N° d'ordre : 37/2012-M/EL

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieure et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Électronique et d'Informatique

MÉMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

Par : BOUCHAKOUR Lallouani

En : ÉLECTRONIQUE
Spécialité : communication parlée

Etude de l'influence du codeur GSM sur la Reconnaissance Automatique de la Parole appliquée à la langue Arabe

Soutenu publiquement le 06/05/2012, devant le jury composé de :

Mr.A.DJERADI	Professeur à l'USTHB	Président
Mr.M.DEBYECHE	Maître de conférences/A à l'USTHB	Directeur de mémoire
Mr.H.TEFFAHI	Professeur à l'USTHB	Examinateur
Mr.A. AMROUCHE	Maître de conférences/A à l'USTHB	Examinateur
Mme.F.MERAZKA	Maître de conférences/A à l'USTHB	Examinatrice

Résumé :

Ces dernières années, le dialogue homme-machine via le réseau de communication téléphonique (mobile GSM, public) est devenu une réalité. Les applications sont nombreuse et diversifiées nous pouvons citer : la consultation de compte bancaire, la réservation dans les aéroports et les hôtels, l'accès aux bases des données…etc.

Notre travail est consacré à l'étude de l'influence du CoDec GSM EFR (Global System for Mobile communications Enhanced Full Rate) sur la performance d'un système de Reconnaissance Automatique de la Parole (RAP). La communication mobile (GSM) s'accompagne de plusieurs contraintes (bruit ambiant, dégradations dues au codeur GSM, fort taux de compression, le canal de transmission, etc.) qui influent négativement sur la performance des systèmes de RAP. Dans ce mémoire, nous avons étudie l'influence du codage GSM sur la performance des systèmes de RAP par deux variantes l'une sur le signal de parole transcodé GSM et l'autre sur le flux bitstream. Les paramètres acoustiques utilisés pour la variante parole transcodée sont les coefficients cepstraux MFCCs (Mel Frequency Cepstral Coefficients). La variante bitstream est basée sur les paramètres extraits directement du flux des bits tels que les coefficients LSF (Line Spectral Frequencies), LPC (Linear Predictive Coding), PCC (Pseudo-Cepstral Coefficients) et PCEP (Pseudo-Cepstrum). Le système de reconnaissance mise en œuvre est basé sur les Modèles de Markov Cachés (HMM : Hidden Markov Models pour l'anglais) en mode indépendant du locuteur. La base de données d'évaluation est la base ARADIGIT composée d'un vocabulaire de 10 chiffres arabes (de zéro à neuf) prononcés par soixante (60) locuteurs (31 locuteurs masculins et 29 locutrices), chaque chiffre est répété trois fois par chaque locuteur. Le taux de reconnaissance obtenu pour la variante parole transcodée est de 77,83%. Pour la variante bistream, il est de 83,67% pour les coefficients LSF, 89,67% pour les coefficients PCC et de 88,50% pour les coefficients PCEP. De ce travail, nous pouvons déduire que l'utilisation du flux des bits (bitstream) comme entrée des systèmes de reconnaissance permet d'avoir de meilleurs taux de reconnaissance.

Remerciements

Je tiens d'abord à adresser mes plus vifs remerciements à Monsieur Mohamed DEBYECHE, Maître de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, pour son aide et l'intérêt avec lequel il a suivi mon travail.

Je tiens à remercier Monsieur Amar DJERADI, Professeur à la Faculté d'Electronique et d'Informatique de l'USTHB, pour l'honneur qu'il me fait de présider ce jury.

Je remercie également Monsieur Hocine TEFFAHI Professeur à la Faculté d'Electronique et d'Informatique de l'USTHB et Monsieur Abderrahmane AMROUCHE et Mademoiselle Fatiha MERAZKA, tous deux Maître de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, pour l'honneur qu'ils me font en acceptant de participer à ce jury.

J'exprime mes remerciements aux membres du laboratoire LCPTS pour m'avoir aidé chacun à sa façon.

Enfin, j'exprime toute ma gratitude à ceux qui, de près ou de loin, chacun à sa manière, ont contribué à l'élaboration de mémoire.

Dédicaces

A ma mère

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

DTW: Dynamic Time Warping .

DAP: Décodage Acoustico-Phonétique.

HMM: Hidden Markov Models.

ASR: Acoustic Speech Recognition.

HTK: Hidden Markov Toolkit.

GSM: Global System for Mobile communication.

FR: Full Rate.

HR: Half Rate.

EFR: Enhanced Full Rate.

ACELP: Algebraic Code Excited Linear Prediction.

CELP: Code Excited Linear Predictor.

LTP: Long Term Prediction .

LPC: Linear Predictive Coding.

LSF: Line Spectrum Frequencies.

LSP: Line Paire Spectral.

PCC : Pseudo Cepstral Coefficients.

PCEP : Pseudo CEPstrem.

MPCC: Mel-Frequency Pseudo Cepstrem.

MPCEP: Mel-Frequency PCEP.

LPCC: Linear Prediction Cepstral Coefficient.

LP_MFCC: Linear Predictive Coding based Mel Frequency Cepstral Coefficient

MFCC: Mel Frequency Cepstral Coefficient.

RAP : Reconnaissance Automatique de la Parole.

DCT: Discrete Cosine Transform.

TFD: Transform Fourier Discrete.

TMC : Taux de Mots Corrects

SM : Station Mobile.

FM : Fréquency Modulation.

FDMA : Fréquence Division Multiple Access.

TDMA: Time Division Multiple Access .

CDMA : Code Division Multiple Access.

NSR: Network Speech Recognition.

ESR: Embedded Speech Recognition.

DSR: Distributed Speech Recognition.

Introduction Générale

La Reconnaissance Automatique de la Parole (RAP) est un processus important rentrant dans la chaine de communication Homme Machine. La RAP permet ainsi la communication orale avec la machine d'une manière directe ou à traves les réseaux de communication tel que le réseau téléphonique (réseau public ou réseau Mobile). Ces applications sont nombreuses et diversifiées nous pouvons citer la consultation à distance des comptes bancaires, l'accès aux bases des données, la détection de mots clef, la sécurité de l'information confidentielle…etc.

Le codeur GSM est le standard pour la téléphonie mobile dans toute l'Europe. La norme propose trois algorithmes de codage de la parole désignés par GSMFR (Full Rate), GSMHR (Half Rate) et GSM EFR (Enhanced Full Rate). Le rôle des codeurs GSM est de compresser le signal de la parole de manière à réduire les nombres des bits nécessaires à sa représentation, tout en maintenant une qualité du signal acceptable en sortie.

La communication mobile s'accompagne de plusieurs contraintes (bruit ambiant, fort taux de compression, influence du canal de transmission, etc.) qui influent négativement sur la performance globale de tout Système de Reconnaissance Automatique de la Parole (SRAP).

Pour l'étude de l'influence du codec GSM EFR sur la performance des systèmes de reconnaissance nous avons dans ce travail simulé par le logiciel MATLAB une ligne de communication GSM EFR et mis en œuvre un système de reconnaissance basé sur la méthode de classification probabiliste utilisant les modèles de Markov Cachés Continus CHMM (Continuous Hidden Markov Model). Des expériences de reconnaissance ont été réalisées sur le signal transcodé GSM et sur le flux de bits (bitstream) en langue arabe.

Ce travail est présenté par quatre chapitres. Apres cette introduction, on présente dans le premier chapitre la technologie GSM. Le deuxième chapitre est consacré à la reconnaissance automatique de la parole avec ces différentes approches. Le troisième chapitre présente le front-end à savoir la partie extraction des paramètres acoustiques utilisés dans les systèmes de reconnaissance de la parole. Dans le quatrième chapitre, sont présentés les expériences et les résultats auxquels nous avons abouti. Enfin, nous terminons par une conclusion générale et les perspectives.

Chapitre : 1

La Technologie GSM

1.1. Introduction

La technologie GSM (**Global System for Mobile communication, anglais**) est de nos jours un objet du quotidien et à l'attrait de la nouveauté succède l'exigence d'une qualité la plus proche possible de celle du téléphone fixe. Nombreux algorithmes de codage ont été normalisés depuis quelques années. On retient notamment, le GSM plein débit (Full Rate GSM), demi-débit (Half Rate GSM) et le GSM plein débit amélioré (Enhanced Full Rate GSM). Un système de codage de la parole comprend une partie codage et une partie décodage. Le codeur analyse le signal pour en extraire un nombre réduit de paramètres pertinents représentés par un nombre restreint de bits pour l'archivage ou pour la transmission. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique.

Dans ce chapitre nous présentons le principe de base de fonctionnement du réseau GSM et les techniques d'accès au réseau GSM pour économisée la bande de fréquence, ensuit en présente le principe de codeur/décodeur GSM EFR, basé sur l'algorithme ACELP (Algebric Code Excited Linear Prediction) .

1.2. La communication GSM

Le GSM, (Global System for Mobile communications), est un système cellulaire de communication mobile. Il a été rapidement accepté et a vite gagné des parts de marché telles qu'aujourd'hui plus de 180 pays ont adopté cette norme et plus d'un milliard d'utilisateurs sont équipés d'une solution GSM. Le réseau GSM offre à ses abonnés, des services qui permettent la communication de station mobile de bout en bout à travers le réseau. On peut citer, par exemple, la possibilité de téléphoner depuis n'importe quel réseau GSM dans le monde. Les services avancés et l'architecture du GSM ont fait de lui un modèle pour la troisième génération de systèmes cellulaires, le réseau UMTS.

Dans le réseau de communication GSM, le téléphone portable est utilisé comme un système acquisition de la parole, une fois la parole enregistrée, elle est codée (norme de codage, EFR) et transmis via un réseau de communication à un récepteur, qui effectué le travaille de décodage. Comme il est illustré dans les Figures **1.1.a** et **1.1.b** suivantes.

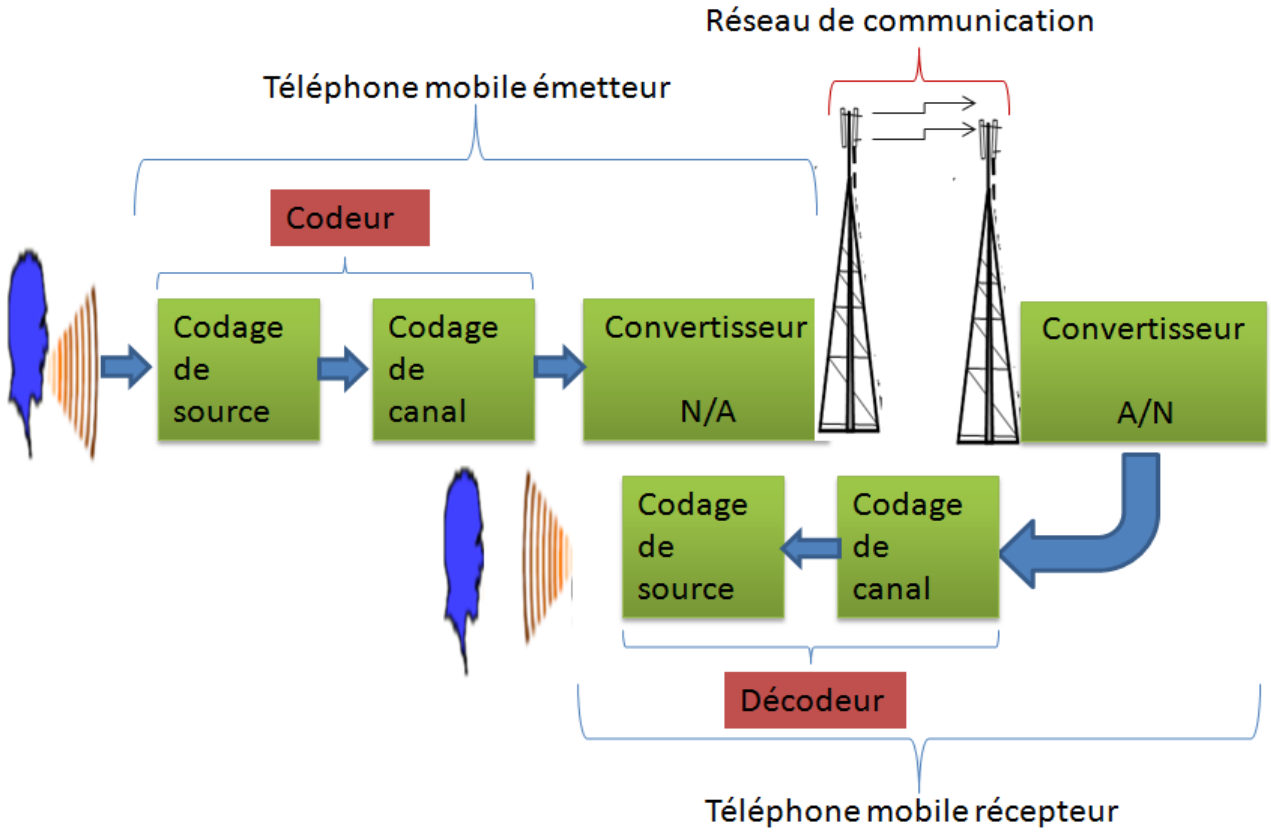


Figure 1.1.a : Architecture du réseau GSM.

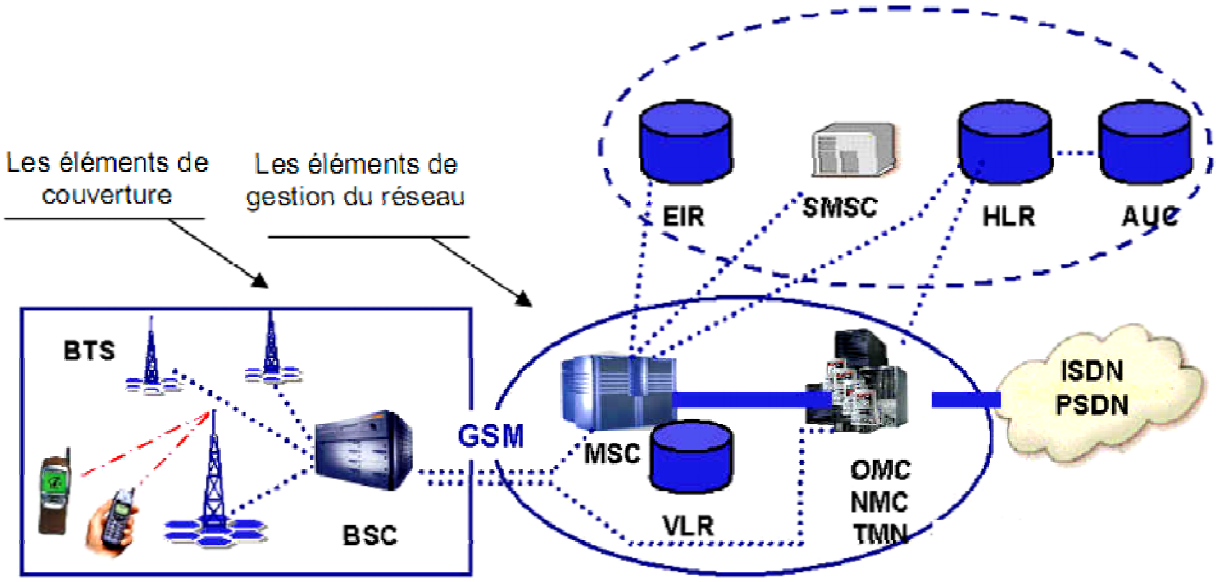


Figure 1.1.b : Architecture du réseau GSM.

1.2.1. L'équipement d'un réseau GSM

- **BTS** (Base Transtevere Station) : Station de base réceptionnant les appels entrant et sortant des **ME** (Mobile Equipment).
- **BSC** (Base Station Contrôler) : Contrôleur des station de base.
- **MSC** (Mobile Switching Centre) : Commutateur de réseau.
- **HLR** (Home Location Register) : Base de données sur l'identité et la localisation des abonnés.
- **AUC** (Authentification Centre) : Centre d'authentification des terminaux sur le réseau.
- **VLR** (Visitor Location Register) : Base de données sur les visiteurs du réseau.
- **OMC** (Operation And Maintenance Centre) : Centre d'exploitation et de maintenance du réseau de l'opérateur.
- **ME** (Moile Equipement) : Terminal de l'abonné.
- **SIM** (Sim Identity Module) : Carte SIM identifiant l'abonné sur un réseau défini.

Un système de radiotéléphonie permet l'accès au réseau téléphonique à partir d'un terminal portatif dans un territoire. Ce service utilise une liaison radioélectrique entre le terminal et le réseau.Ce terminal est aussi appelle SM « Station Mobile » dans le cadre du GSM. L'utilisation de La transmission FM (Fréquency Modulation) reste toujours la colonne vertébrale de la communication par système cellulaire. L'idée cellulaire à commence a apparaitre en début des années 80. Dans plusieurs années, laquelle dérive de la radiodiffusion basée sur le principe d'installation d'un émetteur de forte puissance au sommet du plus haut point de la zone et projetait le signal vers l'horizon. Cela permettait de bien adapter la couverture sur une plus grande surface. Ce qui signifiait que le peu de canaux disponibles était bloqué sur une grande surface pour un petit nombre d'appels (abonnés), on a alors l'idée d'utiliser plusieurs émetteurs de faible puissance chacun spécifiquement conçu pour une cellule (petite surface)**[1] [2]**.

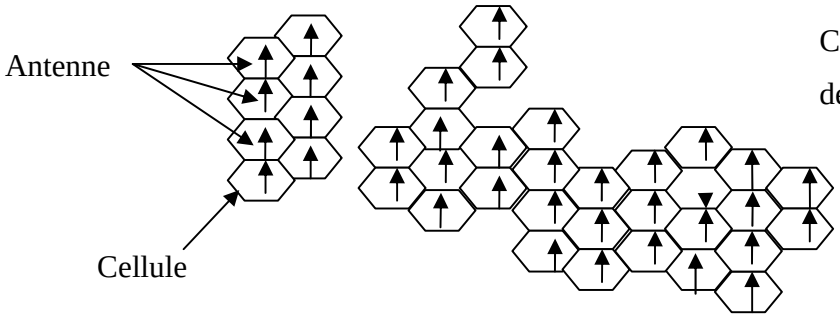


Figure1.2: Architecture cellulaire.

Les premiers calculs montrent que du fait des interférences entre les mobiles opérant sur le même canal dans des cellules adjacentes, on ne pouvait utiliser les mêmes fréquences dans deux cellules voisines. Il était nécessaire de sauter plusieurs cellules avant de réutiliser les mêmes fréquences.

On appelle **motif**, le plus petit groupe de cellules comportant l'ensemble des canaux une et une seule fois. Ce motif est répété sur toute la surface à couvrir. La distance minimale entre deux émetteurs utilisant la même fréquence est appelée « **distance de réutilisation** ».

La taille du motif est déterminée en fonction du seuil **C/(I + N)** (rapport signal sur bruit) au-delà duquel la réception est correcte. **C** est la puissance de la station de base, **I** la puissance totale des interférences et **N** la puissance du bruit. Plus le **C/(I + N)** est bas, plus la distance de réutilisation sera faible. Ainsi, la taille du motif pourra être réduite

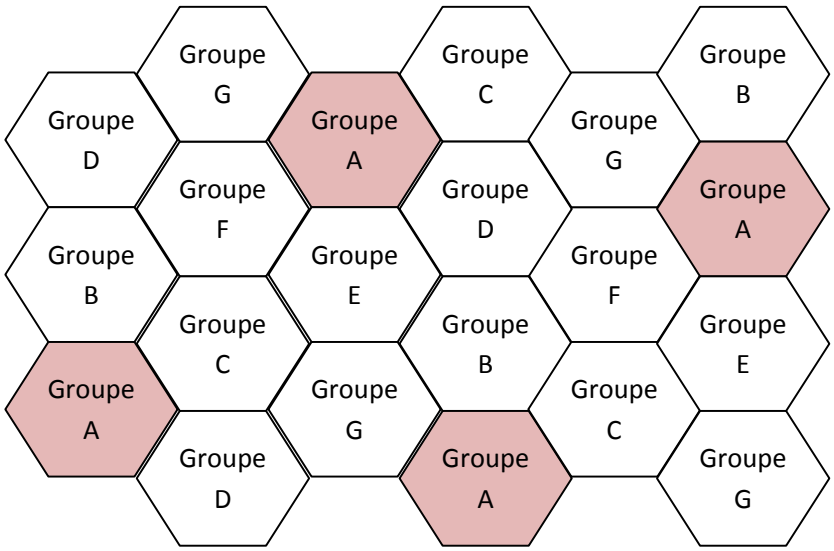


Figure 1.3 : Réutilisation des fréquences pour les différentes cellules (groupes) adjacentes. [**3].**

Mais l'idée de la réutilisation en elle-même semblait valable. L'organisateur du système pouvait en effet de crée plus d'un circuit téléphonique mobile à partir du même réutilisé dans différentes parties de la ville. Il est très cher de construire des system des milliers de cellules. Il apparut cependant que les cellules a grand rayon pouvaient évoluer aisément en cellules de plus petit rayon sur une certain période par le biais d'une technologie appelée division cellulaire, la division cellulaire offre plusieurs avantages. Elle permet en premier lieu une répartition des investissements au fur et à mesure la croissance du système de nouvelles cellules ne sera ajoutée que le nombre d'utilisateurs qui génèrent des revenus croitra (des petites cellules dans des endroits à haut densité de trafic et des cellules à plus grand rayon pour les zones isolées).

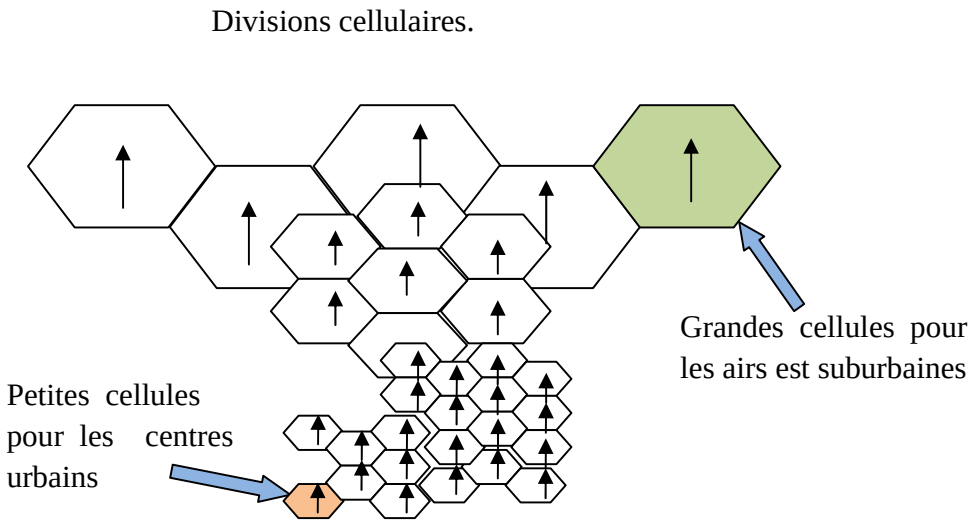


Figure 1.4: La division cellulaires zone urbain et zone suburbaine.

1.2.2. Forme hexagonale des cellules

Une cellule est approximée à un hexagone qui est le polygone le plus proche du cercle permettant de paver le plan.

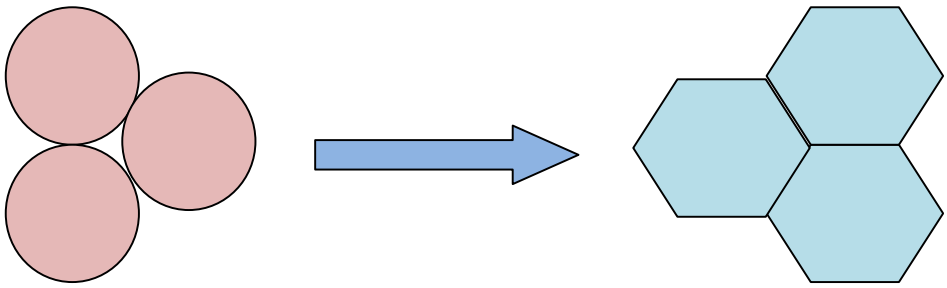


Figure 1.5 : Modèle de la cellule le plus proche.

1.2.3. Les méthodes d'accès

La bande de fréquence est une ressource rare, qu'il faut utiliser à bon escient et partager entre tous les utilisateurs. Il est donc nécessaire de transmettre simultanément sur un même canal la plus grande information possibles (plusieurs abonnés) et pour cela, on doit utiliser plusieurs techniques de multiplexage.

On considère trois techniques de découpage du spectre alloué, ou plus précisément du plan temps fréquence**[4]**.La définition des canaux de communication dépend de la méthode d'accès multiples retenue. Les trois techniques principales d'accès multiples sont:

- Accès Multiple par Répartition en Fréquences (AMRF) ou Fréquence Division Multiple Access (FDMA).
- Accès Multiple Par Répartition dans le Temps (AMRT) ou Time Division Multiple Access (TDMA).
- Accès Multiple par Répartition desCodes (AMRC) ou Code Division Multiple Access (CDMA).

1.2.3.1.Principes des techniques d'accès CDMA/FDMA/TDMA

Pour tout système mobile, il est nécessaire de définir et d'optimiser la façon dont les ressources radio disponibles sont allouées entre plusieurs utilisateurs. C'est à dire, il faut définir la technologie d'accès qui permet une gestion plus efficace de l'interface radio. Les trois techniques d'accès FDMA, TDMA, et CDMA sont définis comme suite:

a- Le FDMA

La technique d'accès FDMA repose sur un multiplexage en fréquence. Un tel procédé divise la bande de fréquence en plusieurs sous bandes. Chacune est placée sur une fréquence dite porteuse (carrier) qui est la fréquence spécifique de canal. Chaque porteuse ne peut transporter que le signal d'un seul utilisateur. La méthode FDMA est essentiellement utilisée les réseaux analogiques **[1].**

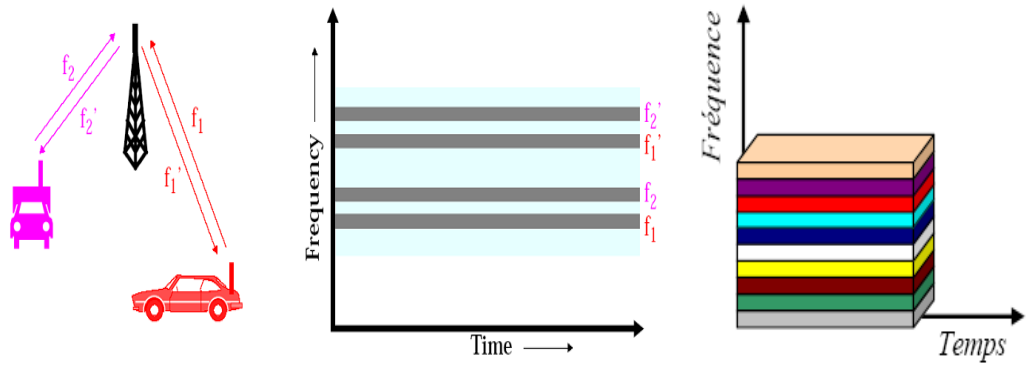


Figure1.6 : Multiplexage fréquentielle (FDMA).

b- Le TDMA

La technique d'accès TDMA offre la totalité de la bande de fréquence à chaque utilisateur pendant une fraction de temps donnée dénommée ***slot*** (portion). L'émetteur de la station mobile stock les informations avant de les transmettre sur le slot autrement dit dans la fenêtre temporelle qui a été réservée. Les différents slot sont regroupés par la suite en trames, le système offrant ainsi plusieurs voies de communication aux différents utilisateurs. La succession des slots dans les trames forme le canal physique de l'utilisateur. Le récepteur enregistre les informations à arrivée de chaque slot et reconstitue le signal à la vitesse du support de transmission. Toutefois la combinaison des deux techniques est possible.

