

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



Mémoire
Présenté pour l'obtention du diplôme de MAGISTER
En : Electronique
Spécialité : communication parlée

Par
BERKANI née SABOUNE Hadjer

Thème

***Reconnaissance du locuteur par une approche
HMM multi-bandes via une ligne de
communication GSMFR***

Soutenu publiquement le 06 Mai 2012, devant le Jury composé de :

Mr	B. HADDAD	Professeur	USTHB	Président
Mr	M. DEBYECHE	Maître de conférences	USTHB	Directeur de mémoire
Mme	L. FERGANI	Maître de conférences	USTHB	Examinatrice
Mme	L. FALEK	Maître de conférences	USTHB	Examinatrice

Remerciements

Mes remerciements les plus sincères vont à mon Directeur de mémoire, Mr M.DEBYECHE, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, pour sa simplicité, ses conseils avisés, son soutien et sa patience. Qu'il trouve ici mes respects et profonde gratitude.

Je tiens à exprimer mes remerciements à Mr B. HADDAD, Professeur à la Faculté d'Electronique et d'Informatique de l'USTHB pour l'honneur qu'il me fait de présider le jury de soutenance.

Mes remerciement s'adresse également à Mme H. FERGANI, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, et Mme L.FALEK, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, membres examinateurs du jury, pour l'intérêt qu'elles ont porté à notre travail et pour leur disponibilité.

Grand merci à mes chers parents, qui m'ont inculqué la soif du savoir.

Je remercie mon mari qui m'a toujours soutenu et encouragé.

Je remercie également tous ceux qui m'ont aidé et encouragé de près ou de loin.

Summary

This work investigates the influence of GSM Full Rate (FR) speech coding on text independent speaker recognition performance. The GSM FR speech coder is based on the GSM 6.10 protocol used by mobile telephone operators which belongs to the class RPE-LTP (regular pulse excitation- long term prediction) coders. The general principle of the speaker recognition system used in this work is to split the whole frequency domain into several sub-bands on which statistical sub-recognizers based on Continuous Hidden Markov Models (CHMMs) are independently applied and then recombined to yield a global score and a global recognition decision. The choice of the sub-band architecture and the recombination strategies are discussed. This technique had been shown to be robust for speech recognition when narrow band noise degradation occurs. We verify this robustness for the speaker identification task in the case of GSM coded speech. We also study which information is really used to recognize speakers. For this, speaker identification experiments on independent sub-bands are conducted for 60 speakers, on ARADIGIT and ARADIGIT GSM FR databases. The ARADIGIT database evaluation consists of 10 Arabic digits (zero to nine) spoken by sixty (60) speakers (30 men and 30 women). Each utterance is repeated three times by each speaker. For our multi-bands architecture we used two systems, the first is with four sub-bands and the second is that of seven sub-bands frequencies. The results show that the performances obtained with this new system, using fuzzy integral recombination strategies, are better in the case of 7 sub-bands than a conventional full band recognizer. But, the addition of the full band information show a better adaptability of this approach for the GSM coded speech.

Résumé

L'étude de l'influence de la parole transcodée GSM FR sur la performance du système de reconnaissance du locuteur est le but principale de ce travail. Le codeur GSM FR (Full Rate) correspond au protocole GSM 6.10 utilisé par les opérateurs de téléphonie mobile, qui est basé sur l'algorithme RPE-LTP (Regular Pulse Excitation- Long Term Prediction). L'approche utilisée pour mettre en œuvre le système de reconnaissance est la modélisation stochastique basée sur les modèles de Markov cachés continus (CHMM: Continus Hidden Markov Models). Cette technique est largement utilisée en reconnaissance de la parole ou du locuteur et a prouvé son efficacité par les résultats de reconnaissance obtenus. La particularité du système de reconnaissance implémenté est son architecture Multi-bandes, le principe est de diviser le signal de parole dans le domaine spectrale en sous bandes de fréquences. Le choix des sous bandes est lié à la perception auditive humaine. Ce type d'approche s'est déjà montré robuste en reconnaissance de la parole lorsque le signal est corrompu par un bruit bande étroite. Nous vérifions cette robustesse pour le cas de la parole transcodée par le codeur GSM FR. Nous avons utilisé deux systèmes, le premier est à 4 sous bandes et le deuxième est celui de 7 sous bandes de fréquences. Les sorties des systèmes des sous bandes sont ensuite combinées par la méthode des intégrales floues (Fuzzy integral). Les expériences de validation ont été réalisées en mode indépendant du texte sur la base de données ARADIGIT 16 kHz composée de chiffres arabes de 0 à 9 prononcés par 60 locuteurs (30 femmes et 30 hommes). Les résultats des expériences effectuées ont montré que l'approche multi-bandes, ainsi que la méthode de recombinaison par les intégrales flous sont efficaces pour la parole transcodée GSM FR dans le cas de 7 sous bandes. En ajoutant la bande totale, les taux de reconnaissance augmente de manière plus considérable pour les deux systèmes 4 et 7 sous bandes).

Table des matières

Introduction générale.....	1
Chapitre 1: Principes généraux de la reconnaissance du locuteur	4
1.1 Introduction	4
1.2 complexité de la tâche de reconnaissance	4
1.2.1 Le mode de reconnaissance.....	4
1.2.2 Notion de variabilité.....	4
1.2.2.1 Variabilité due au locuteur	4
1.2.2.2 Variabilité due à l'environnement.....	5
1.3 Les différentes tâches en RAL	5
1.3.1 Identification automatique du locuteur	6
1.3.2 Vérification automatique du locuteur.....	7
1.3.3 Indexation en locuteur	7
1.3.4 Types d'erreur	7
1.4 Architecture d'un système de RAL	7
1.4.1 Extraction des paramètres	8
1.4.1.1 Mise en forme du signal de parole.....	9
1.4.1.2 Calcul des coefficients	10
1.4.1.3 L'analyse LPC	10
1.4.1.4 L'analyse cepstrale.....	14
1.4.1.5 Calcul des coefficients différentiels.....	17
1.4.2 Approche de reconnaissance	18
1.4.2.1 L'approche vectorielle	18
1.4.2.2 L'approche statistique	19
1.4.2.3 L'approche connexionniste	20
1.4.2.4 L'approche prédictive	21
1.4.3 La prise de décision.....	21
1.5 Modèles parallèles.....	22
1.6 Conclusion.....	23
Chapitre 2 : Paradigme Multi-bandes à base de CHMM.....	24
2.1 Introduction	24
2.2 Motivations et avantages	24
2.3 Définitions des sous bandes de fréquences	25
2.4 Architecture du système multi-bandes	25
2.5 Les modèles de Markov cachés.....	26
2.5.1 Définitions.....	26
2.5.2 Modélisation du locuteur par HMM.....	28
2.5.2.1 Topologie des HMMs utilisés pour la RAL	28
2.5.2.2 Modélisation des observations acoustiques.....	29
2.5.3 Sous-reconnaisseur à base de CHMMs	31
2.5.3.1 Reconnaissance d'une séquence d'observations	31
2.5.3.2 Probabilité d'émission des observations	32
2.5.3.3 Estimation directe.....	33

2.5.3.4 Estimation Backward.....	34
2.5.3.5 Recherche des états cachés.....	35
2.5.3.6 Apprentissage d'un modèle.....	36
2.5.3.7 Problème de mise en ouvre	40
2.6 Recombinaison des scores et décision	40
2.6.1 Les intégrales floues.....	41
2.6.1.1 Définitions	41
2.6.1.2 Algorithme	42
2.7 Conclusion.....	44
Chapitre 3 : Protocole de communication GSM FR	45
3.1 Introduction	45
3.2 Architecture du réseau GSM.....	45
3.2.1 Le principe cellulaire.....	45
3.2.2 la communication dans le GSM	46
3.2.3 Présentation de l'infrastructure d'un réseau	47
3.2.4 Les équipements d'un réseau GSM	48
3.2.5 Canaux de transmission.....	49
3.2.6 Modulation	49
3.3 Codage de la parole	50
3.3.1 Principe.....	50
3.3.2 Codeurs GSM.....	51
3.3.3 Codec RPE-LTP	52
3.3.3.1 Analyse-Synthèse court terme.....	53
3.3.3.2 Détermination de l'excitation.....	55
3.3 Conclusion	58
Chapitre 4 :Evaluation expérimentale.....	59
4.1 Introduction	59
4.2 Description des bases de données utilisées	59
4.3 Implémentation du codec GSM FR.....	60
4.4 Description du système d'IAL Multi-bandes	60
4.3.1 Filtrage en sous bandes	61
4.3.2 Extraction des paramètres	61
4.3.3 Reconnaisseurs à base de CHMM.....	61
4.3.4 Le module de recombinaison.....	62
4.3.5 Décision	62
4.4 Tests et évaluations.....	62
4.4.1 Test 1 : Effet du codeur GSM FR	63
4.4.2 Test 2 : Architecture Multi-bandes.....	64
4.5 Interprétation des résultats.....	69
4.6 Conclusion.....	70
Conclusion et perspectives	71
Bibliographie	

Liste des Figures

Figure 1.1 : La tâche d'identification automatique du locuteur	6
Figure 1.2 : Structure générale d'un système de RAL	8
Figure 1.3 : Schémas synoptique de l'extraction de paramètres acoustiques..	8
Figure 1.4 : Etapes de mise en forme du signal de parole.....	9
Figure 1.5 : Exemple d'échantillonnage d'un signal de parole.....	9
Figure 1.6 : : Le modèle de production de parole avec les tubes acoustiques.....	13
Figure 1.7 : Calcul des coefficients cepstraux.....	14
Figure 1.8 : Calcul des MFCCs.....	10
Figure 1.9 : Les filtres triangulaires passe-bande en Mel-Fréquence (B(f))	16
Figure 1.10 : Modèle typique d'extraction de paramètres standards (MFCC) pour la RAP... ..	17
Figure 2.1 : Schéma synoptique d'un système multi-bandes.....	26
Figure 2.2: Types de modélisation de la loi de probabilité des observations :(a) Modélisation de la loi continue (b) modélisation de la loi discrète.	27
Figure 2.3: Un exemple de HMM à 3 états modélisant un signal contenant des vecteurs acoustiques.....	28
Figure 2.4: Exemple d'un HMM avec une topologie de type Bakis à 3	29
Figure 2.5: Progression de la procédure estimation directe.....	34
Figure 2.6: Progression de la procédure Estimation rétrograde.....	35
Figure 2.8 : Progression de la procédure de Baum-Welch.....	38
Figure 2.9: Fonctionnement de la combinaison par intégrale flou.	44
Figure 3.1: Le découpage cellulaire	46
Figure 3.2 : Architecture du réseau GSM	47
Figure 3.3: Création d'un signal modulé par GMSK au départ d'un train binaire	50
Figure 3.4: Schéma fonctionnel simplifié d'un système de transmission vocale	51
Figure 3.5: Modèle source-filtre de production de la parole	53
Figure 3.6: Prédiction en boucle ouverte.....	54
Figure 3.7 : Prédiction en boucle fermée.....	54
Figure 3.8 : Prédiction long terme en boucle fermée.. ..	56
Figure 3.9: Codeur RPE-LTP (analyse en boucle fermée)	57
Figure 4.1 : Création de la base de données ARADIGIT_GSM	59
Figure 4.2: Système multi-bandes à quatre sous-bandes.....	61
Figure 4.3 : Schéma synoptique du module de recombinaison par FI	62
Figure 4.4: Les résultats comparatifs du taux de reconnaissance du locuteur.....	64
Figure 4.5: Taux de reconnaissances des 4 sous bandes de fréquences pour le même type de bases d'apprentissage et de test.. ..	65

Figure 4.6 : Taux de reconnaissances des 7 sous bandes de fréquences pour le même type de bases d'apprentissage et de test	67
Figure 4.7: Comparaison des taux de reconnaissances des systèmes multi-bandes avec le système de référence (B.T).....	67
Figure 4.8: Comparaison des taux de reconnaissances des systèmes multi-bandes + BT avec le système de référence (B.T)	69

Liste des tableaux

Tableau 3.1 : Distribution des paramètres codés pour chaque trame de 20 ms.....	58
Tableau 4.1: Taux de reconnaissance comparatifs entre bases ARBDIGIT8K et ARADIGIT_GSM, ARADIGIT_GSM_AW.....	63
Tableau 4.2: Taux de reconnaissance pour le système multi-bandes utilisant 4 bandes de fréquences.....	65
Tableau 4.3: Taux de reconnaissance pour le système multi-bandes utilisant 7 bandes de fréquences.....	66
Tableau 4.4 : Taux de reconnaissance pour les deux systèmes en ajoutant la bande totale.....	68

Liste des abréviations

CHMM	Continues Hidden Markov Models
DTW	Dynamic Time Warping (alignement temporel dynamique).
ETSI	European Standard Telecommunications Series.
EFR	Enhanced Full Rate
FR	Full Rate.
GMM	Gaussian Mixture Model.
GPRS	General Packet Radio Service
GSM	Global System Mobile.
HR	Half rate
HMM	Hidden Markov Models (modèles de Markov cachés).
HTK	HMM Toolkit (boîte à outils pour modèles de Markov cachés).
IAL	Identification Automatique du Locuteur.
LAR	Log Area Ratio.
LPC	Linear Predictive Coding.
MLE	Maximum Likelihood Estimation.
MLP	Perceptrons Multi-Couches (Perceptron Multi-Couches).
MFCC	Mel Frequency Cepstral Coefficient (coefficient cepstral en échelle Mel).
QV	Quantification Vectorielle.
RCs	<i>Reflection Coefficients</i>
RAP	Reconnaissance Automatique la parole.
RAL	Reconnaissance Automatique du Locuteur.
RPE-LTP	Regular Pulse Excitation with Long Term Prediction
UMTS	Universal Mobile Telecommunications System
VAL	Vérification Automatique du Locuteur

Introduction générale

L'intégration des technologies de la langue dans les réseaux de communication en particulier le réseau GSM (Global System Mobile) devient un sujet de recherche qui capte l'intérêt aussi bien des chercheurs que des prestataires intervenants dans le marché des technologies de l'information. Parmi les technologies de la langue, la Reconnaissance Automatique du Locuteur via le réseau de communication GSM constitue un problème d'actualité.

La Reconnaissance Automatique du Locuteur (RAL) est interprétée comme une tâche particulière de la reconnaissance de formes. La RAL regroupe les problèmes relatifs à l'identification ou à la vérification du locuteur sur base de l'information contenue dans le signal acoustique : il s'agit de reconnaître une personne à partir de sa voix. Le champ d'application est très large, allant des applications domestiques aux applications militaires en passant par des applications judiciaires. L'application téléphonique (téléphone, Internet), pour des services à distance, est intéressante car la reconnaissance automatique du locuteur est l'unique accès biométrique offrant cette alternative. De plus, elle propose des services dans divers domaines entre autres la validation des transactions bancaires par téléphone (pour venir en complément d'un code d'accès confidentiel et améliorer ainsi le service bancaire). [Grassi, 2000]

Défini en 1988, le codeur GSM FR est le premier codeur utilisé par le réseau GSM (Global System Mobile), ce dernier est le standard pour la téléphonie mobile utilisé en Algérie. Il appartient à la classe des codeurs de prédiction linéaire RPE-LTP (Regular Pulse Excitation with Long Term Prediction). Le rôle du codec GSM est de compresser le signal de parole avant sa transmission, de manière à réduire le nombre de bits nécessaires à sa représentation, tout en maintenant une qualité de signal acceptable en sortie. Il s'agit d'une compression avec pertes, affectant les propriétés spectrales du signal. Dans ce travail, nous nous intéressons à l'étude de l'influence de la parole transcodée GSM FR sur la performance globale du système de RAL dans le problème particulier de l'Identification Automatique du Locuteur (IAL). L'approche de reconnaissance utilisée est de type stochastique à paradigme multi-bandes.

Plus récemment sont apparus des systèmes qui utilisent une architecture multi-bandes regroupant des reconnaisseurs en parallèle. L'hypothèse sous-jacente à la réussite de ces systèmes est que les différents reconnaisseurs n'exploitent pas la même information spécifique du locuteur : si le premier reconnaisseur est défaillant sur un événement acoustique particulier, le second ne l'est pas forcément et la robustesse du système global en est

renforcée. En fait, que les systèmes utilisent un ensemble de classes pour modéliser les locuteurs, ou bien qu'ils emploient un ensemble de reconnaisseurs pour prendre une décision finale, les enjeux essentiels sont les mêmes :

- 1- Sélection des meilleures entités (classes ou reconnaisseurs) participant à la décision finale,
- 2- Choix d'une stratégie pour rassembler les scores ou décisions issus des entités sélectionnées auparavant.

L'idée initiale d'un système unique et intelligent est ainsi abandonnée au profit d'un système supervisant plusieurs modules de reconnaissance travaillant en parallèle. L'intelligence réside alors dans la façon de sélectionner et de faire coopérer les différents modules entre eux. Cette approche multi-bandes (parallèle) a été proposée pour améliorer la robustesse des systèmes de reconnaissance de la parole dans les environnements bruités. Dans ce travail, nous étudions son apport par rapport aux pertes introduites par le codage de la parole (GSM FR).

L'approche utilisée pour mettre en œuvre le système de reconnaissance multi-bande est la modélisation stochastique basée sur les modèles de Markov cachés continus (CHMM : Continuous Hidden Markov Models). Chaque sous reconnaisseur modélise une bande de fréquence indépendamment des autres, les sorties de ces reconnaisseurs seront fusionnées par une méthode de recombinaison des scores fondée sur la théorie de la logique floue.

Ce mémoire est organisé comme suit :

Le premier chapitre concerne les principes généraux de la reconnaissance du locuteur. On présentera la structure générale d'un système de RAL, les paramètres les plus efficaces de caractérisation du signal de parole et les méthodes de reconnaissance les plus utilisées actuellement.

Dans le deuxième chapitre nous présentons le formalisme de l'architecture multi-bandes à base des modèles de Markov cachés, et ceci en explicitant dans un premier temps la théorie et les principes des CHMMs pour la modélisation du locuteur, ensuite nous exposons la structure du système multi-bandes et son module de recombinaison.

Le chapitre trois s'intéresse à la présentation du réseau GSM et notamment le codec GSM FR standardisé par ETSI en exposant les différents étages du codeur ainsi que du décodeur.

Dans le chapitre quatre nous présentons le contexte expérimental et les résultats obtenus avec les différents systèmes mis au point ainsi que l'évaluation objective de ces systèmes.

Nous terminons ce manuscrit par une conclusion générale en mettant en évidence le travail réalisé dans ce mémoire. A la lumière des résultats obtenus nous donnons aussi les perspectives futures envisagées.

Principes généraux de la RAL

1.1 Introduction

La Reconnaissance Automatique du locuteur (RAL) est interprétée comme étant un processus de prise de décision utilisant des caractéristiques de la parole, afin de déterminer si une personne en particulier est à l'origine d'une énonciation [Rose, 2002]. Cette prise de décision porte sur une éventuelle familiarité entre la voix cible et les voix de référence [Yarmey, 1995]. Depuis plus de deux décennies, des recherches intensives dans ce domaine ont été réalisées par de nombreux laboratoires internationaux. Des progrès importants ont été accomplis grâce au développement d'algorithmes de reconnaissance puissants ainsi qu'aux avancées en sciences cognitives, technologie de l'information, l'intelligence artificielle, le traitement du signal, l'algorithmique, etc.

Ce chapitre s'intéresse aux concepts de bases de la RAL. Il présente tout d'abord, les divers problèmes de la RAL, les différentes tâches associées telles que l'Identification et la Vérification Automatique du Locuteur. Les tâches plus récentes comme l'Indexation en Locuteurs, la structure d'un système de RAL ainsi que son paradigme parallèle sont finalement présentés.

1.2 Complexité de la tâche de reconnaissance

1.2.1 Le mode de reconnaissance

Une première classification des systèmes de RAL repose sur le niveau de dépendance au texte. En premier lieu, on distingue généralement les systèmes dépendants du texte des systèmes indépendants du texte.

- ***RAL en mode dépendant du texte*** : la reconnaissance d'une personne est réalisée sur la base d'un message dont le contenu linguistique est connu par le système.
- ***RAL en mode indépendant du texte*** : dans ce mode, il n'existe aucune contrainte sur le message que le locuteur doit prononcer, ni sur la langue qu'il peut utiliser.

1.2.2 Notion de variabilité

La capacité des systèmes de RAL à authentifier une personne repose essentiellement sur la capacité à discriminer les individus grâce à la variabilité du signal vocal

1.2.2.1 Variabilité due au locuteur

- ***Variabilité interlocuteur*** : elle concerne les caractéristiques qui sont propres à chaque locuteur et qu'on ne retrouve pas chez d'autres locuteurs (variabilité

induite par le changement du locuteur). La grande variabilité entre les locuteurs est due, d'une part aux variations anatomiques des organes responsables de la production vocal (âge, sexe, etc....) d'autre part l'héritage linguistique et au milieu socioculturel de l'individu (l'accent régional).

- *Variabilité intra locuteur* : elle se présente lorsqu'un même locuteur prononce la même phrase ou le même mot à plusieurs reprises (parfois à différentes sessions). Plusieurs sources peuvent être attribuées à cette variabilité, telle que la fatigue, le stress, le sommeil, l'horaire de la journée (matin, soir), le débit d'élocution, l'état émotionnel, etc.

1.2.2.2 Variabilité due à l'environnement

Les conditions environnementales (bruit, changement climatique, etc.) peuvent influencer sur la performance globale d'un système de reconnaissance. La plupart des systèmes de reconnaissance fonctionnent relativement bien dans un environnement calme mais les performances se dégradent notablement si les conditions environnementales sont très bruitées (réelles). De plus le matériel utilisé influe aussi sur le signal lors de l'enregistrement du signal vocal (ex : microphone, combiné téléphonique), ou lors de transmission (ex : ligne téléphonique, air ambiant). Ces influences apparaissent le plus souvent sous forme de déformations ou dégradation du signal de parole.

1.3 Les différentes tâches en RAL

La reconnaissance du locuteur est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs sont regroupés en trois catégories principales :

- l'identification de locuteurs,
- la vérification du locuteur,
- l'indexation en locuteurs ou le suivi de locuteurs.

Chacune de ces catégories propose son protocole de reconnaissance selon que l'identité du locuteur à reconnaître soit proclamée, ou que les locuteurs à reconnaître soient connus ou non du système de RAL [Preti, 2008]. Le système de RAL peut valider une identité pour la

vérification du locuteur, proposer une identité à partir d'un ensemble de locuteurs, déterminer les durées de parole d'un locuteur, compter le nombre de locuteurs présents dans un signal.

1.3.1 Identification automatique du locuteur

L'IAL est un processus qui consiste à déterminer parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vue schématique (Figure 1.1), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu, la séquence de parole est comparée à un modèle de référence caractéristique du locuteur. En sortie, on récupère l'identité du locuteur dont le modèle de référence est le plus proche de la séquence présentée à l'entrée du système [Fredouille, 2000] [Sayoud, 2003]. Deux modes d'identification sont proposés :

- ✓ L'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système.
- ✓ L'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu par le système qui doit décider de la fiabilité de son jugement en acceptant en rejetant l'identité qu'il a trouvée.

De par son principe qui est de déterminer une identité parmi les identités potentielles, les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente.

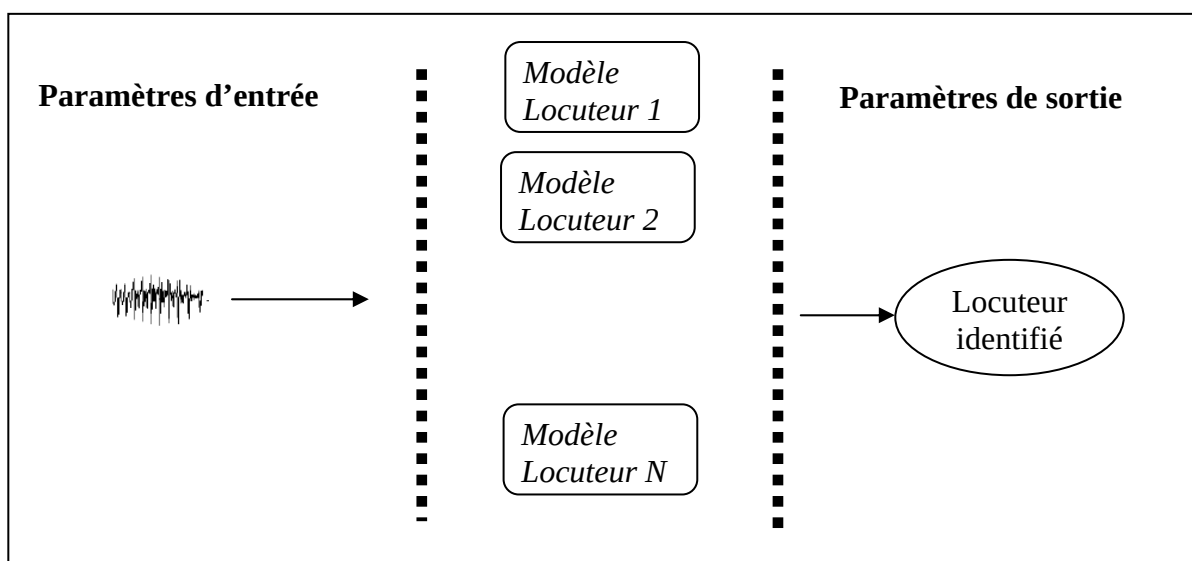


Figure 1.1 : La tâche d'identification automatique du locuteur

1.3.2 Vérification automatique du locuteur

La vérification du locuteur consiste à déterminer si l'identité proclamée d'un message vocal correspond à la véritable identité du locuteur. En pratique la réponse est binaire, acceptation ou rejet. Les éléments mis en jeu sont donc une identité proclamée et la référence associée à un échantillon connu de l'identité proclamée. Une mesure de similarité entre le signal à vérifier et cette référence est calculée. Cette mesure est comparée à un seuil de vérification. Dans le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté [Preti, 2008].

1.3.3 Indexation en locuteur

L'indexation en locuteurs permet de déterminer les temps de parole des individus dans un signal audio. La spécificité de cette tâche réside dans le fait que le système ne détienne pas de référence pour les locuteurs présents dans le signal audio. Le mécanisme d'apprentissage est aveugle et adaptatif [Meignier, 2002]. Il est possible de segmenter un signal audio par prise de parole des intervenants, étiqueter des données audio pour permettre des recherches de documents audio par locuteurs ou, enfin, identifier le nombre de locuteurs présents dans le signal.

Le suivi de locuteur est similaire à l'indexation en locuteur, à ceci près que les locuteurs présents dans le signal sont connus par le système de RAL. Il s'agit donc d'une simplification de la tâche d'indexation en locuteur mais qui reste néanmoins, une tâche très complexe.

1.3.4 Types d'erreur

Chacune des tâches citées précédemment possède un type d'erreurs, le système d'identification en ensemble fermé dont ses performances se mesurent par son taux de mauvaise identification. . Celles d'un système de vérification se mesurent par son taux de fausse acceptation et par son taux de faux rejet. Enfin celles d'un système d'identification en ensemble ouvert se mesurent par son taux de mauvaise identification, son taux de fausse acceptation (c'est à dire un imposteur est accepté comme l'un des locuteurs de la base de référence) et son taux de faux rejet.

1.4 Architecture d'un système de RAL

L'architecture d'un système de RAL est décomposée en différents modules de traitement (figure 1.2). La plupart des approches standards en RAL utilisent une structure similaire :

- ❖ Extraction des paramètres : il extrait du signal les éléments permettant une discrimination des locuteurs,
- ❖ Modélisation de locuteurs : à partir des données du locuteur, extraites par le module de paramétrisation, un modèle de référence du locuteur est créée. Elle sert d'élément de référence pour la RAL, où elle est comparée avec le signal de test.
- ❖ Décision : effectue la comparaison entre le modèle de référence (tâche de vérification) ou les modèles de référence (tâche d'identification) et le signal de test. A partir du résultat de ce test, cet étage rend la décision (le nom d'un locuteur en identification, un rejet ou une acceptation en vérification) en tenant compte éventuellement du niveau de sécurité souhaité.

