeur de parole AMR_WB

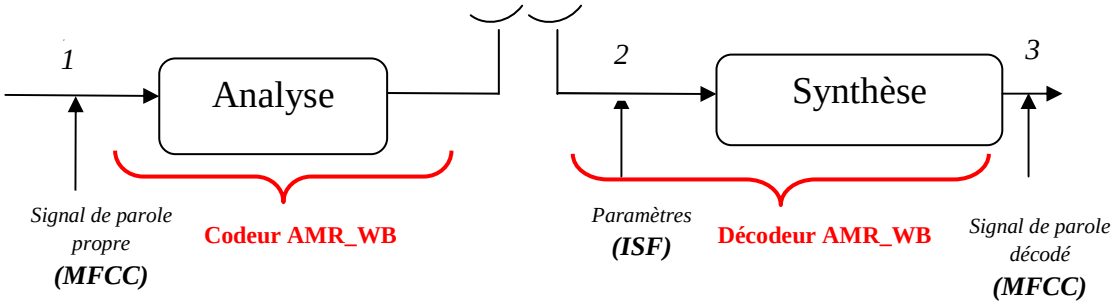


Figure.4.9 *Les méthodes d'extraction des paramètres possibles disponibles aux différents niveaux de l'architecture d'un système de communication mobile.*

- **Evaluation de la qualité de codeur AMR_WB**

Nous évaluons dans un premier temps la qualité de codeur AMR_WB, à travers une comparaison entre le signal original de la base de données ARADIGIT et celui du signal resynthétisé de la base de données ARADIGIT_AMRwb (la figure.4.9 représente les formes d'ondes des signaux de parole), et au travers les tests objectifs, en calculant le PESQ (le résultat est une note de qualité d'écoute).

❖ **PESQ (Perceptual Evaluation Speech Quality)**

C'est un modèle objectif, basé sur le modèle psycho-acoustique humain. Il effectue une mesure comparative entre deux informations : celle injectée à une extrémité et la même information recueillie à l'autre extrémité après avoir traversé le système étudié. Après traitement, le PESQ restitue ses résultats sur une échelle MOS (*Mean Opinion Score*), sous forme d'un scalaire compris entre **-0.5** et **4.5**.

Les résultats obtenus montrent que pour la parole la qualité de codeur AMR_WB s'améliore avec l'augmentation de débit, une note de 4.17 pour un mode de 23.85Kbit/s.

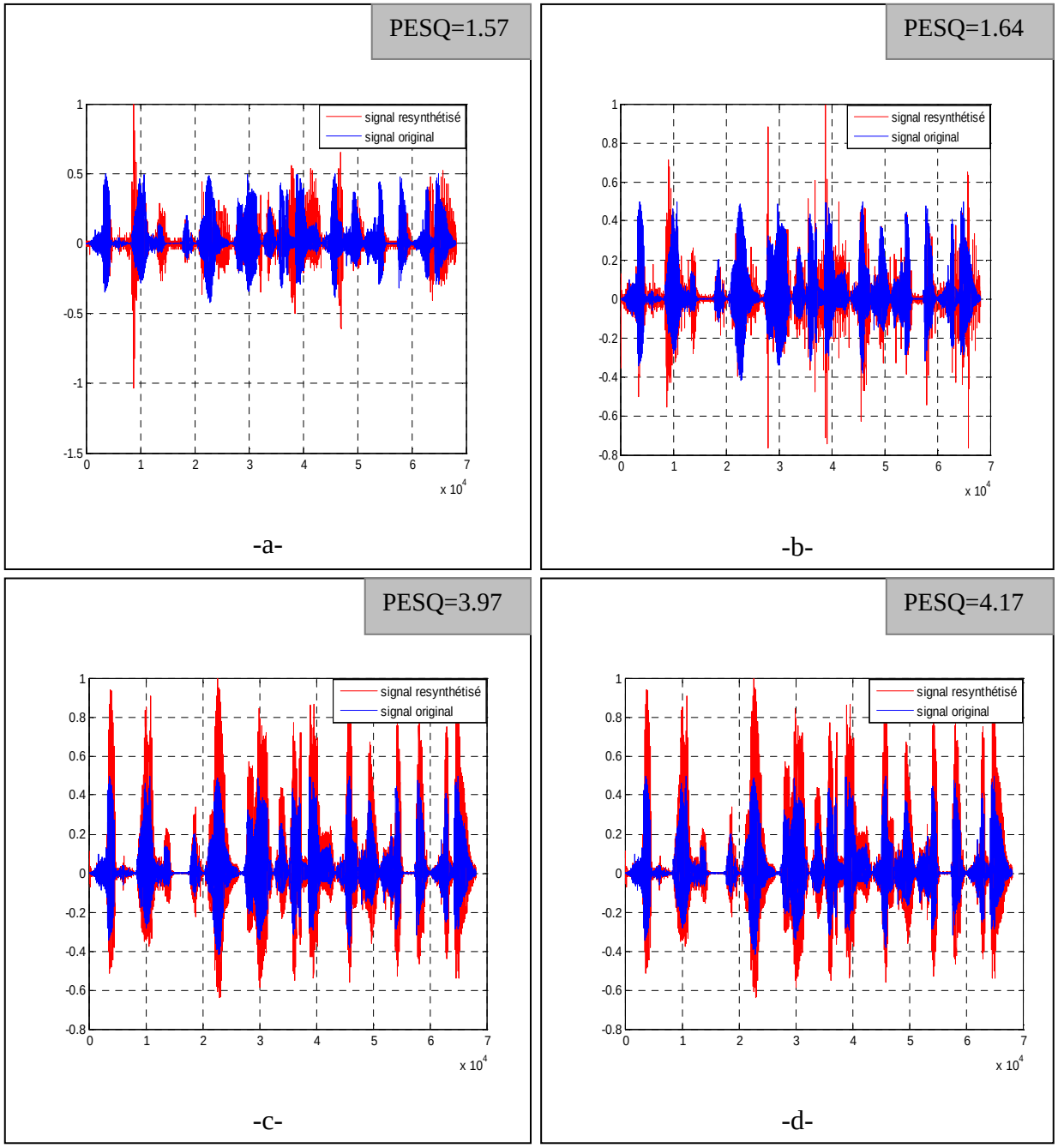


Figure.4.10 *Comparaison entre le signal de parole original et celui resynthétisé pour différents modes* (-a- mode 6.60Kbit/s, -b- mode 8.85 Kbit/s, -c- mode 12.65Kbit/s, -d- mode23.86 Kbit/s)

- **Comparaison entre les différentes méthodes d'extraction des paramètres**

Le Tableau 4.5 et les histogrammes de la Figure.4.11 résument les résultats globaux comparatifs entre les différentes bases de données présentées précédemment.