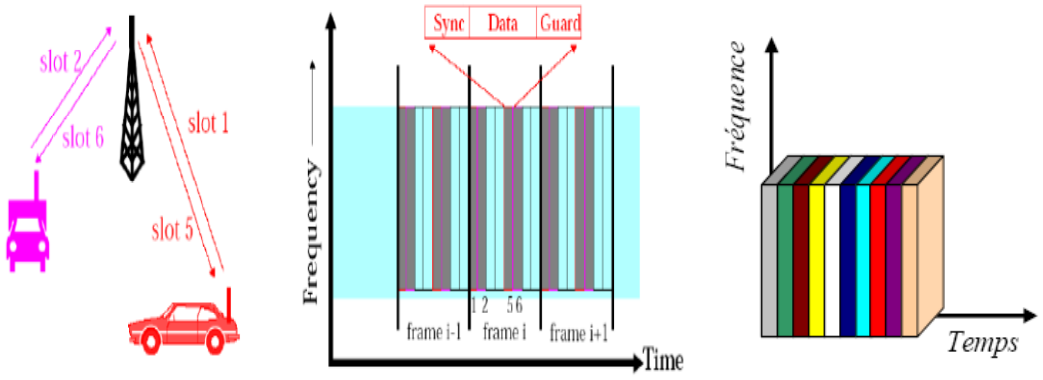


Figure1.7 : Multiplexage temporel (TDMA).

On peut combiner les deux techniques, la technique TDMA et la technique FDMA tel que chaque abonné peut utiliser une fréquence dans un temps bien déterminé.

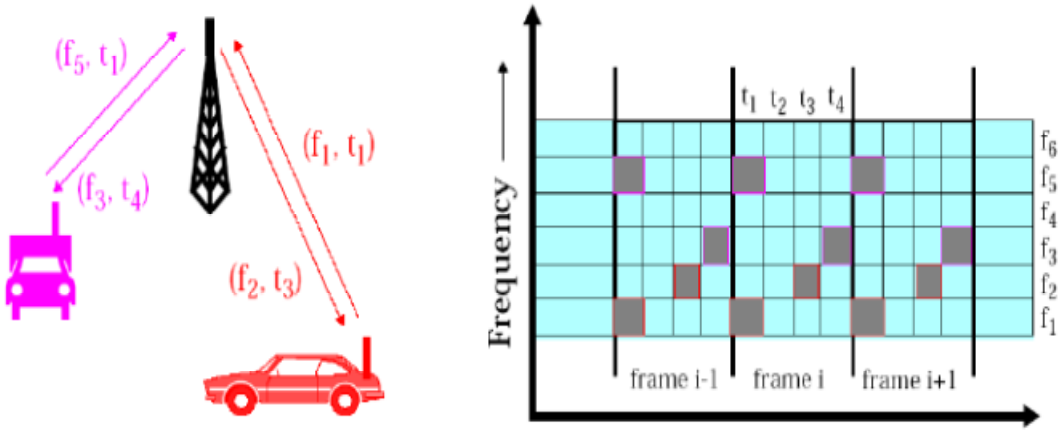


Figure1.8 : Combinaison TDMA et FDMA.

c- Le CDMA

La technique CDMA autorise l'allocation de la totalité de la bande de fréquences de manière simultanée à tous les utilisateurs dans la même cellule. Pour ce faire, un code binaire spécifique est octroyé à chaque utilisateur. Ce dernier se sert de son code pour transmettre l'information qu'il désire communiquer en format binaire d'une manière orthogonale, c'est-à-dire sans interférence entre les signaux ou autres communications. En CDMA l'usage de codes permet une réutilisation de la même fréquence dans les cellules adjacentes.

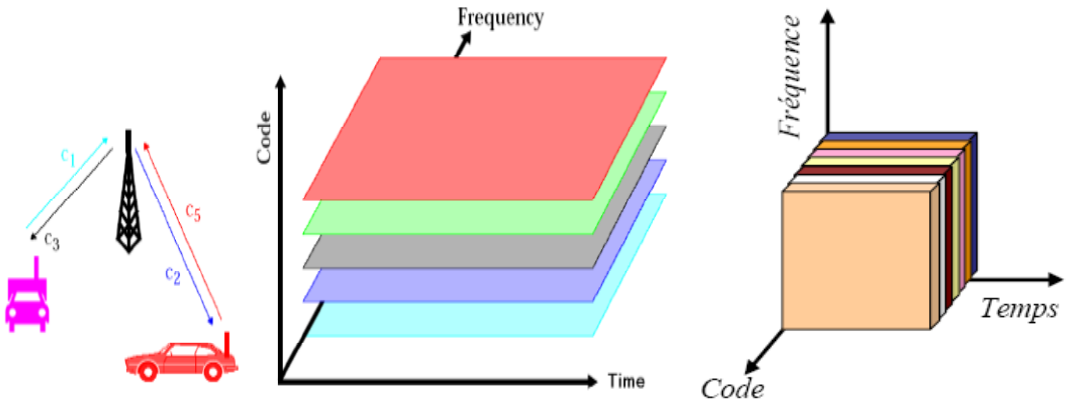


Figure1.9 : Multiplexage de code (CDMA).

1.3. Codeur de parole plein débit amélioré (EFR)

La standardisation du codeur GSM EFR est définie par l'égide ETSI (Européenne Télécommunication Standards Institut) et développé par le laboratoire de recherche sur le codage de la parole et des signaux audio de l'Université de Sherbrooke. Le codeur GSM EFR est basé sur Algorithme ACELP (Algebric Code Excited Linear Prédiction). Dans ce codeur, la parole est codée avec un débit de 12.2 Kbit/s et le débit de canal à 10.6 Kbit/s pour la protection des données contre les erreurs. Le débit de communication est donc de 22.8 Kbit/s, la qualité est équivalente à celle du codeur ADPCM 32 kbit/s (G721) pour la téléphonie publique fixe**[8].**

Le principe d'algorithme ACELP est présenté dans la Figure **1.10** suivante :

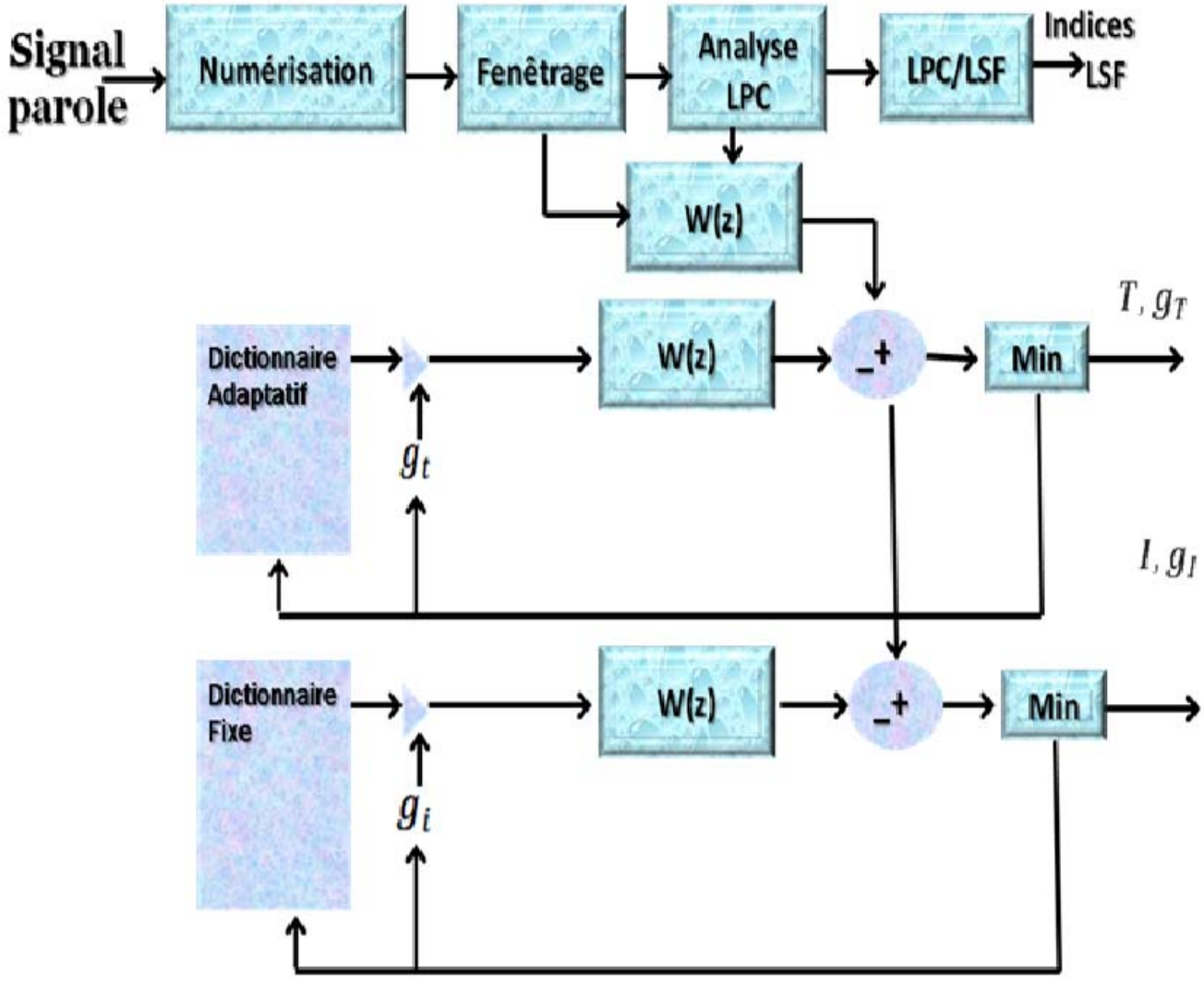


Figure 1.10 : Principe du codeur ACELP**[9].**

Le modèle de codage ACELP est un sujet d'etude et d'optimation depuis de nombreuses années. Jusqu'à maitenant, il s'agit du meilleur modèle pour reproduire un signal de parole à des débits aussi faibles que 16 kbps tout en conservant une qualité comparable à celle offerte par un lien téléphonique standard 64kbps. Ce type de codeur repose fondamentalement sur la modélisation du conduit vocal (prédiction à court terme) et sur la rechereche d'un signal d'excitation permettant de minimiser l'erreur entre le signal original et la signal de synthèse. On utilise pour ce codeur une trame de 20 ms avec une fréquence échantillonnage de 8 KHz c'est à dire 160 échantillons par trame. Pour chaque trame, le signal est analysé pour extraire les paramètres du modèle ACELP (Coefficients du filtre de prédiction Linéaire, les indices et les gains des dictionnaires fixe et adaptatif. Ces paramètres sont ensuite codés et transmis. Au niveau du décodeur, la parole est synthétisée à partir de ces mêmes paramètres.

Dans ce modèle, on cherche à construire un signal synthétique le plus ressemblant possible au signal original sur l'ensemble de la fenêtre d'analyse. La recherche du signal d'excitation à l'entrée du filtre de prédiction linéaire est trouvée en ajoutant deux vecteurs d'excitation des dictionnaires adaptatif et fixe. La séquence d'excitation optimale dans un dictionnaire est choisie en appliquant le principe d'analyse par synthèse. Dans cette recherche l'erreur entre les signaux de parole originale et synthétisé est minimisée suivant une mesure de distorsion.

Le tableau suivant illustre les allocations des bits de parametres du codeur GSM EFR.

Les paramètres	Première et troisième sous trame (bits)	Deuxième et quatrième sous trame (bits)	Le total bit par trame (bits)
2 LSP			38
Indice de dictionnaire adaptatif	9	6	30
Gain de dictionnaire adaptatif	4	4	16
Indice de dictionnaire fixe	35	35	140
Gain de dictionnaire fixe	5	5	20
Totale			244

Tableau1. 1 : Allocation des bits du codeur GSM EFR**[8].**

1.3.1. Prédiction linéaire LPC

Le codage de la parole à bas débit exige des représentations compactes et précises du filtre du conduit vocal et de la source d'excitation. C'est pour cela la prédiction linéaire, qui exploite les redondances du signal de parole en le modélisant par un nombre restreint de paramètres sous la forme d'un filtre linéaire, est un outil important de l'analyse de parole. L'idée de base du Codage Prédictif Linéaire LPC (Linear Predictive Coding) est de considérer que tout échantillon de parole peut être exprimé comme une combinaison linéaire des échantillons précédents. Un ensemble unique des coefficients prédictifs peut alors être déterminé et utilisé pour supprimer les redondances du signal à court terme. La redondance proche entre les échantillons du signal de parole est supprimée par le filtre d'analyse LP, qui est donné par la formule suivante:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1.1}$$

Où G représente le gain et les a_k les coefficients du filtre LP.

Mais le facteur de gain G est généralement égal à **un (1)**, H(z) est dans ce cas donnée par la relation suivante :

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{1.2}$$

L'erreur de prédiction $e(n)$ est l'erreur entre l'échantillon du signal original et son approximation. Le signal résiduel $e(n)$ est l'excitation idéale pour le modèle LPC du conduit vocal.

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1.3}$$

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{1.4}$$

Pour obtenir les coefficients a_k, on minimise l'erreur quadratique moyenne sur un intervalle de temps.

$$\text{Min} \sum e(n)^2 \tag{1.5}$$

Où le signal $s(n)$ peut être considéré comme quasi-stationnaire sur de faibles périodes d'analyse de l'ordre de quelques dizaines de millisecondes. Ainsi une segmentation est effectuée en multipliant le

signal de parole numérique $s(n)$par une fenêtre d'analyse $w(n)$, on obtient le signal de parole $s_w(n)$. L'analyse par prédiction linéaire est faite deux fois en utilisant deux fenêtres asymétriques. La première est centrée sur la seconde sous trame, elle est formée de deux moitiés de fenêtre de Hamming avec des tailles différentes, donnée par l'équation suivante **[11]** :

$$w_1(n) = \begin{cases} 0.54 - 0.46 \cos\left(\dfrac{\pi n}{L_1 - 1}\right) \\ 0.54 + 0.46 \cos\left(\dfrac{\pi (n - L_1)}{L_2 - 1}\right) \end{cases} \quad (1.6)$$

On utilise les longueurs $L_1 = 160$ et $L_2 = 80$ échantillons.

La seconde fenêtre est centrée autour de la quatrième sous trame et elle consiste en une moitié d'une fenêtre de Hamming et un quart d'une fonction cosinus. Elle est donnée par l'équation **1.7** suivante :

$$w_1(n) = \begin{cases} 0.54 - 0.46 \cos\left(\dfrac{\pi n}{L_1 - 1}\right) \\ \cos\left(\dfrac{2\pi (n - L_1)}{4 L_2 - 1}\right) \end{cases} \quad (1.7)$$

Avec$L_1 = 232$ *et* $L_2 = 8$ *echantillons* .

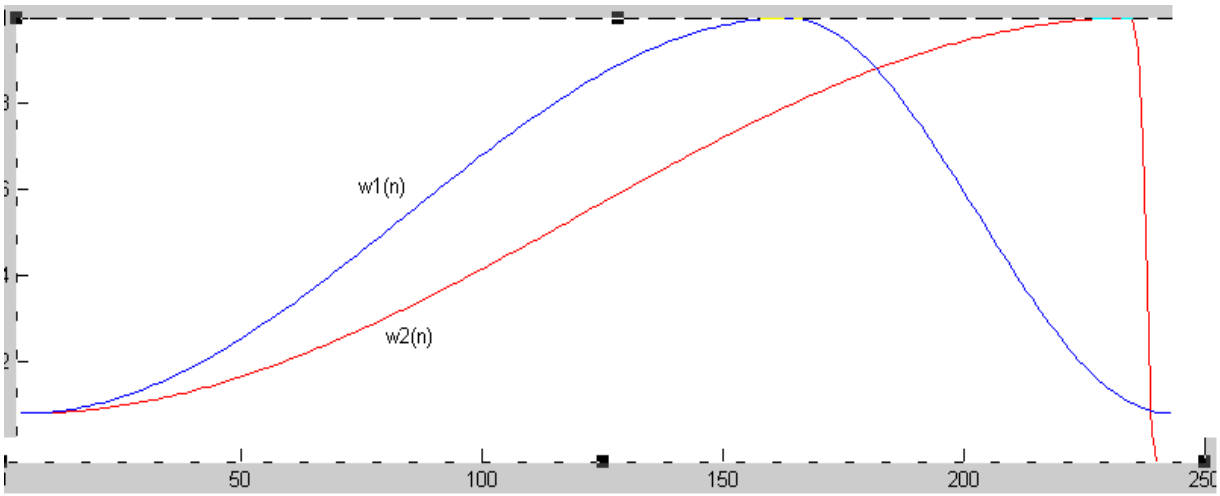


Figure1.11:Les deux fenetres d'analyse du codeur GSM EFR.

Pour trouver les coefficients de prédiction linéaire LP on doit minimisé l'energie de l'erreur $e(n)$ entre le signal original et le signal synthetisé.

$$E = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} (s(n) - \hat{s}(n))^2 = \sum_{n=-\infty}^{\infty} ((s_w(n) + \sum_{k=1}^{p} a_k s_w(n-k))^2 \qquad (1.8)$$

Les valeurs des a_k qui minimisent E sont obtenues en annulant les dérivées partielles de E par rapport à aux coefficients du filtre a_k, en utilisantl'approche autocorrélation.

Nous allons mettre $\frac{\partial E}{\partial a_k}$=0 pour $k = 1, \ldots \ldots p$ on aura p équations avec p inconnues.

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_w(n-i)s_w(n-k) + \sum s_w(n-i)s_w(n) = 0 \qquad 1 \le i \le p \qquad (1.9)$$

La fonction d'autocorrélation est donnée par :

$$R(i) = \sum_{n=i}^{N-1} s_w(n-i)s_w(n) \quad 1 \le i \le p \qquad (1.10)$$

$$\sum_{k=1}^{p} a_k R(\|i-k\|) = -R(i) \quad 0 \le i \le p \qquad (1.11)$$

$$\begin{pmatrix} R(0) & R(1) & R(2) & \ldots & R(p-1) \\ R(1) & R(0) & R(1) & \ldots & R(p-2) \\ R(2) & R(1) & R(0) & \ldots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \ldots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(P) \end{pmatrix} \qquad (1.12)$$

La matrice d'autocorrélation R de dimension ($p \times p$) dont tous les éléments de la diagonalesont égaux est symétrique, elle présente la propriété d'être une matrice de *Toeplitz* symétrique.

Le vecteur des coefficients de prédiction $a_k$peut alors être obtenu par inversion de la matrice Ren utilisant la méthode de *Levinson-Durbin*. On note que la résolution de ce système d'équations permet de garantir le minimum de phase au filtre de prédiction linéaire $A(z)$ (système stable).

1.3.2. Filtre perceptuel

Le masquage fréquentiel est la propriété du système auditif la plus largement exploitée. Elle est employée couramment dans les codeurs de forme d'onde où le bruit de quantification est filtré de telle manière que son spectre soit masqué par le spectre du signal de parole. Le filtre perceptuel est dérivé à partir du filtre $A(z)$ sous la forme générale **[16]** :

$$w(z) = \frac{A(z/y_n)}{A(z/y_d)} \qquad \text{Avec} \quad 0 < \lambda_d < \lambda_n < 1 \tag{1.13}$$

Le filtre $W(z)$ sert à modifier le spectre du bruit de quantification de façon à ce qu'il suive le spectre du filtre $A(z)$(l'enveloppe spectrale du signal de parole). Les régions de l'enveloppe spectrale avec une grande énergie (les formants) sont désaccentuées.

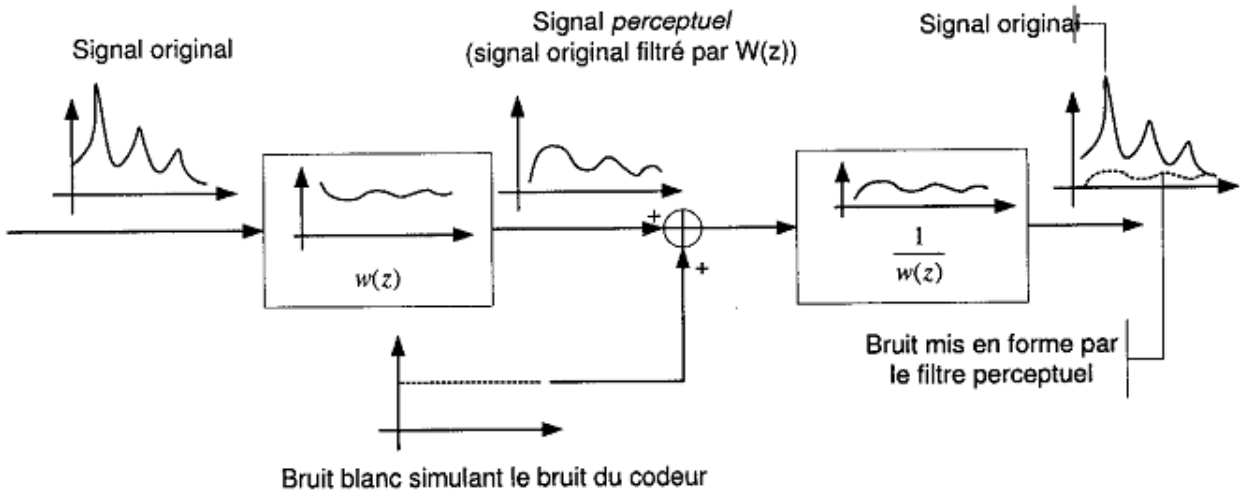


Figure 1.12 : Mise en forme de bruit par un filtre perceptuel.

À bas débit, le filtre perceptuel est utilisé dans les codeurs de type CELP (Code Excited Linear Prédiction). Dans la recherche de la meilleure excitation, l'erreur entre la parole synthétisée et la parole originale est filtrée par le filtre $W(z)$. Cela a pour effet une plus grande tolérance du critère de minimisation dans la région des formants. L'efficacité des filtres perceptuels du type $W(z)$ est limitée par le fait qu'ils ne peuvent que redistribuer le bruit en fonction de la fréquence. Quand le débit du codage est relativement bas, il devient impossible de masquer tout le bruit. Dans ce cas, il est important de garder son niveau suffisamment bas dans la région des formants.

1.3.3. Conversion et interpolation des coefficients de prédiction linéaire LPC

Les coefficients du filtre tout-pôle ne sont pas quantifiés directement car des erreurs de quantification relativement petites de coefficient LP peuvent produire des grandes erreurs dans le spectre et ainsi rendre le filtre instable. Il est donc nécessaire de les transformer dans une autre représentation qui contient la même information mais qui possède de meilleures propriétés pour le codage (quantification).

Des coefficients largement utilisés dans le passé étaient les coefficients de réflexion. Ces coefficients sont moins sensibles aux erreurs de quantification que les coefficients de prédiction linéaire a_k , ils forment un jeu ordonné et la stabilité du filtre peut être facilement contrôlée en vérifiant si chacun de ces coefficients se trouve dans l'intervalle$[-1, 1]$.

Actuellement les coefficients les plus utilisé sont les coefficients LSF (Line Spectrum Frequencies).
Soit les deux polynômes $P(z)$ et $Q(z)$ donnés par la relation suivante **[11]** :

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{1.14}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{1.15}$$

Ces polynômes sont symétrique et antisymétrique respectivement. Il est prouvé que toutes leurs racines sont distribuées sur le cercle unité. On redéfinit les deux polynômes afin d'éliminer les racines $z = -1$ et $z = 1$.

$$F_1(z) = P(z) \Big/ (1 + z^{-1}) \tag{1.16}$$

$$F_2(z) = Q(z) \Big/ (1 - z^{-1}) \tag{1.17}$$

Chacun de ces polynômes possède 5 racines conjuguées sur le cercle unité, ce qui nous permet d'écrire les polynômes sous la forme d'un produit :

$$F_1(z) = \prod_{i=1,3,.....9}(1-2q_i z^{-1}+z^{-2}) \quad (1.18)$$

$$F_2(z) = \prod_{i=2,4,.....10}(1-2q_i z^{-1}+z^{-2}) \quad (1.19)$$

Avec $q_i = cos(w_i), w_i$ une Fréquence de la Ligne Spectrale LSF. Les q_i sont référés comme les LSP (Line Paire Spectral) dans le domaine des cosinus.

En exploitant le fait que ces deux polynômes soient symétriques, seuls les cinq premiers coefficients doivent être calculés. Ils seront trouvés à partir des relations récursives suivantes :

$$f_1(i+1) = a_{i+1} + a_{m-1} - f_1(i) \quad (1.20)$$

$$f_2(i+1) = a_{i+1} - a_{m-1} + f_2(i) \quad (1.21)$$

Pour i = 0, 1, 2, 3,4, et $m = 10$ (l'ordre de prédiction Linéaire).

1.3.4. Conversion des coefficients LSP en LPC

Après la quantification et l'interpolation, les LSP seront reconvertis en coefficient de prédiction linéaire. En connaissant les valeurs de LPS, on peut calculer $f_1(i)$ et $f_2(i)$ lescoefficients de$F_1(z) et F_2(z)$. Ensuite $F_1(z) et F_2(z)$ seront multipliés par $(1+Z^{-1})$ et $(1-Z^{-1})$ respectivement pour obtenir $\acute{F}_1(z)$ et $\acute{F}_2(z)$.