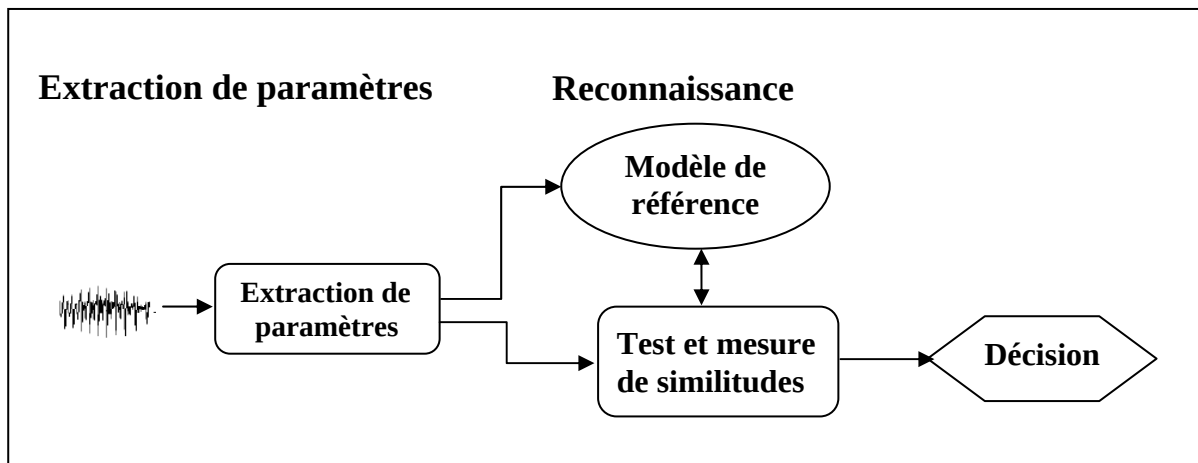


Figure 1.2 : Structure générale d'un système de RAL

1.4.1 Extraction des paramètres

Le signal de parole présente de la redondance et il est porteur de plusieurs types d'informations, ce qui justifie la recherche d'une représentation spécifiquement pertinente pour la reconnaissance. L'extraction des paramètres consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques. Cette extraction débute par une mise en forme du signal de parole suivie par le calcul de coefficients acoustiques comme indiqué sur la figure 1.3 suivante :

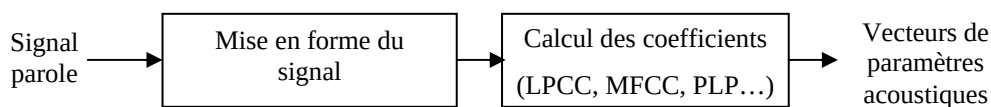


Figure 1.3 : Schémas synoptique de l'extraction de paramètres acoustiques.

1.4.1.1 Mise en forme du signal de parole

L'étape de mise en forme du signal de parole consiste à lui appliquer une série de traitement du signal dont le but est de l'adapter au calcul de coefficients acoustiques. La figure 1.3 illustre ces différentes opérations.

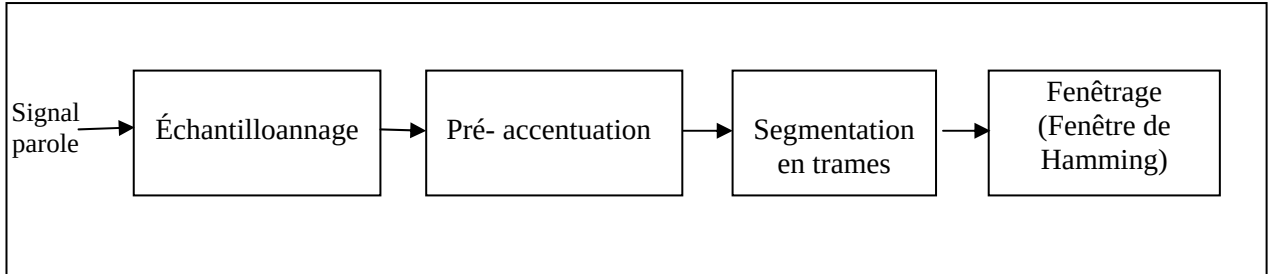


Figure 1.4 : Etapes de mise en forme du signal de parole.

- **Echantillonnage** : effectué à une fréquence f_e compatible avec les exigences du théorème de Shannon, la perte d'information entre le signal continu et le signal discret correspondant est nulle si et seulement si la fréquence d'échantillonnage est au moins supérieure ou égale au double de la fréquence la plus haute contenue dans ce signal.

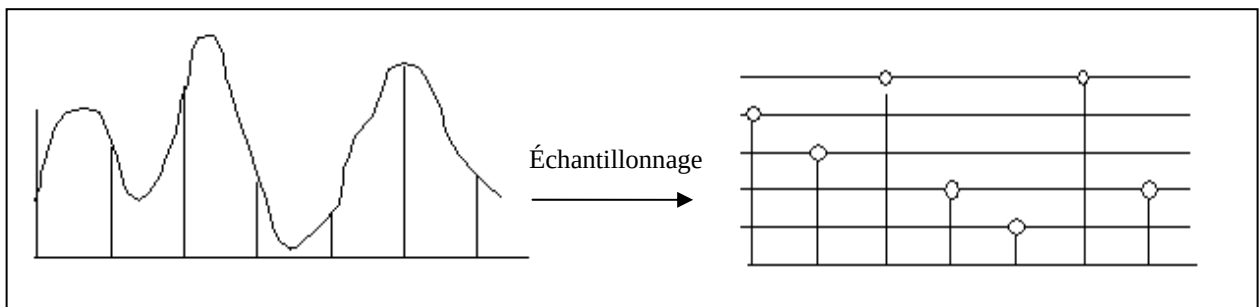


Figure 1.5: Exemple d'échantillonnage d'un signal de parole.

- **Pré- accentuation** : elle consiste à relever les hautes fréquences qui sont moins énergétiques que les basses fréquences, le filtre de pré accentuation est souvent un filtre à réponse impulsionnelle finie (R.I.F) non-récuratif du 1^{er} ordre (donc à un seul coefficient) dont la fonction de transfert est donnée par:

$$H_1(z) = 1 - a_1 z^{-1} \quad (1.1)$$

Les valeurs typiques de « a_1 » sont comprises entre [0.4 et 1.0], une valeur proche de 1 est souvent utilisée pour la parole (de l'ordre de 0.96).

- **Segmentation en trame et fenêtrage** : l'observation du signal de la parole indique qu'il n'évolue pas ou peu sur des durées de quelques millièmes de secondes. On peut donc

considérer ce signal comme étant stationnaire durant ce temps (stationnarité locale). A cause de ces propriétés, le signal de parole est analysé sur des durées successives limitées (on dit qu'il est analysé en trames). Une trame est un prélèvement limité en temps du signal. Ces durées successives s'obtiennent en prélevant les échantillons de parole à l'aide d'une fenêtre temporelle glissante. En général, les fenêtres successives se recouvrent. Le signal $x'(n)$ est fragmenté en trame de N échantillons avec un recouvrement de M échantillons. Le signal de parole est généralement analysé avec un recouvrement des trames d'une durée comprise entre 10ms et 30ms [Calliope, 1989].

Enjeu: Choix de la fenêtre

Une bonne fenêtre qui réduit la convolution dont la largeur du lobe principal (le lobe central) est faible et la prééminence du lobe central forte (de faibles lobes secondaires).

Exemple de fenêtre : *La Fenêtre de Hamming* de longueur N

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi}{T} n\right) \quad n = 0, 1, 2, \dots, N-1 \quad (1.2)$$

1.4.1.2 Calcul des coefficients

Les performances d'un système d'identification du locuteur dépendent essentiellement de la qualité des paramètres choisis. La plus part des systèmes de RAL utilisent les **mel frequency cepstral coefficients (MFCC)** et **Linear Prediction Cepstrum Coefficients (LPCC)** [Pravinkumar, 2001] et ceci pour les raisons suivantes :

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

Il existe aussi d'autres mesures qui peuvent être rajoutés aux vecteurs de paramètres comme l'énergie ou les premières et deuxièmes dérivées des coefficients.

1.4.1.3 L'analyse LPC

La théorie du codage par prédiction linéaire appliquée à la parole figure parmi les techniques les plus utilisées en traitement automatique de la parole [Markel, 1976].

Principe de la prédiction linéaire

La prédiction linéaire est basée sur l'hypothèse que chaque échantillon du signal original $s(n)$ peut être approché par une combinaison linéaire des p échantillons qui le précèdent [Dekker,2003] :

$$s(n) = \left[\sum_{i=1}^P -a_i s(n-i) \right] + e(n) = \hat{s}(n) + e(n) \quad (1.3)$$

Dans cette expression, $\hat{s}(n)$ est l'échantillon prédit, les coefficients $a(i)$ sont appelés coefficients de prédiction d'ordre p et le terme $e(n)$ est l'erreur de prédiction d'ordre p , définie par l'équation (1.5) suivante:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^P a(i)s(n-i) = \sum_{i=0}^P a(i)s(n-i) \quad (1.4)$$

Tel que $a_0 = 1$.

On cherche à estimer les coefficients de prédiction a_i tout en minimisant l'erreur $e(n)$ précédents. Tous ces coefficients sont supposés constants dans toute la trame d'analyse. En d'autres termes, on peut exprimer $s(n)$ par sa transformée en Z comme suit :

$$E(z) = S(z) + \sum_{i=1}^P a_i S(z) z^{-1} = S(z) \left[1 + \sum_{i=1}^P a_i z^{-i} \right] = S(z) A(z) \quad (1.5)$$

$$\text{Où : } A(z) = 1 + \sum_{i=1}^P a_i z^{-i} = \sum_{i=0}^P a_i z^{-i} \quad (1.6)$$

L'équation (1.6) peut être écrite comme :

$$S(z) = E(z) \frac{1}{A(z)} = E(z) H(z) \quad (1.7)$$

Ceci montre que le signal parole peut être vu comme la sortie d'un filtre autorégressif tout pôles (all-pole digital filter) dont la fonction de transfert est définie par :

$$H(z) = \frac{1}{A(z)} \quad (1.8)$$

Et dont l'entrée du filtre est l'erreur $E(z)$.

Estimation des coefficients de prédiction

En résumé, le problème principal du codage prédictif linéaire est de déterminer directement, à partir d'un signal de paroles, un ensemble de coefficients $\{ a_i \}$ de sorte que l'erreur quadratique moyenne de prédiction E soit minimisée dans la fenêtre d'analyse. Pour une trame de longueur $[n_0, n_1]$, l'erreur E est calculée comme suit [Boite, 2000] [Dekker, 2003]:

$$E_n = \sum_{n=n_0}^{n_1} e^2(n) \quad (1.9)$$

Pour minimiser la fonction E , nous calculons sa dérivée partielle par rapport aux coefficients de prédictions a_i (c'est une méthode d'optimisation classique):

$$\frac{\partial E_n}{\partial a_k} = 0 \quad k=0,1,\dots,p \quad (1.10)$$

La résolution de l'équation (1.11) nous donnent un système de p équations et p inconnus dit Yule-walker, qui est représenté si dessus dont sa matrice est une matrice d'autocorrelation :

$$\begin{bmatrix} r_n(0) & r_n(1) & \dots & r_n(p-1) \\ r_n(1) & r_n(0) & \dots & r_n(p-2) \\ \dots & \dots & \dots & \dots \\ r_n(p-1) & r_n(p-2) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ \dots \\ r_n(p) \end{bmatrix} \quad (1.11)$$

La résolution de ce système permettra d'obtenir les coefficients de prédiction a_i . Parmi les méthodes de résolution de ce système, on peut citer la méthode d'auto-corrélation, la méthode de covariance, la méthode de burg [Keiler, 2000], ou la méthode de Levinson-Durbin. Dans ce qui suit, nous étudierons la méthode de Levinson- Durbin.

Algorithme de Levinson-Durbin :

La matrice des valeurs d'auto-corrélation est une matrice Toeplitz. En effet, elle est symétrique et tous les éléments sur chaque diagonale principale sont identiques. L'algorithme récursif Levinson-Durbin se développe comme suit [Haykin, 1986] :

$$E_{(0)} = r(0) \quad (1.12)$$

$$k_i = \left\{ r(i) - \sum_{j=1}^{i-1} a_j^{i-1} r(|i-j|) \right\} / E_{i-1}, k = \overline{1, p} \quad (1.13)$$

$$a_i^{(i)} = k_i \quad (1.14)$$

$$a_j^i = a_j^{(i-1)} - k_i a_{i-j}^{i-1}, j = \overline{1, i-1} \quad (1.15)$$

$$E_{(i)} = (1 - k_i^2) E^{(i-1)} \quad (1.16)$$

Notons qu'à partir des coefficients $a_0, a_1, \dots, a_p$, on peut déduire d'autres coefficients qui sont : *LPCC (Linear Prediction Cepstral Coefficients)*, *LAR (Log Area Ratio)*, et *RC (Reflection Coefficients)*

a. LPCC (Linear Prediction Cepstral Coefficients) :

L'un des ensembles de paramètres les plus importants que nous pouvons déduire en profitant des coefficients **a** **i** sont les coefficients cepstraux **LPCC** en utilisant la procédure récursive suivante [MAKH, 75] :

$$c_0 = r(0) \quad (1.17)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, 1 \leq m \leq p \quad (1.18)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, m > p \quad (1.19)$$

b. LAR (Log Area Ratio):

Les coefficients LAR (Log Area Ratio) sont dérivés à partir des coefficients LPC, le modèle LPC qui caractérise le conduit vocal d'une personne, peut être transformé en autres coefficients appelés LAR. Dans cette analyse, le conduit vocal d'une personne est modélisé comme un tube acoustique non uniforme, formé de **P** tubes cylindriques et uniformes avec plusieurs zones transversales [Joseph, 1997]. La glotte connectée au premier tube n'a pas de zone, par contre les lèvres connectées au dernier tube ont une zone infinie, comme il est illustré par la figure ci contre.

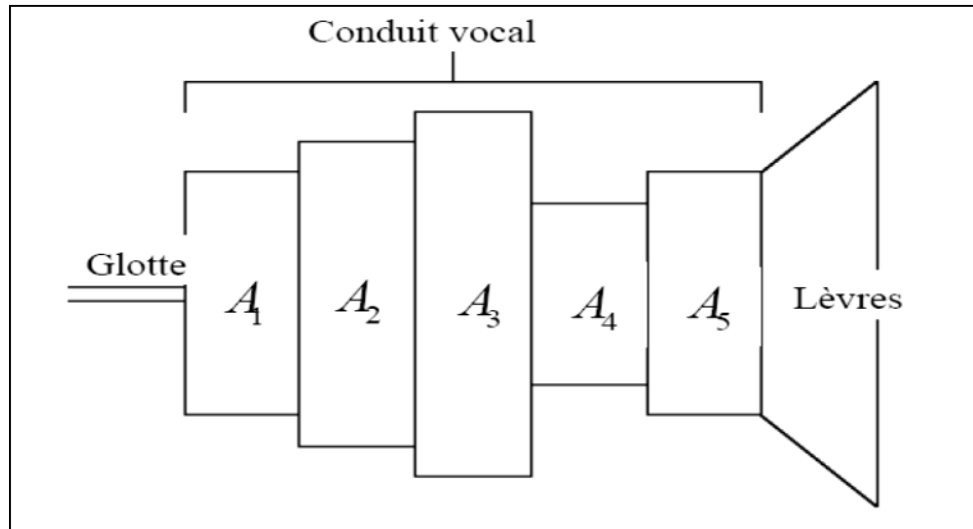


Figure 1.6: Le modèle de production de parole avec les tubes acoustiques.

Les coefficients LAR sont calculés par le logarithme du rapport entre la zone transversale de chacun des deux tubes connectés, le nombre de tubes cylindriques dans le modèle est égal au

nombre de coefficient LAR plus un. La relation entre les coefficients LAR et les coefficients LPC est donnée par la formule suivante :

$$LAR_i = \log\left(\frac{A_i}{A_{i+1}}\right) = \log\left(\frac{1+k_i}{1-k_i}\right), A_{p+1} = 1 \quad (1.20)$$

Avec k_i est le i ème coefficient de parcours [Joseph, 1997]:

c. Coefficients de réflexion

Les coefficients de réflexion sont les coefficients de parcours k_i obtenus dans l'analyse LPC.

1.4.1.4 L'analyse cepstrale

L'analyse cepstrale utilise le modèle source/conduit, avec une source périodique (sons voisés). Le cepstre permet de séparer la structure harmonique de la structure formantique du signal. Le cepstre réel ou coefficient cepstral est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale de puissance du signal, sur la fenêtre d'analyse [Noll, 1964]. Les étapes de calcul des coefficients cepstraux est représenté dans la figure suivante :

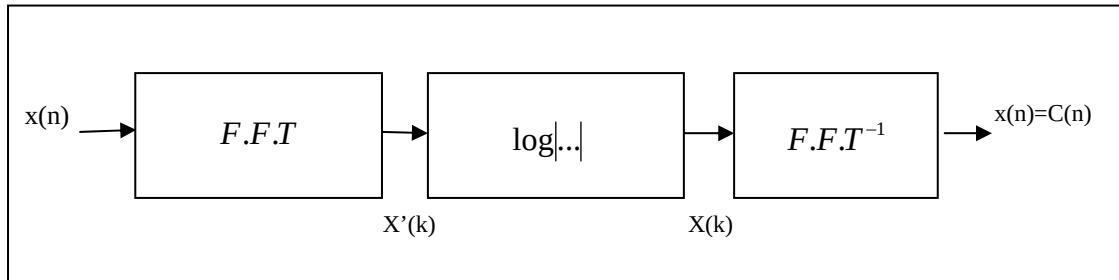


Figure 1.7: Calcul des coefficients cepstraux.

$$C(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \exp(j \frac{2\pi}{N} kn) \quad 0 \leq n \leq N-1 \quad (1.21)$$

$C(n)$: LFCC (Linear Frequency Cepstral Coefficient)

Cette équation $C(n)$ est aussi la DFT inverse du logarithme de l'amplitude. Elle peut être simplifiée en notant que le logarithme du spectre est une fonction réelle symétrique (c.à.d que la DFT inverse est appliquée à un "signal" réel symétrique). On peut donc utiliser au lieu et place une transformée en cosinus.

Nous avons:

$$\hat{x}(n) = C(n) = TFD^{-1}(X(k)) = \sum_{k=0}^{N-1} X(k) \exp(j \frac{2\pi}{N} kn) \quad 0 \leq n \leq N-1 \quad (1.22)$$

Si $X(k)$ est réel et symétrique ($X(N-m) = X(m)$ pour une position m) alors:

$$C(n) = \frac{2}{N} \sum_{k=1}^{N-1} X(k) \cos\left(\frac{2\pi}{N} kn\right) \quad 0 \leq n \leq N-1 \quad (1.23)$$

Les coefficients MFCCs (Mel-scale Frequency Cepstral Coefficients)

Les coefficients MFCC sont une extension des coefficients cepstraux par le passage de l'échelle fréquentielle linéaire à une échelle fréquentielle non linéaire dite l'échelle Mel [Debyeche, 2007]. La fréquence mel-échelle est définie par:

$$B(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (1.24)$$

Où f est la fréquence en Hz, $B(f)$ est la fréquence mel-échelle de f .

L'intérêt de l'échelle Mel est d'être assez proche d'échelles issues d'études sur la perception sonore et sur les bandes passantes critiques de l'oreille [Barras, 1996]. Le calcul des paramètres MFCC se réalise de la façon suivante (figure 1.9):

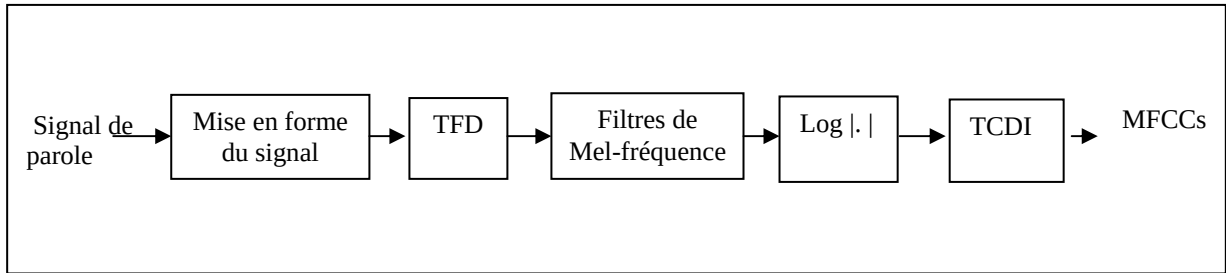


Figure 1.8: Calcul des MFCCs.

Après la mise en forme du signal, une TFD (Transformée de Fourier Discrète) est calculée pour faire passer le signal de parole dans le domaine spectral :

Pour un signal discret $\{s[n]\}$ avec $0 < n < N$, où N est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète $S[k]$ est obtenue par :

$$s[k] = \sum_{n=0}^{N-1} x[n] e^{-j 2 \pi nk / N} \quad (1.25)$$

Le spectre du signal est multiplié avec des filtres triangulaires (figure 1.10) dont les bandes passantes sont équivalentes en domaine mel-fréquence. Les points frontières $B[m]$ des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m \frac{B(f_h) - B(f_l)}{M + 1} \quad 0 \leq m \leq M + 1 \quad (1.26)$$

Où M est le nombre de filtres, f_h est la fréquence la plus haute et f_l est la fréquence la plus basse pour le traitement du signal.

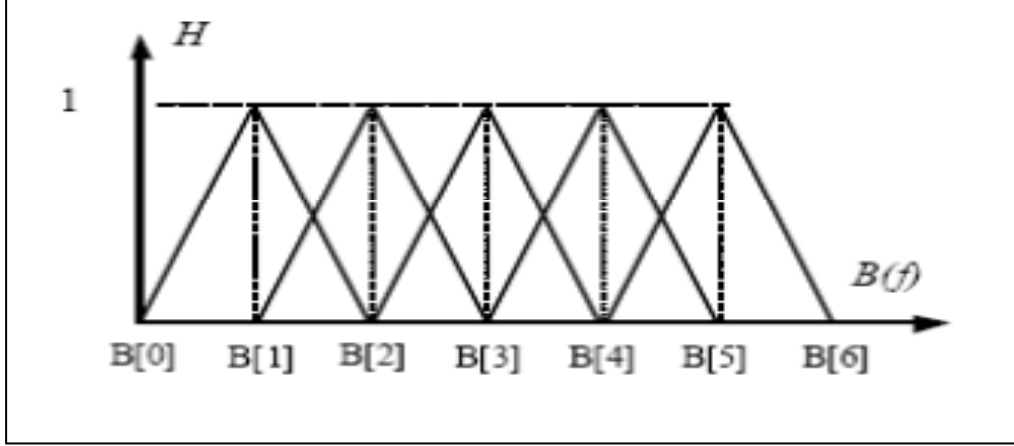


Figure 1.9: Les filtres triangulaires passe-bande en Mel-Fréquence ($B(f)$).

Dans le domaine fréquentiel, les points $f[m]$ discrets correspondants sont calculés par l'équation :

$$f[m] = \left(\frac{N}{F_s} \right) B^{-1} \left(B(f_l) + m \frac{B(f_h) - B(f_l)}{M + 1} \right) \quad (1.27)$$

Où B^{-1} est la transformée de mel-fréquence en fréquence.

$$B^{-1}(m) = 700 * (10^{m/2595} - 1). \quad (1.28)$$

Le coefficient $H_m[k]$ de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & \text{si } k \leq f[m-1] \\ \frac{k - f[m-1]}{f[m] - f[m-1]} & \text{si } f[m-1] \leq k \leq f[m] \\ \frac{f[m+1] - k}{f[m+1] - f[m]} & \text{si } f[m] \leq k \leq f[m+1] \\ 0 & \text{si } k \geq f[m+1] \end{cases} \quad (1.29)$$

Pour un spectre lissé et stable à la sortie des filtres, le logarithme de spectre d'amplitude est calculé :

$$E[m] = \log \left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k] \right] \quad 0 \leq m \leq M \quad (1.30)$$

Les coefficients cepstraux de mel-fréquence (MFCCs) seront obtenus par une TCDI (Transformée en Cosinus Discrète Inverse) qui permet d'obtenir des coefficients peu corrélés à partir des coefficients aux sorties des filtres :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos \left(\frac{\pi m(m+1)}{2M} \right) \quad 0 \leq n \leq M \quad (1.31)$$

Une dizaine de coefficient MFCCs sont généralement considérés comme suffisants pour les expériences de reconnaissance [Barras, 1996].

1.4.1.5 Calcul des coefficients différentiels

Afin de prendre en compte la dynamique du signal, nous ajoutons aux paramètres acoustiques les coefficients différentiels (ou coefficients delta) du premier et du second ordre. Soit le vecteur acoustique à N composantes $C_t = \{c_t^1, c_t^2, \dots, c_t^N\}$. Les coefficients delta de premier ordre sont alors estimés par :

$$\Delta C_t = \frac{\sum_{K=-L}^L K C_{t-K}}{\sum_{K=-L}^L K^2} \quad (1.32)$$

Les coefficients du second ordre sont calculés en itérant deux fois l'expression (1.32). La figure 1.11 schématise un exemple du module typique d'extraction de paramètres acoustiques pour un système de reconnaissance.

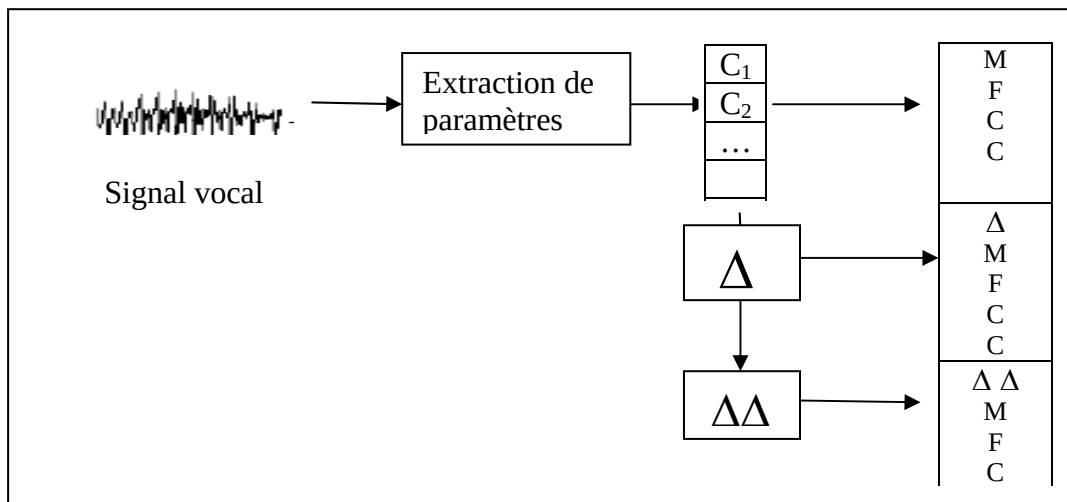


Figure 1.10 : Module typique d'extraction de paramètres standards (MFCC) .

Du point de vue de l'étude bibliographique [Barras, 1996] [Jacob, 1995] [Igounet, 1998] [Young, 1997], à l'heure actuelle, le choix des paramètres MFCCs semble être le plus satisfaisant pour représenter le signal de parole dans le cadre de la reconnaissance (de la parole ou du locuteur). D'où nous avons choisi cette paramétrisation dans notre système de reconnaissance.

1.4.2 Approches de reconnaissance

Les systèmes de reconnaissance automatique de la parole et du locuteur actuels utilisent pour la plupart des algorithmes de comparaison de motifs ("patterns matching" en anglais). Les motifs utilisés sont basés sur les représentations cepstrales. Dans le cadre de la reconnaissance de locuteurs, on modélise les différentes prononciations qu'un locuteur peut avoir produits pour le *même motif*.

De manière à mémoriser des caractéristiques qui dépendent du locuteur, on utilise des algorithmes capables de capturer les points communs entre différentes représentations de motifs spectraux issus du même locuteur (constituant ainsi un *modèle du locuteur*), tout en ayant la possibilité de s'adapter aux variations d'échelles fréquentielles et temporelles liées au signal de parole. Ces motifs peuvent être soit des segments de parole déterminés (mots, phonèmes) si nous travaillons en mode dépendant du texte, soit des segments de parole dont on ne connaît pas le contenu phonétique si l'application fonctionne en mode indépendant du texte. Ces algorithmes sont couplés avec une mesure qui permettra de donner une valeur de distorsion (ou de similitude) entre le modèle du locuteur et un motif inconnu dont on cherche à déterminer la personne qui l'a prononcé. Nous proposons ici un aperçu des approches les plus utilisées en reconnaissance du locuteur.

1.4.2.1 L'approche vectorielle

Dans l'approche vectorielle, un modèle de locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique construit lors de la phase de paramétrisation des signaux d'apprentissage. Lors de la reconnaissance, une distance entre cet ensemble de vecteurs et les vecteurs de paramètres issus des signaux de test est calculée.