Tableau.4.5 *Taux d'identifications comparatifs entre les différentes bases de données et selon le classifieur utilisé.*

Modèle / Type de base	GMM (M=16)	SVM	ANN
ARADIGIT16KHz MFCC (1)	*98.88%*	*92.79%*	*19.88%*
Paramètres ISF (2)	*97.77%*	*76.18%*	*18.22%*
ARADIGIT_AMR$_{wb}$ MFCC (3)	*93.88%*	*89.99%*	*15.83%*

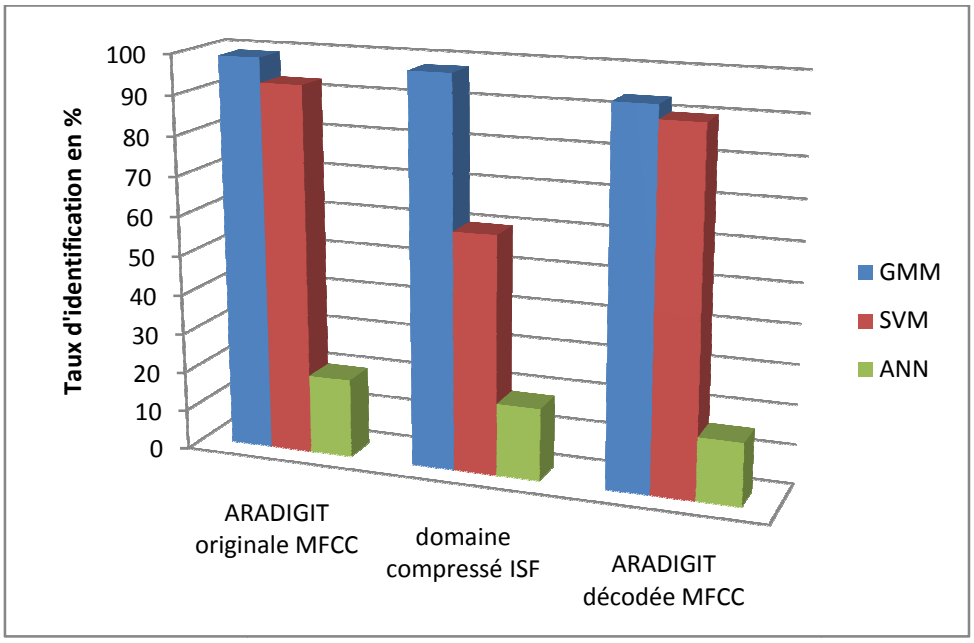


Figure.4.11 *Histogramme de taux d'identification de différentes bases de données utilisées.*

- **Interprétation**

L'analyse de ces résultats démontre que l'extraction des paramètres directement du signal original est une solution robuste (environ 99%), et que lorsque les paramètres ISF sont extraits du train binaire de codeur, le codeur AMR_WB obtient de meilleures performances que la solution utilisant le signal décodé (issu de ce codeur). Le gain s'élève à 4% en terme de TIC par rapport à la méthode 3, soit un TIC de 97.77% contre un TIC de 93.88%.

Par contre le codeur AMR_WB avec l'utilisation des SVMs montre un profil différent il obtient en effet légèrement de meilleurs résultats en travaillant sur le signal décodé qu'en utilisant le bitstream (train binaire), on interprète cela en reliant avec les paramètres utilisés, et donc on peut mettre l'hypothèse que les SVM sont plus adaptés au paramètres MFCC qu'au paramètres ISF.

2. *Expérience 6 : Comparaison entre différents modes de codeur AMR_WB*

Tableau.4.6 *Taux d'identification des différents modèles utilisés en fonction des modes de codeur AMR_WB*

Le type / mode	GMM		SVM	
	Paramètres ISF	Signal décodé MFCC	Paramètres ISF	Signal décodé MFCC
6.60Kbit/s	*85.55%*	*39.44%*	*69.18%*	*68.58%*
8.85Kbit/s	*90.00%*	*81.11%*	*73.96%*	*80.65%*
12.65Kbit/s	*97.77%*	*93.88%*	*76.18%*	*89.99%*
23.85Kbit/s	***98.88%***	*95.55%*	*78.18 %*	*92.32%*

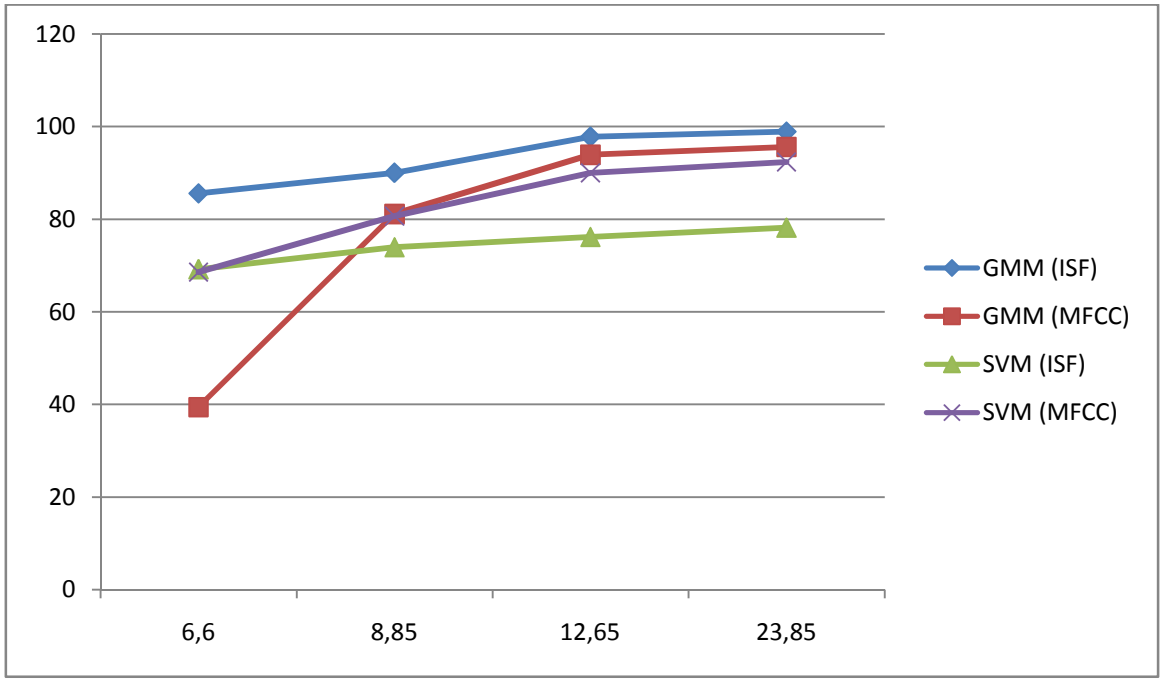


Figure.4.12 *Taux d'identification des différents modèles en fonction du mode utilisé.*

- ***Interprétation***

D'après les résultats présentés dans le tableau, on constate qu'au fur et à mesure que le débit augmente, le taux de reconnaissance augmente ; que ce soit avec l'utilisation du signal codé ou décodé (débit faible donc plus d'impact sur la reconnaissance). L'identification est meilleure en utilisant les GMMs, avec un débit de 23.85Kbit/s on atteint un taux de reconnaissance proche de 99% (résultat identique à celui trouver avec l'utilisation des MFCC extraites du signal original).