Ceci est donné par :

$$f_1'(i) = f_1(i) + f_1(i-1) \qquad i = 1,.....,5 \quad (1.22)$$

$$f_2'(i) = f_2(i) + f_2(i-1) \qquad i = 1,.....,5 \quad (1.23)$$

Enfin, les coefficients de prédiction linéaires sont calculés par :

$$a_i = \begin{cases} 0.5 f_1'(i) + 0.5 f_2'(i) & i = 1,.....5 \\ 0.5 f_1'(11-i) - 0.5 f_2'(11-i) & i = 6,....,10 \end{cases} \quad (1.24)$$

Cela peut conduire à trouver :

$$A(z) = \frac{\acute{F}_1(z) + \acute{F}_2(z)}{2} \quad (1.25)$$

1.3.4.1.Interpolation

Les deux séries des paramètres LSP seront utilisés pour la seconde et la quatrième sous trame. La première et la troisième sous trame utilisent une interpolation linéaire des paramètres des sous trames adjacentes **[11]**.

Soit $p_i^{(n)}$ le vecteur des LSP correspondant à la $i^{éme}$sous trame de la $n^{éme}$ trame.

L'interpolation se déroule comme dans les équations suivantes :

$$p_1^{(n)} = 0.5\, p_2^{(n-1)} + p_4^{(n)} \quad (1.26)$$

$$p_3^{(n)} = 0.5\, p_2^{(n)} + p_4^{(n)} \quad (1.27)$$

$p_2^{(n-1)}$: Vecteur des coefficients de la deuxième sous trame précédente.

L'interpolation sert à calculer un filtre de prédiction linéaire différent pour chacune des sous trames en utilisant les coefficients LP convertis à partir des paramètres LSP.

1.3.5. Recherche d'excitation

1.3.5.1.Dictionnaire adaptatif

Cette recherche est procédée sur la base des sous-trames. Elle consiste en une recherche de délai de pitch par une boucle fermée, et en calculant le vecteur de code adaptatif par interpolation du dictionnaire excitation.

a- Analyse de pitch

L'étude du signal résidu $e(n)$ démontre des impulsions d'énergie périodiques importantes.ces pointes d'énergie sont en fait causées par les impulsions quasi périodiques des cordes vocales. La recherche du délai et du gain optimal se font en deux étapes pour former le dictionnaire adaptatif, ce qui constitue en réalité le filtre $B(z)$**[13]**.

Il peut être montré que l'obtention du gain et du délai optimal s'effectue en maximisant le critère $R(k)$ suivant:

$$R(k) = \frac{\sum_{n=0}^{39} x(n) y_k(n)}{\sqrt{\sum_{n=0}^{n=39} y_k(n) y_k(n)}} \quad (1.28)$$

Ou $x(n)$ est le signal de la cible et $y(n)$ est la séquence d'excitation filtrée.

Et le gain est déterminé par la relation suivante:

$$g = \frac{\sum_{n=0}^{n=39} x(n) y_k(n)}{\sum_{n=0}^{n=39} y_k(n) y_k(n)} \quad (1.29)$$

1.3.5.2.Dictionnaire algébrique

Le vecteur d'excitation dans ce dictionnaire contient dix 10 impulsions non nulles.

Les 40 positions dans une sous trame sont divisées sur **cinq 5** voies **[10][11]**, contenant chacune deux impulsions, comme il est montré dans le tableau ci-dessous:

voie	impulsion	position
1	i_0 ,i_5	0, 5, 10, 15, 20, 25, 30, 35
2	i_1 ,i_6	1, 6, 11, 16, 21, 26, 31, 36
3	i_2 ,i_7	2, 7, 12, 17, 22, 27, 32, 37
4	i_3 ,i_8	3, 8, 13, 18, 23,28, 33, 38
5	i_4 ,i_9	4, 9, 14, 19, 24, 29, 34, 39

Tableau 1.2: Vecteur du dictionnaire fixe

a- Recherche du dictionnaire de code algébrique

Dans un codeur ACELP, le répertoire d'excitation dans le dictionnaire fixe est un ensemble virtuel (dans le sens où il n'est pas stocké), généré algébriquement. Un vecteur ayant très peu de composantes non nulles. Le signal cible est donné par :

$$x_2(n) = x(n) - g_p y(n) \tag{1.30}$$

Le vecteur code d'indice *k* parmi toutes les combinaisons possibles est noté$\boldsymbol{c_k}$. Le dictionnaire de code algébrique est recherché en maximisant le critère suivant **[11][14][15]**:

$$A_k = \frac{(z_k)^2}{E_d} = \frac{(d^t c_k)^2}{c_k{}^t \phi c_k} \tag{1.31}$$

Avec :

$d = H^t x_2$ est la corrélation entre le signal cible $x_2(n)$ et la réponse impulsionnelle$h(n)$.

H est une matrice de convolution triangulaire inférieure, avec $h(0)$ sur la diagonale au centre, et respectivement $h(1)$ ... ... $h(39)$ sur les diagonales inférieures.

$\emptyset = H^t H$est la matrice de corrélations de $h(n)$.

Le calcul du vecteur $\boldsymbol{d}$et de la matrice $\boldsymbol{\Phi}$ sont donnés par les équations suivantes, respectivement :

$$d(n) = \sum_{i=n}^{39} x_2(i) h(n-i) \qquad n = 0,.....,39 \tag{1.32}$$

$$\phi(i,j) = \sum_{n=j}^{39} h(n-i) h(n-j) \quad j > i \tag{1.33}$$

La structure algébrique du dictionnaire de codes permet une recherche très rapide par ce que les vecteurs de code fixe $C_k$ne contiennent qu'un nombre réduit d'impulsions non nulles (dix impulsions parmi 40 impulsions).

L'energie E_d est donnée par la relation suivante :

$$E_d = \sum \phi(m_i, m_i) + 2 \sum_{i=0}^{N_p-2} \sum_{j=i+1}^{N_p-1} v_i v_j \phi(m_i, m_j) \tag{1.34}$$

$$C_k = \sum_{i=0}^{N_p-1} v_i d(m_i) \tag{1.35}$$

Ou m_i est la position de la $i^{éme}$ impulsion , et v_i son amplitude . $N_p = 10$ est le nombre des impulsions.

b- Gain du dictionnaire algébrique

Le gain du dictionnaire fixe est donné par le rapport suivant **[12] [15]:**

$$g_c = \frac{x_2^t}{tt^t} \tag{1.36}$$

Où $x_2$est le vecteur du signal de référence et $\boldsymbol{t}$ le vecteur de code du dictionnaire fixe Cconvolué avec $h(n)$.

$$t(n) = \sum_{i=0}^{n} c(n)h(n-i) \quad n = 0,.....39 \tag{1.37}$$

1.4. Décodeur GSM EFR

La fonction du décodeur consiste à décoder les paramètres transmis qui sont :

- Les coefficients LSP.
- Le vecteur de code du dictionnaire adaptatif.
- Le Gain du dictionnaire de code adaptatif.
- Le Vecteur de code du dictionnaire fixe.
- Le Gain du dictionnaire de code fixe.

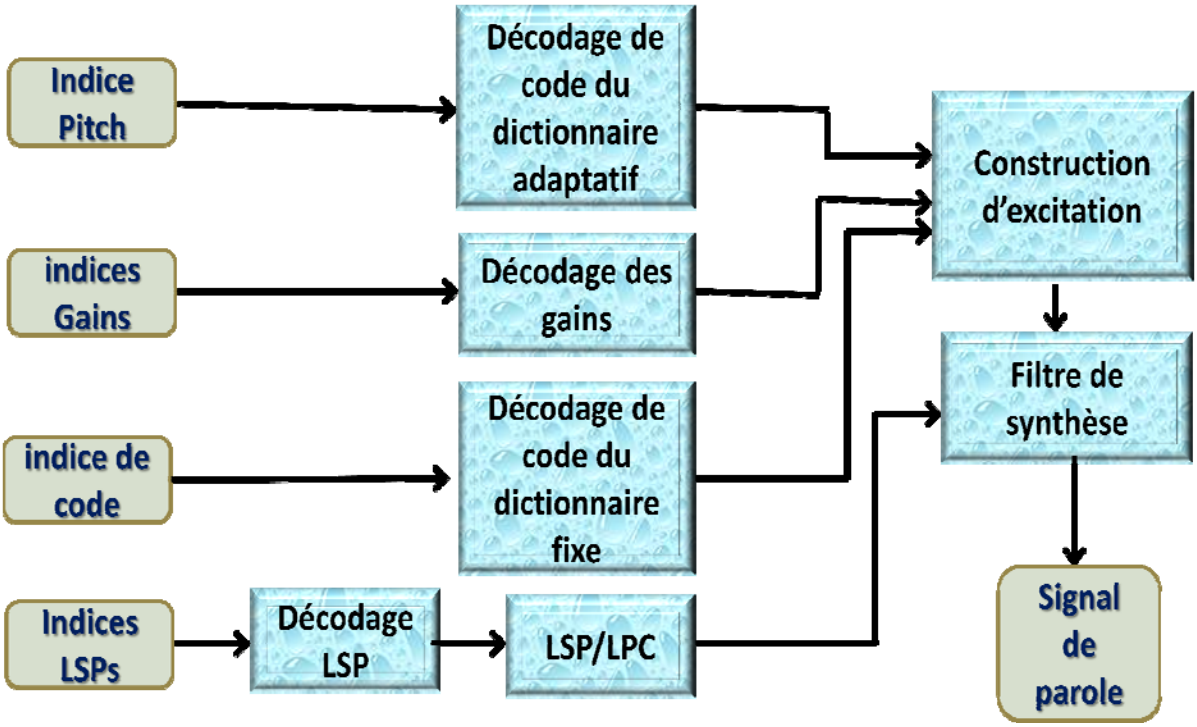


Figure 1.13 : Schéma fonctionnel détaillé du décodeur GSM EFR**[9].**

1.4.1. Décodage et synthèse de parole

Les indices reçus de la quantification des coefficients LSP sont utilisé pour reconstruire les deux vecteurs des LSP quantifiés. L'interpolation décrite dans le codeur précédent est appliquée pour retrouver les 4 vecteurs LSP interpolés, correspondants au quatre (04) sous trames. Pour chacune des sous trame, le vecteur LSP interpolé est converti en coefficients de filtre de prédiction linéaire a_k qui sont utilisés pour synthétiser la parole.

Pour chaque trame on répète les étapes suivantes :

- Décodage du vecteur de code du dictionnaire adaptatif, l'indice de pitch.
- Décodage du gain du dictionnaire de code adaptatif.
- Décodage du vecteur du dictionnaire de code fixe.
- Décodage du gain du dictionnaire de code fixe.

1.5. Conclusion

Dans ce chapitre nous avons présenté le principe de fonctionnement du réseau GSM, les différentes composantes de base ainsi que les méthodes d'accès au réseau. Nous avons décrit les différents étages du CoDec GSM EFR utilisé actuellement par les opérateurs de télécommunication. Les deux parties Codage et Décodage ont été également détaillées. Le codeur GSM EFR est basé sur l'algorithme de codage ACELP.

Chapitre : 2

Reconnaissance Automatique de la Parole.

2.1. Introduction

La Reconnaissance Automatique de la Parole (RAP) s'inscrit dans le domaine général du traitement automatique de la parole dont l'objectif est de faciliter la communication homme-machine en annulant le lien physique qui oblige souvent l'être humain à utiliser ses mains pour interagir avec une machine ou un ordinateur. La RAP est utilisée dans de nombreux domaines. Dans le domaine industriel, la reconnaissance vocale peut être utilisée pour la commande de machines ou le contrôles de processus. Pour les applications grand public, son usage est généralement dédié à l'amélioration de la vie quotidienne nous pouvons citer les logiciels de dictée automatique, la composition vocale des numéros de téléphone, la consultation des comptes bancaire,… etc.

Dans ce chapitre nous commencerons par décrire les principales approches de reconnaissance, l'approche analytique et l'approche globale. Nous insisterons sur l'approche probabiliste basée sur les modèles de Markov cachés et les différentes architectures de reconnaissance automatique de la parole via le réseau de communication Mobile.

2.2. Architecture d'un système de RAP

Les systèmes de Reconnaissance Automatique de la Parole (RAP) se basent généralement sur la même architecture, constituée de plusieurs étapes, que l'on peut visualiser au moyen de la Figure **2.1.**

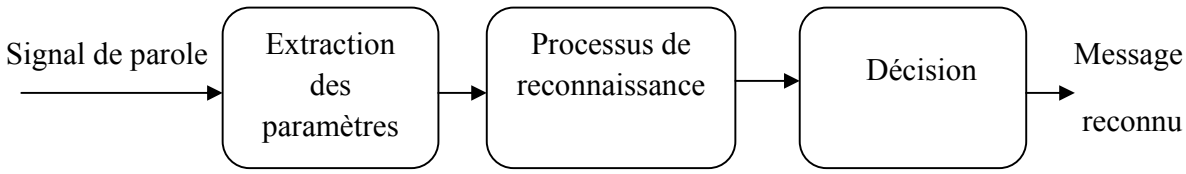


Figure 2.1 : Structure générale d'un système RAP.

Ainsi, après la phase d'extraction des paramètres que nous présenterons dans le chapitre 3, on retrouve le cœur du système à savoir le processus de reconnaissance et la prise de décision. Les différentes approches de reconnaissance citées peuvent être utilisées pour décoder le signal de parole en se basant sur les paramètres extraits auparavant. Nous en retiendrons particulièrement deux approches globale et analytique.

2.3. Approches de reconnaissance

Nous distinguons en reconnaissance des parole deux approches complètement différentes utilisées ces dernières années : l'approche analytique et l'approche globale. La première approche est basée sur les connaissances et utilise les techniques de l'intelligence artificielle. La deuxième approche est basée sur la reconnaissance de forme.

2.3.1. Approche analytique

L'approche analytique cherche à isoler les unités acoustiques courtes comme les phonèmes, les diphonèmes ou les syllabes. Le signal acoustique est d'abord segmenté en unités phonétiques, puis une identification des segments est réalisée. Dans cette approche, appelée aussi approche phonétique, le message est ainsi segmenté en constituants élémentaires (unités acoustiques), ces constituants élémentaires présentent l'avantage d'être en nombre réduit 34 phonèmes pour la langue arabe permettent de décrire l'arabe parlée **[23].**

Le caractère continu du signal vocal complique beaucoup la reconnaissance de la parole et aucun indice acoustique ne permet de localiser les frontières de mots (séparation). Ce problème est abordé, après la phase de prétraitement, d'une part par un décodage acoustique-phonétique (DAP) permettant la transcription de la phrase sous forme d'une suite d'éléments phonétique du langage, et d'autre part par un traitement linguistique faisant appel à diverses sources d'informations (lexicales, syntaxiques, sémantiques) permettant la reconnaissance des mots.
Ce sont donc des systèmes à architectures logicielles complexes à plusieurs sources de connaissances (les systèmes experts outils de l'intelligence artificielle) qui pallient les problèmes de reconnaissances des phrases.

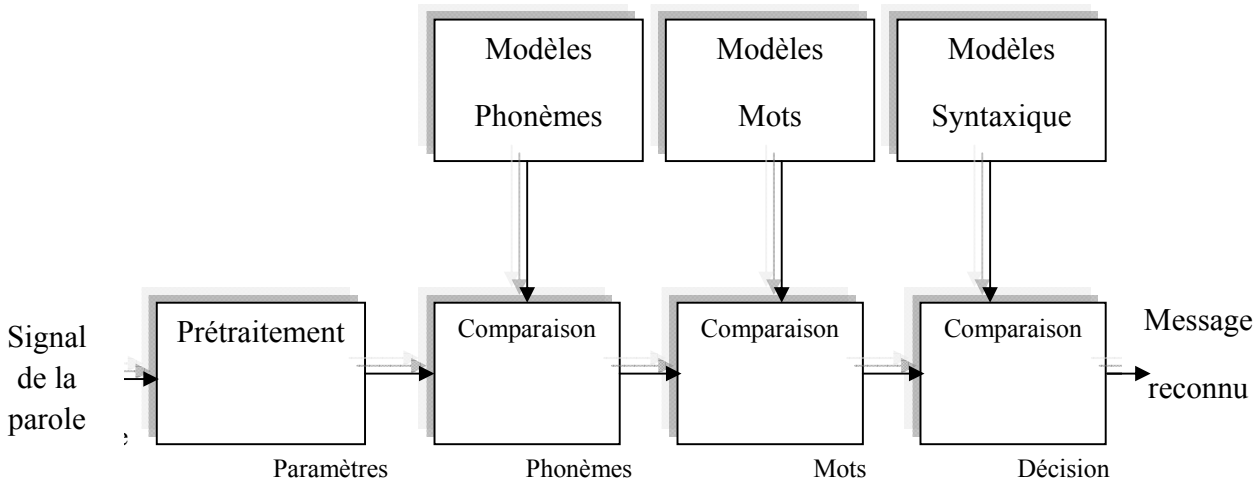


Figure.2.2 : Schéma synoptique d'un système de RAP basé sur l'approche analytique.

2.3.2. Approche globale

L'approche globale évite toute segmentation a priori et ne fait pas d'hypothèses sur le type des éléments à traiter. Elle effectue des comparaisons sur un ensemble de références en traitant les données et les connaissances dans leur globalité. Le principe de base consiste à donner au système de reconnaissance au moins une image de chacune des unités qu'il est censé devoir identifier par la suite. Cette opération est faite lors de la phase d'apprentissage qui permet de constituer la base de données de référence du système. Le processus de décodage consiste alors à comparer l'image de l'unité à identifier avec celles de la base de référence.

La reconnaissance globale comprend deux phases distinctes :

- **La phase d'apprentissage :** pendant laquelle un ou plusieurs locuteurs prononcent une ou plusieurs fois les unités de l'application prévue (dans ce projet nous utilisons la base ARADIGIT **[24]**, composée des chiffres arabes de zéro à neuf prononcés trois fois par soixante locuteurs. Ces prononciations sont toutes prétraitées puis conservées ou bien moyennées dans un dictionnaire de références.
- **la phase de reconnaissance :** où le signal à reconnaître subit le même prétraitement que la phase précédente. Il est ensuite comparé aux références contenues dans le dictionnaire. Le calcul d'une « distance » et sa comparaison à un seuil permet de retenir la référence les plus proches.

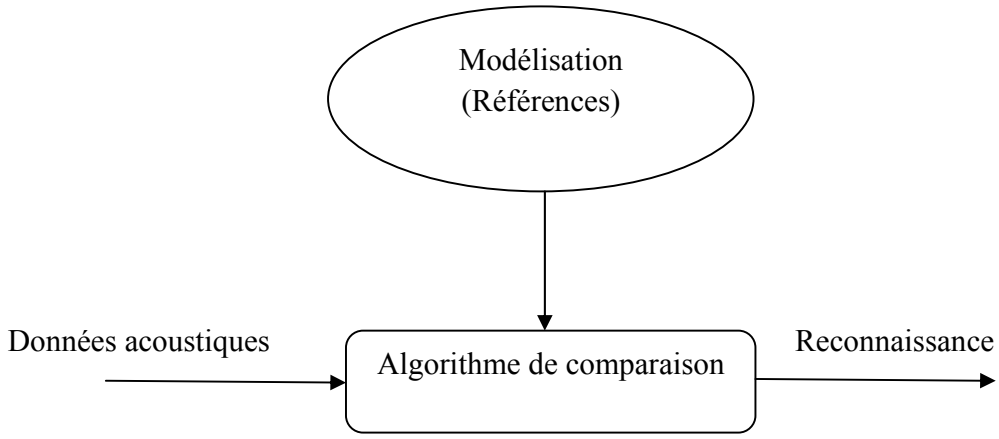


Figure 2.3 : Schéma synoptique d'un système de RAP par l'approche globale

Il existe principalement deux types de modélisation des propriétés d'un signal donné :

- La modélisation déterministe
- La modélisation statistique

2.3.2.1. Reconnaissance basée sur une modélisation déterministe DTW

L'approche DTW (Dynamic Time Warping) est fondée sur un principe de comparaison du signal à reconnaitre à des formes des références. Elle a été utilisée avec succès pour la première fois en reconnaissance de la parole par Sako. Elle consiste principalement à calculer une mesure de la distance entre la forme à reconnaitre (le signal de test) et les formes de références. Le signal identifié correspond à la forme de référence la plus proche en terme de distance calculée **[25].**

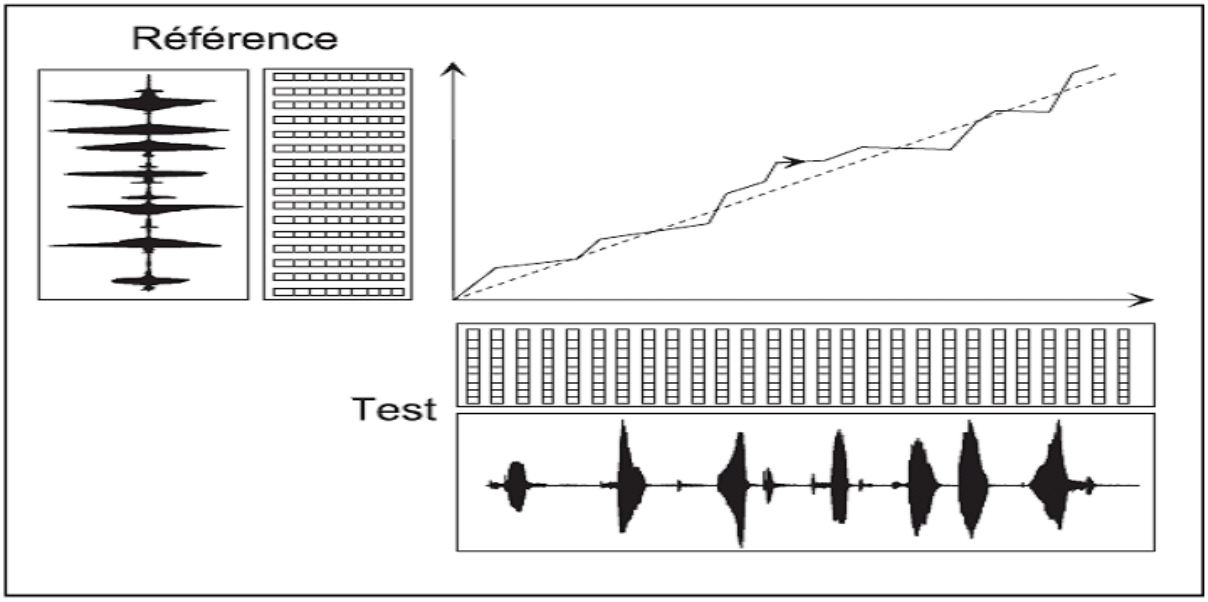


Figure 2.4: Comparaison de formes par la méthode DTW.

Soient $R = (r_1, r_{2,\ldots\ldots\ldots\ldots} r_l)$ une référence de mot et $T = (t_1, t_{2,\ldots\ldots\ldots\ldots} t_l)$ le mot à reconnaitre. Nous appelons $d(r_i, t_j)$ la distance entre les vecteurs acoustique $r_i \; et \; t_j$. Plusieurs types de distances peuvent entre utilisés en fonction des méthodes de paramétrisation retenues, nous pouvons citer à titre d'exemple :

- La distance euclidienne est utilisée dans les systèmes a base d'analyse cepstrale.

$$d(r_i, t_j) = (\sum_{k=0}^{p} |r_k - t_k|^2)^{\frac{1}{2}} \tag{2.1}$$

- La distance d'Itakura est utilisée dans le cadre d'une paramétrisation de prédiction linéaire.

$$d(r_i, t_j) = \log\left[\frac{r_i' R_b r_r}{t_j' R_b t_j}\right] \tag{2. 2}$$

Ou R_b représente la matrice des coefficients d'autocorrélation évalués sur le segment t_j.

2.3.2.2. Reconnaissance basée sur une modélisation statistique

L'approche probabiliste consiste à modéliser la production d'unité linguistique (phonème, syllabe, mot,...) en prenant en considération la statistique des diverses prononciations de cette unité, le but étant d'identifier une unité inconnue en déterminant le modèle ayant la plus forte probabilité de l'avoir produite. Les méthodes stochastiques dans le domaine de la modélisation des signaux présentent actuellement un grand intérêt pour les résultats intéressants qu'elles donnent. Le modèle utilisé dans notre travail est le modèle de Markov caché que nous allons détailler dans ce qui suit.

2.3.2.2.1. Le modèle de Markov caché

Le modèle Markovien correspond à un simple graphe d'états défini par une fonction de transition probabiliste à chaque pas de temps, le modèle subit une transition et modifie son état. Cette transition permet donc au système d'évoluer selon une loi connue par avance. Néanmoins, cette loi de transition est probabiliste. En effet, l'évolution du système peut être incertaine, ou simplement mal connue.

Cette fonction probabiliste permet donc d'exprimer simplement la loi d'évolution du modèle, sous la forme d'une matrice de probabilité.

Nous allons voir comment définir formellement les paramètres des modèles acoustiques c'est-a-dire les probabilités d'émission d'un vecteur acoustique et les probabilités de transition.

Une chaine de Markov cachée est un automate à N états que nous noterons 1,...N. Nous noterons q_t l'etat de l'automate à l' instant t. La probabilité de transition d'un état i a un état j est donnée par a_{ij} **[26] [27]**:

$$a_{ij} = P(q_t = \text{j} / \ \ q_{t-1} = \ \text{i}) \tag{2. 3}$$

Tel que :

$$\sum_{n=1}^{N} a_{ij} = 1 \tag{2.4}$$

Lorsque l'automate passe à l'état q_t il émet l'observation ou le vecteur acoustique O_t ou j prends N valeurs : 1,...N. La probabilité d'émission, qui est la probabilité pour que

l'automate émette une observation O_t (le vecteur acoustique) lorsqu'il est dans l'état q_t elle est notée par $b_j(O_t)$:

$$b_j(O_t) = P(O_t / \ q_t) \tag{2.5}$$

$$\sum_{n=1}^{N} b_j(O_t) = 1 \tag{2.6}$$

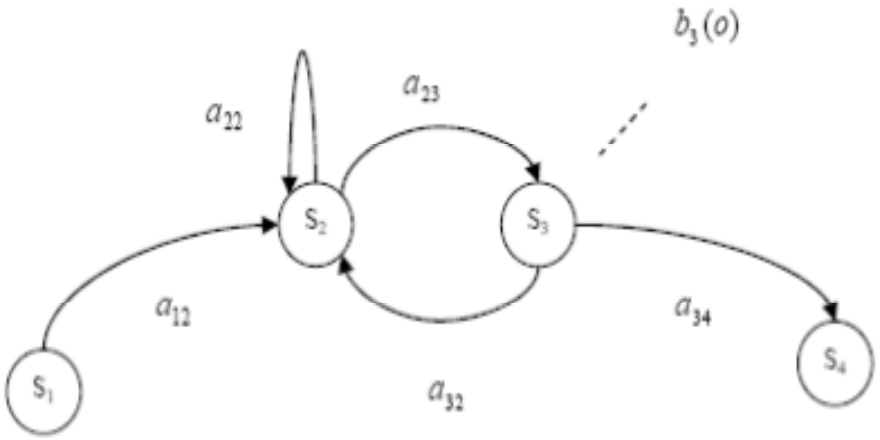


Figure 2.5 : Exemple de Modèle de Markov caché.