L'approche vectorielle compte deux grandes techniques : la programmation dynamique et la quantification vectorielle

❖ programmation dynamique

La programmation dynamique (Dynamic Time Warping : DTW) consiste à aligner temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage, le modèle de locuteur est tout simplement l'ensemble des vecteurs de

paramètres obtenus après paramétrisation des signaux d'apprentissage. Une distance est calculée entre vecteurs d'apprentissage et de test et moyennée sur l'ensemble de la séquence. De par son principe, la programmation dynamique est utilisée exclusivement en mode dépendant du texte [Furui, 1981], [Booth, 1993], [Yu, 1995]. Très rapide et montrant des performances relativement bonnes, la programmation dynamique est toutefois très sensible à la qualité d'alignement et notamment au choix du point de départ.

❖ Quantification vectorielle

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

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

La quantification vectorielle s'applique en mode dépendant ou indépendant du texte [Soong, 1992], [Mason, 1989], [Matsui, 1992]. La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire: plus la taille du dictionnaire augmente, meilleures sont les performances; néanmoins, le processus devient d'autant plus lent.

1.4.2.2 L'approche statistique

• Les mélanges de gaussiennes

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour Gaussian Mixture Models) consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes [Reynolds, 1995]. Ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante de gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble être bien prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes sont considéré comme une extension de quantification vectorielle. [Mami, 2003]

- **Les modèles de Markov cachés**

Les modèles de Markov (ou HMM Hidden Markov Models) ont été initialement introduits en reconnaissance de la parole, puis leurs utilisations s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Dans cette approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'état avec des probabilités de transition d'un état à l'autre, et de loi de probabilité (probabilité d'observation d'un vecteur acoustique dans un état). Nous allons aborder cette méthode avec plus de détaille dans le chapitre suivant.

- **Mesures statistiques du second ordre**

Cette partie présente une famille de mesures de similarité entre locuteurs. Ces mesures repose sur les caractéristiques du second ordre d'une séquence de vecteurs, c'est-à-dire sur le vecteur moyen et la matrice de covariance de cette séquence, plusieurs mesure de distances ont été réalisées : on peut citer , le rapport de vraisemblance, la distance de *Kullbak-leibler* , maximum de vraisemblance, test de sphéricité, déviation absolue des valeurs propres. Ces mesures donnent des résultats très encourageants sur la sur la parole propre, et naturellement, voient leurs performances se dégrader sur la parole téléphonique. De part leur relative simplicité, ces mesures peuvent également servir de références pour évaluer la qualité d'une base de données [bimbot, 1995],[Margin, 1995].

1.4.2.3 L'approche connexionniste

L'approche connexionniste, telle que nous l'entendons ici, repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs clients afin que ce dernier apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle client se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du client concerné ainsi que celles des autres clients du système sont fournies en entrée. Différents types de modèles de réseaux sont proposés dans la littérature MLP7 [Oglesby, 1990], RBF8 [Oglesby, 1991], [Frederickson, 1994] et LVQ9 [Bennani, 1990]. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit produite par un réseau de neurones est calculée. Le principal inconvénient de l'approche connexionniste est la complexité d'apprentissage. En outre, elle pose le problème de l'ajout d'un nouveau client qui nécessite dans la majorité des

prises en œuvre le réapprentissage de tous les modèles. En effet, une nouvelle phase de classification est nécessaire afin de prendre en compte le nouveau client au sein du processus de discrimination entre locuteurs.

1.4.2.4 L'approche prédictive

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

1.4.3 La prise de décision

La prise de décision en RAL est basée sur le formalisme probabiliste. Elle est différente pour l'identification et la vérification du locuteur.

a. Décision pour l'identification automatique du locuteur

Considérons une population de locuteur $i = 1, \dots, N$ avec M_i la référence associée au locuteur i . L'identité retournée M , présente dans le signal X , est alors celle qui maximalise la probabilité :

$$M = \arg \max_i P(M_i / X) \quad (1.33)$$

Sans informations a priori sur l'apparition des locuteurs, $P(M_i)$, et en appliquant la règle de Bayes la relation devient :

$$M = \arg \max_i P(M_i / X) = \arg \max_i \frac{p(X / M_i) \cdot P(M_i)}{P(X)} = \arg \max_i p(X / M_i) \quad (1.34)$$

Où $P(X/M_i)$ est la fonction de vraisemblance du locuteur i qui approxime la densité de probabilité des observations du locuteur i . Lorsque le nombre de locuteurs augmente dans la base de référence, des proximités entre locuteurs apparaissent. Il est plus difficile de

différencier les locuteurs et les performances se dégradent. Il faut aussi noter que, dans ce cas, les ressources nécessaires et les temps de traitement augmentent.

b. Décision pour la vérification automatique du locuteur

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

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

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

1.5 Modèles parallèles

Récemment, de nouvelles techniques sont apparues en vue d'augmenter la robustesse des systèmes de reconnaissance : leur caractéristique commune est l'utilisation de plusieurs reconnaisseurs (travaillant en parallèle) qui sont recombinaisonnés pour prendre une décision finale. Cependant, ces techniques sont divisées en trois classes :

- **Approche multi-classifieurs** : plusieurs méthodes de RAL sont appliquées à un même signal de parole inconnu et la décision résulte de la recombinaison des scores obtenus avec chaque méthode.
- **Approche multimodale** : une ou plusieurs modalités (mouvement des lèvres, image du visage) sont ajoutées à la parole et les résultats issus des reconnaisseurs associés à chaque modalité sont fusionnés.
- **Approche multi-bandes** : cas particulier d'une approche plus générale qui consiste à utiliser plusieurs ensembles de paramètres (correspondant chacun à une information partielle du signal de parole) sur lesquels différents reconnaisseurs sont appliqués puis recombinaisonnés. Les points essentiels dans ces approches concernent la normalisation des scores issus de chaque reconheisseur et le choix de stratégies de recombinaison efficaces.

1.6 Conclusion

Dans ce chapitre, nous avons présenté les méthodes d'analyse du signal et les techniques de reconnaissance les plus utilisées actuellement. Parmi les méthodes développées, nous avons utilisé la modélisation stochastique basée sur les modèles de Markov cachés. Cette modélisation est appliquée à une représentation spectrale du signal sous forme de vecteurs de paramètres (MFCC). Le formalisme *multi-bandes* basé sur les modèles de Markov cachés sera présenté plus en détails dans le chapitre suivant étant donné que nous avons utilisé cette approche dans ce travail.

Paradigme multi-bandes à base de CHMM

2.1 Introduction

Le choix du paradigme multi-bandes par les modèles de Markov cachés continus réside dans l'hypothèse que des reconnaisseurs HMM travaillant indépendamment sur différents sous-ensembles de paramètres acoustiques sont plus robustes qu'un seul reconnaisseur utilisant l'espace acoustique (fréquentiel) dans sa globalité. Cette robustesse est cependant conditionnée principalement par le choix des sous-bandes de fréquences et l'approche de recombinaison.

Dans ce chapitre nous discutons d'abord les motivations et avantages de l'architecture multi-bandes et nous mettons en évidence les bandes de fréquence les plus utiles pour l'identification automatique du locuteur. Nous présentons ensuite l'architecture du système et la stratégie de recombinaison des scores ou des décisions issues de chaque sous-bande fréquentielle basée sur la théorie de la logique floue.

2.2 Motivations et Avantages

L'idée d'appliquer des reconnaisseurs indépendants sur différentes sous-bandes a d'abord été envisagée par [Duchnowski, 1993] qui proposait une nouvelle architecture en quatre bandes pour la reconnaissance de la parole. Des approches multi-bandes ont ensuite été proposées par [Bourlard, 1996] et [Hermansky, 1996] également pour la RAP. En ce qui concerne la RAL, les travaux de L. Besacier ont donné de bon résultats [Besacier, 1998]. La principale motivation vient du fait que les systèmes de reconnaissance classiques traitent chaque vecteur acoustique comme une entité ; en conséquence, même si une seule bande de fréquence est bruitée, tout le vecteur acoustique est contaminé et les performances du reconnaisseur en sont fortement affectées (même si le bruit n'affecte pas les bandes de fréquence les plus importantes pour la reconnaissance). De plus, les travaux de [Fletcher, 1953] et plus récemment de [Allen, 1994] suggèrent que le processus de reconnaissance humain soit basé sur des sous-bandes de fréquences qui sont traitées indépendamment les unes des autres.

Les résultats de plusieurs travaux (*Arai et al*, Miller et Nicelly, Kryter) montrent que les humains perçoivent la parole avec une relative grande précision même avec un nombre limité de paramètres spectraux (pour cause de filtrage). Ils suggèrent donc que le spectre de parole contient une quantité appréciable d'information redondante et que les humains sont dotés de capacité d'intégration facile de paramètres acoustiques provenant de différentes régions fréquentielles [Debyeche, 2007].

Les avantages techniques espérés d'une approche en sous-bandes sont les suivants [Bourlard,1996] [Hermansky, 1996] :

- Meilleure robustesse dans le cas de bruits n'affectant que quelques bandes spécifiques,
- assouplir les contraintes de synchronie inhérentes aux HMM,
- appliquer différentes stratégies de reconnaissance adaptées à chaque bande de fréquence,
- mettre en évidence des sous-bandes plus critiques que d'autres pour la RAL.

2.3 Définition des sous-bandes de fréquences

Plusieurs travaux se sont intéressés à la définition des sous-bandes à utiliser et les limites fréquentielles de chacune de ces sous-bandes. Nous pouvons citer en particulier les travaux de *Bourlard* et les travaux de *Hermansky*. H. Bourlard et S. Dupont ont utilisé dans leurs systèmes 3, 4 et 6 bandes de fréquences. Les meilleurs résultats qu'ils ont obtenus correspondent à l'utilisation de quatre bandes de fréquences dont les limites sont (0-901 Hz, 797-1661 Hz, 1493-2547 Hz et 2298-4000 Hz) [Bourlard,1996]. Ces limites sont calculées à partir des limites des bandes critiques proposées par B.J Allen [Allen,1994]. Par contre H. Hermansky et S. Tiberewala ont utilisé un système à 2,4, et 7 bandes, le meilleur résultat est celui de 7 bandes dont chacune englobe 2 bandes critiques suggéré par Allen (0- 360 Hz, 330-640Hz, 580-950 Hz, 860-1360Hz, 1265- 1920 Hz, 1800-2700 Hz, 2515-4000 Hz) [Hermansky, 1997].

2.4 Architecture du système multi-bandes

Une approche multi-bandes peut être vue comme un cas particulier d'une approche plus générale qui consiste à utiliser plusieurs ensembles de paramètres (correspondant chacun à une information partielle du signal de parole) sur lesquels différents reconnaisseurs sont appliqués puis recombinaisonnés (Figure 2.1). la bande de fréquence totale du signal de parole est divisée en un nombre 'N' de sous-bandes sur les quelles des analyses sont faites afin d'extraire les paramètres acoustiques. Pour chaque sous-bande un sous-reconnaisseur (estimateur de probabilités) est développé, chaque sous-reconnaisseur possède ses propres modèles acoustiques. Les sorties de tous les sous-reconnaisseurs sont par la suite fusionnées afin de produire la décision globale de reconnaissance.

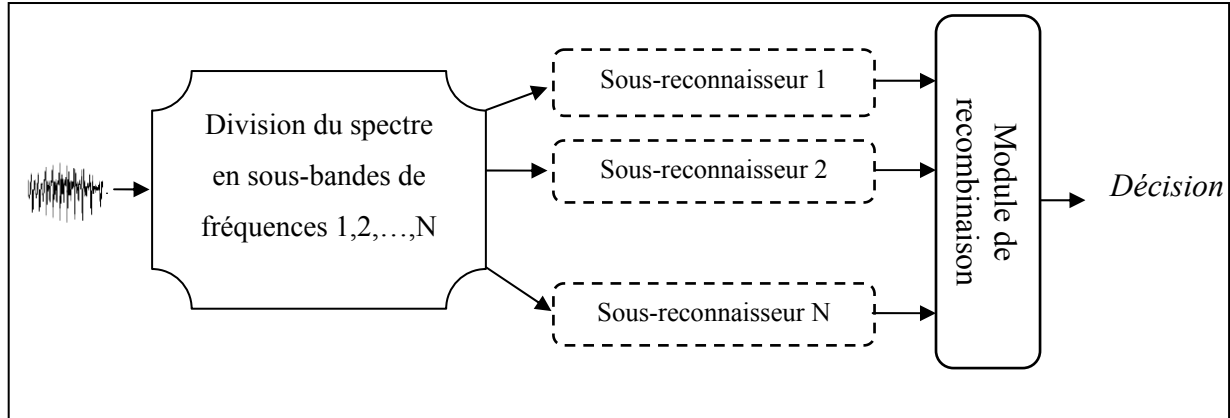


Figure 2.1 : Schéma synoptique d'un système multi-bandes.

Dans notre étude, nous avons choisie un système multi-bandes à base de la modélisation stochastique des HMM. Chaque estimateurs de probabilités (sous-reconnaisseurs) est à base de HMM, la sortie de ces derniers pour une sous-bande b est un vecteur de probabilités dont les éléments sont les vraisemblances $P(Y_b/\lambda_m)$ tel que :

- Y_b : la séquence de vecteurs de paramètres de la même sous-bande b caractérisant une unité de reconnaissance.
- M : le nombre de classes (locuteurs) à reconnaître.
- $Y_1, \dots, Y_b, \dots, Y_M$: les modèles associés respectivement aux M classes.

La sortie désirée du module de recombinaison est alors un vecteur constitué des probabilités $P(Y_1, \dots, Y_b, \dots, Y_M/\lambda_m)$. La décision est basée sur le calcul du modèle qui maximise cette probabilité.

Dans la partie suivante nous allons présenter le principe de base des HMM utilisés dans notre système Multi-bandes.

2.5 Les modèles de Markov cachés

2.5.1 Définitions

Théoriquement, un modèle de Markov caché est un double processus stochastique (Q_t, O_t) $1 \leq t \leq T$. La chaîne interne Q_t non observable, et la chaîne externe O_t observable, s'allient pour générer le processus stochastique. La chaîne interne change d'état en suivant une loi de transition. L'observateur ne peut voir que les sorties des fonctions aléatoires associées aux

états et ne peut pas observer les états de la chaîne Q_t , d'où le terme de Modèles de Markov Cachés [Jacob, 1995].

Le processus $(Q_t) 1 \leq t \leq T$ est une chaîne de Markov d'ordre 1 (la connaissance du passé se résume à celle du dernier état occupé), il doit vérifier:

$$P(Q_{t+1} = q_j / Q_t = q_j, \dots, Q_0 = q_0) = P(Q_{t+1} = q_j / Q_t = q_j) \\ = a_{ij} \quad \text{pour tout } t \geq 1. \quad (2.1)$$

Le processus $(O_t) 1 \leq t \leq T$, est un processus observable qui vérifie :

$$P(O_t = o_t / Q_t = q_i, \dots, Q_1 = q_1, O_{t-1} = o_{t-1}, \dots, O_1 = o_1) = P(O_t = o_t / Q_t = q_i) \\ = b_i(o_t) \quad (2.2)$$

La modélisation de la loi de probabilité $b_i(o_t)$ peut être de nature:

- **Discrète** : $b_i(o_t)$ est une distribution de probabilité discrète : une loi discrète est généralement représentée par les fréquences d'apparitions des observations discrètes (figure 2.1.a).
- **Continue** : $b_i(o_t)$ est une fonction de densité de probabilité définie sur $\mathbb{R}^d$ ($d = |o_t|$): les densités traditionnelles utilisées sont des densités gaussiennes, entièrement définies par le vecteur moyenne et la matrice de covariance, ou des densités de type multi-gaussiennes (sommées pondérées de densités gaussiennes) (figure 2.1.b).

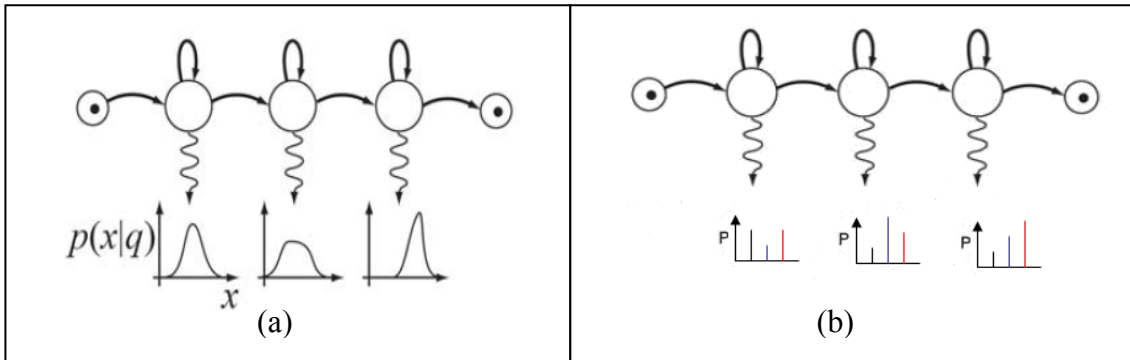


Figure 2.2: Types de modélisation de la loi de probabilité des observations :(a) modélisation de la loi continue (b) modélisation de la loi discrète.

Un Modèle de Markov Caché λ est alors défini par le triplet $\lambda=(\pi, A, B)$ tel que :

- $\pi = \{\pi_i\}$ le vecteur des probabilités initiales :

$$\pi_i = P(q_1 = i), \quad 1 \leq i \leq N \quad (2.3)$$

- $A = \{a_{ij}\}$ la matrice des probabilité de transitions entre les états i et j :

$$a_{ij} = P(q_{t+1} = j | q_t = i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq N, \quad 1 \leq t \leq T$$

$$\sum_{i=1}^N a_{ij} = 1 \quad 1 \leq j \leq N \quad (2.4)$$

- $B = \{b_i(o_t)\}$ la matrice (ou la fonction de densité) de probabilité d'observation selon le type de modélisation d'observations :

$$b_i(o_t) = P(o_t / q_t = i) \quad (2.5)$$

2.5.2 Modélisation du locuteur par HMM

Un modèle HMM va modéliser un locuteur à partir de son signal vocal, de telle façon que chaque segment supposé stationnaire ou pseudo-stationnaire du signal va correspondre à un état dans le modèle HMM. Chaque état HMM est caractérisé par une distribution de probabilité des différents vecteurs acoustiques associés au segment attribué à cet état. La transition d'un segment à un autre segment du signal est modélisée par la transition entre les états, laquelle est supposée être instantanée et caractérisée par la probabilité de transition de l'état (Figure 2.2).

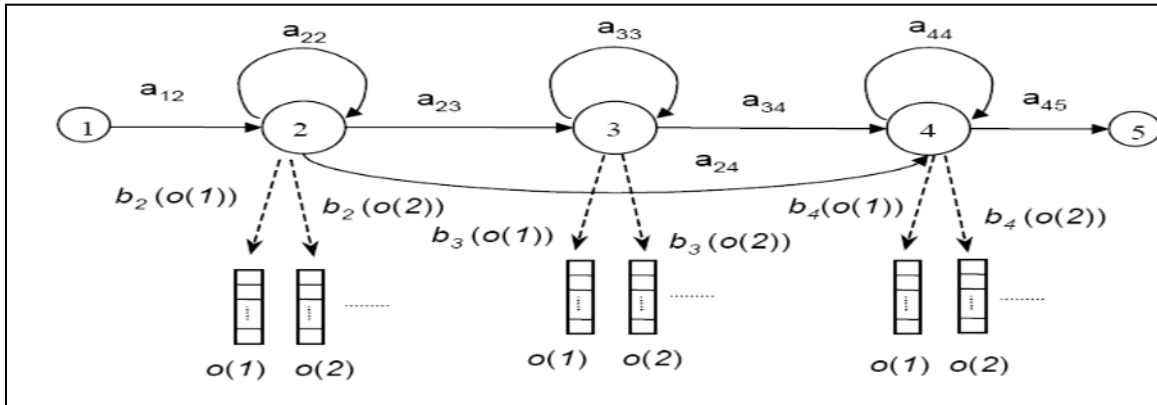


Figure 2.3: Un exemple de HMM à 3 états modélisant un signal contenant des vecteurs acoustiques.

2.5.2.1 Topologie des HMMs utilisés pour la RAL

Le signal de parole est un phénomène dont la dimension temporelle ne peut être ignorée. Les HMM utilisés pour le représenter sont des modèles "gauche-droite" qui ne permettent pas de "retour en arrière" et qui démarrent toujours depuis l'état initial ($i=1$).

C'est-à-dire que leurs probabilités vérifient :

$$i > j \Rightarrow a_{ij} = 0 \quad 2 \leq i \leq N, \quad 1 \leq j \leq N-1 \quad (2.6)$$

$$\pi_i = P(q_1 = i) = \begin{cases} 1 & \text{pour } i = 1 \\ 0 & \text{pour } 1 < i \leq N \end{cases} \quad (2.7)$$

(c-à-d $\pi = \{1 \ 0 \ \dots \ 0\}$)

Dans ce cadre, R. Bakis [Bakis,1976] a proposé un modèle type qui permet le bouclage sur l'état courant (progression acoustique stationnaire) et le passage à l'état suivant (progression acoustique "standard") (figure 2.3). Le nombre d'états du modèle est normalement proportionnel à la durée moyenne du mot prononcé par le locuteur. La pluparts des systèmes de reconnaissance utilisent des modèles à trois états.

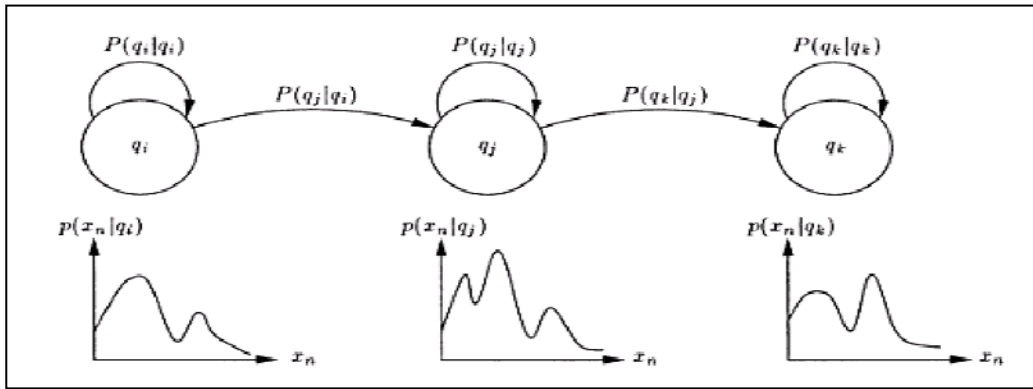


Figure 2.4: Exemple d'un HMM avec une topologie de type Bakis à 3 états.

Ce type de modèle est devenu générique dans le domaine de la reconnaissance de la parole et du locuteur . Il est utilisé dans de nombreux systèmes pour modéliser les unités acoustiques de base.

2.5.2.2 Modélisation des observations acoustiques

Les observations émises lors des transitions représentent la succession des trames acoustiques au cours de la prononciation du mot. Ces observations peuvent être décrites par un nombre fini de symboles au moyen de la quantification vectorielle dans le cas des modèles discrets, ou de modéliser leurs probabilités d'émission par des densités de probabilité continues dans le cas des modèles continus.

a) Observations discrètes

L'espace de représentation du signal de parole est généralement un espace multidimensionnel continu E , et les vecteurs de coefficients calculés sur une trame de signal sont des points de

cet espace. Il est possible de modéliser ces observations par des modèles de Markov à émissions discrètes au moyen de la quantification vectorielle (QV).

La QV permet le passage de l'espace E vers un espace discret, en partitionnant cet espace et en choisissant un représentant pour chaque classe. L'ensemble des représentants constitue un dictionnaire de M prototypes noté $V = \{v_k\}_{1 \leq k \leq M}$. Chaque vecteur $O \in E$ est quantifié par le prototype du dictionnaire dont il est le plus proche au sens d'une distance $d(O, \hat{O})$ définie dans l'espace E :

$$O \xrightarrow{QV} \hat{O} = v_{r'}, \quad \text{avec} \quad r' = \arg \min_{1 \leq r \leq M} d(O, v_r) \quad (2.8)$$

Divers algorithmes ont été proposés pour la réalisation du dictionnaire des prototypes, basés le plus souvent sur une classification hiérarchique descendante ou sur les nuées dynamiques [Gray, 1984]. Cependant, la QV introduit des distorsions, et il est souvent préférable de travailler directement dans l'espace continu E .

b) Observations continues

Le principe de l'émission de symboles discrets peut se généraliser au cas continu. Les probabilités d'émission discrètes $b_j(k)$ sont alors remplacées par des densités de probabilité continues dans l'espace de représentation. Cette solution évite les distorsions introduites par la QV, mais pose le problème du choix des densités de probabilité et de la robustesse de leur estimation. L'utilisation d'une combinaison linéaire de gaussiennes dans l'espace R^d est fréquente :

$$b_j(O) = \sum_{k=1}^M c_{jk} N(O, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^M c_{jk} b_{jk}(O) \quad 1 \leq j \leq N \quad (2.9)$$

Avec M est le nombre de gaussiennes à l'état j . $N(O, \mu_{jk}, \Sigma_{jk})$ où $b_{jk}(O)$ est la distribution gaussienne du $k^{\text{ème}}$ mélange de l'état j . Cette distribution est définie par le vecteur moyen μ_{jk} et la matrice de covariance Σ_{jk} à l'état j . c_{jk} est le coefficient de pondération du $k^{\text{ème}}$ mélange qui satisfait la contrainte stochastique :

$$\sum_{k=1}^M c_{jk} = 1 \quad (2.10)$$

$$c_{jk} \geq 0, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M$$

Nous rappelons que la densité de probabilité d'une loi normale de moyenne μ et de matrice de covariance Σ en dimension d est:

$$N(O, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (O - \mu)' \Sigma^{-1} (O - \mu) \right\} \quad (2.11)$$

L'hypothèse qui est souvent faite d'une indépendance entre les d dimensions de l'espace autorise l'utilisation de matrices de covariance diagonales, cela limite le nombre de paramètres à estimer et simplifie les calculs. D'autres types de densités de probabilité sont possibles, comme le Laplacien, de la densité de probabilité de l'erreur de prédiction par un modèle autorégressif [Juang, 1985]

2.5.3 Sous-reconnaisseur à base de CHMMs

A partir d'une suite d'observations O supposées émises par un modèle, différents problèmes peuvent être posés :

- L'évaluation de la probabilité que la suite des observations ait été émise par un modèle HMM. Lorsque plusieurs modèles existent, cette évaluation permet le choix du modèle le plus probable.
- La recherche de la séquence cachée d'états la plus probable d'un modèle ayant produit les observations.
- L'apprentissage des paramètres d'un modèle. A partir d'un modèle initial et d'observations supposées émises par ce modèle, on cherche les probabilités de transition et d'émission maximisant la vraisemblance des observations.

Nous allons présenter la résolution de ces problèmes dans le cadre de la reconnaissance du locuteur dans laquelle chaque modèle HMM représente un locuteur.

2.5.3.1 Reconnaissance d'une séquence d'observations

Supposons que les observations $O = (o_1 \dots o_T)$ sont les trames acoustiques d'un locuteur inconnu. On cherche à trouver le locuteur qui a prononcé parmi l'ensemble des R locuteurs, ceci revient à désigner parmi les modèles proposés lequel coïncide le mieux avec la séquence d'observations, soit d'après la règle de décision bayésienne (1.34) :

$$M_{best} = \arg \max_M P(O/M)P(M) \quad 1 \leq best \leq R \quad (2.12)$$

Chaque locuteur M est modélisé par un modèle HMM λ_M , d'où l'hypothèse suivante:

$$P(O/M) = P(O/\lambda_M) \quad (2.13)$$

L'équation (2.12) peut alors être reformulée :

$$M_{best} = \arg \max_M P(O/\lambda_M)P(M) \quad 1 \leq best \leq R \quad (2.14)$$

Sa résolution nécessite l'estimation de la probabilité d'émission des observations $P(O/\lambda_M)$ pour chaque modèle HMM modélisant un locuteur.