D'autre part, à un faible débit (6.60Kbit/s), on remarque que l'utilisation de signal resynthétisé dégrade les performances du système RAL, un taux peu significatif d'ordre 39.44%. Et qu'avec l'utilisation des ISF, on obtient un taux de 85.55%. D'après cela, on peut conclure que l'utilisation des paramètres issus de flux binaire est une solution robuste.

IV.5.3 Partie 3 Etude de la robustesse

IV.5.3.1 Effet du canal de transmission

1. Expérience 7 : 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 des niveaux Rapport Signal sur Bruit de 5,10 et 15dB.

Les résultats de cette expérience sont comparés à ceux trouvés avec l'utilisation de la base de données transcodée ARADIGIT_AMR$_{WB}$. Le Tableau 4.7 et la Figure. 4.13 résume ces résultats comparatifs.

Tableau.4.7 *Taux d'identifications comparatifs entre bases ARADIGIT_AMR$_{wb}$ et ARADIGIT_AMR$_{wb}$_AW à différents SNR (signal resynthétisé).*

	Canal idéal	Canal bruité AWGN		
SNR (dB)	-	*15*	*10*	*5*
GMM	*93.88%*	*84.44%*	*42.77%*	*8.88%*
SVM	*89.90%*	*76.07%*	*70.34%*	*59.73%*

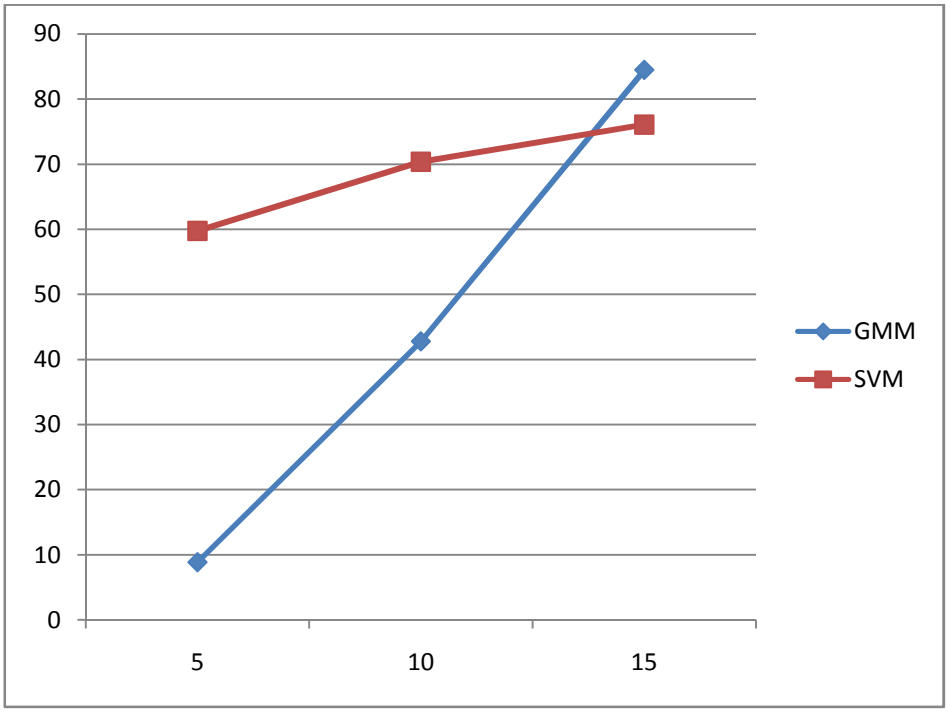


Figure.4.13 *Taux d'identifications comparatifs de la base ARADIGIT_AMR$_{wb}$_AW pour différents SNR (mode 12.65kbit/s).*

2. Expérience 8 : Canal de type Rayleigh

Le canal de transmission est simulé par un canal de type Rayleigh décrit, avec des niveaux Rapport Signal sur Bruit de 5,10 et 15dB.

Le Tableau 4.8 et la Figure.4.14 résume les résultats globaux comparatifs dans le cas d'un affaiblissement avec canal de Rayleigh.

Tableau.4.8 *Taux d'identifications comparatifs pour la base de données ARADIGIT_AMR$_{wb}$_R à différents SNR.*

	Canal idéal	Canal bruité Rayleigh		
SNR (dB)	-	*15*	*10*	*5*
GMM	*93.88%*	*21.11%*	*13.33%*	*3.88%*
SVM	*89.90%*	*62.53%*	*57.16%*	*54.56%*

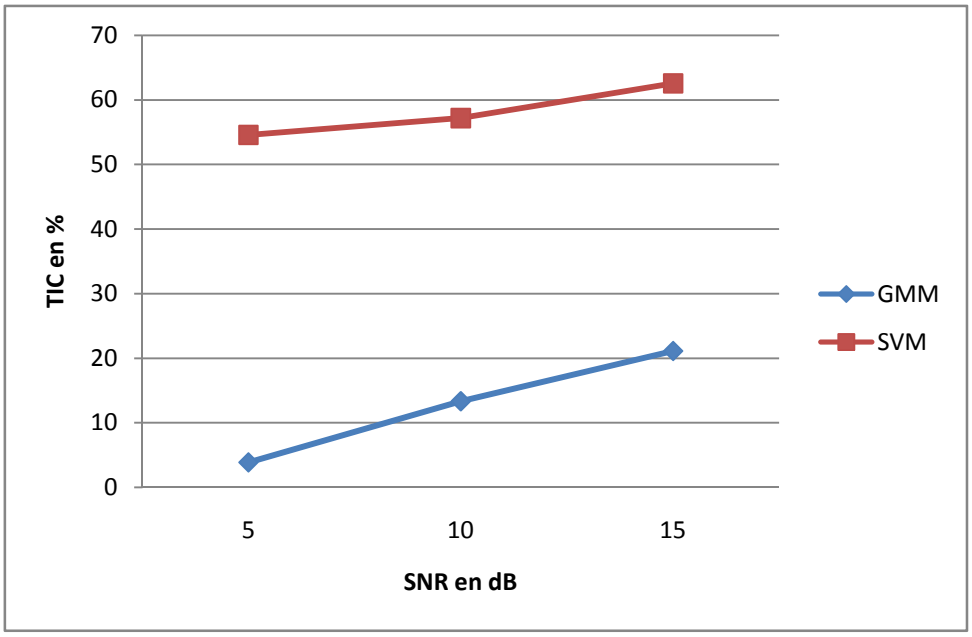


Figure.4.14 *Taux d'identifications comparatifs de la base ARADIGIT_AMRwb_R pour différents SNR (mode 12.65kbit/s).*

3. Expérience 9 : Canal BSC

Le canal de transmission est simulé par un canal binaire symétrique, avec des probabilités de 0.05, 0.1 et 0.3.