2.3.2.2.1.1. Les paramètres du modèle

Un modèle caché de Markov est un quadruplet $\{Q, A, B, O\}$ des ensembles décrits suivants

$Q = \{q_i\}_{1 \le i \le N}$ l'ensemble des N etats, en sachant que le processus part de l'etat initial q_1 à l'instant $t = 0$ et arrive à l'état final q_N à l'instant $t = T$

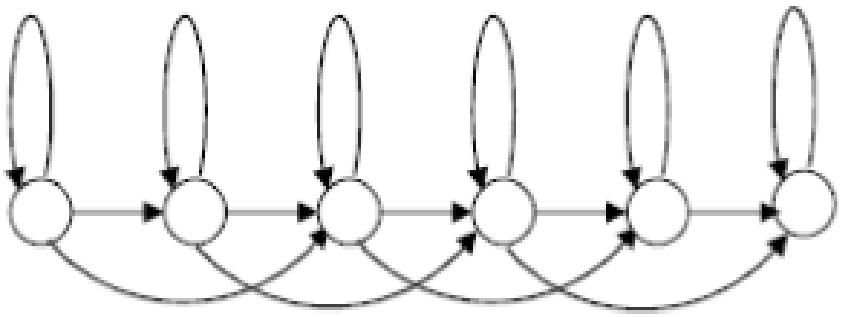
- $A = \{a_{ij}\}_{1 \le i,j \le N}$ l'ensemble des probabilités de transition entre les états i et j.
- $B = \{ b_j(O_t) \}_{1 \le j \le N, 1 \le t \le T}$ l'ensemble des probabilités d'émission du symbole O_t lors d'arrivée dans l'état j.
- $O = \{O_t\}_{1 \le t \le T}$ l'ensemble des O_t symboles observés.

2.3.2.2.1.2. Modèles de mots

Les modèles de Markov utilises pour représenter la parole sont des modèles "gauche droit", ceci traduit par la causalité du processus de production de la parole. Les probabilités de transition vérifient:

$$i > j \rightarrow a_{ij} = 0 \tag{2.7}$$

Des modèles gauche-droit particuliers autorisant le bouclage à l'état courant, le passage à l'état suivant, le saut d'un état ont été introduits par R. Bakis. Pour représenter des mots **(Figure 2.6).**



(2.7)

Figure 2.6. Modèle de Bakis.

2.3.2.2.1.3. Observations continues

Le principe d'émission des symboles discrets peut se généraliser au cas continu. Les probabilités d'émission discrètes $b_j(O_t)$ sont alors remplacées par des densités de probabilité continues dans l'espace de représentation. L'utilisation d'une combinaison linéaire de gaussiennes dans l'espace Rd est fréquente:

$$b_j(o_t) = \sum_{k=1}^{G} g_k \, N(o, \mu_k, \sum_k) \tag{2.8}$$

Avec μ_k, $\sum_k$ sont respectivement la moyenne et la matrice de covariance de la k^{ième} gaussienne, et g_k la pondération affectée de cette gaussienne. Nous rappelons que la densité de probabilité d'une loi normale de moyenne μ et de matrice de covariance Σ en dimension d est:

$$N(o, \mu_k, \Sigma_k) = \frac{1}{2\pi^{d/2}|\sum|^{1/2}} \exp\left(-\frac{1}{2}(o-\mu)^t \Sigma^{-1}(o-\mu)\right)$$

2.3.2.2.1.4. Flux de données

Les vecteurs de paramètres extraits du signal de parole contiennent souvent des données non corrélés.

Il est possible de découper les vecteurs d'observation en plusieurs groupes de coefficients appelés des flux de données et de modéliser ces flux comme des observations indépendantes. La probabilité d'émission pour l'observation complète $o_t = (o_t{}^1, o_t{}^2, \dots \dots o_t{}^f)$ composée de f flux est le produit de la probabilité d'émission pour chacun des flux:

$$b_j(o_t) = \prod_{k=1}^{f} b_j^k\,(o_j^k) \tag{2.10}$$

2.3.2.2.1.5. Les problèmes fondamentaux des modèles

A partir d'une suite d'observations O (suite de vecteurs acoustiques) supposées émises par un modèle, nous sommes amenés à traiter les trois problèmes **[28][29]**.

a- Problème 1

Reconnaissance : c'est l'évaluation de la probabilité $P(O/M)$ que la suite des observations à été émis par un modèle. Lorsque plusieurs modèles existent, cette évaluation permet le choix du modèle le plus probable.

b- Problème 2

Evaluation de la Probabilité d'émission d'une séquence (recherche de la séquence d'état la plus probable), la recherche de la séquence d'états d'un modèle ayant produit les observations. La séquence cachée de plus forte probabilité est déterminée par l'algorithme de Viterbi.

c- Problème 3

Apprentissage des paramètres d'un modèle $a_{ij}, b_j(o_t)$, comment calculer ou plutôt estimé les paramètres du modèle (probabilité de transition a_{ij} probabilité d'émission $b_j(o_t)$) pour maximiser la vraisemblance des observations.

2.3.2.2.2. Application

Le système de reconnaissance automatique de la parole pour les mots isolés est construit par trois étapes que nous allons expliquer en détails.

Soit $W = (\lambda_1, \lambda_2, \ldots . \lambda_n)$,un vocabulaire de n mots λ_i pour chaque mot de W, nous construisions un modèle de markov cachés de N états.

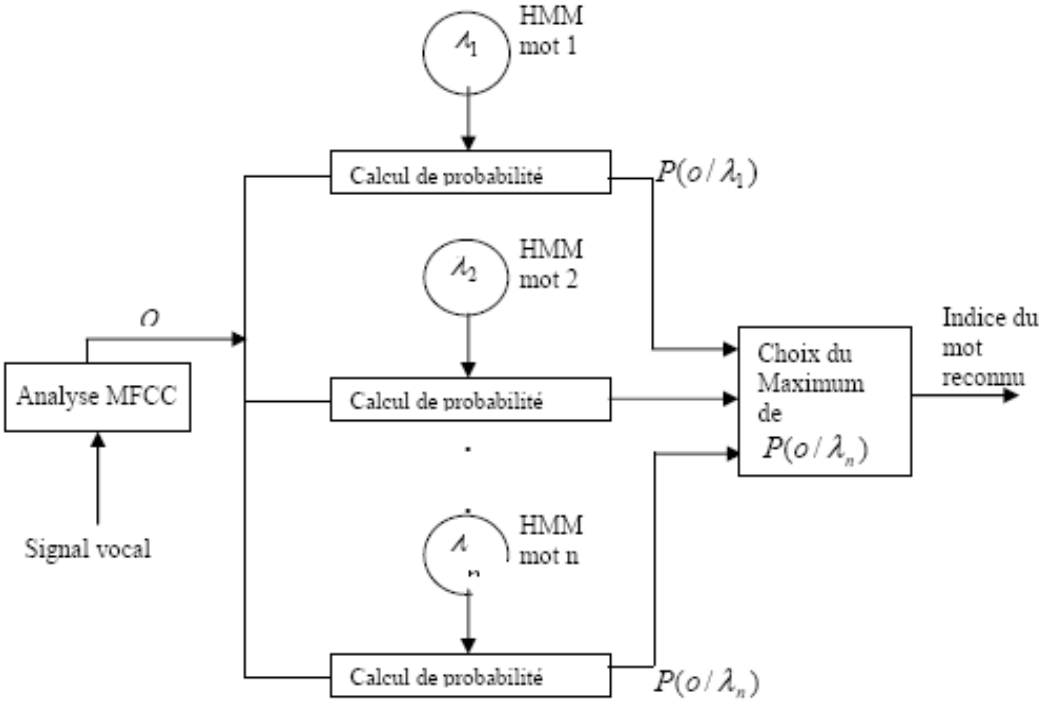


Figure 2.7 : Système de RAP pour les mos isolés.

Le signal de parole pour un mot donné est représenté dans le temps par une suite des vecteurs acoustiques.

Le système de reconnaissance de la parole pour les mots isoles est construit en trois étapes:

- **Etape 1** : Construction de modèle individuel pour chaque mot. Cette étape est réalisée en utilisant la solution du problème 3 pour estimer d'une façon optimale, les paramètres du modèle de chaque mot.
- **Etape 2 :** recherche la séquence optimale des états suivant la séquence d'observation correspondante, on utilise un critère convenable. La solution du problème 2 est utilisée.
- **Etape 3 :** Une fois les n modèles HMM construits et optimises, la reconnaissance du mot inconnu est effectuée en utilisant la solution du problème 1 pour évaluer le modèle du

chaque mot en se basant sur la suite d'observations courantes et de sélectionner le modèle qui génère le meilleur score.

2.3.2.2.3. Solution des problèmes HMMs

❖ **Reconnaissance**

Etant donnés une suite d'observations O=($o_1 o_2, \dots . o_{m-1} o_m$) et un modèle, comment on peut calculer efficacement la probabilité (vraisemblance) que la suite d'observations O soit produite par le modèle M , c'est-à-dire $P(O/M)$. Autrement dit, comment évaluer le modèle afin de choisir parmi plusieurs modèles celui qui génère le mieux la suite d'observations. Plusieurs techniques sont utilisé pour résoudre ce problème : méthode d'évaluation directe, procédure « Forward-Backward » et Algorithme de Viterbi**[29][30]**.

a- Evaluation directe

La probabilité $P(O/M)$ d'une suite d'observations O=($o_1 o_2, \dots . o_{m-1} o_m$) sachant le modèle $M = \{\pi, A, B\}$, est la somme sur tous les chemins de la probabilité que le chemin $Q = (q_1, q_2, \dots . q_T)$ en cours est généré l'observation.

$$P(O/M) = \sum_Q P\left(O/Q, M\right) P\left(Q/M\right) \tag{2.11}$$

Définition de la probabilité d'emprunter le chemin Q:

$$P\left(Q/M\right) = P(s_1 s_2, \dots \dots, s_n/M) = \pi_1 a_{1,2} a_{2,3} \dots \dots a_{n-1,n} \tag{2.12}$$

la probabilité que cette séquence d'états Q émette les signaux observations O:

$$P\left(O/Q, M\right) = P\left(O/s_1 s_2, \dots \dots, s_n, M\right) = b_1(o_1) b_2(o_2) b_3(o_3) \dots \dots b_n(o_n) \tag{2.13}$$

On obtient :

$$P(O/M) = \sum_Q \pi_1 a_{1,2} b_1(o_1) a_{2,3} b_2(o_2) \dots \dots a_{n-1,n} b_n(o_n) \tag{2.14}$$

Ceci générera $(2T-1)N^T$ multiplication et $N^T - 1$ additions, soit environ $2TN^T$ opérations.

On constate que de nombreuses multiplications sont répétées. L'idée est de calculer $P(O/M)$ de manière incrémentale (Forward-Backward).

b- Algorithme Forward-Backward

Dans cet algorithme, on considère que l'observation peut se faire en deux étapes :

1. L'émission de la suite d'observations $(o_1 o_2, \ldots . o_{t-1} o_t)$ et la réalisation de l'état q_t au temps t (**Forward) [28].**
2. L'émission de la suite d'observations $(o_{t+1} o_{t+2}, \ldots . o_{T-1} o_T)$ en partant de l'état q_i au temps t (**Backward) [28].**

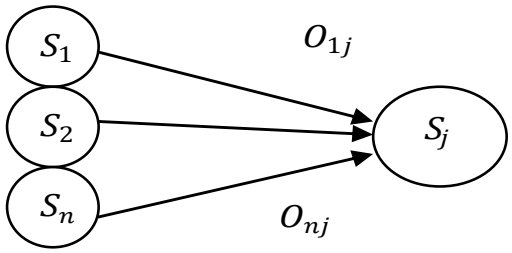
$P(O/M)$ peut être défini à chaque instant $t \in [1, T]$ par :

$$P(O/M) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \tag{2.15}$$

Ou $\alpha_t(i)$ est la probabilité d'émettre la suite $(o_1 o_2, \ldots . o_{t-1} o_t)$ et d'aboutir à l'état q_t à l'instant t et $\beta_t(i)$ la probabilité d'émettre la suite $(o_{t+1} o_{t+2}, \ldots . o_{T-1} o_T)$ en partant de l'état q_t .Le calcul de $\alpha_t(i)$ se fait avec t croissant tandis que celui de $\beta_t(i)$ se fait avec t décroissant, d'où l'expression Forward-Backward.

c- La variable Forward

Soit la probabilité $\alpha_t(i) = P\left(O, q_t = s_i / M\right)$ de générer $O = (o_1 o_2, \ldots . o_{t-1} o_t)$ et de se trouver dans l'état q_t à l'instant.



Forward

$\alpha_t(i)$ ⟶ $\alpha_{t+1}(j)$

Cet algorithme permet de calculer cette probabilité :

1. Initialisation : $\alpha_1(i) = \pi_1 * \beta_i(o_1)$
2. Itération :

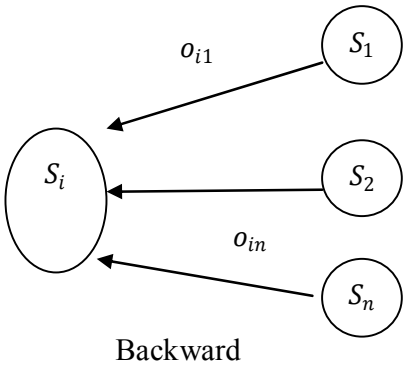
$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] * \beta_j(o_{t+1}) \quad \text{Pour } t \in [1, T-1], j \in [1, N]$$

3. $$P(O/M) = \sum_{i=1}^{N} \alpha_T(i)$$

Ceci générera $N(N+1)(T-1)+N$ multiplications et $N(N+1)(T-1)$ additions, soit environ $2*N^2 * T$ opérations.

d- La variable Backward

Soit la probabilité $\beta_t(i) = P\left(O/q_t = s_i, M\right)$ de générer $O\,(o_{t+1}o_{t+2}, \ldots . o_{T-1}o_T)$ en sachant que l'on est dans l'état q_t à l'instant t.



Backward

$\beta_t(i)$ $\beta_{t+1}(j)$

Cet algorithme est donné par :

1. Initialisation : $\beta_T(i) = 1$, $1 \leq i \leq N$
2. Itération :

$$\beta_t(j) = \left[\sum_{i=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)\right] \text{ Pour } t \in [T-1,1], i \in [1,N]$$

Pour être dans l'état q_t à l'instant t tout en tenant compte de la suite d'observations de $O = (o_{t+1}o_{t+2}, \ldots . o_{T-1}o_T)$ il faut considérer tous les états possibles S_j (toutes les transitions a_{ij}) et l'observation o_{t+1} dans l'état j (les $b_j(o_{t+1})$), puis la suite d'observations partielle restante à partir de l'état $j(\beta_{t+1}(j))$.

Ceci générera $N(N+1)(T-1)+N$ multiplications et $N(N+1)(T-1)$ additions, soit environ $2*N^2 * T$ opérations.

- **Evaluation de la Probabilité**

Etant donnés une suite d'observations $O = (o_1o_2, \ldots . o_{t-1}o_t)$ et un modèle $M = \{\pi, A, B\}$ comment peut-on choisir une suite des états $Q = (q_1q_2, \ldots . q_{T-1}q_T)$, qui soit optimale selon un critère convenable. La difficulté réside dans la suite optimale des états, il existe plusieurs algorithmes utilisés : le critère local, le critère global et l'algorithme de Viterbi.

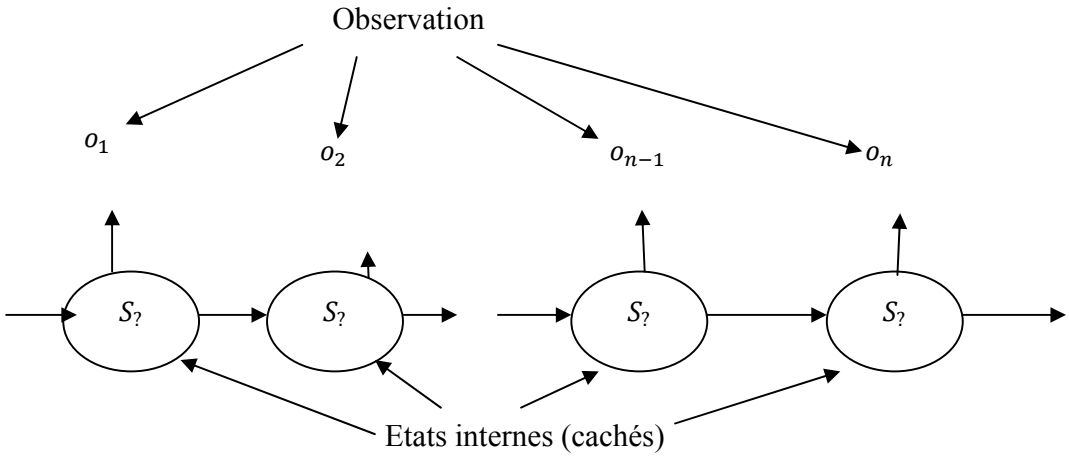


Figure 2.8: La séquence d'observation et la séquence des états.

D'un point de vue mathématique on cherche la séquence qui a la plus grande probabilité d'avoir généré O parmi N^M possibilités.

a- Critère local

Cette technique consiste à choisir l'état q_t qui est le plus probable et ceci indépendamment des autres états individuellement pour chaque t. Soit la probabilité $\gamma_t(i)$ d'être dans l'état q_t au temps t.

$$\gamma_t(i) = P\left(q_t = s_i / O_t, M\right) \text{ alors : } \gamma_t(i) = \frac{P\left(q_t = s_i / O_t, M\right)}{P\left(O / M\right)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \qquad (2.16)$$

Note : on peut calculer γ_t une fois α_t et β_t calculée, tel que $\sum_{i=1}^{m} \gamma_t(i) = 1$.
L'état optimal est donné par :

$$q_t = argmax[\gamma_t(i)] \text{ pour tout } t \in [1, T], i \in [1, N]$$

Cette méthode n'est pas viable car rien ne garantit les transitions entre chaque état de q_t non valides (transition interdit).

b- Critère global

On cherche à trouver le trajectoire optimale de la suite d'états, donc de maximiser

$$P\left(Q/_{O,M}\right)$$

ce qui revient au maximiser :

$$P\left(O,Q/_{M}\right), \text{ car } P\left(O,Q/_{M}\right) = P\left(Q/_{O,M}\right) * P(O/_{M}) \tag{2.17}$$

Pour cela, on peut définir la probabilité maximale d'une séquence au temps t qui se termine dans l'état q_t .

La formule de critère globale :

$$\delta_t(j) = \max_{q_1 q_2, \ldots . q_{t-1}} p(q_1 q_2, \ldots . q_t = s_i, o_2, \ldots . o_t / M)$$

$$\text{D'où } \delta_{t+1}(j) = \left[max\delta_t(j)a_{ij}\right] * b_j(o_{t+1}) \tag{2.18}$$

En conservant pour chaque *t* et chaque *i* l'état ayant amené au maximum : $\delta_t(j), \varphi_t(j)$ on obtient l'algorithme de Viterbi.

c- Algorithme de Viterbi

Cette solution est la plus utilisée. Elle est basée sur la technique de programmation dynamique. C'est un algorithme récursif qui permet de trouver à partir d'une suite 'observations, une solution optimale au problème d'estimation de la suite d'états.

Initialisation :

for $1 \leq i \leq N$

$$\delta_t(j) = \pi_i * b_i(o_1)$$

$$\varphi_1(i) = 0$$

Calcule itératif : for $2 \leq t \leq T$

for $1 \leq j \leq N$

$$\delta_t(j) = \left[max\ \delta_{t-1}(i)a_{ij}\right] * b_i(o_t)$$

$$1 \leq i \leq N$$

$$\varphi_t(j) = \text{argmax}\ [\delta_{t-1}(i)a_{ij}]$$

$$1 \leq i \leq N$$

Terminaison : $P^* = [max\ [\delta_T(i)]$

$1 \leq i \leq N$

$q_T^* = \arg\max \delta_{t-1}(i)a_{ij}$

$1 \leq i \leq N$

Backtracking: for t=T-1 to 1

$$q_t^* = \varphi_t(q_{t+1}^*)$$

End

❖ **Apprentissage**

Comment peut-on ajuster les paramètres du modèle $M = \{\pi, A, B\}$ pour maximiser $p(O/M)$?

Le fait que la longueur de la suite d'observations (données d'apprentissage) soit finie, il n'existe pas de solutions analytiques directes (d'optimisation globale) pour construire le modèle. Néanmoins, nous pouvons choisir $M = \{\pi, A, B\}$, la probabilité $p(O/M)$ est maximum local en utilisant une procédure itérative telle que celle de Baum-Welch.

L'idée de l'application est donc d'utiliser des procédures de ré-estimation qui affinent le modèle, petit à petit, selon les étapes suivantes :

- Choisir un ensemble initial des paramètres M_0.
- Calculer M_1à partir de M_0.
- Répéter ce processus jusqu'à un critère de fin.

a- Algorithme Baum-Welch

Probabilité de passer d'un état à états suivent en générant O :

$$\xi_t(i,j) = p(^{q_t = s_i, q_{t+1} = s_j}/_{O,M}) \quad (2.19)$$

Les variables backward et forward nous permettent d'écrire :

$$\xi_t(i,j) = \frac{p(^{q_t=s_i, q_{t+1}=s_j}/_{O,M})}{p(^{O}/_{M})} \quad (2.20)$$

D'où :

$$\xi_t(i,j) = \frac{\alpha_t(i) * a_{ij} * b_j(o_{t+1}) * \beta_{t+1}(j)}{p(^{O}/_{M})} \quad (2.21)$$

$\gamma_t(i)$ est le nombre espéré de transitions de s_i vers s_j, on peut écrire le nombre espéré de transitions par :

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) = \frac{\alpha_t(i)\beta_t(i)}{p(^{O}/_{M})} \text{ pour } t \in [1, T-1] \quad (2.22)$$

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \zeta_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \le i \le N, \quad 1 \le j \le N \quad (2.23)$$

$$\widehat{b}_j(k) = \frac{\sum_{t=1,\ x_t = V_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \quad 1 \le j \le N, \quad 1 \le k \le M \quad (2.24)$$

Cette méthode de Maximum de Vraisemblance est la plus utilisée dans les applications apprentissage.

L'algorithme **Baum-Welch** peut être représenté sous la forme itérative suivante :

- Initialisation de modèle :$k = 0$

  $M = \{\pi_0, A_0, B_0\}$ telque $A_0 = a_{ij}^0, B_0 = b_j^0(o_t)$

- Calcule, en utilisant les variables Forward Backward :

  $\xi_t(i,j)$ et $\gamma_t(i)$ pour $i \in [1,N]$ et $j \in [1,N]$

- Nouvelle estimation $k = 1,2,3, \dots \dots . K$

$M_k = \{\pi_k A_k B_k\}$ telque $A_k = a_{ij}^0, B_k = b_j^0(o_t)$ pour $i \in [1, N]$ et $j \in [1, N]$

Le test d'arrêt, est généralement un nombre d'itérations qui est fixé à l' avance. Le choix d'un modèle initial influe sur les résultats des paramètres, par exemple toutes les valeurs nulles de A et de B au départ, restent à zéro jusqu'à la fin de l'apprentissage.

Il est noté que l'algorithme converge vers des valeurs de paramètres qui forment un point critique de $P(^{O_t}/_{M^k})$. Nous obtenons un maximum local ou un point d'inflexion. Donc la nécessité de bien choisir le modèle initial.

Pour avoir une estimation convenable du modèle, les ré-estimations se font sur un ensemble de plusieurs suites d'observations appelées corpus d'apprentissage. Donc la taille du corpus d'apprentissage influe sur les résultats d'apprentissage.

Une fois le type de méthode de la reconnaissance sélectionné, viennent ensuite s'ajouter ce que l'on appelle les facteurs de complexité, qui dépendent du type d'application visée :

- ✓ reconnaissance mono-locuteur ou multi-locuteurs.
- ✓ reconnaissance de mots isolés ou de parole continue.
- ✓ reconnaissance d'un vocabulaire restreint ou large.
- ✓ reconnaissance en environnement calme ou bruité.

2.4. Les différentes architectures de reconnaissance dans les systèmes mobiles

Nous avons trois architectures différentes pour la RAP dans les environnements mobiles :

- ➢ La reconnaissance de la parole dans le réseau téléphonique.
- ➢ La reconnaissance de la parole dans le téléphone (système embarquée).
- ➢ La reconnaissance de la parole distribuée (répartie).

2.4.1. La reconnaissance de la parole dans le réseau téléphonique NSR (Network Speech Recognition)

Le principe de la **RAP NSR** consiste à se servir du terminal uniquement (téléphone portable) comme un système d'acquisition de la parole. Une fois le signal enregistré, il est transmis via un réseau, généralement sans fil, à un serveur qui effectue le travail de reconnaissance **[31].**

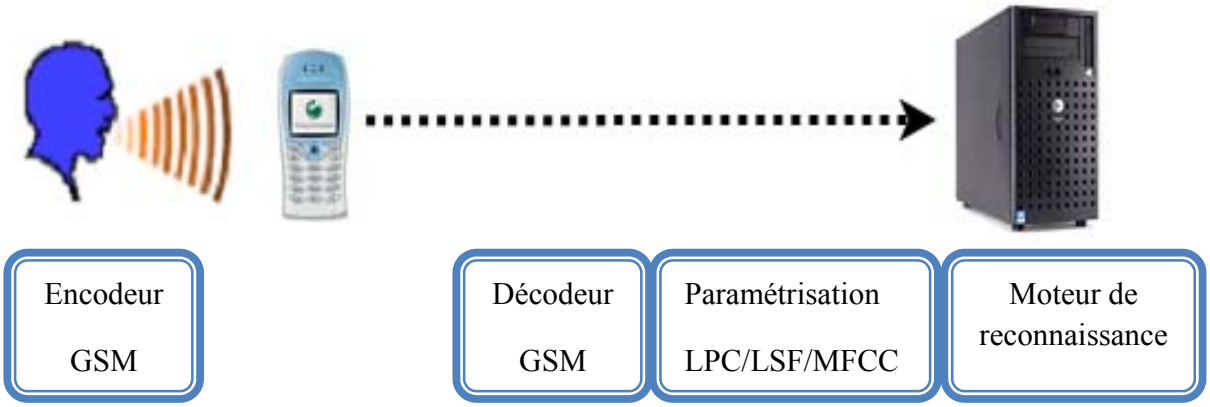


Figure **2.9**: Architecture du système **NSR**.