2.5.3.2 Probabilité d'émission des observations

La probabilité que la suite d'observations $O = (o_1 \dots o_T)$ soit émise par le modèle λ peut être calculée comme la somme des probabilités conjointes de l'observation O et du chemin Q pour l'ensemble de tous les chemins possibles de longueur T :

$$P(O/\lambda) = \sum_Q P(O, Q/\lambda) \quad (2.15)$$

D'après les définitions du modèle, la probabilité de partir à l'instant $t=1$ de l'état d'indice $q_1=1$ et de suivre le chemin $Q = (q_1 \dots q_T)$ est le produit des probabilités de transition sur le chemin:

$$P(Q/\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \quad (2.16)$$

Et la probabilité d'avoir émis les observations O en suivant ce chemin (supposant que les observations sont indépendantes) est:

$$P(O/\lambda) = \prod_{t=1}^T P(o_t / q_t, \lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \dots b_{q_T}(o_T) \quad (2.17)$$

D'où la probabilité conjointe du chemin et des observations :

$$P(O, Q/\lambda) = P(O/Q, \lambda) P(Q/\lambda) \quad (2.18)$$

La probabilité d'émission des observations est finalement :

$$P(O/\lambda) = \sum_Q P(O, Q/\lambda) = \sum_Q P(O/Q, \lambda) P(Q/\lambda) = \sum_Q \prod_{t=1}^T a_{q_{t-1} q_t} b_{q_t}(o_t) \quad (2.19)$$

Avec un modèle HMM à N états, le nombre de chemins possibles est de l'ordre de N^T , et l'ensemble des chemins devient très rapidement impossible à décrire. Il existe heureusement un algorithme rapide, dit "Forward-Pass" (estimation directe), qui permet de calculer récursivement cette quantité [Baum, 1972].

2.5.3.3 Estimation directe

Une variable intermédiaire est introduite pour le calcul de la probabilité d'émission. La variable directe $\alpha_t(i)$ est définie comme la probabilité que les observations jusqu'à l'instant t aient été émises par le modèle λ , et que l'état à cet instant soit l'état d'indice i :

$$\alpha_t(i) = P(o_1, \dots, o_t, q_t = i / \lambda) \quad (2.20)$$

On résout alors $\alpha_t(i)$ par récurrence:

- Initialisation:

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N \quad (2.21)$$

- Récurrence : (figure 2.4)

Pour t allant de 1 à T ,

Pour j allant de 1 à N

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \quad (2.22)$$

Fin

Fin

- Résultat final:

$$P(O / \lambda) = \sum_{i=1}^N \alpha_T(i) \quad (2.23)$$

Puisque par définition : $\alpha_T(i) = P(o_1 o_2 \dots o_T, q_T = i / \lambda)$ et donc $P(O / \lambda)$ ce n'est que la somme des $\alpha_T(i)$.

Le nombre d'opération de calcul effectuer par cette procédure est d'ordre $N^2 T$ opérations, il est claire que c'est mieux de $2TN^T$ opérations effectuer dans le calcul directe.

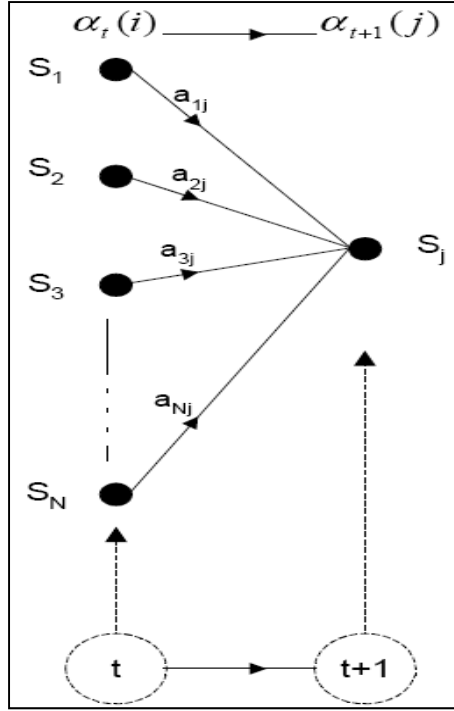


Figure 2.5: Progression de la procédure estimation directe.

2.5.3.4 Estimation Backward

L'estimation directe est suffisante pour obtenir la probabilité d'émission recherchée. Cependant, une estimation rétrograde dans le temps est aussi possible, avec la probabilité $\beta_t(i)$ que les observations après l'instant t soient émises en partant de l'état d'indice i :

$$\beta_t(i) = P(o_{t+1} \dots o_T / q_t = i, \lambda) \quad (2.24)$$

- Initialisation :

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \quad (2.25)$$

- Récurrence : (figure 2.5)

Pour t allant de T à 1 ,

 Pour i allant de 1 à N

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad (2.26)$$

 Fin

Fin

Résultat final :

$$P(O/\lambda) = \beta_1(i) \quad (2.27)$$

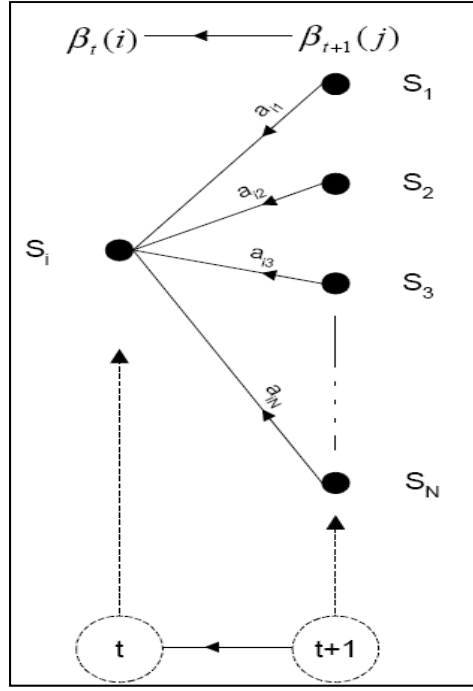


Figure 2.6: Progression de la procédure Estimation rétrograde.

2.5.3.5 Recherche des états cachés

La procédure d'estimation directe ou rétrograde fournit la probabilité d'émission des observations sur toutes les séquences d'états possibles, sans choisir un chemin particulier. Il est parfois utile de connaître la séquence d'états qui a émis les observations. L'algorithme de Viterbi [Viterbi,1967] [Forney,1973] cherche la séquence d'états cachés la plus probable et calcule la probabilité d'émission le long de ce chemin. La procédure peut être exploitée pour l'initialisation des modèles.

On recherche la meilleure séquence d'états, $Q = \{q_1, q_2, \dots, q_T\}$, connaissant la séquence d'observations $O = (o_1 \dots o_T)$. On définit alors $\delta_t(i)$, comme la probabilité maximale que les observations observées jusqu'à l'instant t aient été émises par le modèle λ en suivant un chemin qui arrive à l'état d'indice i :

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(o_1, o_2, \dots, o_t, q_1, q_2, \dots, q_t = i \mid \lambda) \quad (2.28)$$

Alors une récurrence similaire à celle suivie pour le calcul de la probabilité d'émission s'applique, à laquelle s'ajoute la mémorisation du meilleur chemin:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(o_{t+1}) \quad (2.29)$$

Pour conserver la trace de l'indice de l'état qui maximise un $\delta_t(j)$ donné, on emploie le vecteur $\psi_t(j)$. On énonce alors la procédure complète comme suit :

-Initialisation :

$$\begin{cases} \delta_1(i) = \pi_i b_j(o_1) & 1 \leq i \leq N \\ \psi_1(i) = 0 \end{cases} \quad (2.30)$$

Récurrence :

Pour t allant de 2 à T ,

Pour j allant de 1 à N ,

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(o_t) \quad (2.31)$$

on conserve la mémoire du meilleur état précédent par:

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \quad (2.32)$$

Fin

Fin

Résultat final :

$$P(O, Q \setminus \lambda) = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.33)$$

construction du meilleur chemin $\tilde{Q} = (\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_T)$ avec:

$$\tilde{q}_T = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.34)$$

Puis, pour t allant de $T-1$ à 1:

$$\tilde{q}_t = \psi_{t+1}(\tilde{q}_{t+1}) \quad (2.35)$$

2.5.3.6 Apprentissage d'un modèle

La reconnaissance d'un locuteur est rendue possible par l'évaluation de la probabilité d'émission des observations par tous les modèles de locuteurs. Cela suppose l'existence d'un modèle pour chaque locuteur, et l'apprentissage des paramètres de ces modèles. Ces paramètres sont les probabilités de transition entre états et les probabilités d'émission associées aux états. Ainsi, connaissant une suite d'observations O émises par un modèle, il est possible de modifier les paramètres du modèle $\tilde{\lambda}$ de manière à rendre maximale la probabilité

d'émission des observations O par le modèle. Il s'agit d'une estimation par le critère du maximum de vraisemblance (Maximum Likelihood Estimation ou MLE) qui est réalisée par l'algorithme de Baum-Welch [Baum,1972] [Rabiner,1989].

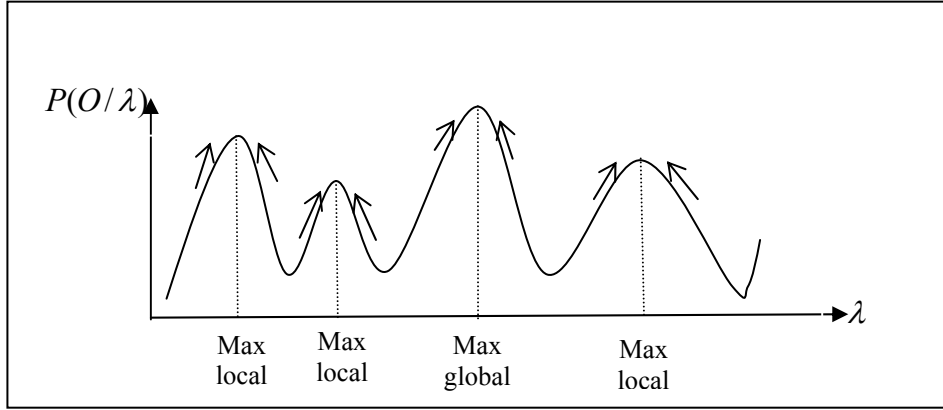


Figure 2.7 : Recherche du maximum de vraisemblance.

Les formules de l'algorithme de Baum-Welch permettent une ré-estimation itérative des paramètres a_{ij} et $b_j(k)$ du modèle $\tilde{\lambda}_n$, le nouveau modèle $\tilde{\lambda}_{n+1}$ vérifie:

$$P(O \setminus \tilde{\lambda}_{n+1}) \geq P(O \setminus \tilde{\lambda}_n) \quad (2.37)$$

La convergence vers un optimum local est démontrée, mais les valeurs initiales des paramètres A et B sont cruciales pour assurer une convergence correcte et rapide le plus près possible du maximum global. L'algorithme de Viterbi réalisant le décodage peut servir à l'initialisation des modèles.

Pour décrire les procédures de réestimation, on va définir deux variables :

On appelle $\gamma_t(i)$ la probabilité d'être dans l'état i au temps t , sachant le modèle λ et la séquence d'observation O :

$$\gamma_t(i) = P(q_t = i \setminus O, \lambda) \quad (2.38)$$

et $\xi_t(i, j)$ la probabilité d'être dans l'état i au temps t et dans l'état j au temps $t+1$, sachant le modèle λ et la séquence d'observation O :

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j \setminus O, \lambda) \quad (2.39)$$

On peut développer ces probabilités en utilisant les variables des algorithmes forward et Backward (figure 2.6) :

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \setminus \lambda)} \quad (2.40)$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(j)}{P(O \setminus \lambda)} = \frac{\sum_{j=1}^N \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O \setminus \lambda)} = \sum_{j=1}^N \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O \setminus \lambda)} = \sum_{j=1}^N \xi_t(i, j) \quad (2.41)$$

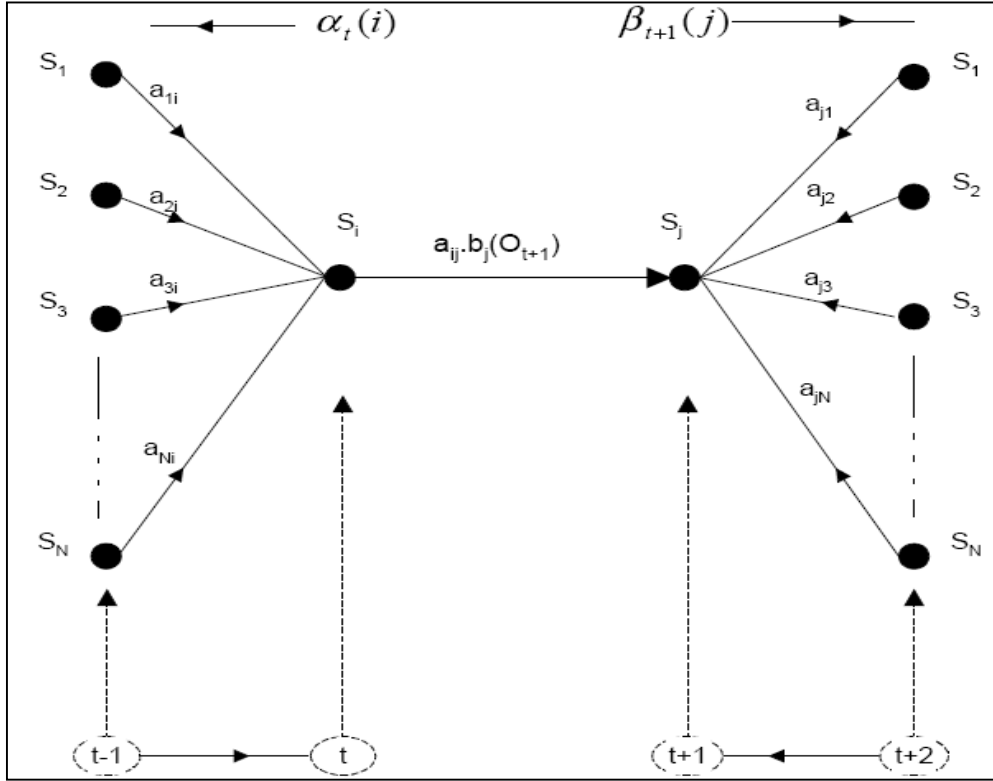


Figure 2.8 : Progression de la procédure de Baum-Welch

La somme de $\gamma_t(i)$ et $\xi_t(i, j)$ au cours de temps t , peut être interprétée comme :

$$\sum_{t=1}^{T-1} \gamma_t(i) = \sum_{t=1}^{T-1} P(q_t = i \setminus O, \lambda) = \text{nombre de transitions depuis l'état } i. \quad (2.42)$$

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \sum_{t=1}^{T-1} P(q_t = i, q_{t+1} = j \setminus O, \lambda) = \text{nombre de transitions depuis l'état } i \text{ vers l'état } j. \quad (2.43)$$

En utilisant ces sommes, les nouveaux paramètres réestimés $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ pour un modèle

$\lambda = (A, B, \pi)$ sont :

$$\bar{\pi}_i = \text{nombre de passage par l'état } i \text{ au temps } (t=1) = \gamma_1(i) \quad (2.44)$$

$$\bar{a}_{ij} = \frac{\text{nombre de transitions de l'état } i \text{ vers l'état } j}{\text{nombre de transitions depuis l'état } i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (2.45)$$

a) Dans le cas d'une modélisation discrète des observations :

$$\bar{b}_j(k) = \frac{\text{nombre d'observations du symbole } v_k \text{ dans l'état } j}{\text{nombre de passages par l'état } j} = \frac{\sum_{t=1}^T \gamma_t(j) \text{ avec } o_t = v_k}{\sum_{t=1}^T \gamma_t(j)} \quad (2.46)$$

b) Dans le cas d'une modélisation continue des observations :

Pour une loi de densité multi gaussiennes définie (la section 2. 3.3.2) par:

$$b_j(O) = \sum_{k=1}^M c_{jk} N(O, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^M c_{jk} b_{jk}(O) \quad 1 \leq j \leq N \quad (2.47)$$

Le coefficient de pondération c_{jk} , le vecteur de moyenne μ_{jk} et la matrice de covariance Σ_{jk} de la densité de probabilité associée au $k^{\text{ème}}$ mélange de l'état j sont recalculés comme suit:

$$\bar{c}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k)}{\sum_{t=1}^T \sum_{k=1}^M \gamma_t(j, k)} \quad (2.48)$$

$$\bar{\mu}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k) \cdot o_t}{\sum_{t=1}^T \gamma_t(j, k)} \quad (2.49)$$

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k) \cdot (o_t - \bar{\mu}_{jk})(o_t - \bar{\mu}_{jk})'}{\sum_{t=1}^T \gamma_t(j, k)} \quad (2.50)$$

Où $\gamma_t(j, k)$ est une généralisation de terme $\gamma_t(j)$, il est défini comme la probabilité d'être à l'état j au temps t pour le $k^{\text{ème}}$ mélange de gaussienne :

$$\gamma_t(j, k) = \left[\frac{\alpha_t(j) \beta_t(j)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)} \right] \left[\frac{c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{k=1}^M c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})} \right] \quad (2.51)$$

Il a été démontré par Baum et ses collègues [Baum, 1968] que $P(O | \bar{\lambda}) \geq P(O | \lambda)$.

2.5.3.7 Problème de mise en œuvre

L'une des particularités des probabilités acoustiques est leur faible valeur numérique. Ce qui implique que la valeur de la probabilité calculée lors de la phase de décodage est très faible au point de sortir des possibilités de codage numérique habituelles des ordinateurs. La méthode employée pour résoudre ce problème est de transposer l'ensemble des calculs dans le domaine logarithmique. Les propriétés du domaine logarithmique permettent de transformer la multiplication de probabilités en addition, ce qui a pour effet une diminution des temps de calcul :

L'équation (2.17) peut alors s'écrire :

$$P(O / \lambda) = \prod_{t=1}^T P(o_t / q_t, \lambda) = \prod_{t=1}^T b_{q_t}(o_t) \Rightarrow \log P(O / \lambda) = \sum_{t=1}^T \log b_{q_t}(o_t) \quad (2.52)$$

Ceci évite également d'avoir à calculer l'exponentielle dans la fonction de densité gaussienne :

$$P(O / \lambda) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp^{-\frac{1}{2}(O - \mu)^T \Sigma^{-1} (O - \mu)} \quad (2.53)$$

$$\Rightarrow \log P(O / \lambda) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(\det(\Sigma)) - \frac{1}{2} (O - \mu)^T \Sigma^{-1} (O - \mu) \quad (2.54)$$

De plus comme le log est une fonction croissante monotoniquement, l'inégalité des vraisemblances (probabilités) reste valable pour le log-vraisemblance :

$$P(O / \lambda_1) > P(O / \lambda_2) \Rightarrow \log P(O / \lambda_1) > \log P(O / \lambda_2) \quad (2.55)$$

2.6 Recombinaison des scores et décision:

Le problème le plus crucial dans la mise en œuvre du modèle multi-bandes est la méthode de fusion à utiliser pour combiner les sorties des reconnaissseurs partiels. La recombinaison de ces sorties multiples est basée sur le critère suivant : les reconnaissseurs doivent produire une information complémentaire afin d'espérer atteindre une meilleure performance que le reconnaissseur unique (modèle classique).

Plusieurs techniques de fusion relatives à la reconnaissance des formes sont proposées dans la littérature. Les techniques les plus utilisées dans le cas qui nous intéresse ici, le modèle HMM multi-bandes, peuvent être classées en deux grandes catégories :

- **Les méthodes linéaires :** Ces méthodes sont basées sur une combinaison des sorties des différents reconnaissseurs par les opérateurs arithmétiques (moyenne simple, moyenne pondérée, Min, Max...).

- **Les méthodes non linéaires :** Les méthodes de fusion linéaires sont basées sur l'hypothèse de l'indépendance des sous-bandes. Néanmoins, cette hypothèse peut ne pas être vraie, une des raisons à cela : les sous-bandes bien qu'elles soient définies pour des régions de fréquences différentes il existe toujours une étendue de recouvrement entre les bandes due au chevauchement naturel des bandes critiques. Il a été montré que dans le cas où les reconnaisseurs d'un système sont dépendants les uns des autres, la règle de fusion optimale est de nature non linéaire [Debyeche, 2007]. Nous citons parmi ces méthodes : le vote majoritaire, la technique de recombinaison sur les rangs, recombinaison neuronale, les intégrales floues etc.

Dans le paragraphe suivant, nous présentons la méthode de fusion que nous avons utilisée pour la recombinaison, elle est basée sur le principe mathématique de la logique floue.

2.6.1 Les intégrales floues

L'approche des intégrales floues a prouvé son efficacité pour la recombinaison des classifieurs dans le domaine de la reconnaissance des formes (Pattern recognition) par plusieurs études [Cho, 1995], [Verikas, 1999] [Kuncheva, 2003]. Nous présenterons seulement l'algorithme de l'approche. La théorie de la logique floue sera détaillée en Annexe A.

2.6.1.1 Définitions

- a. **Profil de décision :** Sachant que les sorties des reconnaisseurs HMM sont des probabilités de vraisemblance étiquetées pour chaque classe (locuteur) nous pouvons définir un profil de décision (*Decision Profile*) sous forme d'une matrice tel que :

$$DP(x) = \begin{bmatrix} d_{1,1}(x) & \dots & d_{1,j}(x) & \dots & d_{1,M}(x) \\ \dots & & & & \\ d_{i,1}(x) & \dots & d_{i,j}(x) & \dots & d_{i,M}(x) \\ \dots & & & & \\ d_{L,1}(x) & \dots & d_{L,j}(x) & \dots & d_{L,M}(x) \end{bmatrix} \quad (2.56)$$

Sortie du reconnaisseur $D_i(x)$

↑
Probabilité donné par $D_1, \dots, D_L$
pour une classe j

L est le nombre de reconnaisseurs CHMM et M est le nombre de classes (Locuteurs) à reconnaître.

Pour chaque entré x du module de recombinaison on calcul M vecteurs de longueur L dont chacun correspond à une classe (locuteur) et contient L valeurs de mesures floues. Chaque colonne de la matrice $DP(x)$ sera normalisée ensuite triée du plus probable au moins probable et fusionnée avec une mesure floue pour donner un nouveau paramètre $\mu_D^j(x)$ définie pour chaque classe (locuteur) appelé support d'une classe j.

La normalisation qu'on a utilisée nous permet d'avoir une somme des probabilités issues de chaque reconnaisseurs égale à 1 :

$$\sum_{j=1}^M d_{i,j}(x) = 1 \quad (2.57)$$

b. Mesure floue : Une mesure floue g est définie sur X par : [Temko, 2008]

$$g : Q(X) \rightarrow [0,1] \text{ V\'erifiant : } g(\emptyset) = 0, g(X) = 1 \quad (2.58)$$

$$A \subseteq B \Rightarrow g(A) \leq g(B)$$

$Q(X)$ indique l'ensemble de tous les sous-ensembles de X

$$g(A \cup B) = g(A) + g(B), A \cap B = \emptyset \quad (2.59)$$

c. Intégrale floue : La notion d'intégrale floue découle de celle de mesure floue. Ce sont des intégrales de fonctions réelles définies par rapport à une mesure floue. On définit l'intégrale floue de Choquet de la fonction $f : X \rightarrow \mathbb{R}^+$ par :

$$C_g(f(x_1), \dots, f(x_n)) = \sum_{i=1}^n (f(x_i) - f(x_{i-1}))g(A_i) \quad \text{Avec } A_i = \{x_i, \dots, x_n\} \quad (2.60)$$

2.6.1.2 Algorithme

Le principe de l'intégration floue est de chercher le degré maximal d'accord entre l'exactitude objective (fournie par la sortie du sous reconnaiseur pour chaque classe) et l'expérience (les L valeurs de mesures floues) ; le problème principal de cette approche est de trouver le vecteur de mesures floues. L'ordre de combinaison des classifieurs privilégie les sources les moins conflictuelles. L'intégrale floue peut donc s'interpréter par une maximisation des décisions pondérées des sous ensembles de classifieurs.

La plupart des travaux qui utilisent les intégrales floue suggère le calcul du paramètre λ définie pour une mesure floue g_λ , ces mesures peuvent être interprétées comme étant l'importance des sous-reconnaisseurs.

Wang [Wang,1998] et Kim [Kim, 1995] ont obtenu de bons résultats. Le premier a utilisé une mesure floue $g^i = \hat{p}_i / 2$ tel que $\hat{p}_i$ est la probabilité d'une classification correcte du reconnaissseurs D_i et le deuxième a utilisé $g^i = \hat{p}_i, i=1 \dots L$.

Pour trouver le support $\mu_D^k(x)$ d'une classe k , on fixe les densités floues $g^1, \dots, g^L$. Par la suite on calcule le paramètre λ (équation 2.61). Les colonnes de la matrice $DP(x)$ seront ensuite trié dans l'ordre décroissant ainsi que leurs densités floues correspondantes. Le calcul de la mesure floue de chaque reconnaissseur est effectué en utilisant le paramètre λ . Enfin, Le support d'une classe (locuteur) est déterminé par la formule de l'intégrale de Choquet (équation 2.60). Ces étapes sont illustrées par l'algorithme suivant :

-Fixer les densités floues $g^1, \dots, g^L$

-Calculer $\lambda > -1$, la seule racine supérieure à -1 de l'équation :

$$\lambda + 1 = \prod_{i=1}^L (1 + \lambda g^i) \quad (2.61)$$

-Pour une entré x trier la k^{eme} colonne de la matrice $DP(x)$ pour obtenir $[d_{i_1,k}(x), \dots, d_{i_L,k}(x)]^T$ tel que $d_{i_1,k}(x)$: Ayant la plus grande probabilité ; et $d_{i_L,k}(x)$: Ayant la plus petite probabilité.

-Trier les densités correspondantes $g^{i_1}, \dots, g^{i_L}$

-Mettre $g(1) = g^{i_1}$

-Pour $t=2$ à L calculer récursivement :

$$g(t) = g^{i_t} + g(t-1) + \lambda g^{i_t} g(t-1) \quad (2.62)$$

-Calculer le degré de support final :

$$\mu_D^k(x) = C_g(d_{i_1,k}(x), \dots, d_{i_L,k}(x)) \quad (2.63)$$

Lors de la combinaison des décisions, le choix de la classe s'effectue en prenant en considération la valeur de l'intégrale la plus grande. Les intégrales floues offrent une gestion souple des fiabilités, combinent les mesures de décisions dans un ordre privilégiant les reconnaissseurs les moins conflictuels. Le processus de fusion n'est cependant pas guidé par un expert, la structure de combinaison (figure 2.) étant imposée a priori .

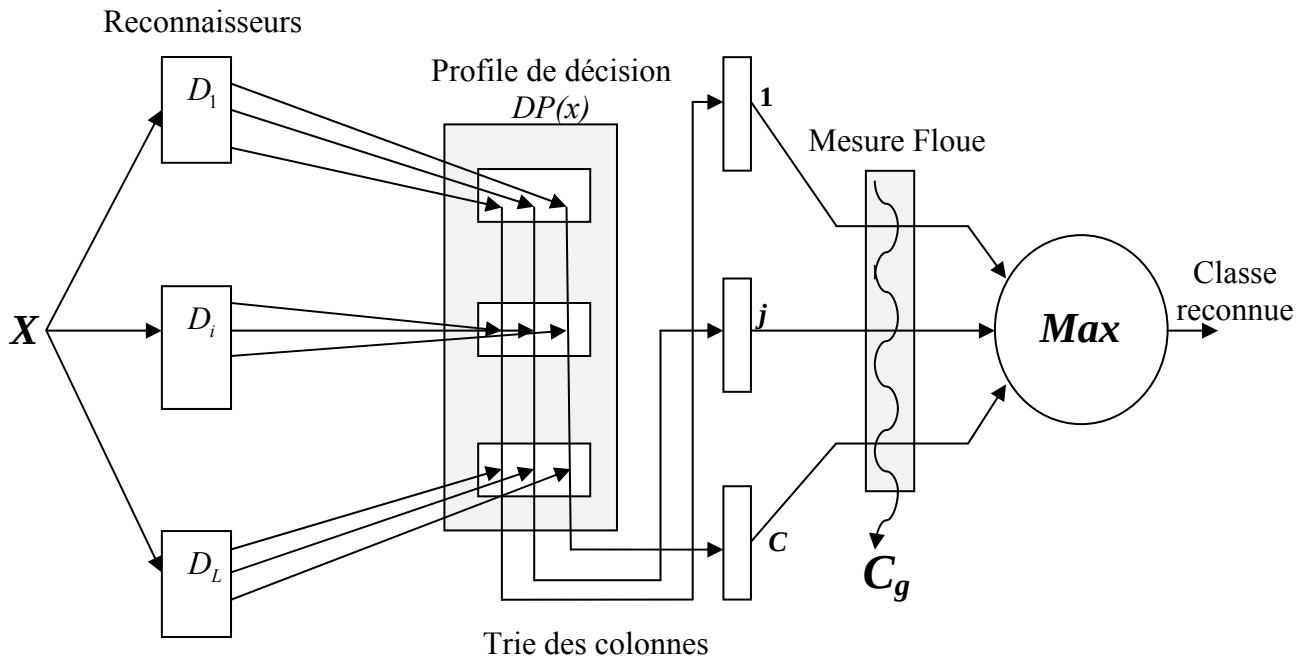


Figure 2.9 : Fonctionnement de la combinaison par intégrale floue.