Les résultats de cette expérience sont comparés à ceux trouvés avec l'utilisation de la base de données transcodée ARADIGIT_AMR$_{WB}$, la base de données ARADIGIT_AMR$_{WB}$_AW et la base de données ARADIGIT_AMR$_{WB}$_R. Le Tableau 4.7 et les Figure.4.15 et 4.16 résument les résultats globaux comparatifs pour une identification à travers un canal BSC.

Tableau 4.9 Les résultats globaux comparatifs en %.

	Canal idéal	*Canal AWGN*			*Canal Rayleigh*			*Canal BSC*		
		SNR=15dB	SNR=10dB	SNR=5dB	SNR=15dB	SNR=10dB	SNR=5dB	P=0.05	P=0.1	P=0.3
GMM	*93.88*	*84.44*	*42.77*	*8.88*	*21.11*	*13.33*	*3.88*	*52.77*	*27.22*	*3.33*
SVM	*89.90*	*76.07*	*70.34*	*59.73*	*62.53*	*57.16*	*54.56*	*70.84*	*66.86*	*56.06*

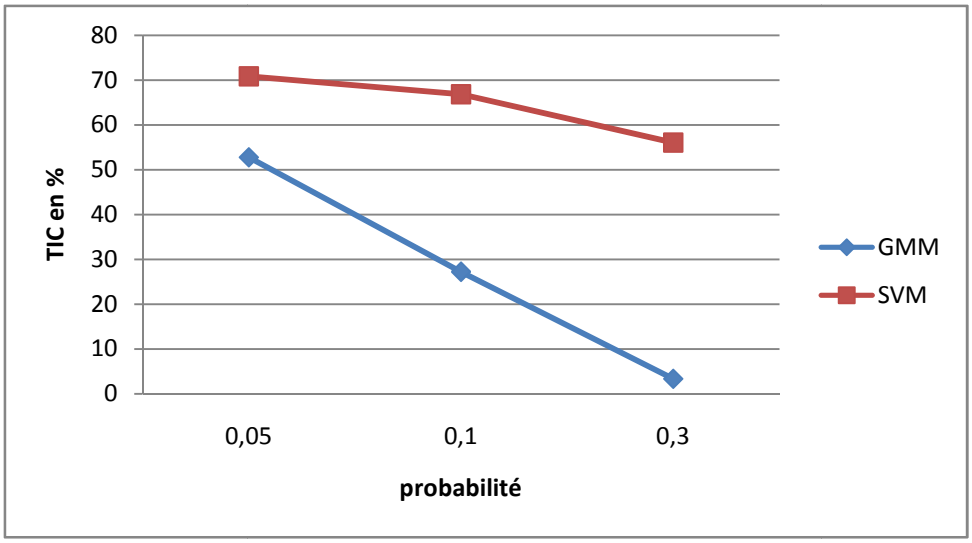


Figure.4.15 *Taux d'identifications comparatifs de la base ARADIGIT_AMR$_{wb}$_BSC pour différentes probabilités (mode 12.65kbit/s).*

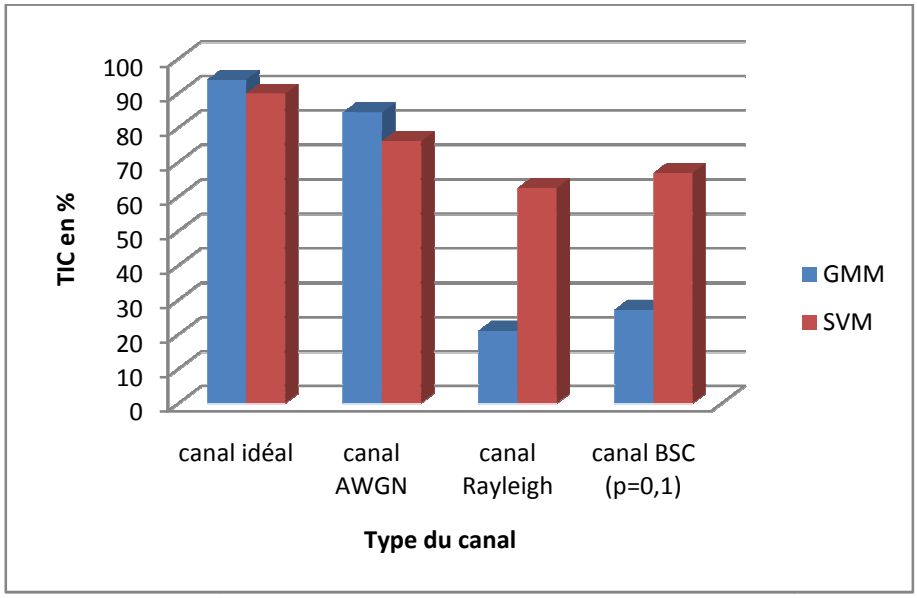


Figure.4.16 *Histogramme de Taux d'identifications comparatifs entre les bases ARADIGIT_AMR$_{wb}$, ARADIGIT_AMR$_{wb}$_AW, ARADIGIT_AMR$_{wb}$ _R et ARADIGIT_AMR$_{wb}$_BSC (mode=12.65kb/s, SNR=15dB et p=0.1).*

4. Expérience 10: *L'influence du type de canal sur les paramètres ISF*

Pour un deuxième cas (concernant la reconnaissance dans le domaine compressé), nous allons voir l'effet du canal de transmission (AWGN, Rayleigh et BSC) sur les paramètres ISF. Les résultats obtenus (mode 12.65kbit/s) sont comparés avec ceux trouvés dans le cas d'un canal idéal. Le Tableau 4.10 et les Figure.4.17 résument les résultats globaux comparatifs pour une identification à travers les différents canaux.

Tableau.4.10 Les résultats comparatifs.