Cette approche permet d'envisager l'utilisation des 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 de la parole telle qu'un système de dialogue ou encore des vocabulaires dédiés à chaque application. Pour résumer, l'avantage majeur de cette architecture est la levée des contraintes liées aux ressources limitées de système téléphonique.

Les inconvénients de ce type d'architecture est la perte d'informations due a la transmission du signal (effet de codage source, effet sur le canal de transmission). En effet, avant d'être envoyé au serveur de reconnaissance, le signal doit être codé par le codeur GSM. La transmission du signal au serveur et du serveur au terminal occupe toujours le canal de transmission.

2.4.2. La reconnaissance embarquée (Embedded Speech Recognition) ESR

Avec cette approche, toutes les étapes de la reconnaissance sont effectuées dans le système embarqué (le téléphone). Aucun serveur à distance est nécessaire, le système est donc entièrement autonome, il réalise lui-même l'acquisition de la parole puis le décodage **[32]**.

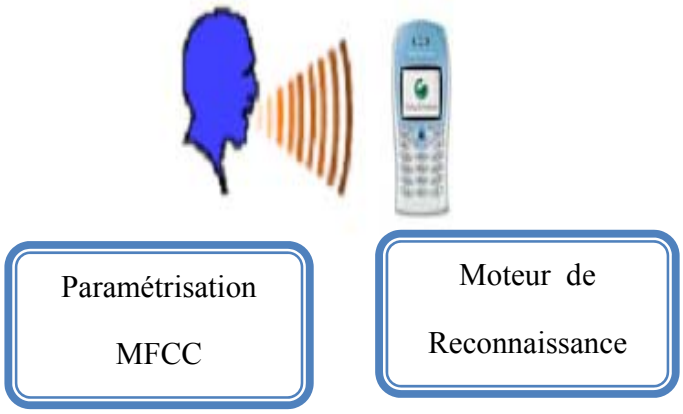


Figure **2.10**: Architecture du système embarquée.

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 GSM (FR, HR, EFR). De plus, les inconvénients de la reconnaissance **NSR** la perte d'informations due à la transmission et le codage bas débit sont évités. La solution présente toutefois un inconvénient majeur : les ressources nécessaires. En effet, toutes les phases paramétrisation du signal de la parole et la reconnaissance sont effectuées au sein du téléphone. Les différentes phases de la reconnaissance nécessitent beaucoup de ressource comparées aux capacités d'un téléphone portable. De plus, l'augmentation des ressources d'un téléphone (mémoire et puissance de calcul), bien que possible, mais elle reste très couteuse.

2.4.3. La reconnaissance distribuée DSR (Distributed Speech Recognition)

Ce mode de reconnaissance se situe entre les deux architectes précédentes **RAP NSR** et **RAP ESR**. En effet, une partie du travail est effectuée sur le système embarqué et une autre partie sur le serveur à distance **[32][33]**.

Généralement, la phase d'extraction des paramètres est réalisée sur le terminal et la partie reconnaissance se trouve sur le serveur.

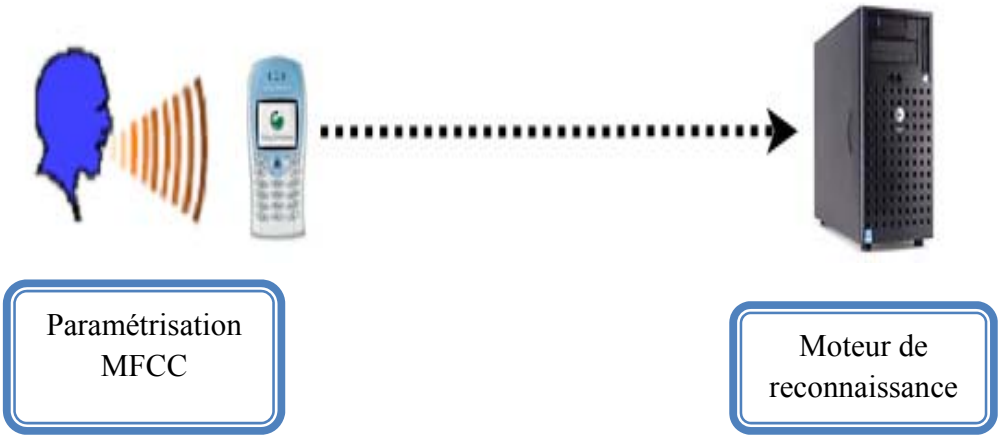


Figure **2.11** : Architecture du système répartie.

Cette approche permet de contourner les problèmes liés à la perte d'informations due au transport. Avec cette architecture, une seule phase (extraction des paramètres de reconnaissance) est nécessaire : les deux premières étapes (codage et décodage) deviennent inutiles. Les pertes engendrées par l'encodage effectué sur le téléphone sont alors évitées (GSM EFR).

Cette approche implique, par ailleurs, que le téléphone s'accorde avec l'opérateur désirant fournir le service de reconnaissance vocal, étant donné que l'extraction des paramètres est effectuée directement dans le téléphone mobile. Ces paramètres sont ensuite utilisés par le système de reconnaissance durant la phase de décodage.

2.5. Conclusion

Dans ce chapitre nous avons présenté les différentes approches de reconnaissance automatique de la parole. Nous avons en particulier insisté sur l'approche probabiliste basée sur les modèles de Markov cachées. Cette dernière approche est la plus utilisée actuellement pour l'efficacité de ses principaux algorithmes de modélisation et de reconnaissance, ces algorithmes sont à la base de son succès. Nous avons également présenté les différentes architectures de reconnaissance en cours actuellement à savoir l'architecture NSR, ESR et DSR.

Chapitre 3
Analyse et extraction des paramètres

3.1. Introduction

Le signal de la parole est trop redondant et de caractère non stationnaire, il est donc complexe et très difficile à utiliser directement dans un système de reconnaissance automatique de la parole (RAP). Le but essentiel du traitement du signal de la parole est de lui substituer une représentation moins redondante avec une extraction des paramètres pertinents pour la reconnaissance automatique de la parole. La plupart des techniques de paramétrisation consistent à décrire l'enveloppe du spectre à court terme dans le domaine fréquentiel. Dans ce chapitre, nous décrivons les principales techniques utilisées en analyse du signal de la parole en vue de la reconnaissance.

3.2. Prétraitement du signal vocal

3.2.1. Numérisation (Echantillonnage et Quantification)

Pour être utilisable par la machine, le signal doit tout d'abord être numérisé. Cette opération tend à transformer un signal analogique en une suite d'éléments discrets des échantillons. Ceux-ci sont obtenus avec une carte spécialisée courante de nos jours dans les ordinateurs depuis l'avènement du multimédia. La numérisation de signal de la parole repose sur deux étapes : l'échantillonnage et la quantification.

L'échantillonnage effectué à une fréquence f_e (Shannon $f_e \geq 2f_m$) compatible avec la capacité de stockage et de calcul. La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore à chaque pas de l'échantillonnage.

3.2.2. Fenêtrage

Le signal de la parole est un processus aléatoire et non stationnaire. On effectue généralement le traitement sur des tranches de la durée de quelques millisecondes (de 10 à 30 ms). Dans cet intervalle, on peut considérer que les mouvements du conduit vocal sont négligeables (la fonction de transfert est constante pendant cette tranche de temps) et donc le signal peut être considérer comme stationnaire. Le fenêtrage consiste alors à pondérer (multiplier) le signal par une fenêtre de pondération, on effectué l'analyse sur des échantillons $S_N(n)$ définis comme suit :

$$S_N(n) = S(n)W(n) \tag{3.1}$$

$W(n)$: La fenêtre de pondération.

$S(n)$: Les échantillons du signal.

$S_N(n)$: Les échantillons multipliés par la fenêtre.

Il existe différents types de fenêtres, nous pouvons citer :

a- Fenêtre Rectangulaire

$$W(k) = \begin{cases} 1 & |k| \leq \frac{N}{2} \\ 0 & ailleurs \end{cases} \quad (3.2)$$

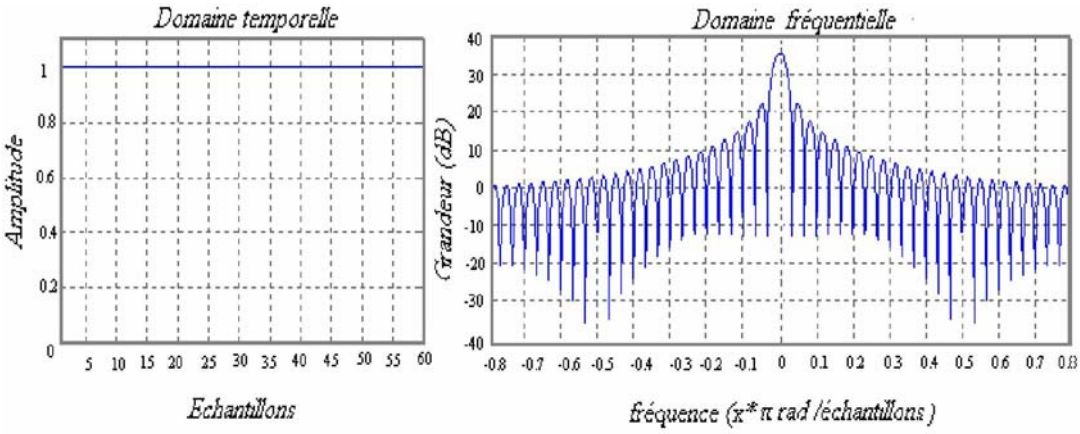


Figure 3.1 : Fenêtre rectangulaire.

b- Fenêtre Triangulaire

$$W(k) = \begin{cases} 1 - \frac{2*|k|}{N} & |k| \leq \frac{N}{2} \\ 0 & ailleurs \end{cases} \quad (3.3)$$

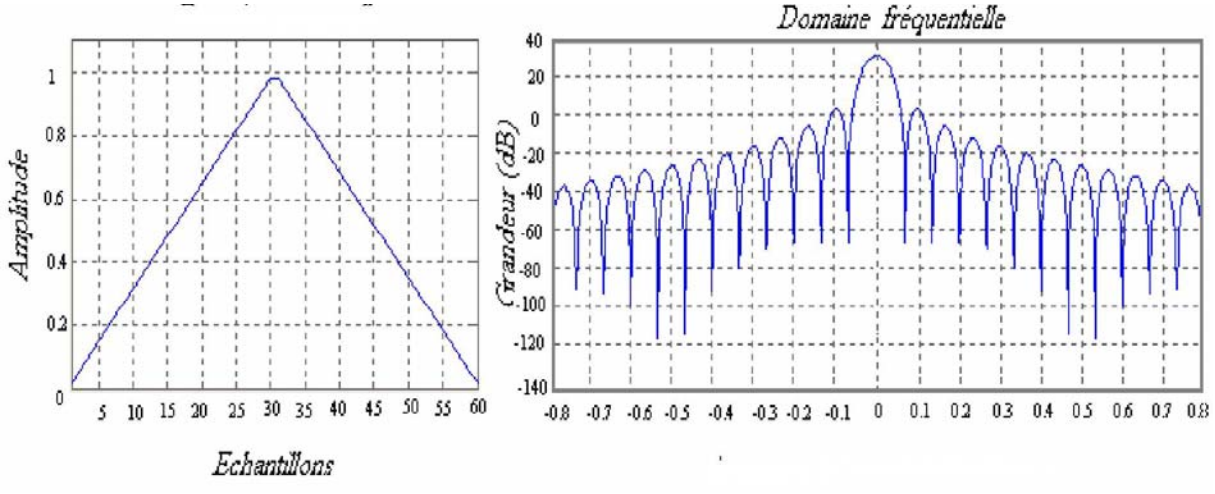


Figure 3.2 : Fenêtre triangulaire.

c- Fenêtre de Hamming

$$W(k) = 0.54 - 0.46\,cos(2x\,k/N) \qquad k = 0.1. \ldots \ldots N-1 \tag{3.4}$$

N : La durée de la fenêtre est donnes par : $N = fe * S$ (S la durée de la trame).

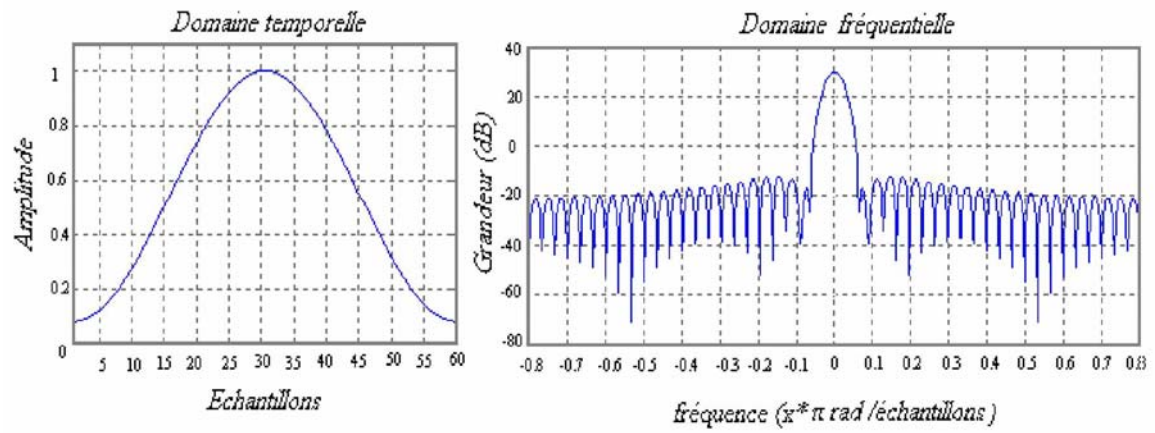


Figure 3.3 : Fenêtre de Hamming.

Nous découpons le signal en fenêtres de 10 à 30 ms en prenant soin que ces fenêtres se chevauchent pour éviter l'apparition de discontinuités à la synthèse (un recouvrement à moitié de fenêtre est généralement utilisé).

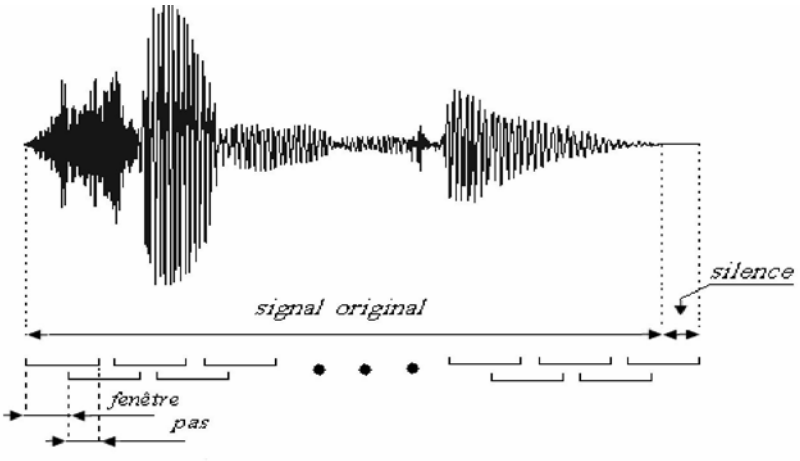


Figure 3.4 : Principe du découpage en fenêtres.

3.3. L'extraction des paramètres

Le signal de la parole véhicule plusieurs types d'informations des informations sur la prosodie, le timbre,…etc. Pour répondre aux besoins des nombreuses applications du traitement de la parole telles que le codage, la synthèse ou la reconnaissance automatique de la parole, il existe un très grand nombre d'outils d'analyse allant du simple calcul d'énergie jusqu'à la modélisation complète du conduit vocal. Au cours de cette partie, nous détaillerons principalement les outils nécessaires à la conception de notre système de reconnaissance automatique de la parole. Dans le système RAP nous avons besoin d'une représentation du signal de la parole par un nombre réduit de paramètres pertinents, robuste et non corrélés entre eux. Les méthodes les plus courantes sont les analyses spectrales réalisées soit par transformée de Fourier, soit par les Coefficients Prédiction Linéaire LPC (Linear Predictive Coding) ou par évaluation des coefficients cepstraux MFCC (Mel-Frequency Cepstrum Coefficients) et PLP (Peceptual Linear Precdictive).

Dans ce travail nous avons utilisé deux variantes de paramétrisation. La première variante parole transcodée, elle est basée sur les coefficients standards MFCC (Mel Frequency Cepstral Coefficients). Ces derniers sont largement utilisés, et donnent une bonne performance des systèmes de Reconnaissance Automatique de la Parole (RAP). La deuxième variante bit-stream utilise les coefficients LSF (Line Spectral Frequencies) décodés directement du flux des bits. Ces coefficients sont dérivés de coefficients LPC pour représenté la contribution spectral de modèle ACELP.

3.3.1. Variante parole transcodée

3.3.1.1. Les Coefficients Cepstraux de type MFCC

L'analyse par les coefficients MFCC consiste en l'évaluation des paramètres Cepstraux à partir d'une répartition fréquentielle dans l'échelle de Mel.

Le calcul de ces coefficients repose sur une suite de transformations de base du traitement de signal qui sont schématisées dans la figure suivante **[53]**:

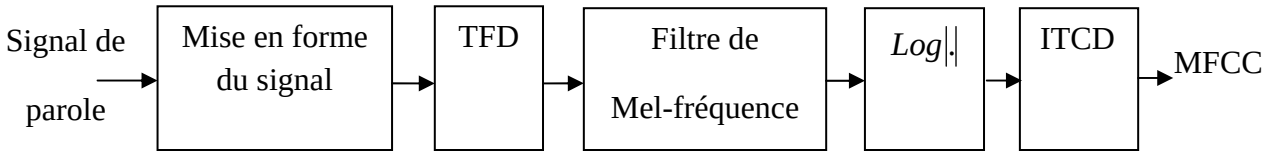


Figure 3.5.a : Schéma bloc d'une extraction de coefficients MFCC.

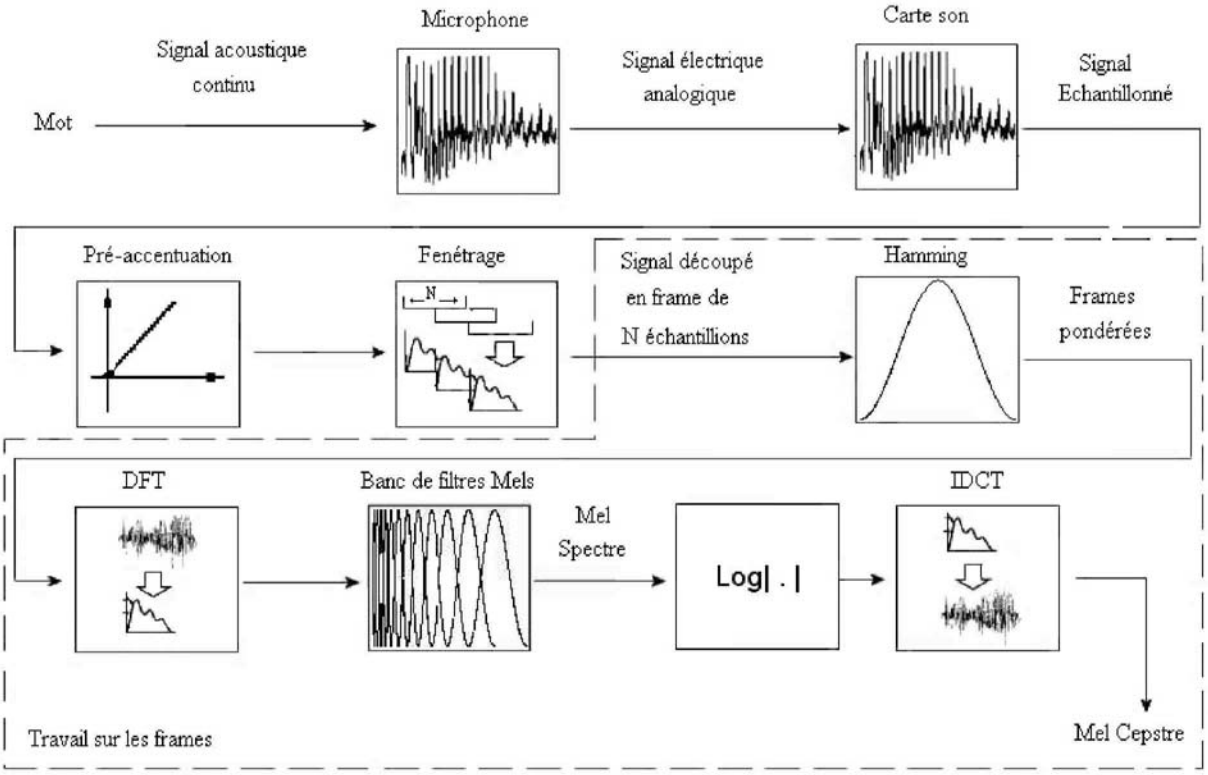


Figure 3.5.b : Les différentes étapes d'extraction des coefficients MFCC.

3.3.1.2.Préaccentuation et fenêtrage

Au niveau des lèvres, une baisse d'énergie et une distorsion de l'onde sonore due à une désaccentuation sur tout le spectre (-6dB/octave).

La préaccentuation consiste en le passage du signal de parole dans un filtre passe-haut dont la fonction de transfert est:

$$H(Z) = 1 - aZ^{-1} \qquad (0.9 \leq a \leq 1) \tag{3.5}$$

a est proche de zéro pour les sons non voisés et de l'unité pour les sons voisés (généralement une valeur de a =0.95 est utilisée). Le filtre de la préaccentuation a pour effet d'accentuer la partie haute fréquence du spectre.

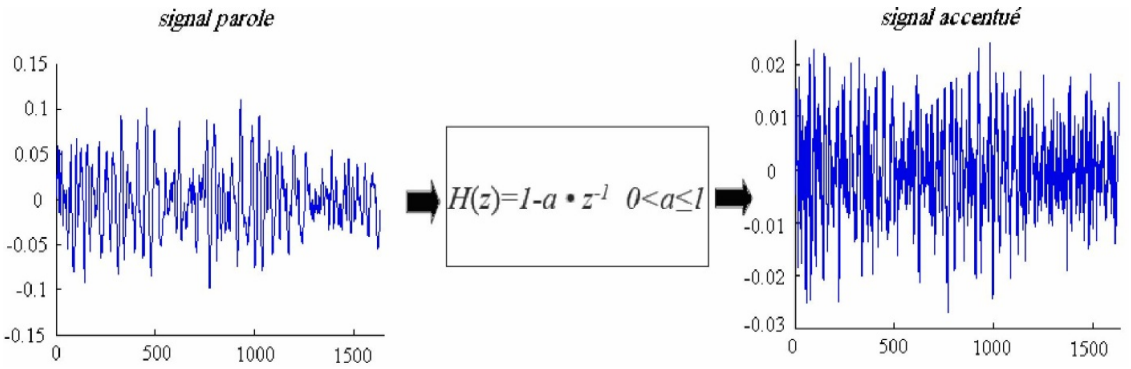


Figure .3.6. Effet de la préaccentuation sur le signal de la parole.

La seconde étape consiste à découper le signal en trames. Ceci est réalisé en multipliant le signal par la fenêtre de pondération. Nous avons utilisé la fenêtre de Hamming donnée par la formule suivante **[52]**:

$$W(n) = 0.54 - 0.46 * \cos\left(2 * \pi * \frac{n}{N-1}\right);\ 0 \leq n \leq N - 1 \tag{3.6}$$

Une transformée de Fourier rapide permet ensuite de passer du domaine temporel au domaine fréquentiel. On obtient le spectrogramme du signal.

La déformation en échelle fréquentielle Mel est à présent calculée afin de prendre en compte la perception de l'oreille humaine **[54].**

$$\text{mel}(f) = 2595 * \log10(1 + f/700) \quad (3.7)$$

Cette transformation est réalisée grâce à l'utilisation d'une banque de filtres pour chaque fréquence Mel désirée. Chaque filtre passe-bande a une réponse en fréquence triangulaire.

La dernière étape a pour but de ramener les coefficients dans le domaine temporel au moyen d'un transformée en cosinus discrète (DCT).

3.3.1.3. La Transformée de Fourier Discrète DFT

Pour passer du domaine temporel au domaine fréquentiel on utilise la transformée de Fourier. La transformation de Fourier peut être considérée comme le produit d'une matrice, appelée la matrice de la transformation par un vecteur formé des échantillons de signal de la parole.

3.3.1.4. L'échelle de Mel

Afin de prendre en compte la perception de l'oreille humaine il faut calculer la déformation en échelle fréquentielle Mel. **[31](Figure. 3.7).**

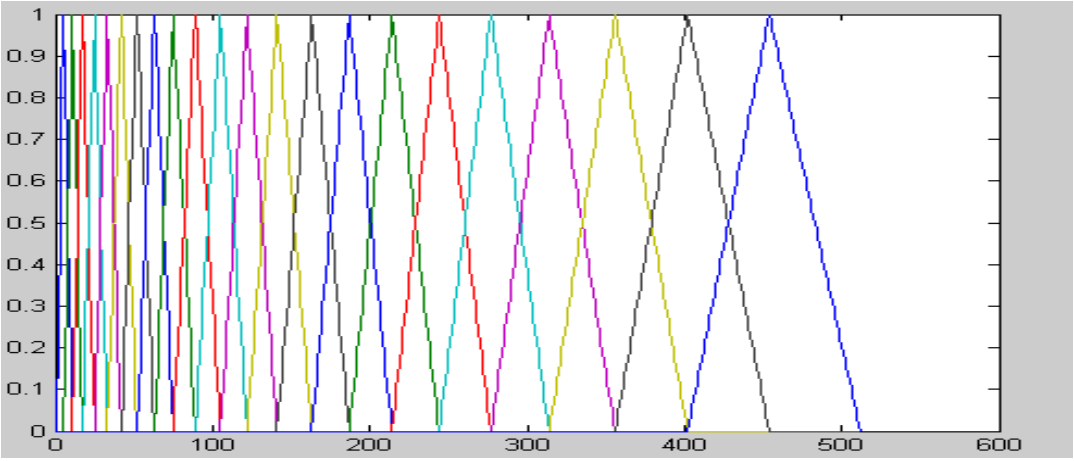


Figure 3.7 : Banc de filtres Mel.