2.7 Conclusion

Dans ce chapitre nous avons présenté le principe d'un système à base de CHMM multi-bandes en expliquant les motivations et avantages d'une telle architecture ainsi que le choix de la stratégie des intégrales floues pour combiner les scores issus de chaque sous-bande fréquentielle. Nous avons étudié aussi les différents algorithmes utilisés pour pallier aux problèmes d'initialisation, apprentissage et décodage pour les sous reconnaisseurs à base de CHMM. Dans le chapitre suivant, nous introduirons la notion de la norme de communication GSM et plus particulièrement le codeur que nous avons utilisé dans notre système de reconnaissance : Le codeur GSM Full Rate.

Le protocole de communication GSM FR

3.1 Introduction

Les réseaux de transmission de données fixes et mobiles sont actuellement en pleine expansion, ils offrent des débits de plus en plus élevés et un service de plus en plus fiable. Le GSM Global System for Mobile communication est la première norme de téléphonie cellulaire de seconde génération, tout numérique il est idéal pour les communications de type voix, le réseau étant commuté, les ressources ne sont allouées que pour la durée de la conversation comme lors de l'utilisation téléphoniques fixes. Sa normalisation a débuté en 1982, la première norme est sortie en 1991.

Depuis son utilisation plusieurs protocoles de codage de la parole ont été créés pour le GSM, comme le GSMFR (Full Rate) ,GSMHR(Half Rate), GSMEFR (Enhanced Full Rate) et le dernier est le GSMAMR (Adaptive Multi Rate). Dans ce chapitre nous présentons l'architecture du réseau de communication GSM et son premier protocole de codage de la parole GSMFR.

3.2 Architecture du réseau GSM

3.2.1 Le principe cellulaire

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

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

Deux cellules adjacentes ne peuvent utiliser la même bande de fréquence afin d'éviter les interférences la distance entre des cellules ayant la même bande doit être de 2 à 3 fois le diamètre d'une cellule la taille des cellules peut atteindre jusqu'à 35km et dépend de la densité d'utilisateur et de la topographie. La taille limitée des cellules permet de limiter la puissance d'émission nécessaire pour la liaison donc augmenter l'autonomie des mobiles [Krobba,2008].

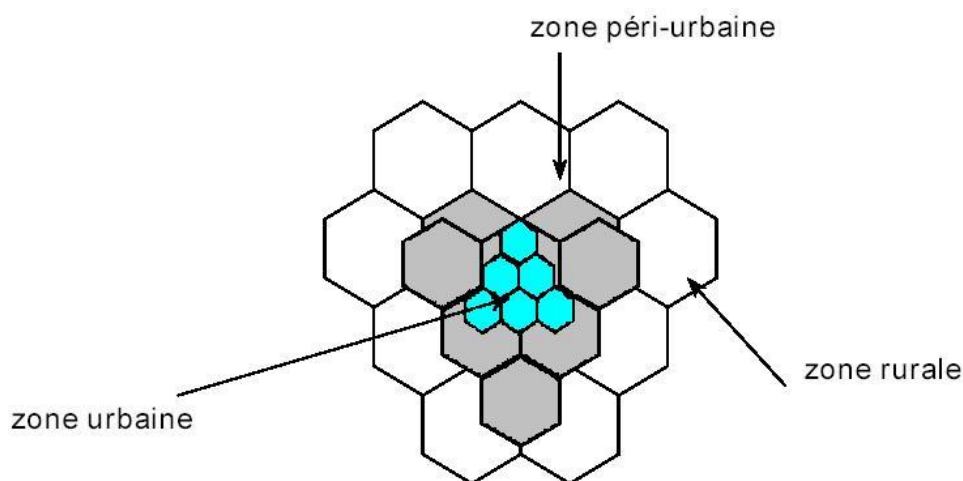


Figure 3.1 : Le découpage cellulaire

3.2.2 La communication dans le GSM

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

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

La norme GSM prévoit que la téléphonie mobile par GSM occupe deux bandes de fréquences aux alentours des 900 MHz :

1. la bande de fréquence 890-915 MHz pour les communications montantes (du mobile vers la station de base) et
2. la bande de fréquence 935-960 MHz pour les communications descendantes (de la station de base vers le mobile).

Comme chaque canal fréquentiel utilisé pour une communication a une largeur de bande de 200 kHz, cela laisse la place pour 124 canaux fréquentiels à répartir entre les différents opérateurs. Mais, le nombre d'utilisateurs augmentant, il s'est avéré nécessaire d'attribuer une bande supplémentaire aux alentours des 1800 MHz. On a donc porté la technologie GSM 900

MHz vers une bande ouverte à plus haute fréquence. C'est le système DCS-1800 (*Digital Communication System*) dont les caractéristiques sont quasi identiques au GSM en termes de protocoles et de service. Les communications montantes se faisant alors entre 1710 et 1785 MHz et les communications descendantes entre 1805 et 1880 MHz.

3.2.3 Présentation de l'infrastructure d'un réseau

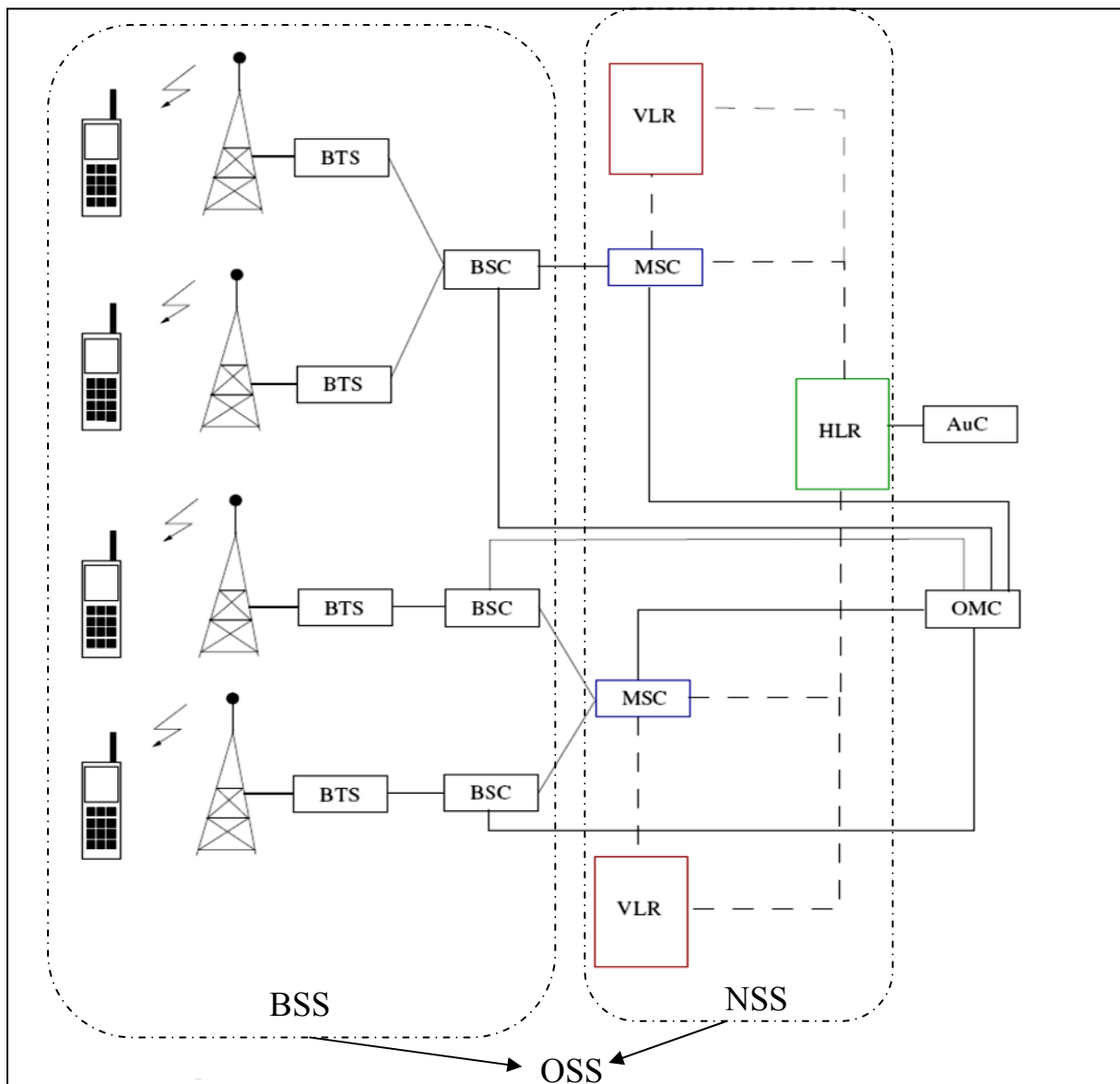


Figure 3.2 :Architecture du réseau GSM.

Le réseau GSM est composé de trois sous ensembles (figure 3.2):

- Le sous système radios : BSS Base Station Sub system qui assure et gère les transmissions radios.

- Le sous système d'acheminement : NSS Network Sub System (ou SMSS Switching and Management Sub-system). Il comprend l'ensemble des fonctions nécessaires pour appels et gestion de la mobilité.
- Le sous système d'exploitation et de maintenance OSS Operation Sub System qui permet à l'opérateur d'exploiter son réseau.

3.2.4 Les équipements d'un réseau GSM

a. Le téléphone GSM : ou équipement mobile (Mobile Equipment) est caractérisée par deux identités :

- ▶ Le numéro d'équipement, IMEI (International Mobile Equipment Identity) mis dans la mémoire du mobile lors de sa fabrication.
- ▶ Le numéro d'abonné IMSI (International Mobile Subscriber Identity) se trouvant dans la carte SIM (Subscriber Identity Module) de l'abonné.

b. La station de base (BTS) :

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

c. Le contrôleur de station de base (BSC) :

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

d. Le centre de commutation mobile(MSC) :

Le **MSC** est relié au sous-système radio via des interfaces, son rôle principal est d'assurer la commutation entre les abonnés du réseau mobile et ceux des autres réseaux. De plus, il participe à la fourniture des différents services aux abonnés tels que les services supplémentaires et les services de messagerie.

e. Le registre des abonnés nominaux (HLR) :

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

- ✓ Toutes les informations relatives aux abonnés: le type d'abonnement, le numéro de l'abonné (**IMSI**), etc.

- ✓ Ainsi qu'un certain nombre de données dynamiques telles que la position de l'abonné dans le réseau en fait, et l'état de son terminal (allumé, éteint, en communication, libre,...).

f. Enregistreur de localisation des visiteurs (VLR) :

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

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

3.2.5 Canaux de transmission

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

- **TCH** : Transport Channel, le canal de transport de communication. Deux sont nécessaire par communication pour assurer le full-duplex. Le TCH permet de transmettre la parole à 13 kbps en Full Rate, 12 kbps en EFR et 5.6 kbps en HR et des données jusqu'à 12 kbps mais en grande partie à 9kbps.
- **BCCH** : Braodcast Control Channel, canal de diffusion des informations sur la cellule. Un par cellule.
- **FCH** : Frequency Channel, même voix que le BCCH, il permet au BTS de diffuser la fréquence du signal auquel il fonctionne.
- **SCH** : Signal Channel, même voix que le BCCH, signalisation.
- **SDCCH**: Stand-alone Dedicated Control Channel. Canal de control dédié à une communication. A chaque Canal de Transport est associe 1/8 de canal pour le SDCCH.

3.2.6 Modulation

En raison de la forte variabilité de l'amplitude des signaux dans un environnement mobile, on préfère recourir à une technique de modulation angulaire pour ce type d'environnement. La technique de modulation utilisée pour porter le signal à haute fréquence est la modulation

GMSK (*Gaussian Minimum Shift Keying*). Comme le suggère son nom, il s'agit d'une variante d'une modulation MSK appartenant à la famille des modulations de fréquence (FM) numériques. On utilise la GMSK car, en raison de la transition rapide entre 2 fréquences ($f_c - \Delta f$ et $f_c + \Delta f$), la modulation par MSK aurait nécessité une trop large bande de fréquences. La modulation GMSK consiste en une modulation de fréquence à deux états portant non pas sur la séquence originale mais sur une nouvelle séquence dont le bit n est produit comme le résultat de la fonction du OU exclusif (XOR) entre le bit courant et le bit précédent. Après application du XOR, le signal est filtré. La figure 3.3 montre la création d'un signal modulé par GMSK.

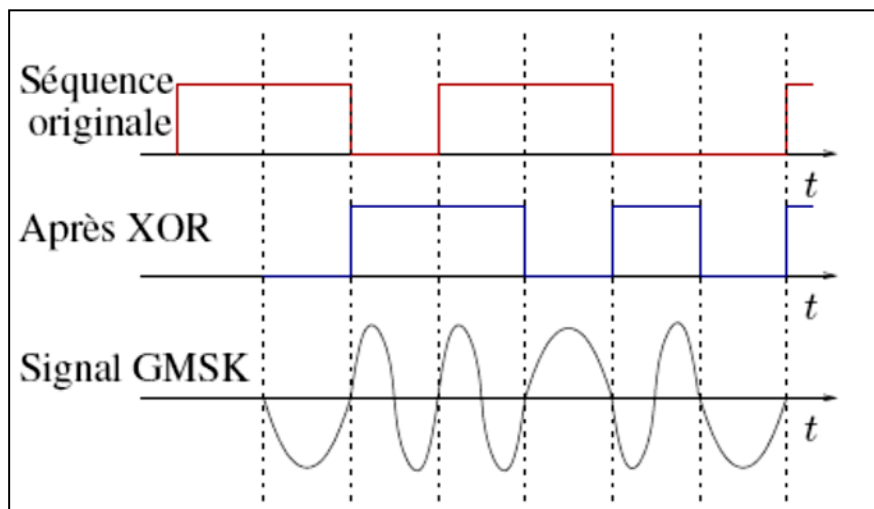


Figure 3.3 : Création d'un signal modulé par GMSK au départ d'un train binaire.

Au bout du compte, il faut une largeur de 200 [kHz] par fréquence porteuse. Sachant que le débit atteint 270 kb/s, on atteint un rapport du débit à largeur de bande, appelé efficacité spectrale, proche de 1. Cette valeur est typique pour des environnements mobiles, ce qui signifie que, pour doubler le débit, il n'y a d'autre solution que de doubler la largeur de bande.

3.3 Codage de la parole

3.3.1 Principe

Le codage de parole permet la réduction de débit de transmission du signal et des communications dans des canaux à largeur de bande limitée. La largeur de bande d'une transmission devra être minimisée tout en préservant la qualité du signal vocal reconstruit [OShaughnessey,1987] et en répondant aux autres exigences liées au réseau GSM.

Tout codage de parole réalise dans un premier temps un prélèvement d'échantillons du signal vocal analogique à un taux d'échantillonnage et une résolution suffisamment grande pour

préserver la qualité désirée. Généralement, un taux de prélèvement de 8 kHz pour échantillonner la parole est employé. Le signal est alors analysé en utilisant des techniques de traitement numérique du signal pour en extraire une représentation compacte sous forme d'un nombre restreint de bits qui sont envoyés à travers le canal de transmission. Les transformations majeures ajoutées pour la transmission sont :

1. Codes correcteurs d'erreur qui permettent à la réception de réparer les erreurs de transmission
2. Algorithmes de cryptage qui assurent une certaine confidentialité des communications.

Au niveau du décodeur, les informations reçues sont employées pour reconstruire le signal de parole, qui est converti de nouveau en un signal analogique en utilisant des filtres de reconstitution et un convertisseur numérique/analogique [Madre,2004].

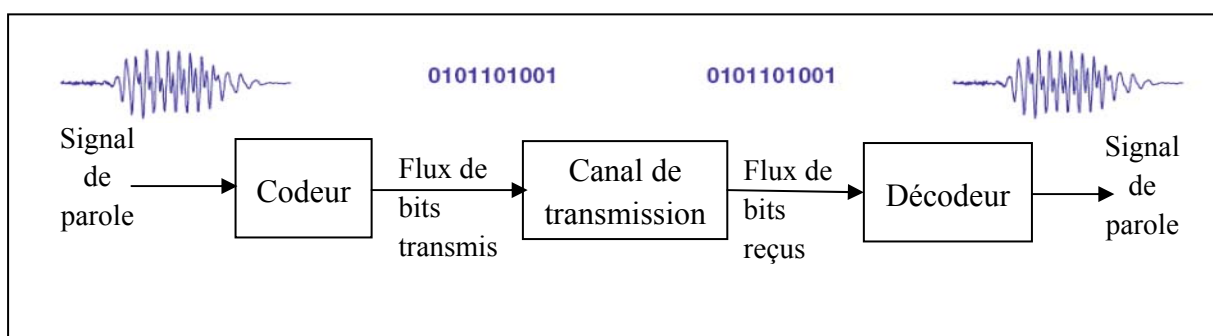


Figure 3.4 : Schéma fonctionnel simplifié d'un système de transmission vocale.

3.3.2 Codeurs GSM

De nombreux algorithmes de codage ont été normalisés par ETSI (European Telecommunications Standards Institute) depuis quelques années pour les systèmes de communications (FR, HR, EFR, AMR). Ces différentes versions de codeurs prennent en entrée des fichiers audio numériques au format PCM (Pulse Code Modulation), échantillonnés à 8 kHz et représentés sur 13 bits. Chaque trame en entrée, d'une durée de 20 ms, est traitée séparément. Ce type de signaux sont couramment appelés signaux de parole dans la bande téléphonique et sont utilisés par la deuxième génération de téléphonie mobile (2G, en Europe GSM phase 2) [Besacier,1999].

- **Codeur de parole Plein Débit (FR):** Le codeur plein débit a été standardisé en 1988. Il appartient à la classe des codeurs de prédiction linéaire RPE-LTP (Regular Pulse

Excitation with Long Term Prediction). A l'entrée du codeur, une trame de 160 échantillons est encodée en bloc de 260 bits à un débit de 13 kb/s. Le décodeur ordonne les séquences de 260 bits en blocs de 160 échantillons. Le canal GSM plein débit peut supporter un trafic de 28,8 kb/s.

- **Codeur de parole demi-débit (HR):** Le codeur demi-débit a été établi afin de contrer l'augmentation du nombre d'utilisateurs. C'est un codeur VSELP (Vector Sum Excited Linear Prediction) standardisé en 1995 par Motorola et ayant un débit de 5,6 kb/s. Le canal supporte un trafic de 11,4 kb/s conformément au fait de doubler la capacité du système. Ainsi, 5,8 kb/s sont utilisés en gestion d'erreurs. Le niveau de qualité du signal de parole en sortie est comparable au codeur plein débit et offre aussi des performances moyennes dans les communications mobiles à mobile.
- **Codeur de parole Enhanced Full Rate (EFR):** Le codeur plein débit amélioré fut standardisé en 1996 par Nokia dans l'intention de proposer une alternative au codeur plein débit (FR). Il apporte une certaine amélioration de la qualité par rapport aux codeurs FR et HR, notamment dans les communications mobile à mobile. Le codeur EFR utilise 12,2 kb/s pour le codage de la parole et 10,6 kb/s pour la gestion des erreurs. Le codage de la parole suit un modèle ACELP (Algebraic Code Excited Linear Prediction).
- **Codeur Adaptive Multi rate (AMR):** Ce codeur a été normalisé en 1998. Son but est d'améliorer la capacité et la qualité du service. Différents débits sont utilisés (12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 et 4.75 kb/s) pour différentes techniques, y compris le ACELP de compression, transmission discontinue (DTX), détection d'activité vocale (VAD) et la génération de bruit de confort (CNG). Son principe réside dans l'adaptation optimisée du débit au canal de transmission radio HR ou FR.

Dans notre travail, nous avons utilisé la première version des codeurs qui est le GSM 6.10 connue sous le nom de GSM FR. Le succès de ce codeur est une combinaison de quelques avantages, comme le temps de calcul faible, une synthèse de parole de haute qualité, robustesse contre les erreurs de canal, et l'efficacité de codage. Nous présentons dans la partie suivante le codeur RPE-LTP qui est le principe de base du GSM FR avec ces différents étages [Degener,1992].

3.3.3 Codec RPE-LTP

L'algorithme RPE-LTP (regular pulse excitation- long term prediction) utilise deux filtres dont l'un fonctionne à court terme (short term prediction), reconstitue le rôle des cordes

vocales et autres cavités résonnantes humaines. L'autre filtre excite le précédent et reconstitue un mélange d'ondes et de bruit par prédiction à long terme. Ce type est appelé codeur par analyse-synthèse .

3.3.3.1 Analyse-Synthèse court terme

Dans cette étape, la parole est d'abord analysée et ses paramètres d'analyse quantifiés. Elle est ensuite filtrée par un prédicteur dont les coefficients correspondent aux paramètres d'analyse quantifiés. Le résidu décorrélé en sortie du prédicteur est enfin quantifié. Ainsi dans l'étape d'analyse et synthèse, la quantification est locale à chaque élément. L'analyse peut être en *boucle ouverte* ou en *boucle fermée*. [Chu, 2003]

L'analyse effectuée est une analyse par prédiction linéaire (voir chapitre 1 §1.3.1.3) son fonctionnement est illustré dans la figure suivante :

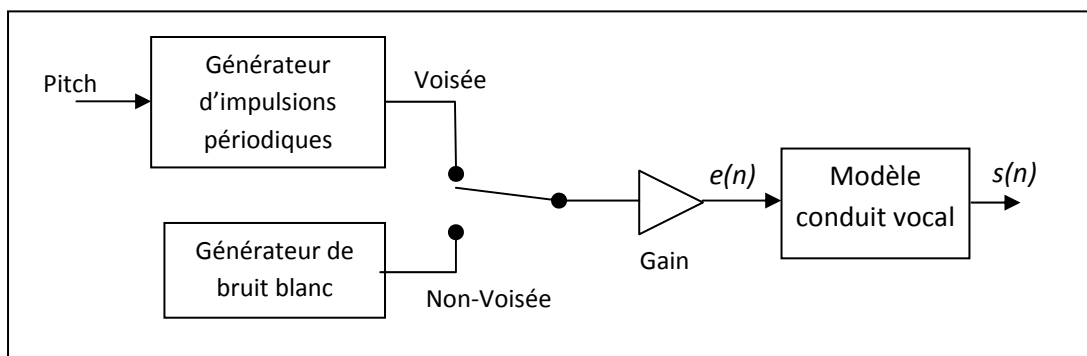


Figure 3.5: Modèle source-filtre de production de la parole.

- **Analyse en boucle ouverte**

La structure en boucle ouverte, illustrée par la Figure 3.4, est la plus immédiate. On notera $s(n/n-1, \dots, n-p)$, la prédiction du signal $s(n)$ à l'instant n d'après les P échantillons précédents :

$$s(n/n-1, \dots, n-p) = \sum_{k=1}^p a_k s(n-k) \quad (3.1)$$

Cependant, cette configuration « boucle ouverte » peut entraîner une propagation d'erreur au décodage. Considérons l'erreur de quantification du résidu $\varepsilon(n) = e(n) - \hat{e}(n)$, on a pour la structure en boucle ouverte ;

$$\hat{s}(n) - s(n) = \varepsilon(n) + \sum_{k=1}^p a_k (s(n-k) - \hat{s}(n-k)) \quad (3.2)$$

Autrement dit, une erreur de quantification constante au codeur peut faire diverger le

décodeur. Cette propagation d'erreur sera d'autant plus forte que le gain du prédicteur linéaire est élevé.

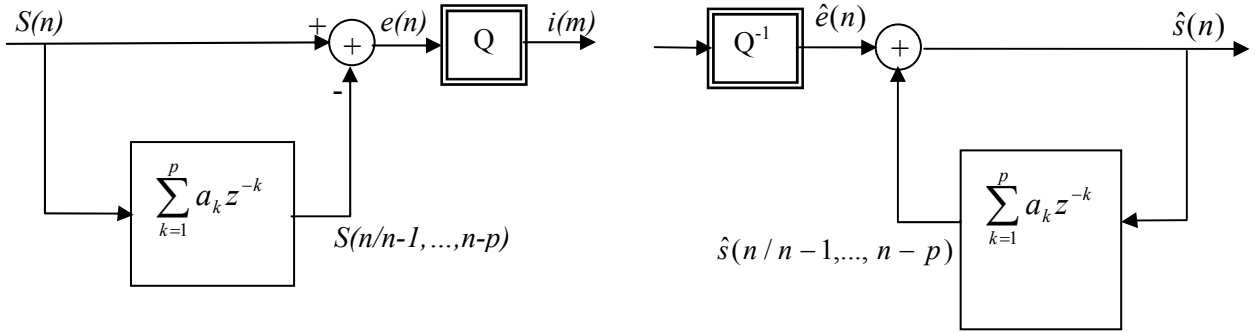


Figure 3.6 : Prédiction en boucle ouverte.

- **Analyse en boucle fermée**

Cette propagation d'erreur peut être évitée en considérant une structure en « boucle fermée » comme illustrée Figure 3.5. Dans une structure en boucle fermée, un *décodeur local* est intégré au codeur. On peut ainsi prédire l'échantillon courant du signal de parole $s(n)$ à partir des échantillons reconstruits précédents $\hat{s}(n-1), \dots, \hat{s}(n-p)$. On a alors ;

$$\hat{s}(n) - s(n) = \varepsilon(n) = \hat{e}(n) - e(n) \quad (3.3)$$

Cette relation montre qu'une erreur de quantification dans le domaine du résidu entraîne une erreur identique dans le domaine de la parole. Elle démontre aussi l'intérêt de la quantification sur le résidu de la prédiction linéaire plutôt que directement sur le signal de parole, puisque l'erreur de quantification est proportionnelle à la puissance du signal à quantifier et que la puissance du résidu est inférieure à celle du signal de parole.

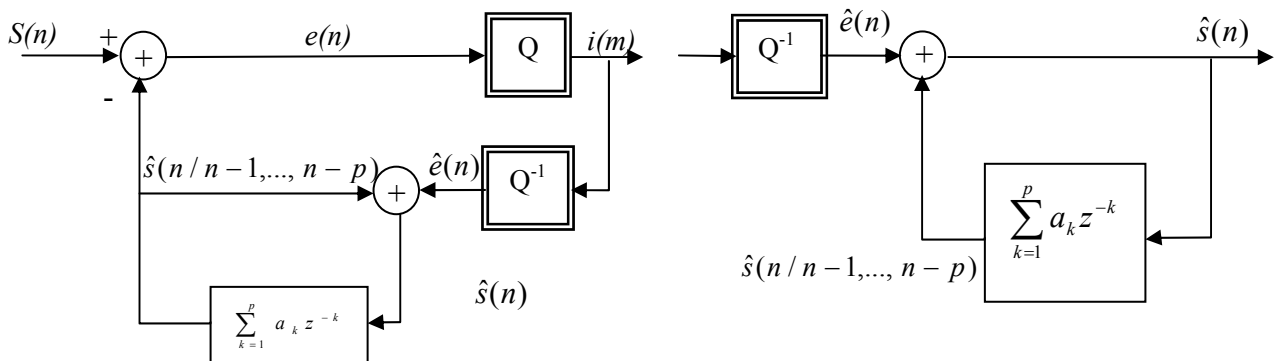


Figure 3.7 : Prédiction en boucle fermée.

3.3.3.2 Détermination de l'excitation

a. Modélisation de la périodicité à long-terme de l'excitation

Lors des sons voisés, le signal d'excitation présente une forte périodicité à long-terme. Le codeur RPE-LTP utilisent un *filtre prédicteur* afin d'extraire cette redondance à long-terme avant quantification.

Dans cette approche, la modélisation de la corrélation à long-terme du signal d'excitation $e(n)$ est similaire à celle employée pour la corrélation à court-terme de la parole $s(n)$. Commençons par décrire l'analyse en boucle ouverte. Selon le schéma de la Figure 3.6 où le signal d'entrée à analyser est désormais le résidu de prédiction à court-terme $e(n)$ et le prédicteur utilisé, un prédicteur à long-terme de la forme :

$$p(z) = 1 - \beta z^{-D} \quad (3.4)$$

Où β est le coefficient de prédiction et D le délai du prédicteur. Le délai D correspond au pitch estimé et β indique le niveau de voisement associé. Ils s'obtiennent en minimisant l'erreur de prédiction :

$$E_{LTP} = \sum_{-\infty}^{\infty} [e_{\omega}(n)\beta - e_{\omega}(n-D)]^2 \quad (3.5)$$

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

$$\beta(D) = \frac{\sum_{-\infty}^{\infty} e_{\omega}(n) - e_{\omega}(n-D)}{\sum_{-\infty}^{\infty} e_{\omega}^2(n-D)} \quad (3.6)$$

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

$$E_{LTP} = \sum_{-\infty}^{\infty} e_{\omega}^2(n) - \frac{\left[\sum_{-\infty}^{\infty} e_{\omega}(n) - e_{\omega}(n-D) \right]^2}{\sum_{-\infty}^{\infty} e_{\omega}^2(n-D)} \quad (3.7)$$

On considère souvent comme constant le numérateur de cette expression. L'estimation du délai D se limite alors à la *recherche du maximum de l'auto-corrélation* de e . En pratique, le pitch de la parole n'est pas forcément un multiple de la période d'échantillonnage. On utilise alors un délai D avec une *résolution fractionnaire* [Kleijn, 1995], en le décomposant en une composante entière et une composante fractionnelle.