Type de canal / Le modèle	canal idéal	Canal AWGN 15dB	canal Rayleigh 15dB	Canal BSC P=0.1
GMM	*97.77%*	*93.33%*	*41.66%*	*47.77%*
SVM	*76.18%*	*75.22%*	*64.68%*	*67.76%*

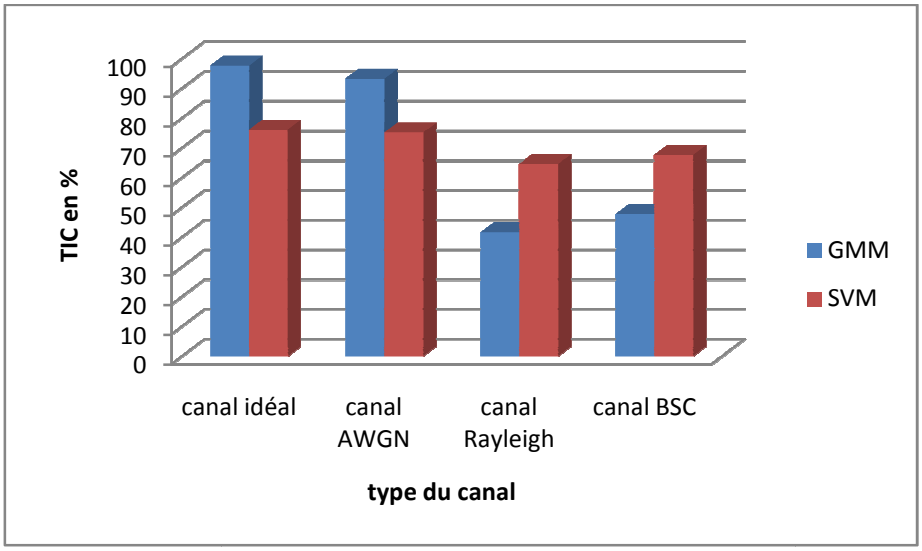


Figure.4.17 *Histogramme de Taux d'identifications comparatifs avec l'utilisation des paramètres ISF et selon le type du canal utilisé.*

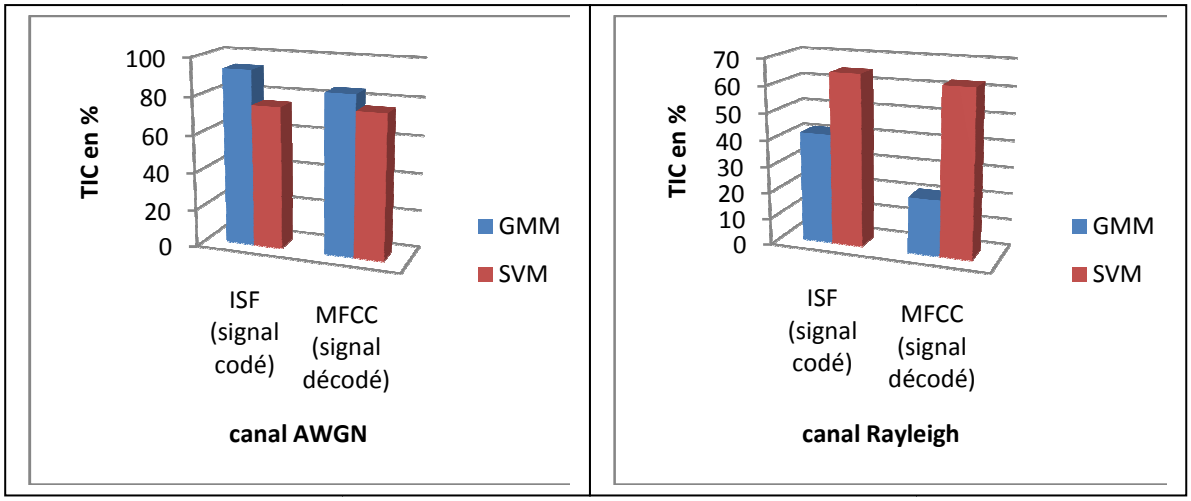


Figure.4.18 : *comparaisons entre l'utilisation des paramètres ISF du signal codé et les paramètres MFCC du signal décodé et selon le type du canal utilisé (SNR=15dB).*

- **Interprétation**

Concernant la modélisation par les GMM, nous avons remarqué une dégradation des performances du système RAL, cette dégradation est très remarquable en simulant le canal par un bruit du Rayleigh. Les GMM ne sont pas adapter à la reconnaissance dans ce cas.

D'autre part (toujours avec la modélisation par les GMM) nous constatons que l'utilisation des ISF permet d'obtenir des performances significativement meilleures que celles obtenues avec l'utilisation des MFCC issus du signal resynthétisé (pour un canal de Rayleigh un taux de 41.66% par les ISF contre un taux de 21.11% par les MFCC).

Les SVM montrent un profil différent, par exemple pour 15dB que se soit dans le canal AWGN ou Rayleigh, le taux de reconnaissance est au-dessus de 60%, Ils commettent plus que 30% moins d'erreurs que les GMM. D'autre part, la différence entre l'utilisation des ISF ou des MFCC n'est pas très importante dans ce cas.

IV.5.3.2 bruit de l'environnement

1. ***Expérience 11 : Identification du locuteur dans un milieu bruité (base de données originale)***

Dans cette partie, nous bruitons notre base de test ARADIGITS qui contient 60 locuteurs (suivant la procédure de bruitage décrite dans la Fig. 4.16) par des bruits issus de la base NOISEX'92 NATO (Varga) ; et à différents niveaux Rapport Signal sur Bruit (RSB : 0dB, 5dB, 10dB et 20dB) suivant la Figure.4.19.

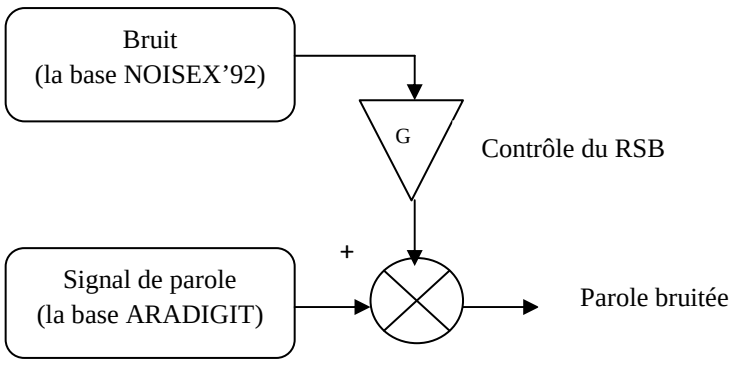


Figure.4.19 *la procédure de bruitage.*

Les étapes de bruitage de la base de données sont :

- Addition des signaux de bruit et de parole au niveau temporel.
- Un contrôle du gain pour obtenir le RSB considéré pour l'expérience.