Les paramètres MFCC sont basés sur la connaissance de la variation des bandes critiques de l'oreille humaine en fonction de la fréquence. En effet, la majeure partie de l'information du signal de parole se trouve dans les basses fréquences, les filtres sont donc espacés linéairement en-dessous de 1000Hz et de manière logarithmique (Echelle de Mel) au-

dessus de 1000Hz. L'échelle de Mel est une échelle biologique. C'est une modélisation de l'oreille humaine. A noter que le cerveau effectue en quelque sorte une reconnaissance vocale complexe avec filtrage de sons. Prenons l'exemple suivant, où vous êtes à table avec compagnie de nombreuses personnes, l'ensemble de ses personnes parle en même temps et vous discutez avec votre voisin. Malgré le bruit, vous arrivez à discerner clairement ce que vous dit votre voisin, vous ignorez de façon naturelle le bruit de fond et vous amplifier le son qui vous paraît le plus important. Vous pouvez répéter cette expérience avec chacun des convives. Le cerveau ne se contente non pas seulement de filtrer les sons et de les amplifier mais aussi de prédire. Prenons l'exemple suivant où une personne discute avec vous par un volume sonore très bas, vous n'avez pas entendue une certaine partie de la phrase mais vous arrivez à la reconstituer et à la comprendre. A partir de l'étude du cerveau nous pouvons faire une idée de la complexité de la reconnaissance vocale et nous pouvons rapprocher d'un modèle de plus en plus puissant. Par la conception de Mel on considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence.

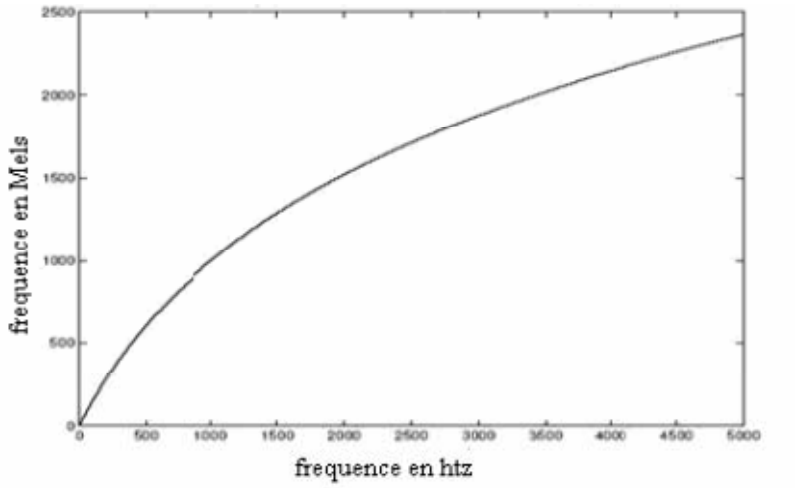


Figure 3.8 : Graphe de conversion de la fréquence Hz-Mel.

On remarque qu'avant 1000 Hz, la courbe est à peu près droite, ce qui traduit bien l'équivalence entre Hz et Mels.

3.3.1.5. Transformée de Cosinus Discrète (DCT : Discrete Cosine Transform)

La dernière étape a pour calculer les coefficients MFCC est de ramener les valeurs des énergies des filtres dans le domaine temporel au moyen d'une transformée en cosinus discrète (DCT).

On peut résumer les étapes de calcul les coefficients cepstraux dans la relation suivante **[55]:**

$$c(n) = \sum_{j=1}^{M} \log S(j) \cos\left(\frac{n\pi}{M}\left(m - \frac{1}{2}\right)\right) \qquad n = 0,1, \dots \dots \ N \tag{3.8}$$

$S(j)$: est l'énergie du filtre Mel correspondant.

$c(n)$:Les coefficients MFCC.

M : nombres de filtres. M=24.

3.3.1.6. Motivation pour le choix des coefficients MFCC

Les représentations spectrales, autorégressives ou cepstrales du signal de la parole sont les plus usuelles mais d'autres représentations sont utilisées en traitement automatique de la parole comme les représentations temps-fréquence. Les représentations du signal basées sur la perception suscitent beaucoup d'espoirs. En effet, l'étude de la production des sons et de leur perception est complémentaire. Le système auditif est une référence naturelle à toute l'analyse acoustique devant servir au RAP. Les échelles fréquentielles Mel ou Bark issues d'études psycho-acoustiques ont déjà été intégrées avec succès dans les chaînes d'analyse.

L'utilisation des coefficients MFCC est ainsi donc motivée par :

- Les échelles fréquentielles Mel ou Bark issues des études sur la perception sont préférables à une échelle linéaire, l'intégration des modèles d'audition plus complets dans les systèmes de reconnaissance automatique reste expérimentale.
- L'analyse cepstrale permet d'obtenir un nombre réduit des coefficients peu corrélés à partir des sorties d'un banc de filtres.
- Les coefficients différentiels permettent de prendre explicitement en compte la dynamique du signal dans les systèmes RAP.

3.3.2. La variante bit-stream

Dans cette variante au lieu de travailler sur le signal transcodé (reconstitué), le traitement se fait directement sur le flux de bits à la sortie du codeur. Ainsi, à partir du train de bits le décodage des coefficients LSFs (Line Spectral Frequencies) permet d'extraire plusieurs coefficients **[58][59].**

Le schéma ci-dessous représente la structure permettant d'obtenir les différents paramètres de la variante bit-stream. Ces paramètres sont calculés directement des coefficients de codage LSF et utilisés directement dans le système de reconnaissance.

Figure 3.9 : Extraction des paramètres bit-stream.

3.3.2.1. Calcul des coefficients Pseudo-Cepstral Coefficients (PCC)

Soient les deux polynômes *P(z)* et *Q(z)* donnés par les équations suivantes **[60]:**

$$p(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{3.9}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{3.10}$$

$$P(z)Q(z) = A_p^2(z)\left[1 - R^2(z)\right] = \left(1 - z^{-2}\right)\prod_{i=1}^{p}(1 - e^{jw_i} z^{-1})(1 - e^{-jw_i} z^{-1}) \tag{3.11}$$

Tels que w_i représentent les coefficients LSF.

$$R(z) = z^{-(p+1)} A_p\left(z^{-1}\right) \Big/ A_p(z) \tag{3.12}$$

Le logarithme de l'équation **(3.12)** permet d'obtenir l'équation suivante :

$$2\log A_p(z) + \log\left(1 - R^2(z)\right) = \log\left(1 - z^{-2}\right) + \sum_{i=1}^{p}\left(\log\left(1 - e^{jw_i} z^{-1}\right) + \log\left(1 - e^{-jw_i} z^{-1}\right)\right) \tag{3.13}$$

Sachant que :

$$\log A_p\left(e^{jw}\right) = -\sum_{n=1}^{\infty} C_n e^{-jnw} \tag{3.14}$$

$$-2\sum_{n=1}^{\infty} C_n e^{-jnw} + \sum_{n=1}^{\infty} R_n e^{-jnw} = -\sum_{n=1}^{\infty}\frac{1}{n}\left(1 + (-1)^n\right)e^{-jnw} - \sum_{n=1}^{\infty}\frac{1}{n}\sum_{i=1}^{p}\left(e^{jnw_i} + e^{-jnw_i}\right)e^{-jnw} \tag{3.15}$$

$$c_n = \frac{(1 + (-1)^n)}{2n} + \frac{\sum_{i=1}^{p}\cos(nw_i)}{n} + \frac{R_n}{2} \tag{3.16}$$

Les n^{ième} paramètres de PCC sont donnés par la formule suivante :

$$c_n^{PCC} = \frac{(1+(-1)^n)}{2n} + \frac{\sum_{i=1}^{p} cos\,(nw_i)}{n} \tag{3.17}$$

3.3.2.2. Calcul des coefficients Pseudo-CEPstrum (PCEP)

On peut extraire ces paramètres à partir des paramètres PCC d'une manière simple par élimination du terme $\frac{(1+(-1)^n)}{2n}$ **[59].**

La formule qui nous donne ces coefficients est ainsi :

$$d_n^{PCEP} = \frac{\sum_{i=1}^{p} \cos(nw_i)}{n} \tag{3.18}$$

3.3.2.3. Calcul des coefficients Mel-Frequency PCC (MPCC)

Les coefficients MPCC sont calculés par la même équation que les coefficients PCC. On transforme seulement les coefficients LSF en Mel-scaled, ils sont donnés par la formule suivante :

$$w_i^{Mel} = w_i + 2\tan^{-1}\left(\frac{\alpha \sin w_i}{1-\alpha\cos w_i}\right), 1 \le i \le M \tag{3.19}$$

Tel que α constant égale à 0.45.

Les coefficients MPCC sont donc calculés par l'équation suivante :

$$C_n^{MPCC} = \frac{(1+(-1)^n)}{2n} + \frac{\sum_{i=1}^{p} \cos(nw_i^{Mel})}{n} \tag{3.20}$$

3.3.2.4. Calcul des coefficients Mel-Frequency PCEP (MPCEP)

Les coefficients MPCEP (Mel Pseudo CEPstrum) sont calculés par la même procédure que les coefficients MPCC.

Les coefficients MPCEP (C_n^{MPCEP}) sont donc calculés par l'équation suivante :

$$C_n^{MPCEP} = \frac{\sum_{i=1}^{p} \cos(nw_i^{Mel})}{n} \tag{3.21}$$

3.3.2.5. Calcul des coefficients LPCC

La méthode d'obtention des coefficients LPCC (Linear Prediction Cepstral Coefficient) sont calculés après la conversion des coefficients LSF en coefficients LPC. Les coefficients cepstres (LPCC) sont calculés par une procédure récursive selon les formules suivantes:

$$\ln\left[1/A(z)\right] = \sum_{n=1}^{\infty} C_n z^{-n} \tag{3.22}$$

$$C_n = \left\{ -a_n - \frac{1}{n}\sum_{k=1}^{n-1} kC_k a_{n-k}, \ldots\ldots 1 \le n \le M \right\} \tag{3.23}$$

Généralement, l'ordre de la représentation cepstrale est plus grand que l'ordre de prédiction.

Ainsi, les coefficients LPCC n supérieur à m, sont calculés selon la formule (3.24) suivante :

$$C_n = \left\{ -\frac{1}{n}\sum_{k=1}^{M} kC_k a_{n-k} \ldots\ldots\ldots\ldots n > M \right\} \tag{3.24}$$

3.3.2.6. Calcul des coefficients LP_MFCC

Les coefficients LP_MFCC (Linear Predictive Coding based Mel Frequency Cepstral Coefficient) sont obtenus à partir d es coefficients LSF**[58].** Le processus de calcul des coefficients LP_MFCC peut être décrit par la **Figure 3.10** suivante :

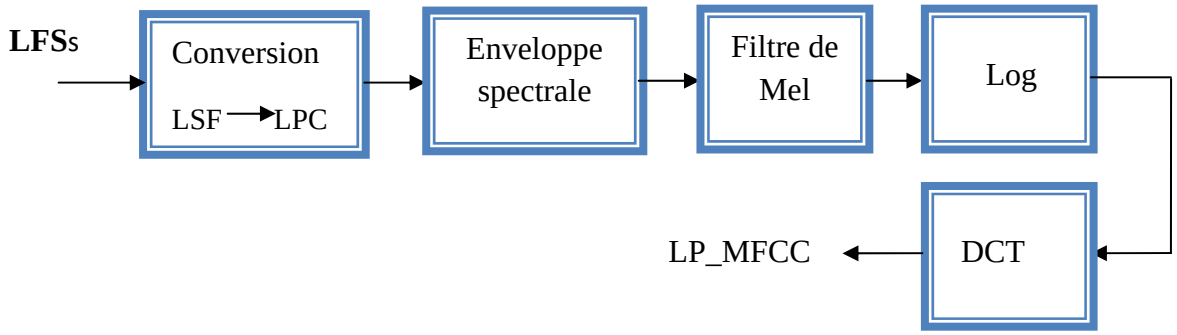


Figure 3.10 : Processus de calcul des coefficients LP_MFCC.

Le module de filtre A(z) en fonction de la fréquence est donnée par :

$$\left|A(e^{jw})\right|^2 = 2^M \left[\sin^2(w/2) \prod_{i=2,4;\ldots,M} (\cos w - \cos w_i)^2 + \cos^2(w/2) \prod_{i=1,3,\ldots,M-1} (\cos w - \cos w_i)^2 \right] \tag{3.25}$$

Donc, les n^{ieme} coefficients LP_MFCC sont calculés par la transformation du spectre *A(z)* dans une échèle non linéaire Mel. Ce spectre est passé à travers le banc de filtres triangulaires, les sorties sont transformées dans le domaine temporel grâce à la transformée de Fourier inverse.

3.4. Conclusion

Dans ce chapitre nous avons présenté l'étape indispensable dans tout système de reconnaissance automatique de la parole à savoir l'étape d'extraction des paramètres. Cette étape consiste en l'extraction de l'information contenue dans le signal de la parole pertinente pour la reconnaissance. Nous avons utilisé deux variantes de paramétrisation, la variante parole transcodée et la variante bit-stream. Dans la variante parole transcodée nous avons décrit le calcul des paramètres cepstraux MFCCs (Mel Frequency Cepstral Coefficients) qui sont les plus utilisés dans le domaine de reconnaissance de la parole offline. Dans la variante bit-stream, nous avons présenté les différents paramètres calculés à partir du flux des bits. Ces paramètres sont utilisés directement comme entrée dans le système RAP.

Chapitre 4

Résultats Expérimentaux

4.1. Introduction

Dans ce chapitre, nous présentons l'évaluation du système de Reconnaissance Automatique de la Parole (RAP) mis en œuvre pour les deux variantes : variante parole transcodée et variante bit-stream. L'objectif étant l'étude de l'influence du codeur GSM EFR sur la performance du système de reconnaissance appliquée à la langue arabe.

4.2. Description de la base de données

La base de données utilisée est constituée des chiffres arabes prononcés par des locuteurs algériens âgés entre 18 et 50 ans de niveau universitaire. Elle a été enregistrée par 60 locuteurs dans un large auditorium (auditorium de l'USTHB) dont la capacité est de 1800 places avec une hauteur de plafond de 15 mètres. L'enregistrement ayant été effectué alors que l'auditorium était vide. Cet espace s'est révélé être un endroit calme avec un niveau de bruit ambiant inférieur à 35 dB. Les sons enregistrés à 22050 Hz ont été ensuite sous échantillonnés à 16 000 Hz et à 8 000 Hz. La fréquence d'échantillonnage 16KHz permet de diminuer l'encombrement mémoire tout en gardant une parole de bonne qualité. La fréquence 8KHz est celle utilisée pour le codeur GSM EFR. Pour éviter l'effet musical, les locuteurs et locutrices doivent prononcer des suites de chiffres dans un ordre aléatoire. Dans notre présente étude, nous avons utilisé seulement la partie chiffres standard prononcé dans le mode broadcast ARADIGITS **[24]** de la base étendue. Dans la phase d'apprentissage, nous avons utilisé 40 locuteurs des deux sexes donc 1200 chiffres et dans la phase de test, nous avons utilisé 600 chiffres prononcés par 20 locuteurs des deux sexes. Les expériences ont été menées dans le cadre de la reconnaissance de la parole en mode indépendant du locuteur ainsi les locuteurs ayant servi à la base d'apprentissage sont différents de ceux ayant servi à la base de test

4.3. Système de RAP mis en œuvre

Le système de reconnaissance mis en œuvre utilise l'approche probabiliste basée sur les modèles de Markov cachés continus (MMCC). La **Figure 4.1** suivante représente la structure de ce système.

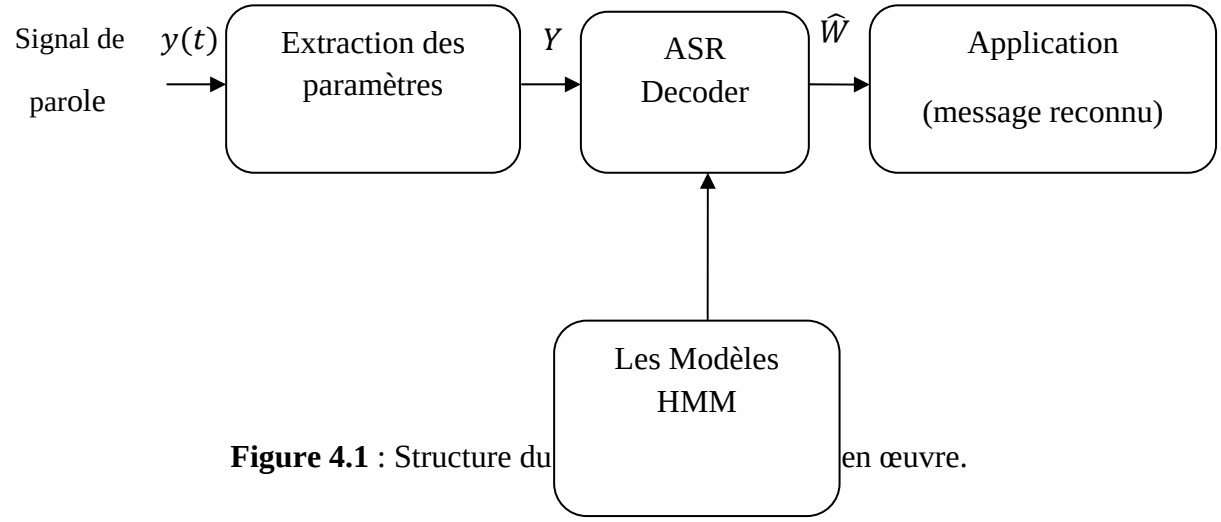


Figure 4.1 : Structure du en œuvre.

y(t) : signal de la parole.

Y : Vecteur acoustique (paramètres de reconnaissance).

$\widehat{W}$: Mot reconnu (modèle).

Le système de reconnaissance utilisé pour validé notre résultats est la plate-forme HTK (Hidden Markov Model Toolkit, ou "boîte à outils de modèles de Markov cachés")**[67]** version **3.3,** a été développée à l'Université de Cambridge par S.J Young et son équipe. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole à base de modèles de Markov cachés. Les étapes de réalisation d'un système de reconnaissance HTK sont présentées dans la **Figure 4.2** ci-dessous:

- Le calcul des paramètres du signal.
- L'apprentissage des modèles.
- La reconnaissance.

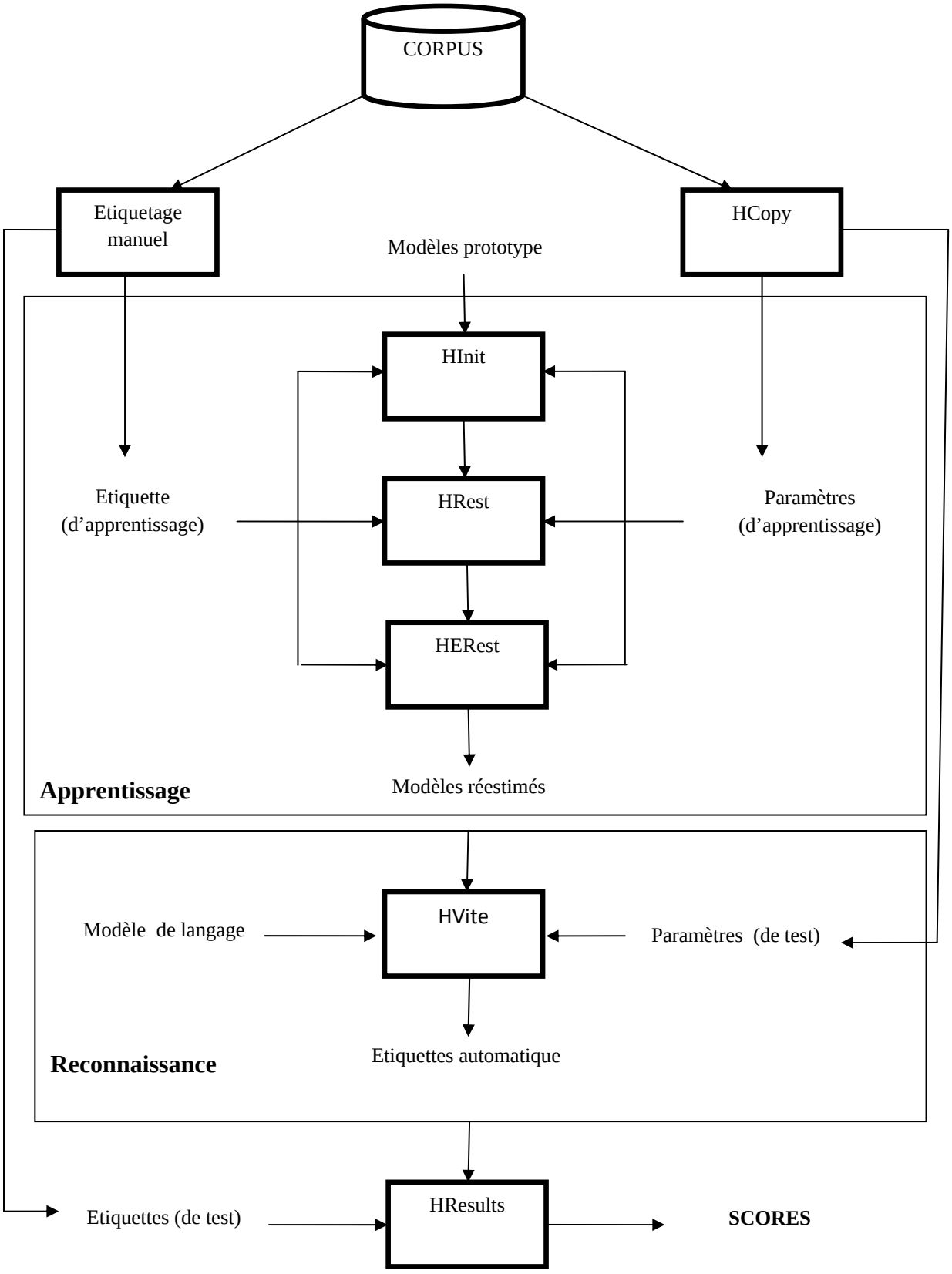


Figure 4.2 : Structure d'un système de reconnaissance à l'aide de HTK.

Paramétrisation : Cette phase sert à représenter le signal par des vecteurs de paramètres (MFCC). Cette représentation est obtenue en utilisant l'outil **HCopy** à partir la configure suivante.

```
SOURCEFORMAT = WAV

WINDOWSIZE = 250000.0 # = 25 ms = length of a time frame

TARGETRATE = 100000.0 # = 10 ms = frame periodicity

USEHAMMING = T # Use of Hamming function for windowing frames

PREEMCOEF = 0.97 # Pre-emphasis coefficient

NUMCHANS = 24 # Number of filterbank channels

TARGETKIND = MFCC

NUMCEPS = 12
```

Figure 4. 3: Fichier de configuration pour l'outil **HCopy**.

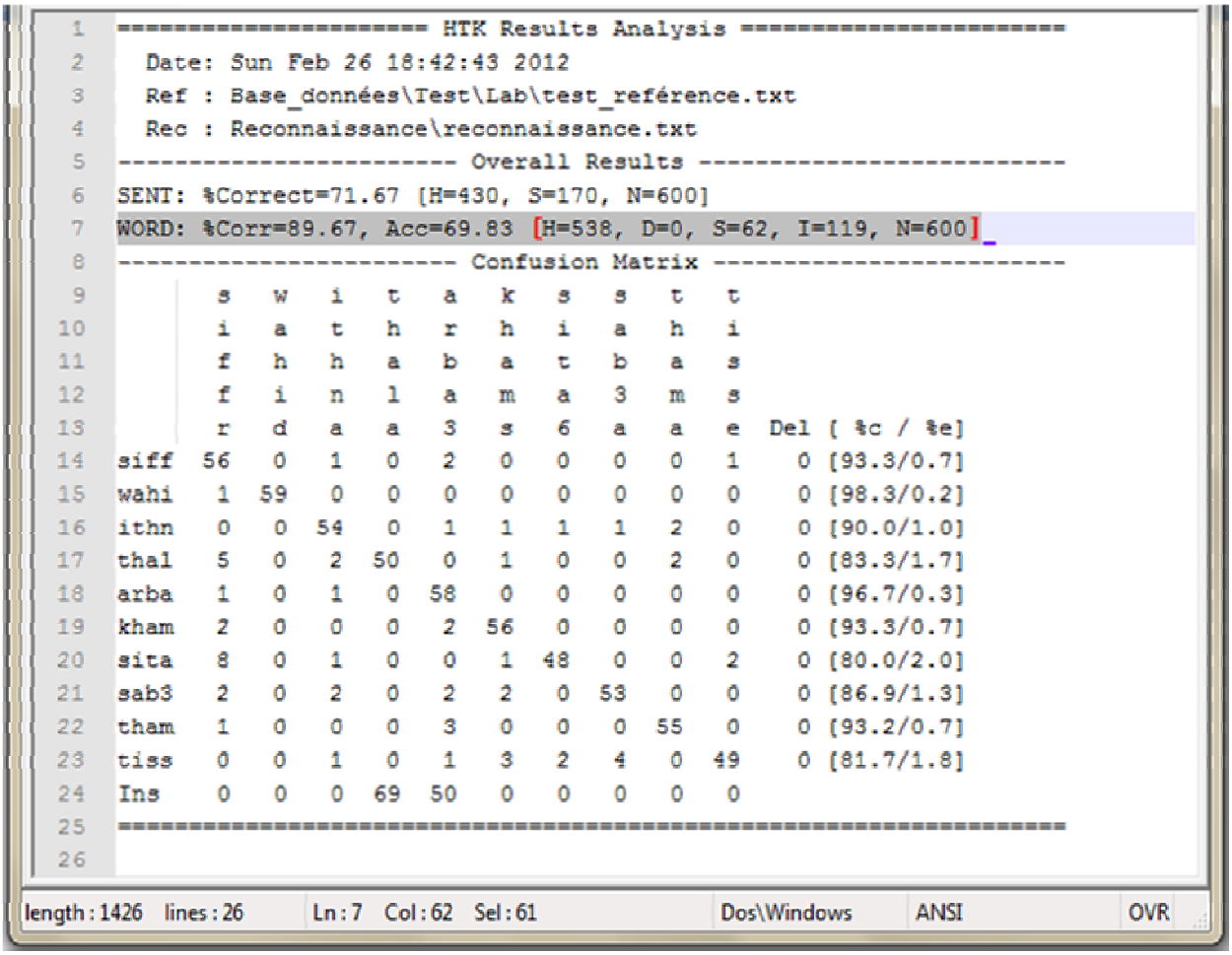
- **Apprentissage des modèles :**

Pour chaque unité acoustique, il faut définir un modèle prototype contenant la topologie choisie. Le nombre d'états du modèle (trois états) et la loi de probabilité associée à chaque état (probabilité continu jusqu'à 12 gaussiennes).

L'initialisation des modèles se fait par l'outil **HInit** fondée sur l'algorithme de Viterbi. L'estimation des paramètres d'un modèle est affinée avec **HRest** qui applique l'algorithme de Baum-Welch jusqu'à la convergence et réestime les probabilités d'émission et de transition.