Le gain du prédicteur à long-terme peut être très élevé, ce qui rend instable le schéma en boucle ouverte. Aussi, le prédicteur à long terme $p(z)$ est toujours implanté selon le schéma en

boucle fermée illustré par le figure 3.8. Le critère d'erreur doit donc être modifié pour intégrer le signal d'excitation reconstruit $\hat{e}(n)$:

$$E_{LTP} = \sum_{-\infty}^{\infty} [e_{\omega}(n) - \beta \hat{e}_{\omega}(n - D)]^2 \quad (3.8)$$

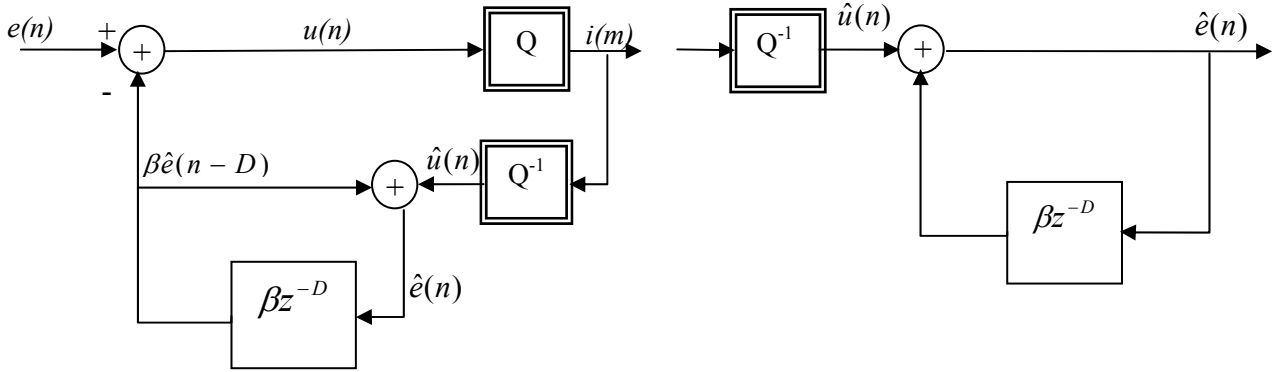


Figure 3.8 : Prédiction long terme en boucle fermée.

Dès lors un problème se pose si le délai D est inférieur à la longueur L de la fenêtre d'analyse $\omega(n)$. En effet, compte-tenu de la mémoire du filtre, l'équation (3.8) devient non linéaire en β . Ceci *restreint la recherche du délai* aux valeurs supérieures à la longueur de la trame d'analyse. Pour contourner cette contrainte, les paramètres du prédicteur à long-terme $P(z)$ sont généralement estimés sur des sous-trames de 5ms, c'est-à-dire 40 échantillons à 8kHz. Une autre motivation du calcul par sous-trames vient du critère d'erreur quadratique moyenne utilisé (3.5), ce critère est mal adapté au pitch puisqu'une faible variation du délai D peut se traduire par une déviance très forte de l'erreur. [Veaux, 2005]

b. Modélisation de l'excitation résiduelle

La quantification de l'excitation résiduelle est très importante pour restituer le naturel de la parole. Elle nécessite cependant l'allocation d'un grand nombre de bits puisqu'elle ne contient pas de redondance modélisable. Les codeurs par analyse-synthèse utilisent simplement des modèles d'excitation (qu'on peut considérer comme des dictionnaires implicites). On choisit la première impulsion mais le nombre d'impulsions et leur espacement est entièrement déterminé.

En résumé

Le codeur RPE-LTP [GSM, 06.10] analyse des trames de parole de 20 ms (160 échantillons à 8 kHz). Les coefficients du prédicteur à court-terme LPC sont calculés sur l'ensemble de la trame de 20 ms puis transformés en coefficients LAR avant d'être quantifiés. Les paramètres du prédicteur à long-terme LTP et de son signal d'excitation sont réactualisés sur des sous-trames de 5 ms (40 échantillons). Le modèle utilisé pour chaque sous-trame de 40 échantillons du signal d'excitation est un modèle RPE (Regular Pulse-Excited) formé d'impulsions régulièrement espacées tous les 3 échantillons. Les paramètres de ce modèle sont la position de la première impulsion dans la sous-trame (offset ou grille RPE) qui est comprise entre 0 et 3 échantillons, ainsi que l'amplitude de chaque impulsion.

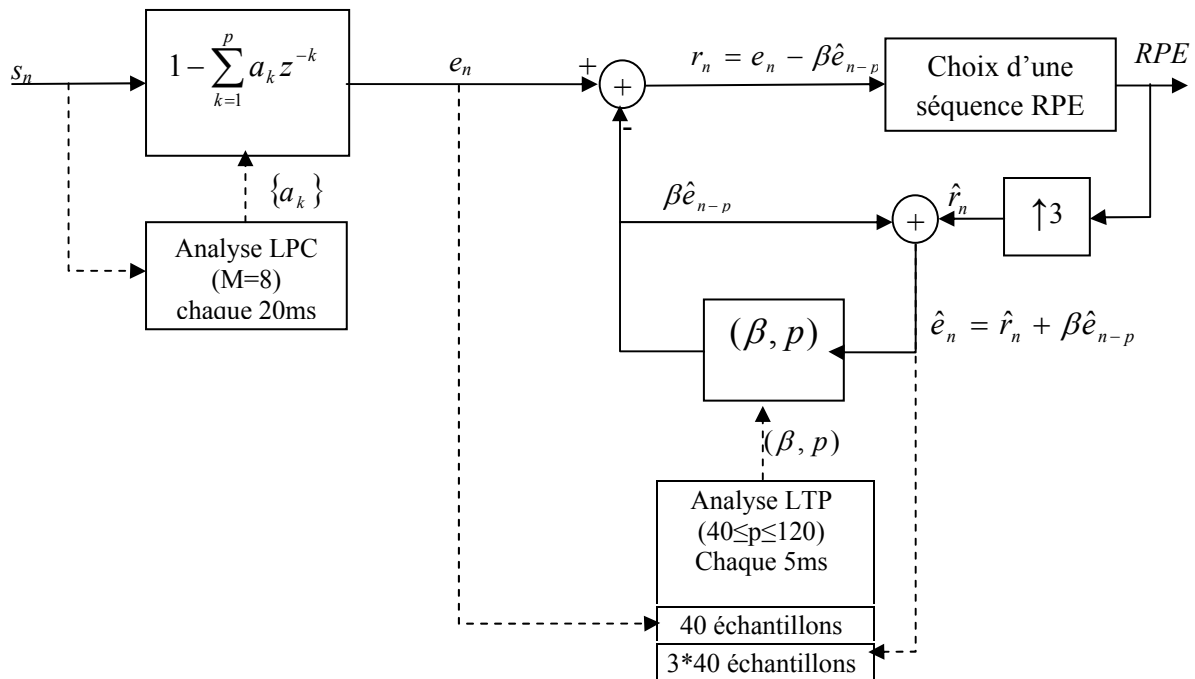


Figure 3.9 : Codeur RPE-LTP (analyse en boucle fermée)

L'ensemble des paramètres RPE-LTP définissant 1 jeu de coefficients du filtre LPC, les 4 jeux successifs de coefficients du filtre LTP et les 4 séquences de 5 ms d'excitation RPE associées sont quantifiés sur un total de 260 bits et transmis toutes les 20 ms au décodeur. Ceci correspond à un débit de 13kbit/s. Pour tous les coefficients, la quantification est *scalaire*.

Tableau 3.1 :Distribution des paramètres codés pour chaque trame de 20 ms

Paramètres	Par sous-trame (5ms)	Total par trame (20ms)
8 coefficients LAR		36
Gain β (filtre LTP)	2	8
Délai p (filtre LTP)	7	28
Gain de calibration	6	24
Grille RPE	2	8
13 amplitudes	39	156
Total		260

Nous avons présenté dans cette partie le principe du codec RPE-LTP, pour voir les étapes détaillées du codeur GSM FR implémenté par ETSI veuillez consulter l'annexe B.

3.4 Conclusion

La parole transmise par un réseau radiomobile subit de nombreuses dégradations à des niveaux successifs. La première d'entre-elles est la distorsion introduite par le codeur de parole qui résulte d'un compromis posé dès la conception du système entre la qualité vocale escomptée et les ressources du réseau allouées à sa transmission. Dans notre système de reconnaissance du locuteur, nous étudions l'influence de cette distorsion sur le taux de reconnaissance par le biais du codeur GSM FR. Ceci fera l'objet du chapitre suivant où nous allons décrire les différentes étapes du système de RAL implémenté.

Evaluation expérimentale

4.1 Introduction

Ce chapitre présente le contexte expérimental et l'évaluation de notre travail. Notre objectif était de développer un système de reconnaissance fondé sur les modèles de Markov cachés Multi-bandes, et ce afin d'évaluer sa performance sur de la parole transcodée par le codeur GSM Full Rate. Le codeur est compilé sous l'environnement Linux et les sous reconnaissances du système de RAL sont mis au point à partir de la plate-forme HTK (Hidden Markov Model Toolkit, ou "boîte à outils des modèles de Markov cachés") de l'Université de Cambridge. L'évaluation du système de reconnaissance est faite sur la base de données de mots isolés ARADIGIT [Amroche,2007] dans un premier temps ensuite sur une base ARADIGIT transcodée par le codeur GSM FR implémenté.

4.2 Description des bases de données utilisées

Cette base a été conçue au laboratoire LCPTS de la faculté d'Electronique et d'Informatique de l'USTHB [Amrouche,2007]. Elle est constituée de prononciations isolées des 10 chiffres de la langue arabe de 0 jusqu'à 9. La base de données est constituée de 1800 fichiers, prononcés par 60 locuteurs des deux sexes qui répètent le même chiffre 3 fois.

La base de données a été enregistrée par des locuteurs algériens âgés entre 18 et 50 ans dans un environnement calme avec un niveau de bruit ambiant inférieur à 35 dB, sous le format WAV, avec une fréquence d'échantillonnage de 16 KHz.

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

- La base de données ARADIGIT 16K sous échantillonnée à 8KHz donnant ainsi la base de données ARADIGIT 8K.
- La base de données ARADIGIT8K est passée ensuite à travers le codec GSM FR que nous avons simulé pour aboutir à la base de données transcodée GSM (source uniquement) à savoir la base nommée ARADIGIT_GSM. La figure 4.1 donne le processus de création de cette base.



Figure 4.1 : Création de la base de données ARADIGIT_GSM

Pour les expériences d'évaluation chacune des bases de données citées précédemment est divisée en deux bases :

- ✓ La base d'apprentissage : elle est composée de 1440 fichiers vocaux prononcés par 60 locuteurs comprenant les deux sexes (30 hommes et 30 femmes). Chaque locuteur répète les chiffres (0, 1, 2, 3, 4, 5, 6,7) trois fois.
- ✓ La base de test : elle est composée de 360 fichiers vocaux prononcés par 60 locuteurs, comprenant les deux sexes (30 hommes, 30 femmes). Chaque locuteur répète les chiffres (8,9) trois fois.

4.2 Implémentation du codec GSM FR

Dans notre mémoire, comme déjà mentionné on s'intéresse à l'influence du codage GSM sur un système de reconnaissance du locuteur. Pour cela on a besoin d'une base de données issue d'un système de communication **GSM**.

Dans ce cas, on a deux solutions :

- ✚ Soit utiliser des données réellement issues de téléphonie mobile mais elles sont rares et coûteuses à acquérir.
- ✚ Soit simuler le fonctionnement du canal **GSM**, qui est en réalité le codage source plus codage du canal. Cette deuxième solution a été adoptée.

En effet, notre simulation est basée sur le codeur de parole de **GSMFR** fondé sur la technologie **RPE-LTP** (codage source seulement), car on a supposé que toutes les dégradations de la qualité des signaux de paroles et qui influent sur le taux de reconnaissance sont dues seulement au codage de la parole (le canal de transmission radio est supposé parfait).

Nous avons utilisé l'implémentation réalisée par J. Degener [Degener,2001] et ceci en la compilant sous l'environnement Linux. Le résultat de cette compilation est une librairie à trois outils. L'outil *toast* est utilisé pour coder un fichier audio de type PCM linéaire (accepte aussi les codages A-law et mu-law) donnant en sortie un fichier d'extension « .gsm » codé sur 260 bits. Le processus de décodage est réalisé par l'outil « *untoast* ». L'implémentation de ce codec est réalisée par un programme écrit en langage script *Perl*.

4.3 Description du système d'IAL Multi-bandes

Le système HMM multi-bandes mis en œuvre est donné par la figure 4.2 suivante. Après la description des différentes étapes de ce système dans le chapitre 2, nous présenterons ici les deux implémentations réalisées avec ce modèle. La première utilise 4 sous bandes avec les limites suivantes 0-901 Hz, 797-1661 Hz, 1493-2547 Hz et 2298-4000Hz et la deuxième est

configurée pour 7 sous bandes dont les limites sont 0- 360 Hz, 330-640Hz, 580-950 Hz, 860-1360Hz, 1265- 1920 Hz, 1800-2700 Hz, 2515-4000 Hz

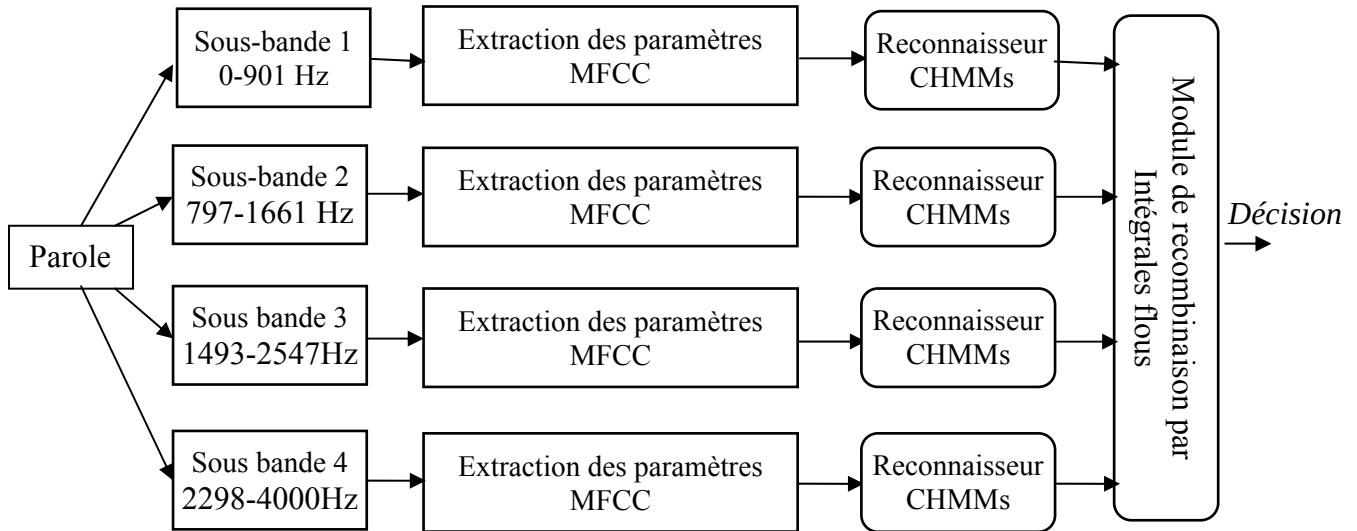


Figure 4.2 : Système multi-bandes à quatre sous-bandes.

4.3.1 Filtrage en sous bandes

La division en sous bande des fichiers de parole est réalisé par un filtre Butterworth d'ordre 6 sous la boîte à outils Matlab.

4.3.2 Extraction des paramètres

La représentation du signal est obtenue avec l'outil HCopy de HTK (plus de détails sont donnés dans l'ANNEX C) qui produit en particulier des coefficients MFCC ainsi que leurs coefficients différentiels du premier et du second ordre. Le réglage des paramètres d'entrée de HCopy est réalisé par :

La pré-accentuation acoustique qui consiste à filtrer le signal vocal est réalisé par un filtre passe haut de transmittance $H(z) = 1 - 0.97z^{-1}$.

L'étape de fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération glissante, nous avons utilisé une fenêtre de Hamming glissante de durée de 25 ms avec un déplacement de 10 ms.

4.3.3 Reconnaisseurs à base de CHMMs

les reconnaissances utilisés dans chaque bande de fréquence sont basés sur les modèles de Markov cachés continus à 3 états émetteurs de même topologie expliqué dans le chapitre 2. Le système fonctionne en deux étapes principales.

- l'apprentissage des modèles de locuteurs qui consiste à estimer les paramètres des modèles de Markov. Ces modèles sont initialisés avec l'algorithme de Viterbi, puis ré-estimés par l'algorithme optimal de Baum-Welch. L'entraînement des modèles est effectué séparément sur chaque bande de fréquence et ceci en utilisant les outils HInit et HRest de HTK
- Le décodage qui est réalisé par l'algorithme de Viterbi (par l'outil HVit), sous la contrainte d'un réseau syntaxique défini par l'utilisateur en donnant la liste des locuteurs de la base de données utilisée.

4.3.4 Le module de recombinaison

La recombinaison par l'intégrale floue (FI :Fuzzy Integral) utilise les scores retournés pour chaque locuteur par tous les reconnaissseurs c'est donc une recombinaison sur les scores issus de tous les reconnaissseurs en sous-bandes en fin d'analyse de ces derniers. Ces scores sont regroupés dans la matrice de profil de décision pour être fusionnés avec une mesure floue en utilisant l'intégrale de Choquet (chapitre2 §2.6). Ce module est implémenté sous Matlab Version R2009b.

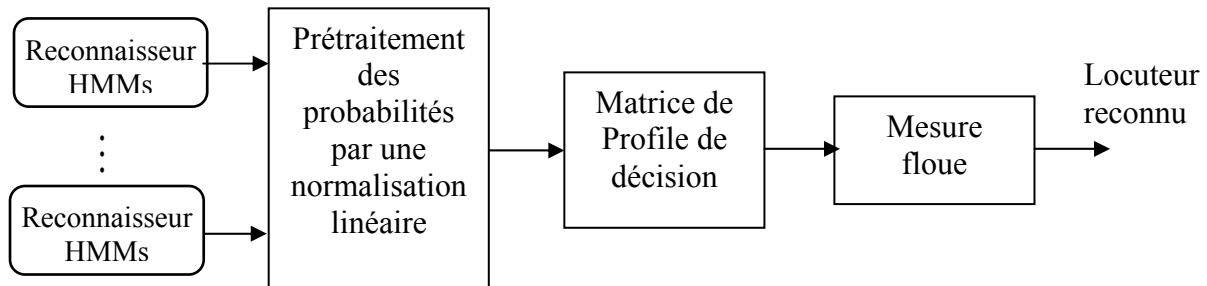


Figure 4.3 : Schéma synoptique du module de recombinaison par FI

4.3.5 Décision

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

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

4.4 Tests et évaluations

Nous avons effectué plusieurs expériences pour évaluer le système d'identification du locuteur mis en œuvre à base de CHMMs multi-bandes. Toutes les expériences que nous avons réalisées sont effectuées en mode indépendant du texte. Pour les bases d'apprentissage et de test, on a deux cas :

- ✓ La base d'apprentissage utilisée est celle de la base originale ARADIGIT_8K, par contre les bases de tests sont transcodées (ARADIGIT_GSM).
- ✓ Les bases d'apprentissage et de tests sont toutes les deux transcodées.

Le système de référence utilisé par chaque sous-reconnaisseur du système multi-bandes possède les caractéristiques suivantes :

- La topologie des HMM est établit par un modèle de type gauche-droite à 3 états émetteurs (type Bakis).
- La dimension du vecteur acoustique qui modélise chaque locuteur est de 36 (12 MFCC, 12 Δ MFCC et $\Delta\Delta$ MFCC).
- Le nombre de gaussienne optimal trouvé est égal à 12.

4.4.1 Expérience 1 : Effet du codeur GSM FR

La première expérience est exécuté pour des bases de données en bande totale pour situer l'effet du codeur GSM Full rate sur le taux d'identification du locuteur, et ceci en comparant les résultats des expériences effectuées avec la bases de test ARADIGIT_GSM en utilisant dans premier temps comme base d'apprentissage la base ARADIGIT_GSM ensuite la base de données ARADIGIT 8Khz à ceux trouvés avec l'évaluation de la base ARADIGIT 8Khz. Le tableau 4.1 résume ces résultats comparatifs obtenus.

Tableau4.1 : Taux de reconnaissance comparatifs entre bases ARBDIGIT8K et ARADIGIT_GSM.

Types de données		Taux d'identification
Apprentissage	Test	
ARADIGIT 16K	ARADIGIT 16K	95 .28%
ARADIGIT 8K	ARADIGIT 8K	89 .72%
ARADIGIT_GSM	ARADIGIT_GSM	88.33%
ARADIGIT 8K	ARADIGIT_GSM	78 .89%

La notation employée dans les parties qui suivent pour indiquer l'utilisation du même type de bases pour l'apprentissage et le test est MA (Matched conditions). Par contre, l'utilisation de deux types différents des bases est notée par MI (Mismatched conditions).

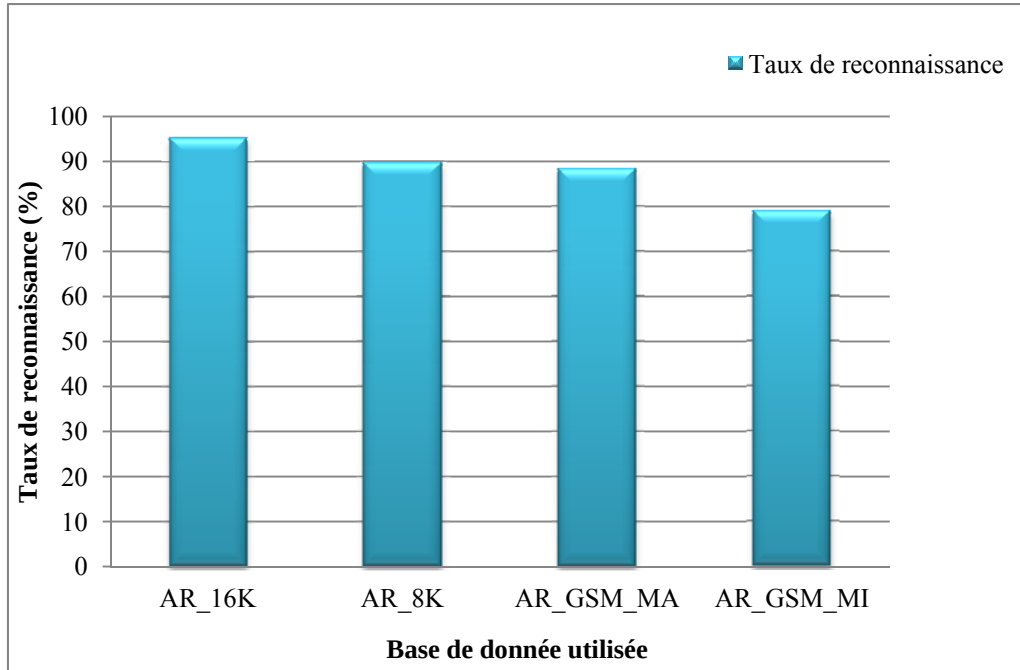


Figure 4.4 : Les résultats comparatifs du taux de reconnaissance du locuteur.

La figure. 4.4 montre les variations du taux d'identification en fonction du type de base de données utilisée pour le test. Elle montre une dégradation du taux d'identification du locuteur de l'ordre de 10% pour une base de données transcodée GSM. Ceci pour la base d'apprentissage originale ARADIGIT_8K mais pour une similitude entre la base de test et la base d'apprentissage (ARADIGIT_GSM) la dégradation du TIC est marginale.

4.4.2 Expérience 2 : Architecture Multi-bandes

Ces expériences concernent l'architecture multi-bandes utilisant 4 et 7 sous bandes de fréquences. Les expériences du système multi-bandes sont réalisées pour les bases ARADIGIT_8K, ARADIGIT_GSM afin d'évaluer l'apport de cette architecture dans le cas de la parole transcodée GSM. Nous avons effectué deux catégories d'expériences concernant la base de données de la parole transcodée, la première concerne l'utilisation du même type de base de données d'apprentissage et de test. Le deuxième type de tests est effectué pour le cas d'une base d'apprentissage ARADIGIT_8K et une base de test ARADIGIT_GSM. Les résultats de ces expériences sont données par les tableaux 4.2 et 4.3 respectivement.

Tableau 4.2 : Taux de reconnaissance pour le système multi-bandes utilisant 4 bandes de fréquences.

Types de données	Bandes de fréquences	Taux d'identification par bande	Taux d'identification après recombinaison par FI
ARADIGIT_8K	Bande 1	86.39%	88.89%
	Bande 2	65.83%	
	Bande 3	55.83%	
	Bande 4	66.39%	
ARADIGIT_GSM_MA	Bande 1	83.89%	85%
	Bande 2	62.22%	
	Bande 3	48.33%	
	Bande 4	55.83%	
ARADIGIT_GSM_MI	Bande 1	77,78%	77.50%
	Bande 2	61,39%	
	Bande 3	37,5%	
	Bande 4	43,89%	

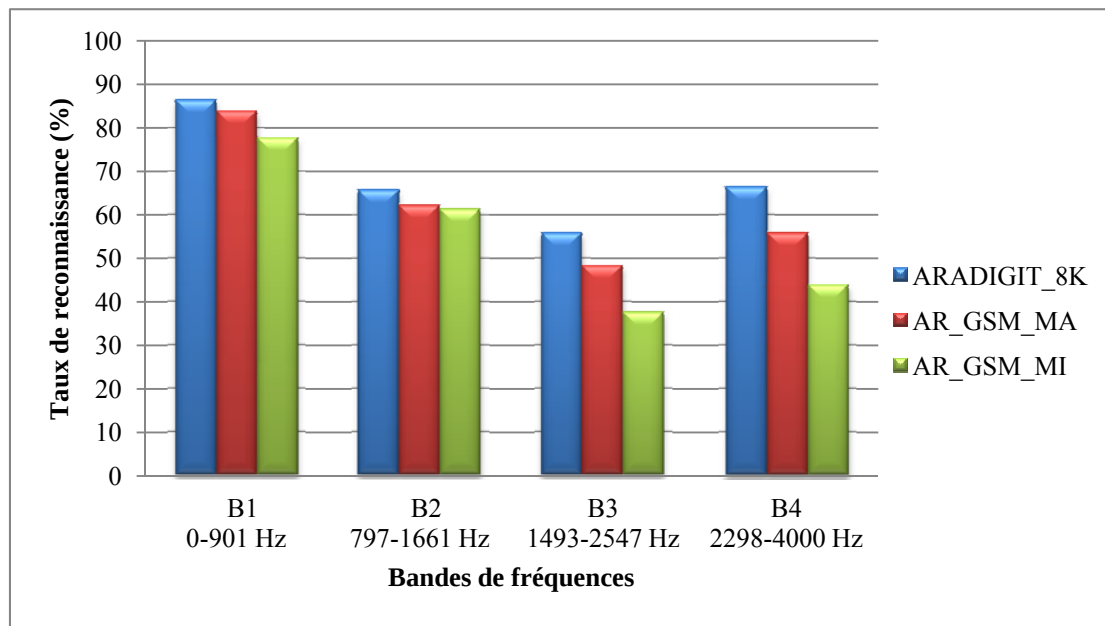


Figure 4.5 : Taux de reconnaissances des 4 sous bandes de fréquences pour le même type de bases d'apprentissage et de test.

Tableau 4.3 : Taux de reconnaissance pour le système multi-bandes utilisant 7 bandes de fréquences.