Les résultats expérimentaux obtenus sont résumés dans le tableau 4.11 et les Figure.4.20. ci-dessous

Tableau.4.11 *Comparaison des taux d'identification des différents modèles utilisés dans un milieu bruité (babble speech) et pour différents SNR.*

SNR / modèle	*20dB*	*10dB*	*5dB*	*0dB*
GMM	*96.11%*	*69.44%*	*38.33%*	*14.44%*
SVM	*90.97%*	***80.60%***	***71.66%***	***63.17%***

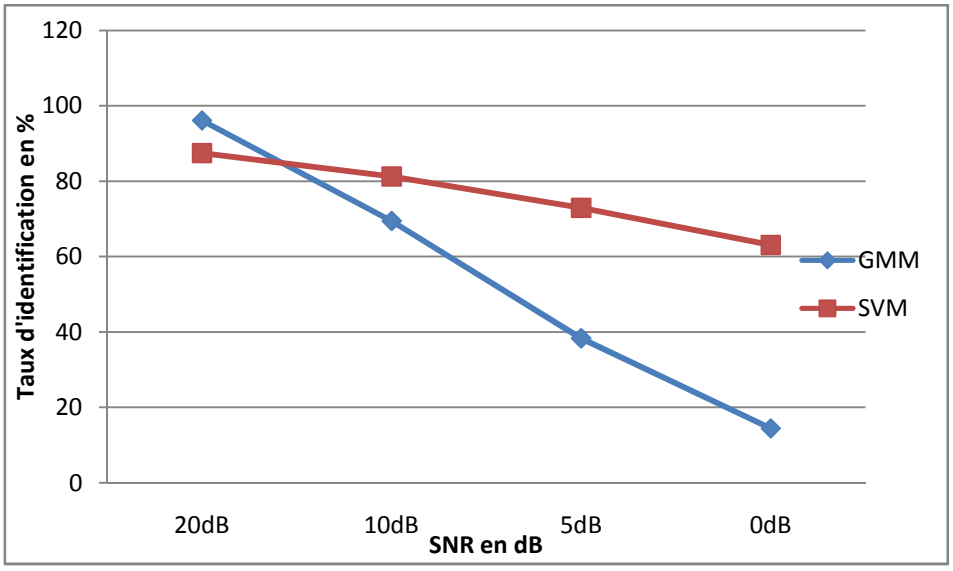


Fig.4.20 *Histogramme de taux d'identification des différents modèles utilisés dans un milieu bruité (babble speech) et pour différents SNR.*

- **Interprétation**

Un facteur important de dégradation du signal de parole réside dans le bruit additif de l'environnement extérieur. Des expériences menées sur la base ARADIGIT, démontrent que dans des conditions d'enregistrements très favorables, sans aucun bruit d'ambiance, le taux d'identification correcte de notre système de RAL de référence en utilisant les GMMs est de 98.88%. Le même système, utilisé sur des signaux de test bruités (RSB de 5dB) à l'aide d'un bruit de babble, présente un TIC de 38.33 %.

Cette expérience démontre la très forte corrélation entre les performances d'un système de RAL et le RSB des signaux de test.

D'autre part et d'après les résultats présentés dans le tableau, on constate que les SVMs sont plus robustes au bruit (Exp : un taux d'identification égale à 63.17% pour un SNR de 0db) que d'autre système (GMM). Ce qui implique que les SVMs sont plus adaptés à la reconnaissance du locuteur dans un milieu bruité.

2. *Expérience 12 : Identification du locuteur dans un milieu bruité (base de données transcodée ARADIGIT_AMR$_{wb}$)*

Dans cette expérience, Nous évaluons la robustesse aux bruits des paramètres, calculés dans le domaine compressé (ISF) et après la resynthèse du signal de parole. Les résultats obtenus pour un bruit de foule de personne (babble speech) et un SNR de 10dB, et avec le choix de mode 12.65Kbit/s sont comparés à ceux trouvés précédemment avec l'utilisation du signal original. Les résultats expérimentaux obtenus sont résumés dans le tableau 4.12 et les Figure.4.21 ci-après.

Tableau.4.12 *Comparaison des taux d'identification des différentes bases de données utilisées dans un milieu bruité (babble speech, SNR=10dB).*

modèle / Base de données bruitée	GMM	SVM
ARADIGIT (MFCC)	*69.44%*	*80.60%*
ARADIGIT_AMR$_{wb}$ (MFCC)	*82.77%*	***83.97%***
Paramètres ISF	*66.66%*	*73.18 %*

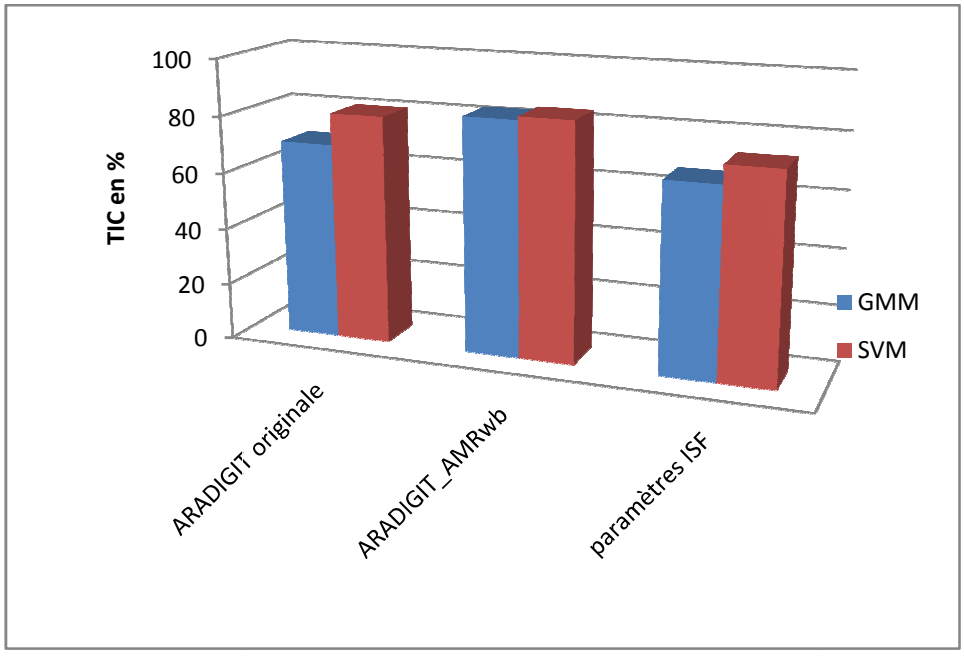


Fig.4.21 *Histogramme de taux d'identification des différentes bases de données utilisées dans un milieu bruité (babble speech, SNR=10dB).*

- **Interprétation**

Ces résultats montrent un profile différents comparés à ceux trouvés pour un milieu propre. Le taux d'identification obtenu pour un signal bruité resynthétisé est meilleur que celui obtenu avec un signal original bruité que ce soit avec l'utilisation des GMMs ou les SVMs. Le décodeur AMR_WB réduit significativement l'influence des bruits ambiants sur les performances de RAL. Et cela parce que l'AMR_WB au niveau de décodeur dispose des correcteurs de trames corrompues qui permettent de réduire l'effet des erreurs et de bruits.

IV.6 Conclusion

Dans ce chapitre, nous avons décrit les expériences réalisées, ainsi que les résultats obtenus au cours de notre travail. A travers ces expériences, nous pouvons dire que le modèle GMM est puissant, et le bon choix de l'ordre du modèle GMM est très important. Ainsi, la performance du système d'identification du locuteur est fortement dépendante de la durée de l'enregistrement vocal. En effet, les durées courtes posent un problème pour modéliser et adapter correctement le modèle de locuteur à l'environnement.