- **Reconnaissance :**

Le module de décodage de la parole **HVite** utilise l'algorithme de Viterbi pour trouver la séquence d'états la plus probable correspondant aux paramètres observés dans un modèle composite et en déduire les unités acoustiques correspondantes. Les résultats du décodage sont obtenus par l'outil **Hresults** (alignement dynamique entre les fichiers de résultats et de références).



```
====================== HTK Results Analysis =======================
  Date: Sun Feb 26 18:42:43 2012
  Ref : Base_données\Test\Lab\test_référence.txt
  Rec : Reconnaissance\reconnaissance.txt
------------------------ Overall Results --------------------------
SENT: %Correct=71.67 [H=430, S=170, N=600]
WORD: %Corr=89.67, Acc=69.83 [H=538, D=0, S=62, I=119, N=600]
------------------------ Confusion Matrix -------------------------
         s   w   i   t   a   k   s   s   t   t
         i   a   t   h   r   h   i   a   h   i
         f   h   h   a   b   a   t   b   a   s
         f   i   n   l   a   m   a   3   m   s
         r   d   a   a   3   s   6   a   a   e   Del [ %c / %e]
siff    56   0   1   0   2   0   0   0   0   1     0 [93.3/0.7]
wahi     1  59   0   0   0   0   0   0   0   0     0 [98.3/0.2]
ithn     0   0  54   0   1   1   1   1   2   0     0 [90.0/1.0]
thal     5   0   2  50   0   1   0   0   2   0     0 [83.3/1.7]
arba     1   0   1   0  58   0   0   0   0   0     0 [96.7/0.3]
kham     2   0   0   0   2  56   0   0   0   0     0 [93.3/0.7]
sita     8   0   1   0   0   1  48   0   0   2     0 [80.0/2.0]
sab3     2   0   2   0   2   2   0  53   0   0     0 [86.9/1.3]
tham     1   0   0   0   3   0   0   0  55   0     0 [93.2/0.7]
tiss     0   0   1   0   1   3   2   4   0  49     0 [81.7/1.8]
Ins      0   0   0  69  50   0   0   0   0   0
===================================================================
```

length : 1426 lines : 26 Ln : 7 Col : 62 Sel : 61 Dos\Windows ANSI OVR

Figure 4.4. Exemple des résultats de taux de reconnaissance pour les coefficients PCC.

4.3.1. Taux de reconnaissance

Dans une application de reconnaissance de la parole, nous pouvons trouver trois types d'erreurs:

- **Erreur d'insertion** : le système reconnait un mot qu'il n'est pas prononcé.
- **Erreur de substitution** : le système confond deux mots différents c'est-à-dire pour un mot prononcé le système reconnait un autre mot.
- **Erreur d'omission** : le système ne reconnait pas un mot prononcé.

Dans la littérature, plusieurs définitions ont été introduites afin de calculer et de comparer les performances des systèmes RAP. Parmi ces définitions, le Taux de Mots Corrects (TMC) qui est la mesure la plus intuitive pour mesurer les performances d'un système RAP. Elle est basée sur le nombre de mots correctement reconnus et le nombre de mots total à reconnaitre. La formule de calcul du TMC est donnée par la relation suivante :

$$TMC = \frac{Nombre\ de\ mots\ correctement\ reconnus}{Nombre\ total\ des\ mots\ \text{à}\ reconnaitre} \times 100 \qquad (4.1)$$

4.4. Paramétrisation du signal vocal

L'objectif de la paramétrisation est d'extraire les coefficients représentatifs du signal de parole. Ces coefficients sont calculés à l'intervalle de temps réguliers. En simplifiant les choses, le signal de parole est transformé en une série de vecteurs acoustiques, ces derniers représentent au mieux ce qu'ils sont censés modéliser et véhiculent le maximum d'information utile pour le système de reconnaissance.

Pour étudier l'influence du codec GSM EFR sur la performance du système de RAP, nous avons utilisé deux variantes différentes. La première utilise la parole reconstituée (transcodée), elle est basée sur les coefficients standard MFCC (Mel Frequency Cepstral Coefficients) et la deuxième bit-stream utilise les coefficients LSF (Line Spectral Frequencies) décoder directement du flux des bits et un ensemble de ses dérivées : LPC (Linear Predictive Coding), PCC(Pseudo-Cepstral Coefficients) et PCEP (Pseudo-Cepstrum) décrits dans le chapitre 3.

4.4.1. La variante parole transcodée

Cette variante utilise les coefficients MFCC après le décodage et la reconstitution de signal de la parole. La figure suivante ci-dessous illustré la stratégie de cette variante**.**

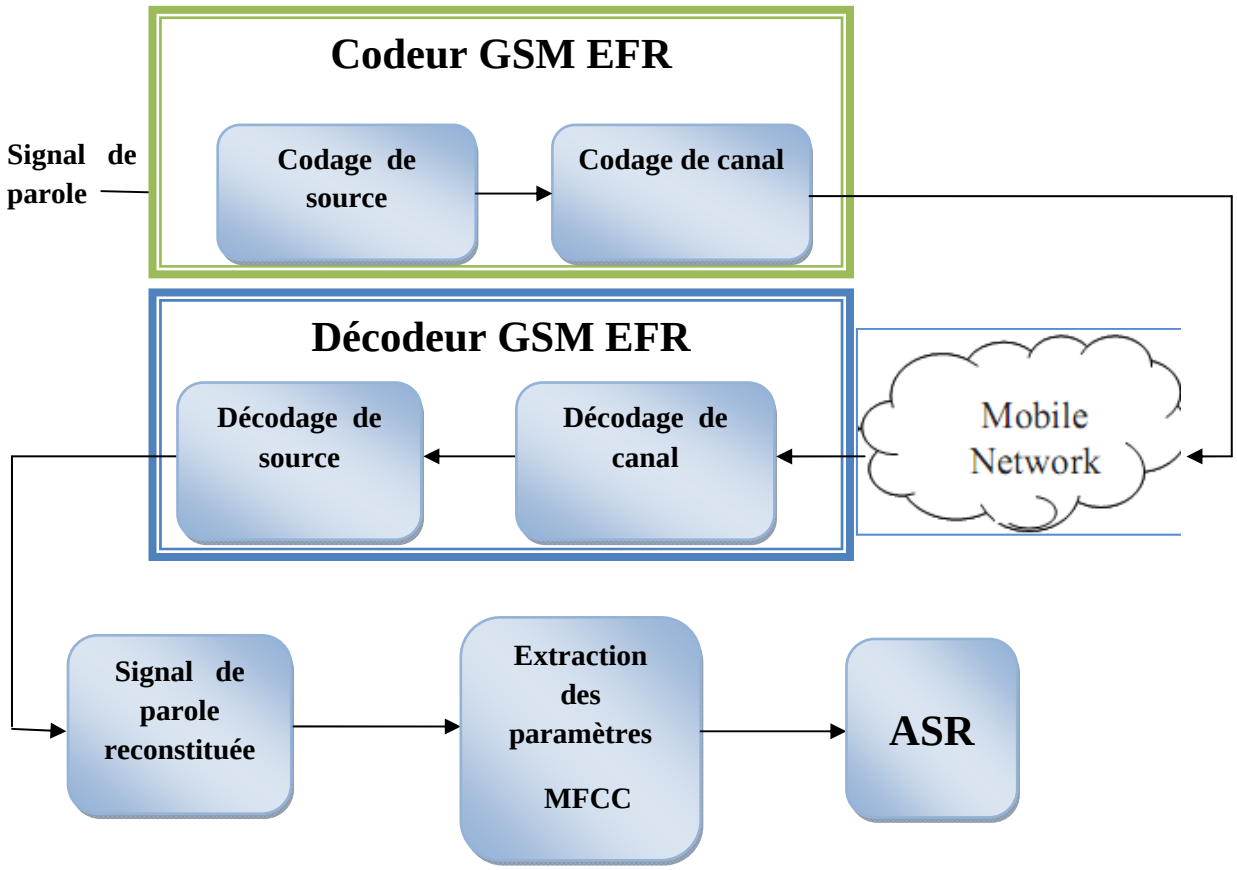


Figure 4.5: Organigramme de l'approche de reconstitution du signal de la parole

4.4.1.1. Calcul des coefficients MFCC

L'extraction des coefficients MFCC est basée sur l'analyse par banc de filtres qui consiste à filtrer le signal par un ensemble de filtre passe-bande. L'énergie en sortie de chaque filtre est attribuée à sa fréquence centrale. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales sont réparties uniformément sur l'échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole. Le nombre de filtres utilisés dans une telle analyse est choisi de manière empirique 24 filtres.

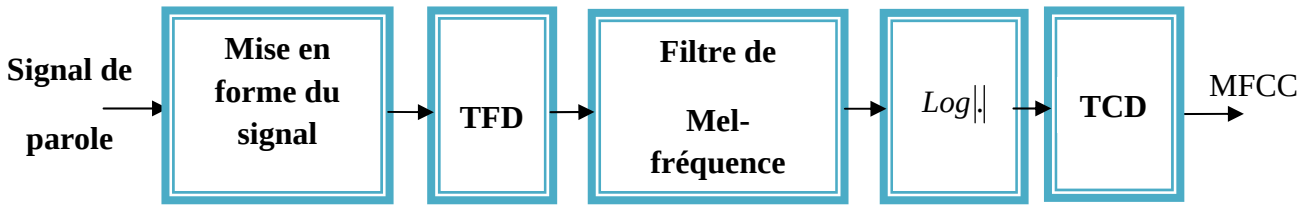


Figure 4.6 : Extraction des coefficients MFCCs.

L'expression qui calcule les coefficients MFCC est donnée par :

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \quad \text{i} = 1,\ldots\ldots,\text{P} \tag{4.1}$$

Ou :

m_j : Valeurs logarithmiques des énergies à la sortie des filtres.

N : Le nombre de filtres (N=24).

P : Le nombre de coefficient MFCC, dans ce travail nous prenons P=12.

4.4.2. La variante bit-stream

Cette variante utilise directement les coefficients décodés (LSF) dans le système de reconnaissance automatique de la parole, comme indiqué par la figure suivante**:**

Figure 4.7 : Extraction des paramètres LSF, LPC, PCC, PCEP.

A partir du train de bits, le décodage des coefficients LSFs (Line Spectral Frequencies) permet d'extraire plusieurs coefficients tels que les coefficients LPC, PCC

(Pseudo Cepstral Coefficients) et les coefficients PCEP (Pseudo Cepstrem), qui sont utilisés directement comme entrée du système de reconnaissance automatique de la parole.

4.5. Les expériences d'évaluation des résultats

4.5.1. Première expérience «variante parole transcodée »

Dans cette première expérience, on compare les résultats de reconnaissance obtenus en utilisant la parole originale et la parole reconstituée (transcodée GSM) en fonction du nombre de Gaussiennes. Les resultats sont résumés sur le tableau **4.1** ci-dessous :

Nombre de gaussiennes	1	2	4	6	8	10	12
MFCC_O_O	83,00%	**92,67%**	91,33%	90.67%	92.00%	90.83%	92.50%
MFCC_O_T	66,00%	77,50%	**77,83%**	74,50%	75,83%	76,67%	74,67%

Tableau 4.1 : Taux de reconnaissance comparatifs en fonction du nombre de gaussiennes.

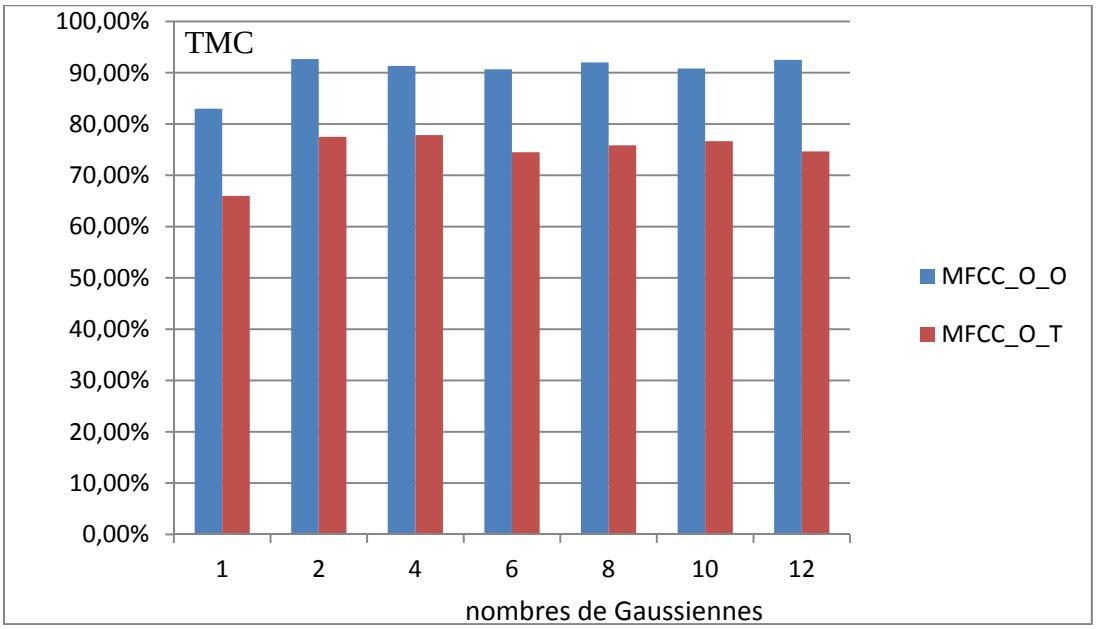


Figure 4.8 : Histogrammes comparatifs en fonction du nombre de gaussiennes.

Sachant que :

MFCC_O_O: Base d'apprentissage Originale (O) et base de test Originale (O)

MFCC_O_T: Base d'apprentissage Originale (O) et base de test Transcodée (T)

Les taux de reconnaissance obtenus avec la base de données transcodée sont légèrement inférieurs par rapport à ceux de la parole originale. Ceci peut s'expliquer par la dégradation de qualité du signal de la parole engendrée par le codeur GSM. Cette dégradation est crée par la convolution des paramètres spectraux et la source d'excitation quantifiée (quantification de la contribution des deux dictionnaires excitation fixe et adaptatif, et la quantification des

paramètres spectraux LSF qui modélisent le filtre de conduit vocale). Comme l'indique à la figure ci-dessous.

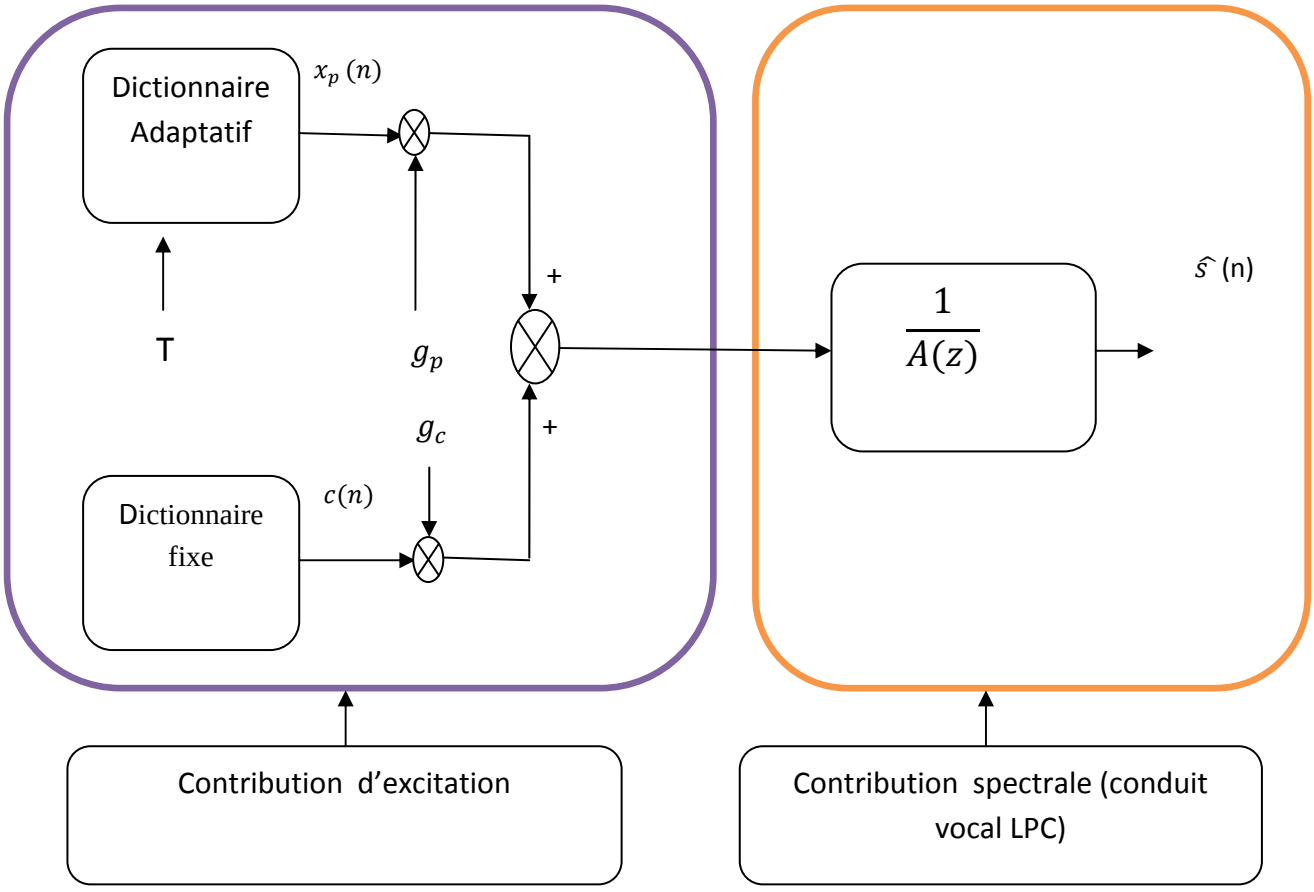


Figure 4.9 : Reconstitution de signal de parole codé GSM (ACELP, excitation, filtre).

4.5.2. Deuxième expérience «variante bit-strem »

Dans cette expérience, nous avons étudié la performance du système de reconnaissance avec différents paramètres utilisés pour cette variante **bit-stream** (LPC, LSF, PCC, PCEP).

a- Résultat pour les coefficients LPC

Nombre de gaussiennees	1	2	4	6	8	10	12
MFCC_O_T	66,00%	77,50%	**77,83%**	74,50%	75,83%	76,67%	74,67%
LPC	66,17%	72,33%	68,50%	70,50%	**73,17%**	70,83%	70,33%

Tableau 4.2 : Taux de reconnaissance comparatifs entre les coefficients MFCC et LPC.

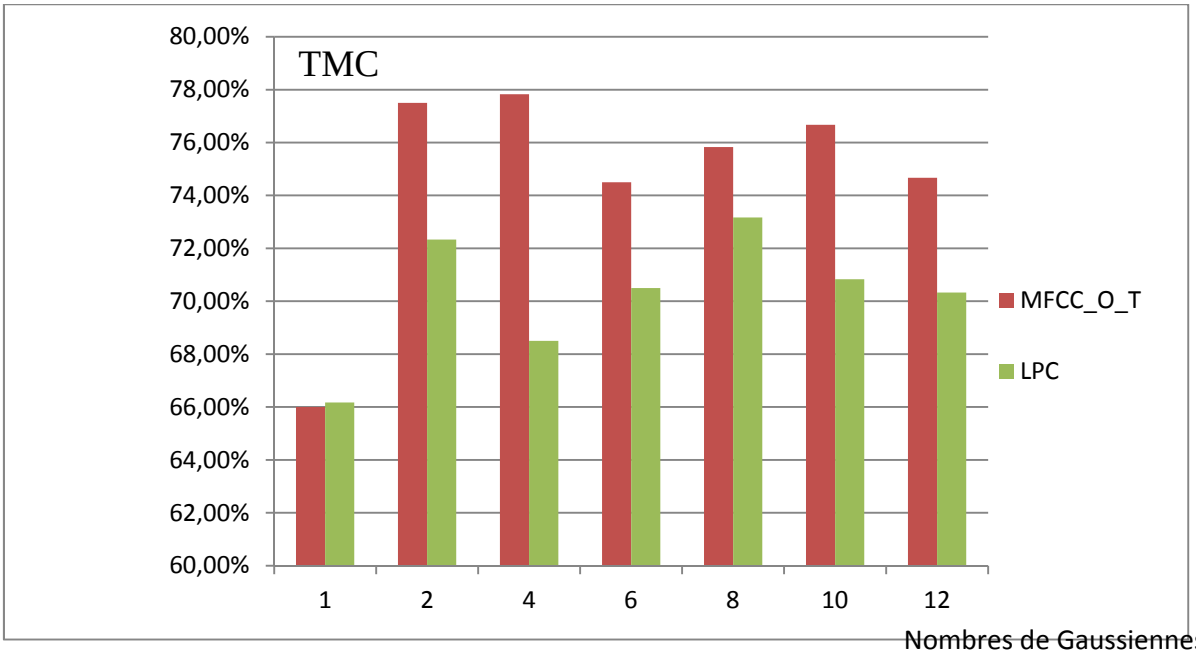


Figure 4.10 : Taux de reconnaissance comparatifs entre les coefficients MFCC et LPC.

Sachant que :

MFCC O_T : résultat pour la base d'apprentissage originale et la base de teste transcodée.

b- Résultat pour les coefficients LSF

Nombres de gaussiennees	1	2	4	6	8	10	12
MFCC_O_T	66,00%	77,50%	**77,83%**	74,50%	75,83%	76,67%	74,67%
LSF	71,17%	79,17%	82,50%	**83,67%**	82,33%	81,83%	82,50%

Tableau 4.3 : Taux de reconnaissance comparatifs entre les coefficients MFCC et LSF.

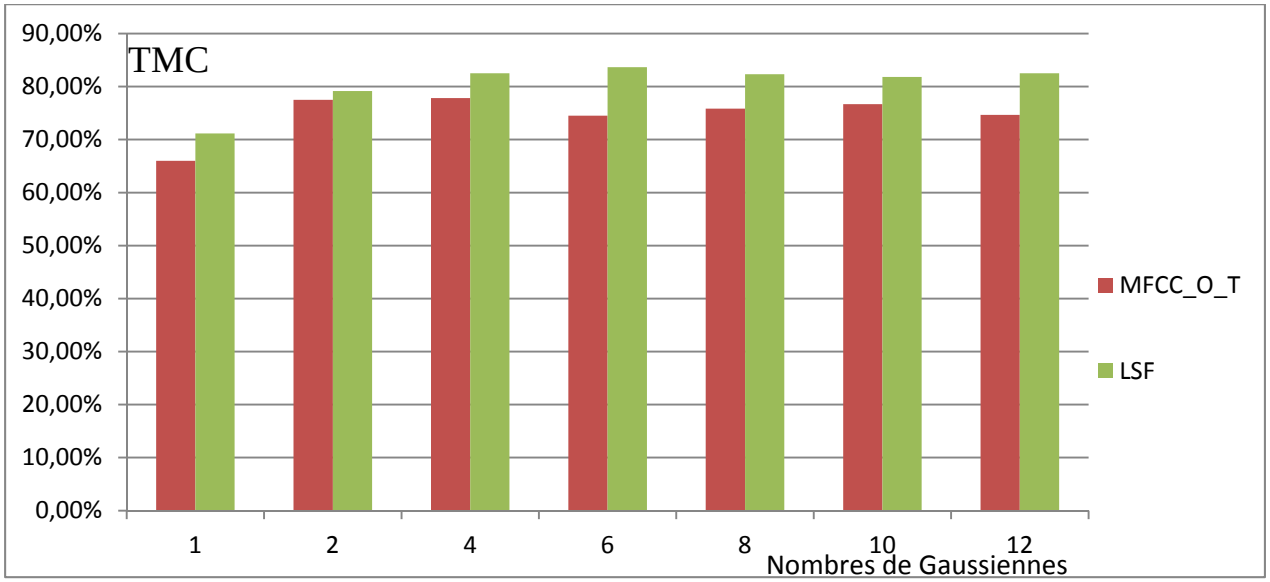


Figure 4.11 : Taux de reconnaissance comparatifs entre les coefficients MFCC et LSF.

c- Résultat pour les coefficients PCC :

Nombres de gaussiennees	1	2	4	6	8	10	12
MFCC_O_T	66,00%	77,50%	**77,83%**	74,50%	75,83%	76,67%	74,67%
PCC	80,33%	86,33%	87,83%	89,00%	89,17	88,50%	**89,67%**

Tableau 4.4 : Taux de reconnaissance comparatifs entre les coefficients MFCC et PCC.

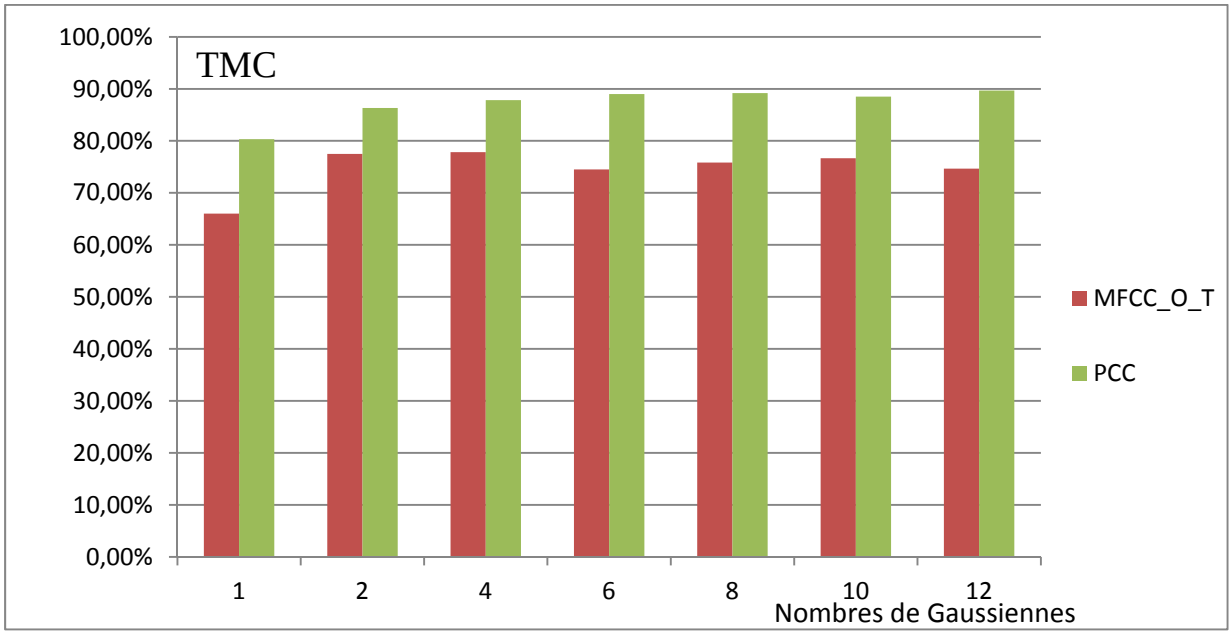


Figure 4.12 : Taux de reconnaissance comparatifs entre les coefficients MFCC et PCC.