Types de données	Bandes de fréquences	Taux d'identification par bande	Taux d'identification après recombinaison par FI
ARADIGIT_8K	Bande 1	75.56%	91.67%
	Bande 2	66.94%	
	Bande 3	63.61%	
	Bande 4	55%	
	Bande 5	46.94%	
	Bande 6	50%	
	Bande 7	66.39%	
ARADIGIT_GSM_MA	Bande 1	73.61%	88.33%
	Bande 2	62.50%	
	Bande 3	56.11%	
	Bande 4	48.33%	
	Bande 5	40.28%	
	Bande 6	43.33%	
	Bande 7	73.61%	
ARADIGIT_GSM_MI	Bande 1	58.89%	81.39%
	Bande 2	58%	
	Bande 3	56.39%	
	Bande 4	46.67%	
	Bande 5	37.22%	
	Bande 6	31.39%	
	Bande 7	38.06%	

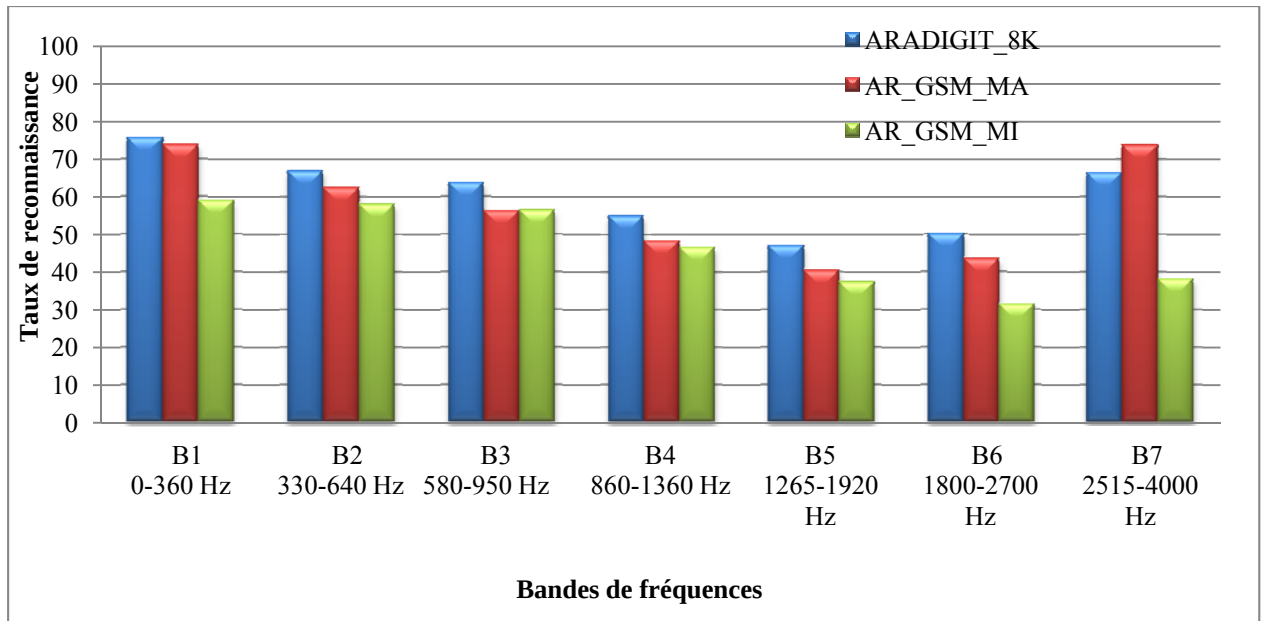


Figure 4.6 : Taux de reconnaissances des 7 sous bandes de fréquences pour le même type de bases d'apprentissage et de test.

Les deux figures 4.4 et 4.5 présentent les taux d'identification du locuteur pour les différentes bandes de fréquences des deux systèmes 4 et 7 sous bandes, nous remarquons que l'information utile à la caractérisation du locuteur est surtout présente en basse fréquence ($f < 500 \text{ Hz}$) et dans les hautes fréquences ($f > 2500 \text{ Hz}$) alors que les performances d'identification baissent nettement pour les bandes intermédiaires.

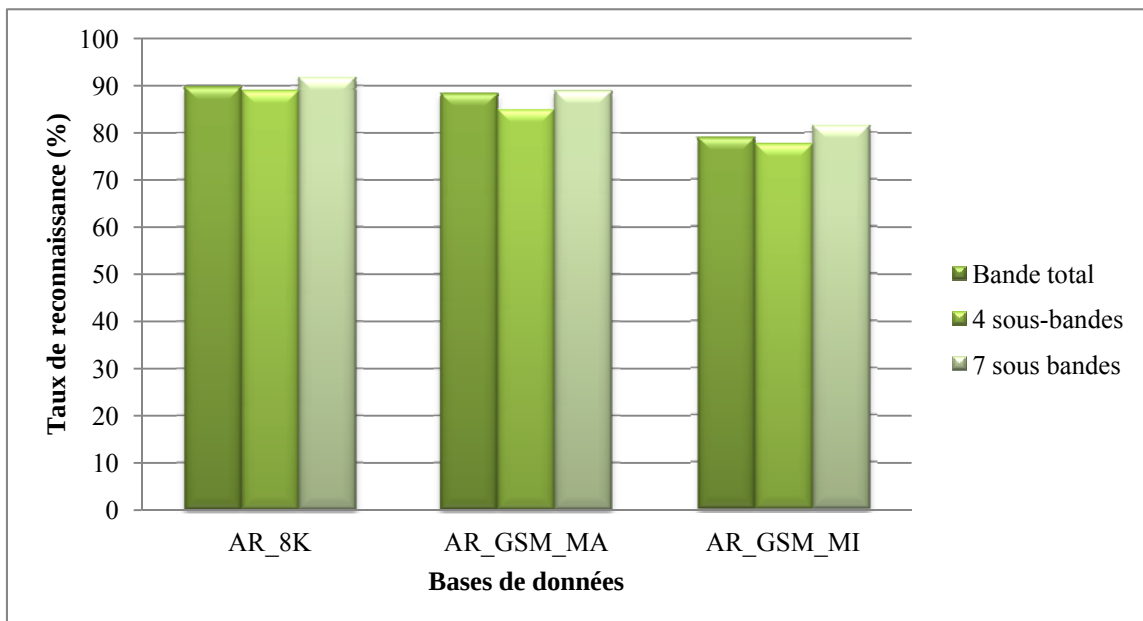


Figure 4.7 : Comparaison des taux de reconnaissances des systèmes multi-bandes avec le système de référence (B.T).

Les performances obtenues après recombinaison dans le cas du système multi-bandes à 4 sous-bandes de fréquences sont inférieures à ceux du système de référence (Bande Totale), par contre le système de 7 sous-bandes apporte une amélioration marginale. Comme certains travaux l'ont suggéré, nous avons pensé d'ajouter la bande totale aux deux systèmes (4 et 7 sous-bandes), ceci nous amène donc à deux modèles le premier est à cinq bandes et le deuxième englobe 8 bandes. Les résultats issus de ces nouveaux modèles sont présentés dans le tableau 4.4.

Tableau 4.4 : Taux de reconnaissance pour les deux systèmes en ajoutant la bande totale.

Type de système	Types de données	Taux d'identification après recombinaison par FI
4 Sous-Bandes + Bande Totale	ARADIGIT_8K	91.94%
	ARADIGIT_GSM_MA	90.83%
	ARADIGIT_GSM_MI	83.33%
7 Sous-Bandes + Bande Totale	ARADIGIT_8K	93.61%
	ARADIGIT_GSM_MA	94.72%
	ARADIGIT_GSM_MI	86.39%

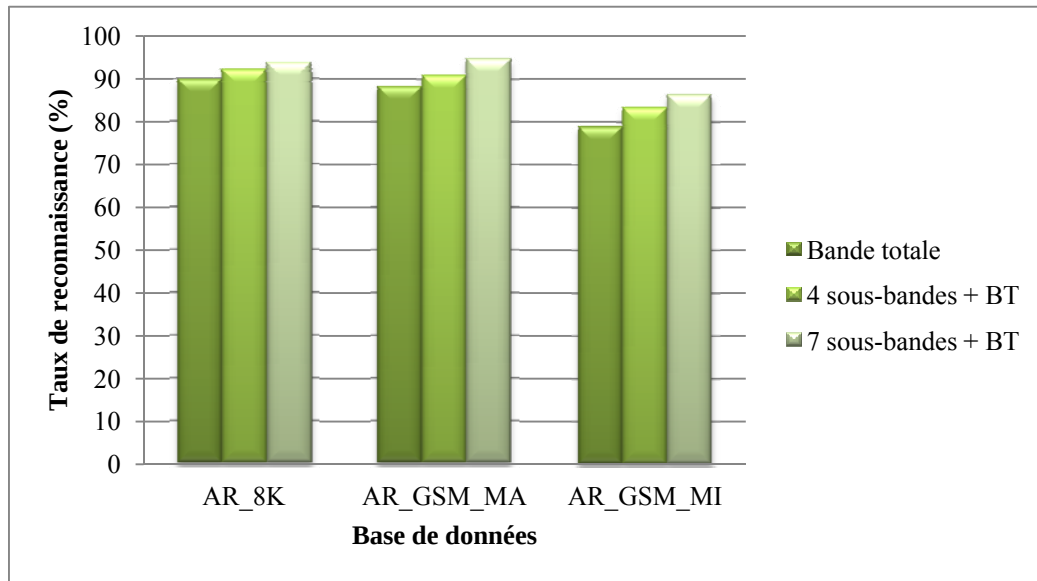


Figure 4.8 : Comparaison des taux de reconnaissances des systèmes multi-bandes + BT avec le système de référence (B.T).

4.5 Interprétation des résultats

Comme nous l'avons déjà remarqué dans le 2^{ème} test, le fait de diviser le spectre en plusieurs parties réduit considérablement la quantité d'information disponible dans chacune des sous bandes de fréquence. En conséquence, les taux de reconnaissance dans chaque sous bande sont bien plus bas que si tout le spectre est utilisé. L'introduction du module de recombinaison par la méthode des intégrales floues n'a pas pallié ce problème pour le cas du système à 4 sous bandes. Le deuxième système (celui de 7 sous-bandes) apporte une amélioration marginale. D'une manière plus générale, le système multi-bandes implémenté reste toujours incapable d'élever le taux de reconnaissance comme nous l'avons espéré.

Le problème cité a été traité dans la littérature par l'ajout de l'information apporté par le spectre totale, cela a augmenté le taux de reconnaissance pour la parole transcodée de 4.5% dans le cas de l'architecture à 4 sous bandes et de 7.5% pour celle de 7 sous bandes, donc l'amélioration apporté par l'architecture multi-bandes pour le cas de la parole transcodée par le codeur GSM Full Rate est assez remarquable.

4.6 Conclusion

Dans ce chapitre, nous avons présenté l'aspect expérimental de notre travail qui consiste à implémenter un système de reconnaissance du locuteur à base des modèles de Markov cachés multi-bandes. Les différentes expériences effectuées pour la parole transcodée via le codeur GSM FR ont montré l'efficacité ainsi que l'apport de l'architecture multi-bandes pour les deux cas étudiés 4 et 7 sous bandes de fréquences. La méthode de recombinaison par les intégrales floues a aussi prouvé son efficacité par l'accroissement du taux de reconnaissance d'une manière considérable. Cette étude nous a permis aussi de mettre en évidence l'importance du spectre complet (Bande Totale) dans le système multi-bandes.

Conclusion et perspectives

L'objectif principal du sujet abordé dans ce présent mémoire est d'étudier et d'évaluer la performance de la Reconnaissance Automatique du Locuteur (RAL) en mode indépendant du texte via une ligne de communication GSM FR. Essentiellement axée sur les problèmes d'authentification et de confidentialité, la RAL est très souvent sollicitée dans un contexte de téléphonie mobile GSM, pour lequel la voix demeure le seul élément disponible pour reconnaître une personne. Les signaux transmis par réseau GSM montrent une dégradation importante causée par la perte d'information due au codage de la parole à un débit relativement bas.

Pour mener cette étude, l'idéal serait de pouvoir utiliser des données réellement issues de téléphonie mobile mais elles sont rares et coûteuses à acquérir. Par conséquent, nous avons opté pour la simulation ainsi nous avons commencé notre travail par simuler le codec GSM FR sous l'environnement linux.

Le codec GSM FR simulé, nous avons mis au point le système de reconnaissance de base utilisant les modèles de Markov cachés continus. Les résultats de reconnaissance obtenus ont montré une nette dégradation des taux d'identification pour la parole transcodée GSM. Le paradigme multi-bandes basé essentiellement sur la perception de l'oreille humaine nous a permis de mettre en œuvre le système de RAL multi-bandes. Les sous-reconnaisseurs du système multi-bandes implémentés sont basés également sur les modèles de Markov cachés continus. Des approches de recombinaison étudiées, nous avons retenu l'approche basée sur la théorie de la logique floue. Les résultats des expériences effectuées ont montré que l'approche multi-bandes, ainsi que la méthode de recombinaison par les intégrales flous sont efficaces pour la parole transcodée GSM FR dans les deux cas étudiés 4 et 7 sous bandes en ajoutant la bande totale. Nous avons mis en évidence les bandes de fréquence les plus utiles pour l'identification automatique du locuteur. Cependant, il n'est pas ressorti de cette étude une configuration optimale (taille et nombre de sous-bandes, recouvrement entre les bandes) permettant d'aborder le problème de la fusion des reconnaisseurs avec une architecture fixée une fois pour toutes. Nous avons également constaté la contribution du spectre complet lorsqu'il est ajouté aux sous bandes.

Malgré les améliorations remarquables des taux de reconnaissance obtenus pour la parole transcodée GSM, le travail proposé dans ce mémoire reste à notre avis encore perfectible. Ainsi, comme perspectives à ce travail nous pouvons citer :

En reconnaissance automatique du locuteur, nous avons vu qu'une grande partie de l'information se situe dans les hautes fréquences ($f > 3000\text{Hz}$) mais aussi dans les basses fréquences ($f < 500\text{Hz}$). Par conséquent, l'analyse classique avec une fenêtre de taille fixe ne permet qu'un compromis pour capturer l'information contenu dans le signal. Il serait peut être plus judicieux de réaliser une analyse du signal de parole en sous-bandes avec des fenêtres de résolution de tailles différentes suivant la sous-bande analysée.

Nous proposons également d'utiliser l'approche de reconnaissance la mieux adaptée pour chacune des sous-bandes de fréquences. Cela peut être réalisé par une implémentation de plusieurs sous-reconnaisseurs utilisant différentes approches de reconnaissance.

Enfin, limité par son débit faible le réseau GSM commence à perdre sa place dans le marché de la télécommunication au profit des nouveaux réseaux GPRS et UMTS. Ces réseaux offrent de meilleurs services avec un débit beaucoup plus élevé en utilisant la technique de paquets. Tout l'intérêt serait donc d'étendre ce travail à la transmission GPRS/UMTS et étudier l'influence de ce type de codage sur la performance des systèmes de RAL.

Bibliographie

[Allen,1994] J.B. Allen « How do humans process and recognize speech ? » IEEE transactions on ASSP, Vol 2,n°4, pp 567-577, 1994.

[Alliche, 1991] A. Alliche « Conception et réalisation d'un logiciel d'un de traitement de la parole et étude de l'emphase sur la structure formantique des voyelles de l'arabe standard ». Thèse de magister. Institut Electronique d'USTHB. 1991.

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

[Bakis,1976] R. Bakis « Continuous speech recognition via centisecond acoustic states» 91th. Meeting of the Acoust.Soc. Am., avril 1976.

[Barras, 1996] C. Barras «Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés » Thèse de doctorat en informatique, Université de Paris VI. Mai 1996.

[Baum, 1972] L. Baum « An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes » Inequality, vol. 3, 1972.

[Baum, 1968] L.E. Baum et G.R. Sell « Growth functions for transformations on manifolds » Pac.J.Math; vol 27, no.2 pp.211-277, 1968.

[Bennani, 1990] Y. Bennani, F. F. Soulie, P. Gallinari «A connectionist approach for automatic speaker identification ». International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1990.

[Besacier ,1998] L. Besacier «Un modèle parallèle pour la reconnaissance automatique du locuteur» », thèse de doctorat, Université d'Avignon et des pays Vaucluse, 1998.

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

- [Bimbot, 1995] F. Bimbot, I. Magrin Chagnolleau, L. Mathan « Second-order statistical measures for text-independent speaker identification ». Speech Communication, volume 17, Août 1995.
- [Boite, 2000] :R. Boite et H. Boulard et autres « Traitement automatique de la parole » Collection électricité. Lausanne : Presses Polytechniques et Universitaires Romandes, 2000.
- [Booth et al., 1993] I. Booth, M. Barlow, B. Watson « Enhancements to DTW and VQ decision algorithms for speaker recognition ». Speech Communication, volume 13(3-4), Décembre 1993.
- [Boulard,1990] H. Boulard et C.J. Wellekens, « Links between Markov models and multilayer perceptrons» IEEE Trans. on Pattern Anal. and Machine Intell., vol. 12, no. 12, pp. 1-12, 1990.
- [Boulard,1996] H. Boulard, et S. Dupont « A new ASR approach based on independent processing and recombination of partial frequency bands » In proceedings ICSLP, Philadelphia, USA, Octobre 1996.
- [Calliope, 1989] « La parole et son traitement automatique », Collection technique et scientifique des télécommunications, Editions MASSON, 1989.
- [Cho, 1995] S. B. Cho et J. H. Kim «Multiple network fusion using fuzzy logic», IEEE Transactions on Neural Networks, 6:497–501, 1995.
- [Chu, 2003] WAI C. CHU « Speech Coding Algorithm Foundation and Evolution of Standardized Coders» A JOHN WILEY & SONS, INC., PUBLICATION, 2003
- [Kim, 1995] J.H. Kim, S.-B. Cho « Combining multiple neural networks by fuzzy integral and robust classification», IEEE Trans.Systems Man Cybernet. 25 (1995) 380}384.1995 .
- [Debyeche, 2007] M. Debyeche « Reconnaissance automatique de la parole appliqué à la langue arabe », Thèse de Doctorat d'Etat, USTHB, 2007.
- [Degener,1992] J. Degener et C. Bormann «GSM 06.10 13 kbit/s RPE/LTP speech compression software», Technische Universitaet Berlin, 1992.
- [Degener,2001] J . Degener « Digital Speech Compression Putting the GSM 06.10 RPE-LTP algorithm to work », 2001

- [Dekker, 2003] M. Dekker « Speech Processing, a Dynamic and Optimization-Oriented Approach» Series: Signal Processing and Communications Series . ISBN-13: 9780824740405, Juin 2003.
- [Demoulin, 2004] C. Demoulin, M. V. Droogenbroeck « PRINCIPES DE BASE DU FONCTIONNEMENT DU RÉSEAU GSM » *Revue de l'AIM*, pages 3.18, N°4, 2004.
- [Duchnowski,1993] P. Duchnowski, «A new structure for automatic speech recognition» *PhD thesis of the Massachusetts institute of technology (MIT)*, September 1993.
- [ETSI,2000] «Digital cellular telecommunications system (Phase 2+);Full rate speech; Transcoding» (GSM 06.10 version 8.1.1 Release 1999), ETSI EN 300 961 V8.1.1 (2000-11) *European Standard (Telecommunications series)*.
- [Fletcher,1953] H. Fletcher, «Speech and Hearing in Communication». *Krieger*. New-York, 1953.
- [Forney,1973] G.D. Forney « The Viterbi algorithm» *Proc. of the IEEE*, vol. 61, no. 3, pp. 268-278, 1973.
- [Frederickson et al., 1994] S. E. Frederickson, L. Tarassenko «Radial basis functions for speaker identification ». Workshop on Automatic Speaker Recognition, Identification, verification, Martigny (Suisse), Avril 1994.
- [Fredouille, 2000] C. Ferdouille « Approche statistique pour la reconnaissance automatique du locuteur : information dynamique et normalisation Bayesienne des vraisemblances », thèse de doctorat, Université d'Avignon et des pays Vaucluse, 2000.
- [Furui, 1981] S. Furui «Cepstral analysis technique for automatic speaker verification» *IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP)*, volume 29(2), Avril 1981.
- [Grassi,2000] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, F. Pellandini, «Influence of GSM Speech Coding on the Performance of Text-Independent Speaker Recognition », *Proc. Of EUSIPCO, European Signal Processing Conference*, Tampere, Finland, September 4-8, pp. 437- 440, 2000.
- [Gray,1984] R.M. Gray « Vector quantization » *IEEE ASSP Mag.*, vol. 1(2):4-29, 1984.
- [Haykin, 1986] S. Haykin, « Adaptive Filter Theory », Prentice Hall, New Jersey, 1986.

[Hermansky, 1996] H. Hermensky, S. Tibrewala, M. Pavel « Towards ASR on partially corrupted speech » In proceedings ICSLP, Philadelphia, USA, Octobre 1996.

[Hermansky, 1997] H. Hermensky, S. Tibrewala « Sub-band based recognition of noisy speech » IEEE, Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997.

[Igounet, 1998] S. Igounet « Eléments pour un système de reconnaissance automatique de la parole continue du français » Thèse de doctorat en Informatique, Université d'Avignon et des Pays de Vaucluse. juillet 1998.

[Jacob, 1995] B. Jacob « Un outil informatique de gestion de Modèles de Markov Cachés : expérimentations en reconnaissance automatique de la parole » Thèse de doctorat, Université Paul Sabatier. Septembre 1995.

[Joseph, 1997] Joseph P. Campbell, JR. «Speaker Recognition: A Tutorial», *Proc. Of IEEE*,VOL.85, no.9, Sept. 1997.

[Juang, 1985] B.-H. Juang et L.R. Rabiner, « Mixture autoregressive hidden Markov models for speech signals », *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1404-1413, 1985.

[Kleijn, 1995] B. Kleijn et K. K. Paliwal « Speech Coding and Synthesis » Elsevier Science, 1995.

[Keiler, 2000] F. Keiler , D. Arfib et U. Zolzer « Efficient Linear Prediction for Digital Audio Effects» Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy. Décembre 2000.

[Krobba, 2008] A. Krobba « Reconnaissance Automatique du Locuteur Via une Ligne de Communication GSM » Mémoire de Magistère, USTHB, 2008.

[Kuncheva, 2003] L. I. Kuncheva « “Fuzzy” vs “non-fuzzy” in combining classifiers designed by boosting» *IEEE Transactions on Fuzzy Systems* 11:729–741, 2003.

[Madre,2004] G. Madre « Application de la transformée en nombres entiers à l'étude et au développement d'un codeur de parole pour transmission sur réseaux IP » Thèse de Doctorat, Université de Bretagne Occidentale, 2004.

[Mami, 2003] Y. Mami « Reconnaissance de locuteurs par localisation dans un espaces de locuteurs de références » Thèse de Doctorat, Ecole Nationale supérieure des télécommunications de Paris, 2003.

[Markel, 1976] J.D. Markel et A.H. Gray « Linear Prediction of Speech » Communication and Cybernetics, Springer-Verlag, Berlin Heidelberg New York ,1976.

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

[Margin, 1995]I. Margin-chagnolleau, J.-F. Bonastre, et F. Bimbot «Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods» *European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, 1995.

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

[Meignier, 2002] S. Meignier « Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection ». Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse, 2002.

[Preti, 2008] A. Preti « Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur », thèse de doctorat, Université d'Avignon et des pays Vaucluse, 2008.

[Noll, 1964] A. Noll «Short-time spectrum and 'cepstrum' techniques for vocalpitch detection ». *Journal of the Acoustical Society of America* 36(2), 430–451, 1964.

[Oglesby et al., 1990] J.Oglesby, J. S. Mason «Optimisation of neural models for speaker identi_cation». *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 261{264, 1990.

[Oglesby et al., 1991] J.Oglesby, J. S. Mason « Radial basis function networks for speaker recognition». *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 393{396, 1991}, Toronto (Canada).

- [OShaughnessey, 1987] D. O'Shaughnessey, « Speech Communication - Human and Machine» Addison Wesley Publishing Company, 1987.
- [Pravinkumar, 2001] P. Pravinkumar and B. M Wasfy « Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review», *Proc. IEEE*, 2001.
- [Rabiner,1989] L.R. Rabiner, «A tutorial on hidden Markov models and selected applications in speech recognition» *Proc. of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.
- [Reynolds, 1995] D. Reynolds« Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* », 17(1): 91-108, 1995.
- [Rose, 2002] P. Rose «Forensic Speaker Identification», *Forensic Science Series*, Taylor & Francis London, New- York, 2002.
- [Sayoud,2003] H. Sayoud, « Reconnaissance de locuteur – Approche connexionniste », thèse de doctorat, USTHB, 2003.
- [Soong, 1992] F. K. Soong, A. E.Rosenberg, L. R. Rabiner, B. H. Juang « A vector quantization approach to speaker recognition ». *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tampa (USA), 1992.
- [Temko, 2008] A. Temko, D. Macho, C. Nadeu « Fuzzy integral based information fusion for classification of highly confusable non-speech sounds» *Pattern Recognition* Volume: 41, Issue: 5, Pages: 1814-1823, 2008.
- [Verikas,1999] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis «Soft combination of neural classifiers: A comparative study». *Pattern Recognition Letters*, 20:429– 444, 1999.
- [Viterbi,1967] A.J Viterbi « Error bounds for convolutional codes and an asymptotically optimal decoding algorithm» *IEEE Trans. Informat. Theory*, vol. IT-13. 1967.
- [Yarmey, 1995] A Yarmey « Earwitness Speaker Identification», *Psychology, Public Policy, and Law*, 1995,
- [Veaux, 2005] C. Veaux «Étude de traitements en réception pour l'amélioration de la qualité de la parole - Application au GSM », thèse de doctorat, l'Ecole Nationale Supérieure des Télécommunications, Paris, 2005.

[Yu, 1995] K. Yu, J. S. Mason, J. Oglesby « Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation». IEE vision, image and signal processing, 1995.

[Young, 2005] S. Young, J. Odell, et d'autres « The HTK Book Version 3.3» Speech group, Engineering Department , Cambridge University. April 2005.

[Young, 1989] S.J. Young, N.H. Russel et J.H. Thornton, « Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems» Cambridge University Engineering Departement rapport technique, 1989.

[Wang, 1998] D. Wang, J.M. Keller, C.A. Carson, K.K. McAdoo-Edwards, C.W. Bailey «Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion », IEEE Trans. Systems Man Cybernet. 28B (4) (1998) 583}591.1998.

ANNEXE A

Annexe A : Ensembles et intégrales flous.

Introduction

La logique floue est une extension de la logique booléenne créé par Lotfi Zadeh en 1965 en se basant sur sa théorie mathématique des ensembles flous, qui est une généralisation de la théorie des ensembles classiques. En introduisant la notion de degré dans la vérification d'une condition, permettant ainsi à une condition d'être dans un autre état que vrai ou faux, la logique floue confère une flexibilité très appréciable aux raisonnements qui l'utilisent, ce qui rend possible la prise en compte des imprécisions et des incertitudes.

1. Ensemble floue :

La théorie des sous-ensembles flous est une théorie mathématique du domaine de l'algèbre abstraite. Elle a été développée afin de représenter mathématiquement l'imprécision relative à certaines classes d'objets et sert de fondement à la logique floue.

Définition :

Une partie A d'un ensemble E est usuellement associée à sa fonction caractéristique. Celle-ci s'applique sur les éléments x de E . Elle prend la valeur 0 si x n'appartient pas à A et 1 si x appartient à A .

On souhaite définir une partie A floue de E en attribuant aux éléments x de E un degré d'appartenance, d'autant plus élevé qu'on souhaite exprimer avec certitude le fait que x est élément de A . Cette valeur vaudra 0 si on souhaite exprimer que x de façon certaine n'est pas élément de A , elle vaudra 1 si on souhaite exprimer que x appartient à A de façon certaine, et elle prendra une valeur comprise entre 0 et 1 suivant qu'on estime plus ou moins certain l'appartenance de x à A . On est donc amené à définir une partie floue de la façon suivante :

Une partie floue (ou sous-ensemble flou) d'un ensemble E est une application de E dans $[0,1]$.

Plus généralement, si L est un treillis complet, distributif et complémenté, on définit une partie L -floue comme étant une application de E dans L . Si $L = [0,1]$, on retrouve la définition précédente de partie floue, et si $L = \{0,1\}$, on retrouve la notion usuelle de partie de E .