D'autre part, nous avons évalué les performances de différentes méthodes d'extraction des paramètres, et démontré que l'extraction des paramètres dans le domaine compressé est une solution robuste (surtout si on parle du temps d'exécution).

Nous avons montré aussi que les SVM peuvent être très efficaces et intéressantes surtout pour des tâches de reconnaissance dans un milieu bruité.

Conclusion et perspectives

Pour mettre en place une authentification sure, pendant la communication, la voix seule est disponible. Pour répondre à ce besoin, nous avons traité dans ce mémoire de l'utilisation de la reconnaissance automatique de locuteur dans les systèmes de communications mobiles.

La recherche dans le domaine de la reconnaissance automatique du locuteur devrait donc viser le développement d'applications qui permettrait une meilleure extraction des paramètres caractéristiques de la voix et leur modélisation, tout en réduisant les influences des systèmes de codage et de transmission des réseaux téléphoniques.

L'objectif principal du sujet abordé dans ce présent mémoire est d'étudier et évaluer les différents types de modélisation (GMM, SVM), et les différentes méthodes d'extraction des paramètres à savoir les paramètres MFFC issus du signal de parole originale (cette méthode est considérée comme la référence car le signal de parole n'est pas altéré par le codage), les paramètres ISF issus du codeur AMR_WB, et enfin celle issus du signal resynthétisé.

A travers les expériences que nous avons effectuées sur la base de données ARADIGIT, nous avons constaté que la performance d'un système d'identification du locuteur est fortement dépendante des données utilisées lors de la phase d'apprentissage et la phase des tests. En effet, la durée des fichiers joue un rôle très important dans l'évaluation de ce type de système. Plus la durée est grande, plus les performances du système sont meilleures.

D'autre part, nous pouvons dire que le modèle GMM (en environnement calme) peut représenter des distributions aléatoires d'une manière fidèle. Mais le bon choix de l'ordre de GMM est très important. En effet comme déjà mentionné plus haut le choix du 16 gaussiennes est largement suffisant pour représenter la distribution des vecteurs acoustiques dans notre étude.

Nous avons vu aussi que l'utilisation des paramètres issus du signal codé permet de rapprocher les résultats de la référence (signal originale).

L'autre point important qui a motivé notre étude, concerne l'étude de la robustesse des systèmes RAL en présence du bruit (bruit de l'environnement, bruit du canal). Les résultats obtenus montrent que le canal de transmission du signal de parole et le bruit de l'environnement influencent sur la performance du système et que dans cette situation, les SVM sont plus robuste que les GMM. Et enfin, l'utilisation du signal resynthétisé donne des meilleurs résultats que l'utilisation des paramètres ISF issus du codeur, le décodeur ajoute des informations utiles, plus des paramètres donc augmentation de la robustesse (la redondance du signal de parole).

En conclusion, la reconnaissance de locuteur en utilisant des paramètres internes de codeur est une solution robuste, Ceci permet notamment d'éviter les distorsions dues à la resynthèse du signal et la consommation de ressources supplémentaires, mais aussi de minimiser le temps d'exécution.

Perspectives

La reconnaissance automatique du locuteur est en plein développement, et elle comporte plusieurs axes de recherche qui peuvent être explorés dans des futurs travaux. Parmi ces axes, nous pouvons mentionner :

- ✓ Pour éviter au maximum les perturbations, une solution simple est d'extraire les paramètres au plus près de la source (signal de parole originale). Et donc l'utilisation du standard Aurora pour une architecture distribuée, entre le terminal mobile et un serveur d'applications distant, permet notamment de s'affranchir des problèmes dus au codage de la parole.

- ✓ L'utilisation d'une approche hybride de type SVM/GMM pour allier le pouvoir discriminant des SVM au pouvoir de modélisation des GMM.

- ✓ L'utilisation de la méthode MAP (Maximum à posteriori) pour estimer un modèle GMM en adaptant un modèle UBM. Cette méthode est très intéressante dans le cas où nous ne disposons pas suffisamment de données lors de la phase d'apprentissage.

- ✓ Au niveau de la paramétrisation du signal, une solution est d'extraire les paramètres aux différents endroits du codeur par exemple avant/après quantification, interpolation, etc.

- ✓ Les systèmes RAL ont souffert jusqu'ici de leur faible robustesse lorsqu'ils sont utilisés en situation réelle, en particulier dans des environnements perturbés. Pour améliorer la performance du système de reconnaissance du locuteur dans un environnement mobile, il faut prévoir des méthodes de réduction du bruit voir le supprimer comme celles basées sur le filtrage de Wiener, technique de masquage de bruit, la notion de missing feature theory qui consiste à ne pas utiliser les trames considérées comme trop bruitées...etc.

Bibliographie

[1] ITU-T G.722.2. Recommandation, " Codage vocal à large bande à 16 kbit/s environ par codage adaptatif multi débit à large bande (AMR-WB) ", 2003.

[2] D.A.Reynolds, "Speaker Identification and Verification using Gaussian Mixture Models", ESCA Workshop on Automatic Speaker Recognition Identification and Verification, pp.27- 30, 1994.

[3] T. Ganchevand, D. K Tasoulis, M. N. Vrahatis et N. Faktoakis, "Locally recurrent probabilistic neural network for text-independant speaker verification", Dans EUROSPEECH, 2003.

[4] V. Wan, "Speaker Verification using Support Vector Machines", Thèse de Doctorat. University of Sheffield, United Kingdom, Juin 2003.

[5] J. G. Rodríguez, J. O. García, J. Luis and S. Bote, "Forensic Identification Reporting using Automatic Biometric Systems", chapter 7 from the book "Biometric Solutions for Authentication in an e-World" Editor: David D.Zhang Kluwer Academic Publishers, pp 169-185, 2002.

[6] P.Rose, F.Clermont, "A comparison of two acoustic methods for forensic speaker identification", Acoustics Australia, 2001.

[7] J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition", the journal of the Acoustical Society of America, 1972.

[8] 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.

[9] C. Fredouille, "Reconnaissance du locuteur et approche statistique : Information dynamiques et normalisation bayésienne des vraisemblances", Thèse de l'Université d'Avignon, 2000,

[10] H. Harb, "Classification du signal sonore en vue d'une indexation par le contenu des documents multimédias", Thèse de Doctorat, LIRIS, Ecole Centrale de Lyon, Décembre 2003,

[11] B.S. ATAL, "Automatic speaker recognition based on pitch contours", The journal of the Acoustical Society of America, n°.52, pp 1687-1697, 1972.

[12] L. Besacier, J.F. Bonaster, "Frame Pruning for Speaker Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing* 12-15 May 1998. Seattle (USA).

[13] P. Pravinkumar, B.M Wasfy, "Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review", *Proc. IEEE*, 2001.