Sachant que :

PCC : Pseudo-Cepstral Coefficients qui dérivé par les paramètres LSF.

d- Résultat pour les coefficients PCEP :

Nombres de gaussiennees	1	2	4	6	8	10	12
MFCC_O_T	66,00%	77,50%	77,83%	74,50%	75,83%	76,67%	74,67%
PCEP	81,83%	87,33%	88,17%	86,00%	87,50%	87,67%	**88,50%**

Tableau 4.5 : Taux de reconnaissance comparative coefficients MFCC avec PCEP.

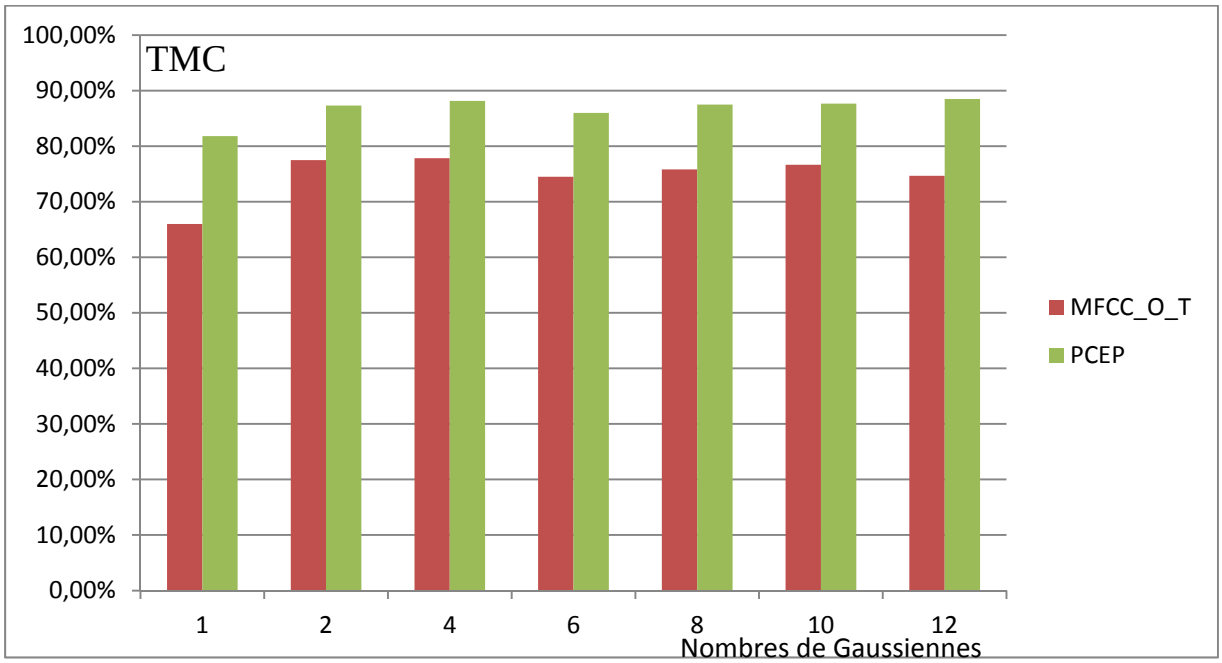


Figure 4.13 : Taux de reconnaissance comparatifs entre les coefficients MFCC et PCEP.

Sachant que :

PCEM : Pseudo-Cepstrum (PCEP) qui dérivé par le paramètres LSF.

e- Résultats globaux:

Nombres des gaussiennees	1	2	4	6	8	10	12
MFCC_O_T	66,00%	77,50%	**77,83%**	74,50%	75,83%	76,67%	74,67%
LPC	66,17%	72,33%	68,50%	70,50%	**73,17%**	70,83%	70,33%
LSF	71,17%	79,17%	82,50%	**83,67%**	82,33%	81,83%	82,50%
PCC	80,33%	86,33%	87,83%	89,00%	89,17	88,50%	**89,67%**
PCEP	81,83%	87,33%	88,17%	86,00%	87,50%	87,67%	**88,50%**

Tableau 4.6 : Les résultats globaux comparatifs du taux de reconnaissance.

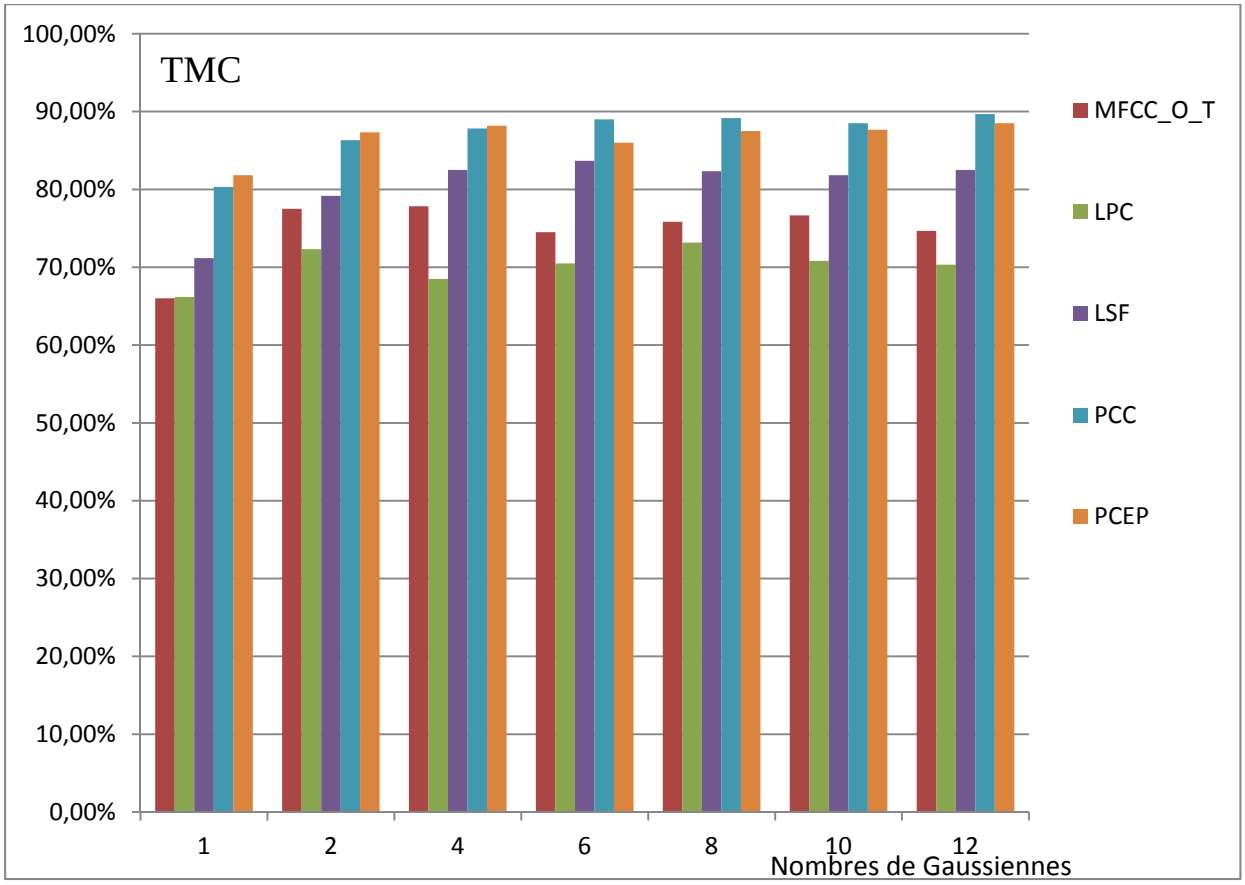


Figure 4.14.a: Les résultats globaux comparatifs du taux de reconnaissance.

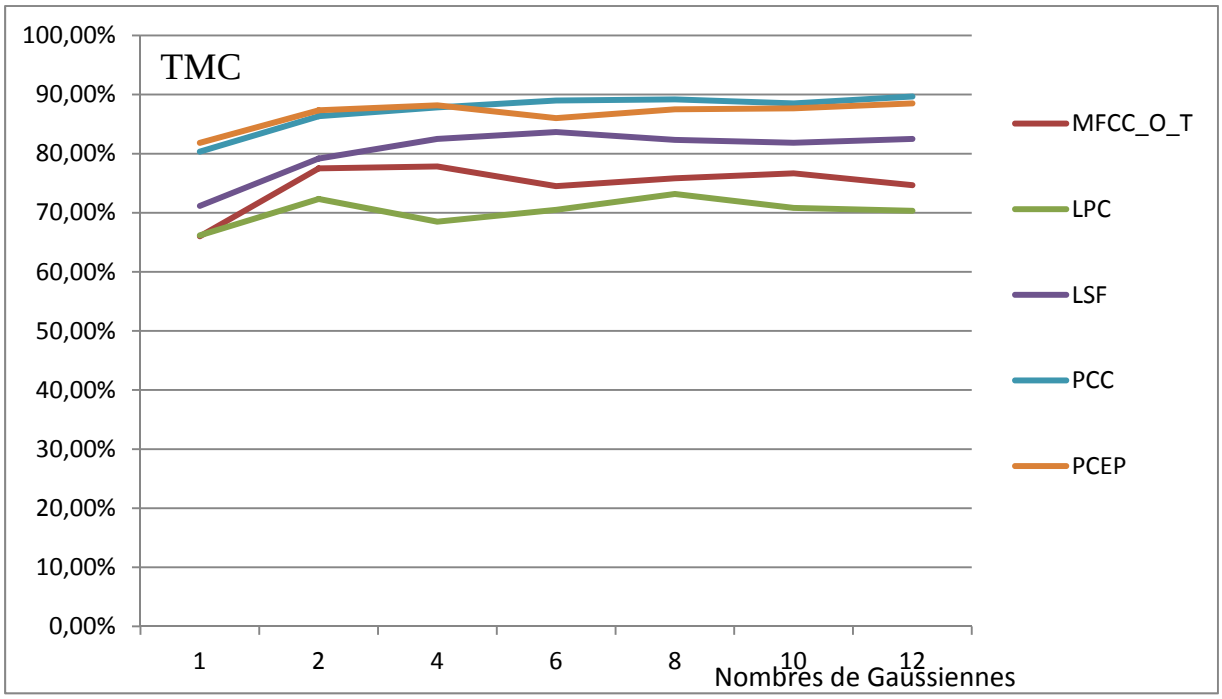


Figure 4.14.b : Les résultats globaux comparatifs du taux de reconnaissance.

Dans cette expérience nous avons comparé les résultats obtenus pour les différents paramètres (LSF, LPC, PCC, PCEP) avec les paramètres standards MFCCs de la variante parole transcodée, paramètres sont largement utilisés dans les systèmes RAP.

D'après l'analyse des résultats présentés dans le tableau **(4.6)** ci-dessus, il apparait que les paramètres de la variante **bit-stream** les coefficients (LSF, LPC, PCC, PCEP) améliorent légèrement la performance du système de reconnaissance automatique de la parole par rapport à la variante parole transcodée (MFCC). En effet, le résultat obtenu pour la variante bit-stream est **89.67%** pour douze (**12**) gaussiennes alors que pour la variante parole transcodée est égale **77,83%** pour quatre (**4**) gaussiennes.

La variante bit-stream basé sur les paramètres spectraux LSF/LPC (modélisent les conduits vocale) offre la possibilité d'éviter influence de la contribution d'excitation pour conserver la forme de l'enveloppe spectrale, cela permet d'améliorer la performance du système RAP.

4.6. Conclusion

Nous avons étudié et réalisé un système de reconnaissance automatique de la parole basé sur les modèles de Markov cachés continus en utilisant la plateforme HTK. Cette étude a concerné deux variantes de parole reconstituée et bit-stream. Les résultats de reconnaissance expérimentaux obtenus ont montré que les taux de reconnaissance avec la deuxième variante (bitstream, LSF, LPC, PCC, PCEP) sont supérieurs à ceux obtenus par la première variante (reconstitution de signal de la parole, MFCC). De plus, la variante bit-stream nécessite moins de temps de calcul et occupe moins d'espace mémoire (10 à 12 valeur pour chaque vecteur acoustique de durée de 20 ms de signal de parole). Elle s'adapte mieux pour la reconnaissance dans le réseau de communication mobile de type GSM.

❖

Conclusion générale

Dans ce travail, nous avons étudié l'influence de codeur GSM EFR sur la performance du système de reconnaissance automatique de la parole appliqué à la langue Arabe. Nous avons fait d'abord une étude sur les principes du réseau de communication mobile GSM aspect réseau et aspect codage. Le codeur utilisé actuellement dans la communication mobile est le GSM EFR (Globale System for Mobile Enhanced Full Rate). Ce codeur est basé sur l'algorithme ACELP (Algébrique Code Excitation Linéaire Prédiction), principe d'analyse par synthèse.

Le système de reconnaissance mis en œuvre utilise les Modèles de Markov Cachées (MMCs). Son principe est de transformer les paramètres acoustiques par des modèles probabilistes qui sont décrits par des vecteurs initiaux, des matrices de transition et des matrices d'émissions. Pour la représentation du signal de la parole nous avons utilisé deux variantes de paramétrisation. La première variante **parole transcodée** utilise les coefficients cepstraux MFCC (Mel-Frequency Cepstrum Coefficients). La deuxième variante **bit-stream** les coefficients utilisés sont les coefficients LSF, LPC, PCC et PCEP qui sont calculés directement du train des bits.

Nous avons constaté que la performance du système de reconnaissance diminue pour la parole transcodée GSM malgré que nous avons utilisé une représentation du signal de la parole par les coefficients MFCC qui ont prouvé leur fiabilité en environnement calme. Ces coefficients s'avèrent donc sensible à la variation du signal de la parole (parole bruitée, parole transcodée). L'utilisation de la variante bit-stream améliore la performance des systèmes de reconnaissance par rapport à la variante parole transcodée. En effet, les résultats de reconnaissance obtenus pour la variante bit-stream sont de l'ordre de **89.67%** alors que pour la variante parole transcodée ils ne sont que de **77,83%**. De plus, cette variante nécessite moins de temps de calcul et un espace mémoire réduit pour la représentation des données acoustiques. Nous concluons que la variante bit-stream s'adapte mieux pour la reconnaissance en environnement mobile.

Comme perspectives à ce travail nous proposons :

- La communication orale avec la machine via le réseau de communication mobile est un rêve humain, elle reste un champ d'étude très important. Pour améliorer encore la performance des systèmes de Reconnaissance Automatique de la Parole dans l'environnement mobile GSM, il faut associer des algorithmes de codage et des méthodes de réduction de bruit (bruit de source, bruit de canal) ou des méthodes de rehaussement de la parole à l'entrée du système de reconnaissance.
- Les systèmes de RAP actuels reposent majoritairement sur des approches probabilistes comme les modèles de Markov cachées. Pour améliorer la classification nous proposons la combinaison des méthodes probabilistes comme les MMCs et les méthodes discriminantes comme les SVMs ou les réseaux de neurones. .

Bibliographie

[1] J. Eberspächer, H.-J. Vögel.(GSM – Architecture, Protocols and Services)Third Edition. John Wiley & Sons. 2009.

[2] Friedhelm Hillebrand (GSM and UMTS: The Creation of Global Mobile Communication).John Wiley & Sons. 2002.

[3] Saeed V. Vaseghi (Multimedia Signal Processing Theory and Applications in Speech, Music and Communications). Wiley; 1 edition. PP 353-525. 2007.

[4] Jerry D. Gibson (Multimedia Communications) Academic Press Series in Communications Networking, And Multimedia.vol.pp.25-42.2001.

[5] WAI C. CHU. (Speech Coding Algorithms Foundation And Evolution Of Standardized Coders). Wiley-Interscienc; 1 edition. pp 423-449. 2003.

[6] Randy Goldberg, Lance Riek (A Practical Handbook of Speech Coders). CRC Press. 1 edition .2000.

[7] AERNOUTS Ludovic (Le réseau GSM). Cnam de Lille 1999.

[8] T. Honkanen , J. Vainoi, Jarvinen, P . Haavisto, R. Salami, C. Laflamme and J-P. Adoul (Enhanced Full Rate Speech Code For Is-136 Digital Cellular System) IEEE International Conference. Volume 2, p.731 -734 vol.2. 2010.

[9] Milan JELINEK. (Modélisation Spectrale Et Compression De Parole A Bas Débit). Thèse de doctorat Université Sherbrooke (Québec), CANADA.PP 11-72. Octobre 1998

[10] K.Jarvinen,J. Vainio,P.Kapanen,T.Honkanen,P.Haavisto.R. Salami,C.Lajlamme, J-P. Adoul . (Gsm Enhanced Full Rate Speech Codec). IEEE International Conference . 725 - 729 vol.2 . 2010.

[11] R. Salami, C. Laflamme, B. Bessette, J-P. Adoul. Canada K. Jawinen, J. Vainio, P. Kapanen, T. Honkanen, P. Haavisto. (Description Of Gsm Enhanced Full Rate Speech Codec). IEEE .1997.

[12] Francis Labonté (Etude, Optimisation Et Implémentation D'un Quantificateur Vectoriel Algébrique Encastré Dans Un Codeur Audio Hybride ACELP/TCX). Mémoire de maîtrise, Université de Sherbrooke Canada, juin 2002.

[13] Venkatraman Atti and Andreas Spanias. (A Simulation Tool For Introducing Algebraic Celp (Acelp) Coding Concepts In A Dsp Course), Frontiers in Education Conference, Boston, Massachusetts, 2002.

[14] Fu-Kun Chen and ,Jur-Feru Yang (Maximum-Take-Precedence Acelp: A Low complexity Search Method), Acoustics, Speech, and Signal Processing, (ICASSP '01). IEEE International Conference on 693-696 vol.2. 2001.

[15] F.-K.Chen, J.-F.Yang and Y.-L.Yan. (Candidate Scheme For Fast ACELP Search). Vision,Image and Signal Processing, IEE Proceedings-pp 10 – 16. 2002.

[16] Francis Labonté (Etude, Optimisation Et Implémentation D'un Quantificateur Vectoriel Algébrique Encastré Dans Un Codeur Audio Hybride ACELP/TCX). Mémoire de maîtrise, Université de Sherbrooke Canada, juin 2002.

[17] Kyung Jin Byun, Hee Bum Jungl, Minsoo Hahn2, Kyung So0 Kim.(A Fast Acelp Codebook Search Method).Signal Processing, 6th International Conferenceon 422 - 425 vol.1. 26-30 Aug. 2002 .

[18] A. Falahati, M. Soleimani (A Proposed Fast ACELP Codebook Search). Communications, APCC 14th Asia-Pacific Conference on 1 - 5. Oct. 2008.

[19] Eung-Don Lee. Korea and Jae-Min Ahn .Div. (Excient Fixed Codebook Search Method For ACELP Speech Codecs). ICHIT 2006, LNAI 4413, pp. 178–187,c_Springer-Verlag Berlin Heidelberg 2007

[20] M. DEBYECHE,A. KROBBA and A. AMROUCHE, "Effect of GSM speech coding on the performance of Speaker Recognition System", IEEE , the 10th International Conference on Information Sciences Signal Processing and their Applications, May 10-13, 2010, Malaisya, pp. 137-140.

[21] A.KROBBA, M. DEBYECHE and A. AMROUCHE, "Evaluation of Speaker Identification System Using GSM-EFR Speech Data" IEEE International Conference on Design and Technology of Integrated Systems, March 23-25, 2010, Tunisia. pp.1-5.

[22] Saeed V. Vaseghi. (Theory And Applications In Speech, Music And Communications). Edition .2001.

[23] M. Debyeche « Reconnaissance Automatique De La Parole Appliqué A La Langue Arabe » Thèse de Doctorat d'Etat, USTHB, 2007.

[24] Abderrahmane Amrouche . "Reconnaissance Automatique De La Parole Par Les Modeles Connexionnistes".These Doctorat d'Etat . USTHB.2007.

[25] CHARPENTIER, Christophe. (Conception D'une Methode Robuste De Reconnaissance De La Parole Pour Un Système Embarqué). UNIVERSITÉ DU QUÉBEC. MONTRÉAL, LE 14 JANVIER.PP.63-70. 2008.

[26] Matthias Woelfel and John McDonough (Distant Speech Recognition). John Wiley & SonsPP.283-301. 2009.

[27] Mark Gales. Steve Young. (The Application of Hidden Markov Models in Speech Recognition). Foundations and Trends in Signal Processing .2008.

[28] Jérôme Dequier. (Chaînes de Markov et applications).GRENOBLE.2005.

[29] Djamel Addou · Sid-Ahmed Selouani . Kaoukeb Kifaya ·Malika Boudraa · Bachir Boudraa·(A Noise-Robust Front-End For Distributed Speech Recognition In Mobile Communications). Int J Speech Technol.2007.

[30] Gernot A. Fink (MarkovModels for Pattern Recognition). Springer. PP.61-92. 2008.

[31] Christophe LEVY.(Modèles Acoustiques Compacts Pour Les Systèmes Embarqués). These de Doctorat . Université Avignon.. PP. 19-26. PP. 38-41. 2006.

[32] M.DEBYECHE, A.AMROUCHE and J.P HATON (2008), “Improved Hidden Markov Models for Speaker Independent Arabic Speech Recognition” 5ème Conférence Internationale JTEA’08, 2-4 Mai, 2008. Tunisie

[33] A.I AMROUS, M.DEBYECHE and A.AMROUCHE, "Robust Arabic speech recognition in noisy environments using prosodic features and formant", International Journal of Speech Technology, Springer-Verlag Ed, ISSN 1381-2416. Vol.14, N°3, 2011.

[34] Imen AMROUS A,Mohamed DEBYECHE and Abderrahmen AMROUCHE“Integration of Auxilliary features in Hidden Markov Models for Arabic Digit Recognition” IEEE International Conference on System, Circuits and Signal, November 05-08, 2009, Tunisia.

[35] Zheng-Hua Tan. Imre Varga (Network, Distributed and Embedded Speech Recognition). Aalborg University,Denmark .2008.

[36] David Pearce. (Distributed Speech Recognition Standards).Springer.2008.

[37] Stephen E. Levinson. (Mathematical Models for Speech Technology). USA. 2005.

[38] Zheng-Hua Tan. Børge Lindberg. (Speech Recognition on Mobile Devices). Aalborg University,Denmark. 2008.

[39] André Berton, Alfred Kaltenmeier, Udo Haiber, and Olaf Schreiner.(Speech Recognition).springer.2006.

[40] Gerasimos Potamianos, Lori Lamel, Matthias Wölfel.(Automatic Speech Recognition). Springer.2008.

[41] Nemat. S. Abdel Kader. (Effect Of Gsm System On Text-Independent Speaker Recognition Performance). Journal of Theoretical and Applied Information Technology . 2008.

[42] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini, (Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition), Proc. Of EUSIPCO, European Signal Processing Conference, pp. 437- 440. , 2000

[43] Amy Neustein. Judith Markowitz . Bill Scholz .(Advances in Speech Recognition).Springer.2010.

[44] Joseph Keshet. Samy Bengio.(Automatic Speech and Speaker Recognition).John Wiley & Sons. PP.101-107..2009.

[45] K. J. Ray Liu. Digital. (Speech Processing, Synthesis, and Recognition).

[46] ohn Holmes. Wendy Holmes (Speech Synthesis and Recognition). Taylor & Francis e-Library, Second Edition . 2003.

[47] Djamel Addou · Sid-Ahmed Selouani . Kaoukeb Kifaya ·Malika Boudraa · Bachir Boudraa·(A Noise-Robust Front-End For Distributed Speech Recognition In Mobile Communications). Int J Speech Technol.2007.

[48] Sergios Theodoridis. Konstantinos Koutroumbas. (Pattern Recongnition). Fouth Edition .PP.512-554.

[49] Zheng-Hua Tan and Børge Lindberg. (Automatic Speech Recognition on Mobile Devices and over Communication Networks).Springer 2008.

[50] Antonio M. Peinado. José C.Segura. (Speech Recognition Over Digital Channels).

John Wiley & Sons Ltd.PP. 10-101.2006.

[51] WAI C. CHU. (Speech Coding Algorithms Foundation And Evolution Of Standardized Coders). John Wiley & Sons. 2003.

[52] Stephen E. Levinson.(Mathematical Models for Speech Technology) . John Wiley & Sons .2005.

[53] Sadaoki Furui. (Digital Speech Processing, synthesis and recognition). Marcel Dekker.2001.

[54] Loïc Barrault. (Diagnostic Pour La Combinaison De Systèmes De Reconnaissance Automatique De La Parole). These doctorat université d'Avignon. PP. 42-47. 2008.

[55] Dominique Vaufreydaz. (Modélisation Statistique Du Langage A Partir d'Internet Pour La Reconnaissance Automatique De La Parole Continue). Université JOSEPH FOURIER - GRENOBLE I. PP. 49-51. 2002.

[56] Ben Milner. (Speech Feature Extraction and Reconstruction).

[57] Zheng-Hua Tan and Børge Lindberg . (Speech Recognition On Mobile Devices). Aalborg University, Denmark. PP 226-229.2008.

[58] Hong Kook Kim and Richard C. Rose. (Speech Recognition over Mobile Networks)Springer, mars 2008
.

[59] Vladimir Fabregas Surigué de Alencar and Abraham Alcaim, (Transformations of LPC and LSF Parameters to Speech Recognition Features). Lecture Notes in Computer Science, springer , Volume 3686.2005.

[60] H. Kook Kim. H. Soo Lee. (On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients). IEEE Transactions On Speech And Audio Processing, Vol. 8, No. 2, March 2000.

[61] V.F.S. de Alencar and A. Alcaim. (Digital filter interpolation of decoded LSFs for distributed continuous speech recognition). Electronics Letters. 14th August 2008 Vol. 44 No. 17.

[62] V. F. S. Alencar and A. Alcaim. (LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese). IEEE. 2008.

[63] Vladimir Fabregas Surigué de Alencar and Abraham Alcaim. (On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese). IEEE. 2009.

[64] Ke Chen, Senior Member. Ahmad Salman. (Learning Speaker-Specific Characteristics with a Deep Neural Architecture). IEEE transactions on neural networks, vol. 22, no. 11, november 2011.

[65] Seung Ho Choi. Hong Kook Kim. Hwang Soo Lee. (Speech recognition using quantized LSP parameters and their transformations in digital communication). Speech Communication 30. 223±233. 2000.

[66] Carmen Pelea-Moreno.Fernando Diaz-de-Mari. Ascension Gallardo-Antolın. Diego F. Gomez-Cajas. (A comparison of front-ends for bitstream-based ASR over IP). Signal Processing 86 .1502–1508.2006.

Steve Young. Gunnar Evermann (The HTK Book). 2006.