[14] R. Blouet, "Approche probabiliste par arbres de décision pour la vérification automatique du locuteur sur architectures embarquées", Université de Rennes 1, 2002.

[15] A.M. Ariyaeeinia, P. Sivakumaran, M. Pawlewski, M.J. Loomes, “Dynamic weighting of the Distortion Sequence in Text-Independent Speaker Verification”, Eurospeech’99, pp.967-970, 1999.

[16] I. Booth, M. Barlow, B. Waston, “Enhancements to DTW and VQ decision algorithms for speaker recognition”, Speech Communication, 13(3-4):427-433, 1993.

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

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

[19] M. Savic, S.K. Gupta, “Variable prameter speaker verification system based on hidden marcov modeling”, Dans *International Conference on Acoustics, Speech and Signal Processing* (ICASSP), volume 1, pages 281-284, 1990.

[20] E.L. Rissanen, J.J. Webb, “Speaker identification experiments using HMMs”, Dans *International Conference on Acoustics, Speech and Signal Processing* (ICASSP), volume 2, pages 387-390, 1993.

[21] D.A. Reynolds, R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans on speech and audio processing*, Vol 3, No 1, Jan 1995.

[22] P.Ding, Y.Lui, B.Xu, “Factor Analyzed Gaussian Mixture Models for Speaker Identification”, ICSLP’02, pp.1341-1344, 2002.

[23] T. Hasan, J. H. L. Hansen, “universal background model training in speaker verification,” in *Proc. IEEE* August 13, pp. 4494–4497, 2010.

[24] Y.Bennany, P.Gallinari, ‘Neural Networks for Discrimination and Modelization of Speakers”, Speech Communications, vol. 17, pp. 159-175, 1995.

[25] 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.

[26] C. Burges, “*A* tutorial on support vector machines for pattern recognition, volume2, Data Mining and Knowledge Discovery, 1998.

[27] J. P. Egan, “Signal Detection Theory and ROC analysis”, Academic Press, New York, 1975.

[28] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, M. A. Przybocki, “The DET Curve in Assessment of Detection Task Performance”, Dans les actes de European Conference on Speech Communication and Technology (Eurospeech), 1997.

[29] A. Morris, D.Wu, J. Koreman, “GMM based clustering and speaker separability in the Timit speech database”, “*Special Section/Issue on Corpus-Based Speech Technologies IEICE TRANS*, VOL. E85-A/B/C/D, No. 1 pp 1-13, March 2005.

[30] 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, 1977.

[31] 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.

[32] Y. Linde, A. Buzo, R. Gray, “An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, vol. 28, pp. 84-94, 1980.

[33] N. Scheffer, “Structuration de l’espace acoustique par le modèle générique pour la vérification du locuteur", Thèse de Doctorat, Université d’Avignon et des Pays de Vaucluse, 2006.

[34] Encyclopédie Wikipédia http//fr.wikipédia.org/wiki/Support_vecteor_machine, 2010.

[35] http://www. ETSI.org

[36] C. Levy, “Modèles acoustiques compacts pour les systèmes embarqués", Thèse de Doctorat, Université d’Avignon et des Pays de Vaucluse, 2006.

[37] ETSI-HR, Half Rate (HR) speech transcoding (GSM 06.20 version 8.0.1), 2000, URL http://www.3gpp.org/ftp/Specs/archive/06 series/06.20/0620-801.zip.

[38] ETSI-FR, Full Rate (FR) speech transcoding (GSM 06.10 version 8.2.0), 2000, URL http://www.3gpp.org/ftp/Specs/archive/06 series/06.10/0610-820.zip.

[39] ETSI-EFR, Enhanced Full Rate (EFR) speech transcoding (GSM 06.60 version 8.0.1), 2000, URL http://www.3gpp.org/ftp/Specs/archive/06 series/06.60/0660-801.zip.

[40] ETSI-AMR, Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1), 2000, URL http://www.3gpp.org/ftp/Specs/archive/06 series/06.90/0690-721.zip.

[41] B. Lilly, K. Paliwal, “Effect of speech coders on speech recognition performance", dans Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP’1996), tome 4, pages 2344–2347, Philadelphia, Pennsylvania, USA, 1996.

[42] S. Euler, J. Zinke, “The influence of speech coding algorithms on automatic speech recognition", dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP’1994), tome 1, pages 621–624, Adelaide, Australia, 1994.

[43] H. Kim, R.V, Cox, “A bitstream-based front-end for wireless speech recognition on IS-260 communication system", IEEE Transactions on Speech and Audio Processing, 9:558–568, 2001.

[44] B. Raj, J. Migdal, R. Singh, “Distributed speech recognition with codec parameters", Dans les actes de ASRU, 2001.

[45] A. Gallardo-Antolin, F. D. Maria, F. Valverde-Albacete, “Recognition from GSM digital speech", dans Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP’1998), pages 584–587, Sydney, Australia, 1998.

[46] S.H. Maes, G. Cohen R. Hoory, D. Chazan, “Conversational networking : Conversational protocols for transport, coding and control", dans Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP’2000), tome 2, pages 198–201, Beijing, China, 2000.

[47] G.N. Ramaswamy, P.S. Gopalakrishnan, “Compression of acoustic features for speech recognition in network environments", dans Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP’1998), tome 2, pages 977–980, Seattle, Washington, USA, 1998.

[48] N. Srinivasamurthy, A. Ortega, S. Narayanan, “Efficient scalable encoding for distributed speech recognition", IEEE Transactions on Speech and Audio Processing, 2005.

[49] ETSI, 2005a. ETSI ES 202 212 v1.1.2 (11-2005).

[50] S. Grassi, M. Ansorge, F. Pellandini, P.A. Farine, “Distributed speaker recognition using the ETSI Aurora standard", Dans les actes de Proc. Of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, 2002.

[51] S. Kuroiwa S. Tsuge, Blind equalization techniques for ETSI standard DSR front-end", Dans les actes de ICASSP, 2003.

[52] L. Mauuary, Blind equalization in the cepstral domain for robust telephone based speech recognition", Dans les actes de EUSIPCO, 359–362, 1998.

[53] M. Petracca, J. Servetti, A. De Martin, Performance analysis of compressed-domain automatic speaker recognition as a function of speech coding technique and bit rate", Dans les actes de ICME, 2006.

[54] Alexandre PRETI, « Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur ", Thèse de doctorat, l’Université d’Avignon et des Pays de Vaucluse, 2008.

[55] http://www.3gpp.org

[56] http://www.UTI.org

[57] A.J. Viterbi, J.K.Omura, principales of Digital Communication and Coding, McGraw Hill Kogakusha, Ltd, 1979.

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

[59] A. Krobba, “Reconnaissance Automatique du Locuteur Via une Ligne de Communication GSM“, Mémoire de Magister, faculté d’électronique et d’informatique, USTHB, 2009.