Propriétés :

- Une partie floue A de E est caractérisée par une application de E dans $[0,1]$. Cette application, appelée **fonction d'appartenance** et notée μ_A représente le degré de validité de la proposition « x appartient à A » pour chacun des éléments x de E . Si $\mu_A(x) = 1$, l'objet x appartient totalement à A , et si $\mu_A(x) = 0$, il ne lui appartient pas du tout. Pour un élément x donné, la valeur de la fonction d'appartenance $\mu_A(x)$ est appelée **degré d'appartenance** de l'élément x au sous-ensemble A .
- L'ensemble E est donné par la fonction d'appartenance identiquement égale à 1. L'ensemble vide est donné par la fonction d'appartenance identiquement nulle.
- Le **noyau** d'une partie floue A est l'ensemble des éléments qui appartiennent totalement à A c'est-à-dire dont le degré d'appartenance à A vaut 1.

$$n(A) = \{x \in E / \mu_A(x) = 1\}$$

- Le **support** d'une partie floue A est l'ensemble des éléments appartenant, même très peu, à A c'est-à-dire dont le degré d'appartenance à A est différent de 0.

$$Supp(A) = \{x \in E / \mu_A(x) > 0\}.$$

- La **hauteur** d'un sous-ensemble flou A de E est définie par

$$h(A) = \sup\{\mu_A(x) \mid x \in E\}.$$

- Un ensemble fini possède un nombre fini de sous-ensembles L-flous si et seulement si le treillis L est fini. Si $L = [0,1]$, un ensemble fini possède une infinité de sous-ensembles flous.

2. Mesures et intégrales floues

- Mesures floues :

La notion de mesure floue est une extension de celle de mesure de probabilité. Les mesures floues servent à mesurer le degré ou au même titre que les mesures de probabilité servent à mesurer l'aléatoire.

Définition : Soit X un ensemble, B un sous ensemble floue de X . On appelle mesure floue toute fonction g de B à valeurs entre 0 et 1, satisfaisant :

- **Bornée :** $g(\phi) = 0, g(x) = 1$

- **Monotone :** A, B deux parties de B : $A \subseteq B \Rightarrow g(A) \leq g(B)$

Le triplet (X, B, g) est dit : espace mesuré flou.

La notion de fonction mesurable se définit de manière analogue à la définition classique.

Différentes versions de l'intégration floue ont été proposées dans la littérature. On présente ici les deux types les plus répandus : l'intégrale floue de Sugeno et celle de Choquet.

Soit X un ensemble ordinaire, A une partie crisp de X , m une mesure floue définie sur X et f une fonction réelle μ -mesurable et bornée.

Définition : L'intégrale floue de Choquet de f sur A , par rapport à m est définie par :

$$\int_A f d\mu = \int_{-\infty}^0 [\mu(A \cap \{f \geq x\}) - \mu(A)] dx + \int_0^{+\infty} \mu(A \cap \{f \geq x\}) dx$$

Le signe d'intégration dans le terme droit désigne l'intégrale impropre de Reimann.

Soit (X, B, g) un espace mesuré flou, f une fonction réelle mesurable sur X .

Définition : L'intégrale floue de Sugeno de la fonction f sur une partie mesurable A de X , par rapport à la mesure floue m est définie par :

$$\int_A f(x) \circ \mu(.) = \sup_{0 \leq \alpha \leq 1} [\alpha \wedge \mu(A \cap F_\alpha)]$$

Où $F_\alpha = \{x / f(x) \geq \alpha\}$

Applications : Diverses applications ont été proposées : Synthèse et agrégation d'informations floues, théorie des possibilités, ...

ANNEXE B

Annexe B : Codeur GSM 6.10

I. Principe du codeur GSM 6.10

L'algorithme RPE-LTP (regular pulse excitation- long term prediction) de cette norme (GSM 6.10) pour téléphones mobiles construit des trames de 260 bits à partir de 160 échantillons PCM à 13 bits, à 8 kHz. Une seconde ne nécessite que 1625 octets, et un méga Octet suffit pour 10 mn. Une trame couvre donc 20 ms (160 éch.), ce qui correspond à une période pour une voix très grave, et à 10 pour une voix très aiguë. Deux filtres sont utilisés dont l'un fonctionne à court terme (short term Prediction), reconstitue le rôle des cordes vocales et autres cavités résonnantes humaines. L'autre filtre excite le précédent et reconstitue un mélange d'ondes et de bruit par prédiction à long terme. (Figure B.1)

I.1. Prédiction linéaire par coefficients de réflexion (prédiction à court terme sur 20 ms)

Au lieu de prédire $s[n]$ par combinaison linéaire $\sum lpc[k]s[n-k]$ avec optimisation de p coefficients $lpc[1], lpc[2] \dots lpc[p]$, (p est en général compris entre 8 et 14), on utilise des coefficients de réflexion sur des portions de tube à l'image d'une propagation dans des portions de tube aux parois rigides et sans pertes.

Les parois de la bouche ne sont pas rigides, ne sont pas immobiles, elles peuvent s'ouvrir sur la cavité nasale. Le modèle ne ressemble pas tout à fait au modèle physique. La prédiction à court terme calcule les coefficients et le résidu qui excite le filtre dans le décodeur. Le filtre sur la boucle est utilisé pour la synthèse avec un parcours en sens inverse du modèle des réflexions. Sont extraits de cette partie du traitement les coefficients qui seront transmis sur 36 bits, et le résidu (les impulsions génératrices) qui va lui-même être généré par une prédiction à long terme.

I.2. Prédiction long terme

L'algorithme de prédiction à long terme travaille sur des blocs 40 échantillons de signal résiduel issu de la prédiction à court terme, produisant 56 bits (soit $4 \times 56 = 224$ pour 160 éch.) de la trame GSM. La prédiction à long terme consiste à trouver une ressemblance avec 40 échantillons passés. Le décalage temporel entre les séquences (courante et passée) est calculé par corrélation : la durée fournissant la corrélation maximum pour un son simple est l'inverse de la fréquence fondamentale. Ce décalage (variant entre 40 et 120 = 3 séquences). Ceci ainsi que le gain à appliquer pour retrouver le signal sont les deux paramètres de la prédiction à long terme. Le gain est calculé par division de la corrélation maximale par l'énergie du signal résiduel reconstitué.

Pour soustraire le signal prédit à long terme (les 40 échantillons) de l'entrée, le signal est mis à l'échelle. On espère alors que le signal résiduel soit faible et aléatoire, donc peu coûteux à coder et transmettre. Si la trame courante est une voyelle, la prédiction aura prédit correctement le phonème, et le signal résiduel est faible. Si le son est non voisé, le résidu est aléatoire et ne nécessitera pas de codage précis.

La séquence des 40 résidus est amputée des deux tiers (un échantillon sur trois est conservé) avec premier échantillon choisi de façon que l'énergie soit maximale. Ce choix est transmis grâce à deux bits. Les treize échantillons sont transmis sur trois bits chacun, et un gain général sur 6 bits est transmis.

Enfin, le codeur prépare la prochaine analyse à long terme, en reconstituant un résidu à court terme.

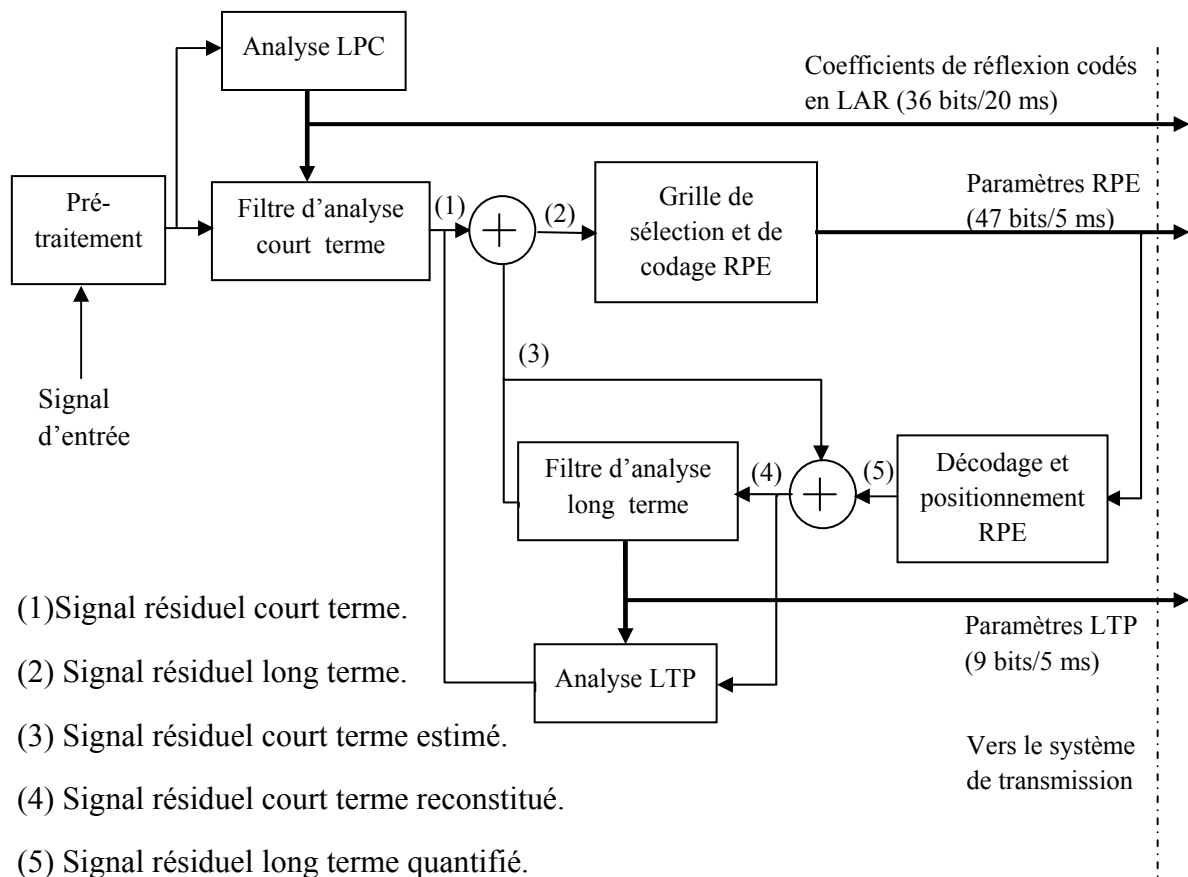


Figure B.1 : schéma simplifié du codeur GSM 6.10 [ETSI,2000]

II. Décodeur

Les 13 échantillons à 3 bits sont multipliés par le facteur d'échelle, est étalé sur 40 échantillons. Ceux-ci sont soumis au filtre de synthèse à long terme. L'algorithme reprend une séquence de 40 valeurs passées reconstituées, la met à l'échelle et l'ajoute

à la séquence de résidus. Le résultat fait partie des séquences conservées pour les trois prochaines prédictions.

En résumé, la trame comprend 4 blocs chacun de $47+9=56$ bits de prédiction à long terme correspondant à 5 msec + 36 bits de réflexions pour 20 msec, soit 260 bits

Enfin le résidu reconstitué est soumis au prédicteur à court terme, passant à travers les filtres à réflexions, et synthétise la parole.

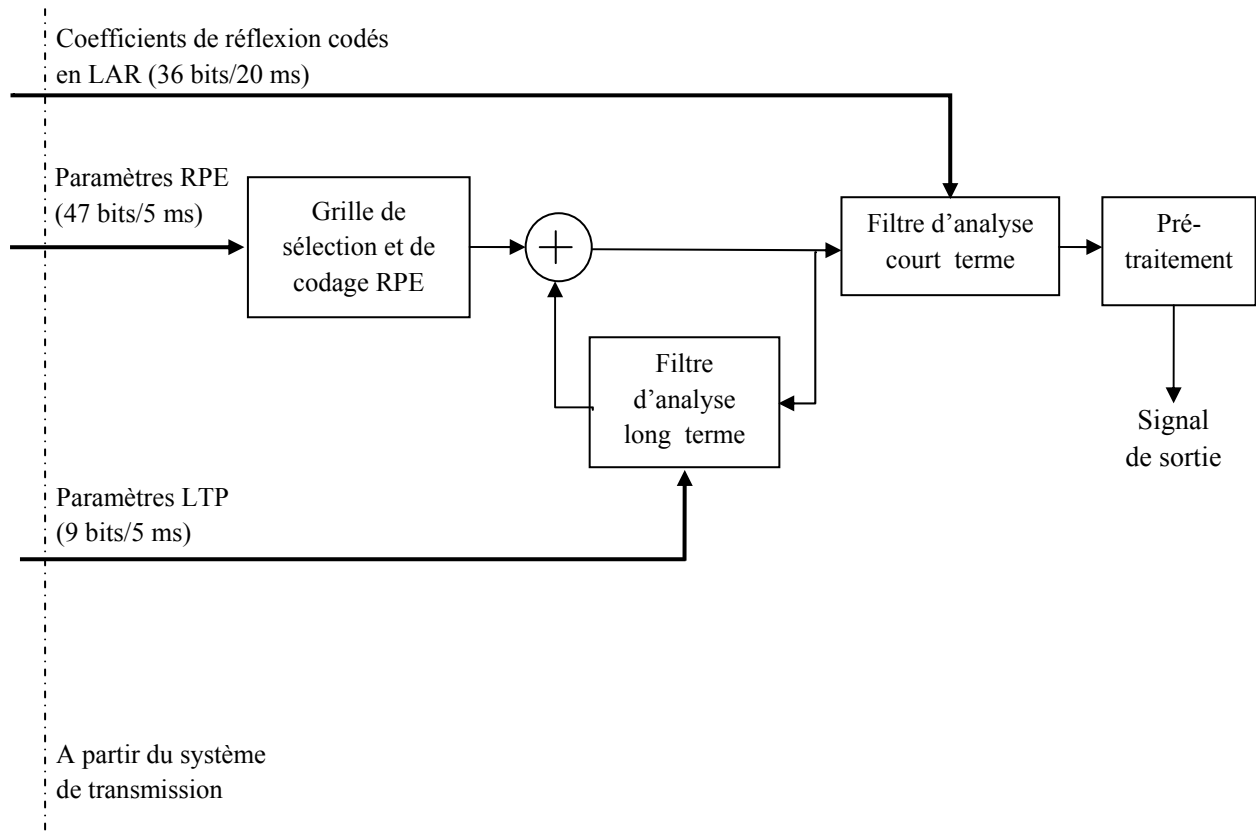


Figure B.2 : schéma simplifié du décodeur. [ETSI,2000]

III. Description fonctionnel du codeur GSM 6.10

III.1 prétraitement

- a.** Compensation d'offset : Avant d'entamer le codage de la parole, on applique un filtrage coupe bande au signal d'entrée afin d'éliminer l'offset.

$$S_{of}(k) = S_0(k) - S_0(k-1) + \alpha \cdot S_{of}(k-1) \quad \text{avec } \alpha = 3275 \cdot 2^{-15} \approx 0,998 \quad (\text{B.1})$$

- b.** Préaccentuation :

$$S(k) = S_{of}(k) - \beta \cdot S_{of}(k-1) \quad \text{avec } \beta = 28180 \cdot 2^{-15} \quad (\text{B.2})$$

III.2 Analyse LPC

- a. *Segmentation* : Le signal de parole $s(k)$ est divisé en trames de longueur $T_0=20\text{ms}$ (sans chevauchement). Une analyse LPC d'ordre ' $p=8$ ' est effectuée pour chaque trame.
- b. *Autocorrélation* : Les premières valeurs ($p+1=9$) de la fonction d'autocorrélation sont calculés par :

$$ACF(k) = \sum_{i=k}^{159} S_{of}(I) \cdot S(i-k) \quad k = 0, 1, \dots, 8 \quad (B.3)$$

- c. *Algorithme récursif de Schur* : Les coefficients de réflexion (RCs) sont calculés (figure B.3) en utilisant l'algorithme de récurrence de Schur. Ces coefficients proviennent de l'analyse LPC du signal de parole où le conduit vocal est modélisé comme un tube non uniforme formé de P tubes cylindriques et uniformes avec plusieurs zones transversales.

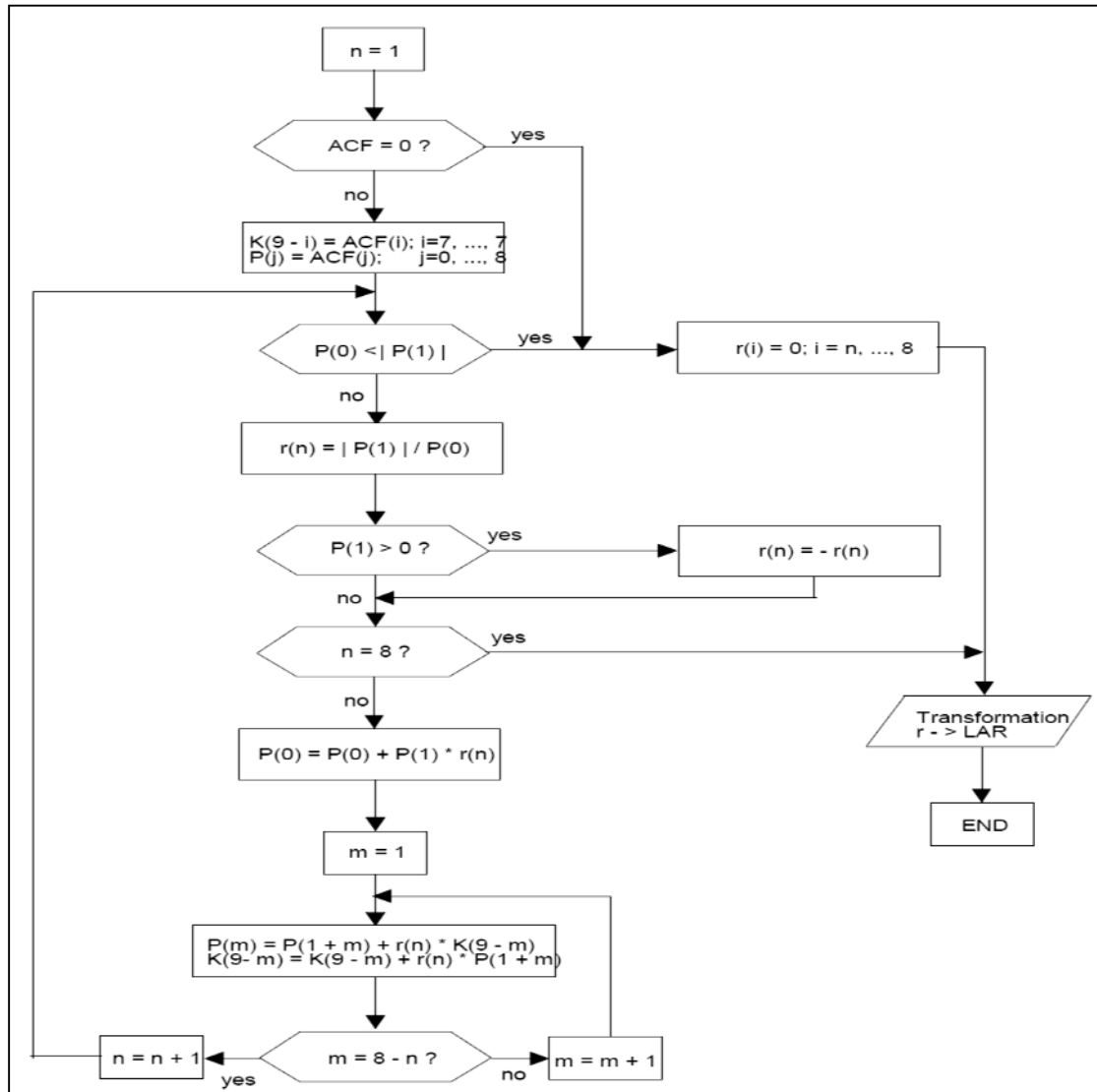


Figure B.1 : Algorithme récursif de Schur.

d. Transformation des RCs en LARs :

Les CR $r(i)$ ($i=1\dots 8$) calculé par l'algorithme de Shur $-1 \leq r(i) \leq +1$. Compte tenu des caractéristiques de quantifications favorables. Les RCs sont converti LARs qui sont défini par :

$$\log_{10}area(i) = \log_{10} \left(\frac{1+r(i)}{1-r(i)} \right) \quad (B.4)$$

Le calcul des LARs est effectué en utilisant l'approximation suivante :

$$LAR(i) = \begin{cases} r(i) & ; |r(i)| < 0,675 \\ Sign[r(i)] \cdot (2|r(i)| - 0,675) & ; 0,675 \leq |r(i)| \leq 0,950 \\ Sign[r(i)] \cdot (8|r(i)| - 6,375) & ; 0,950 \leq |r(i)| \leq 1 \end{cases} \quad (B.5)$$

La transformation inverse est calculée comme suit :

$$r'(i) = \begin{cases} LAR'(i) & ; |LAR'(i)| < 0,675 \\ Sign[LAR'(i)] \cdot (0,500|LAR'(i)| - 0,3375) & ; 0,675 \leq |LAR'(i)| \leq 1,225 \\ Sign[LAR'(i)] \cdot (1,125|LAR'(i)| - 0,796875) & ; 1,225 \leq |LAR'(i)| \leq 1,625 \end{cases} \quad (B.6)$$

e. Quantification et codage des LAR :

Les LAR ont des plage dynamiques et des densités de distribution asymétriques. Pour cette raison, les coefficients LAR sont quantifiés suivant l'équation :

$$LAR_C(k) = N_{int} \{A(i) \cdot LAR(i) + \beta(i)\} \quad (B.7)$$

$$\text{Avec } N_{int}(k) = \text{int}\{z + Sign\{z\} \cdot 0,5\} \quad (B.8)$$

LAR_C : désigne la version quantifiée et codée de LAR.

N_{int} : fonction permet d'arrondir à la plus proche valeur entière par les coefficients $A(i)$ et $B(i)$.

Tableau B.1 : Quantification des LAR

i	$A(i)$	$B(i)$	Minimum $LAR_c(i)$	Maximum $LAR_c(i)$
1	20.000	0.000	-32	+31
2	20.000	0.000	-32	+31
3	20.000	4.000	-16	+15
4	20.000	-5.000	-16	+15
5	13.637	0.184	-8	+7
6	15.000	-3.500	-8	+7
7	8.334	-0.666	-4	+3
8	8.824	-2.235	-4	+3

III.3 Filtrage par analyse court terme :

La trame du signal de parole est conservée en mémoire jusqu'à ce que le calcul des paramètres LPC ($LAR(i)$) est terminé.

La trame est ensuite filtrée par un filtre d'analyse à court terme ($P=8$). Cependant avant l'opération de filtrage, les coefficients sont décodés et prétraités par interpolation.

a. Décodage des LAR quantifiés :

Dans cette partie les LARs quantifiés sont décodés

$$LAR''(i) = [LAR_c(i) - \beta(i)] / A(i) \quad (B.9)$$

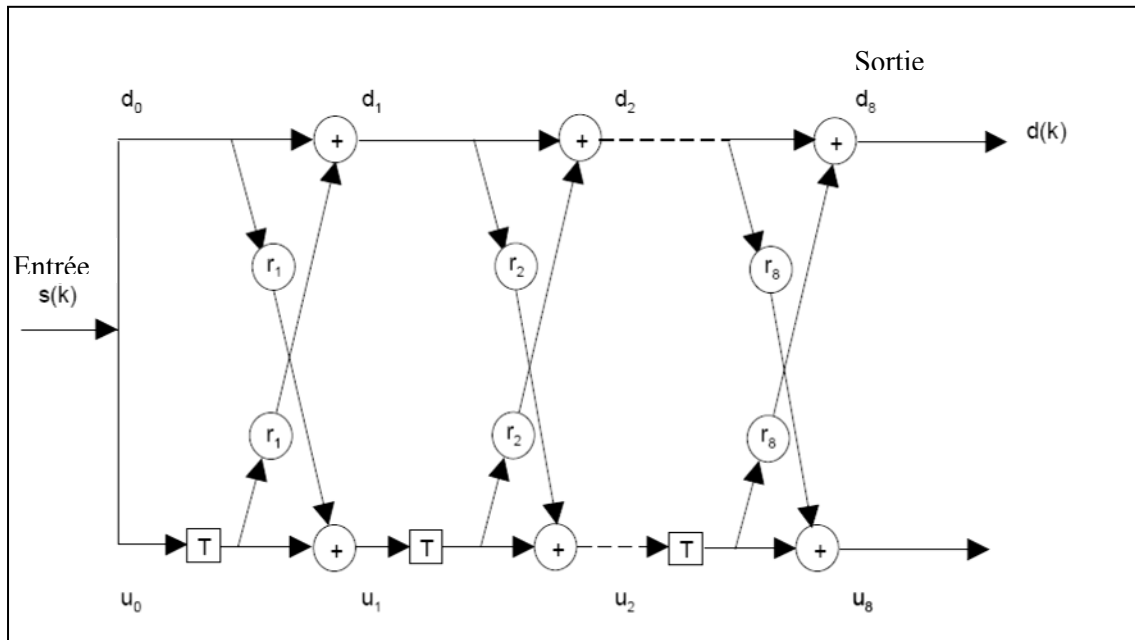
b. Interpolation des LAR :

Pour éviter les parasites transitoires qui peuvent se produire lors du changement brusque des coefficients, deux ensembles successifs sont interpolés linéairement.

Tableau B.2 : Interpolation des LAR (j : indique le segment actuel)

k	$LAR'_j(i)$
$0 \dots 12$	$0.75 \times LAR''_{j-1}(i) + 0.25 \times LAR''_j(i)$
$13 \dots 26$	$0.50 \times LAR''_{j-1}(i) + 0.50 \times LAR''_j(i)$
$27 \dots 39$	$0.25 \times LAR''_{j-1}(i) + 0.75 \times LAR''_j(i)$
$40 \dots 159$	$LAR''_j(i)$

- c. Transformation des LARs en RCs : Les RCs sont déterminés par la transformation inverse (équation B.6)
- d. Filtrage par analyse court terme : Le filtrage par analyse court terme est implémenté en utilisant le trellis de la figure (B.4).

**Figure B.4 :** Filtre d'analyse court terme.

$$d_0(k) = s(k) \quad (B.10)$$

$$u_0(k) = s(k) \quad (B.11)$$

$$d_i(k) = d_{i-1}(k) + r'_i u_{i-1}(k-1) \quad i = 1, \dots, 8 \quad (B.12)$$

$$u_i(k) = u_{i-1}(k-1) + r'_i d_{i-1}(k) \quad i = 1, \dots, 8 \quad (B.13)$$

$$d(k) = d_8(k) \quad (B.14)$$

III.4 Analyse par prédiction long terme (LTP)

a. Sous-segmentation :

Chaque trame du signal résiduel court terme contient 160 échantillons, qui correspondent à 20 ms.

La corrélation long terme est évaluée 4 fois pour chaque trame donc pour un segment de 5 ms. On note $j=0, \dots, 3$ l'indice de la sous trame tel que les échantillons de la $j^{\text{ème}}$ sous trame du signal résiduel est noté par : avec $d(k_j + k)$ avec $j=0, \dots, 3$; $k_j = k_0 + j \times 40$ et $k=0, \dots, 39$. où k_0 correspond à la 1^{ère} valeur de l'échantillon de la trame courante.

b. Calcul des paramètres LTP :

Pour chaque sous-trame on détermine le décalage de la corrélation long terme N_j ($j=0, \dots, 3$) et le gain b_j ($j=0, \dots, 3$), l'implémentation est effectuée en trois étapes :

1°) La 1^{ère} étape concerne l'évaluation de la corrélation $R_j(\lambda)$ du signal résiduel court terme (de la sous trame courante) et du signal résiduel court terme reconstitué à partir des échantillons précédents $d'(k_j+i)$, $i=-120, \dots, -1$.

$$R_j(\lambda) = \sum_{i=0}^{39} d(k_j + i).d'(k_j + i - \lambda) \begin{cases} j = 0, \dots, 3 \\ k_j = k_0 + j.40 \\ \lambda = 40, \dots, 120 \end{cases} \quad (\text{B.15})$$

La corrélation est évaluée pour un décalage $\lambda \geq 40$ et $\lambda \leq 120$.

2°) La seconde étape est de trouver la position du N_j qui correspond au maximum de la fonction de corrélation.

$$R_j(N_j) = \text{MAX} \{ R_j(\lambda), \lambda = 40, \dots, 120 \} \quad \text{avec } j = 0, \dots, 3 \quad (\text{B.16})$$

3°) La 3^{ème} étape est le calcul du gain b_j par :

$$b_j = R_j(N_j) / S_j(N_j) \quad j = 0, \dots, 3 \quad (\text{B.17})$$

$$\text{Avec } S_j(N_j) = \sum_{i=0}^{38} d'^2(k_j + i - N_j) \quad j = 0, \dots, 3 \quad (\text{B.18})$$

Il est évident que les 120 échantillons du signal résiduel court terme reconstitué $d'(k_j+i)$, $i=-120,\dots,-1$ doivent être conservés jusqu'à la prochaine sous trame pour permettre l'évaluation des relations (B.15) et (B.18).

c. Codage /décodage du décalage LTP :

Les N_j peuvent avoir les valeurs d'intervalle [40, 120] et seront codé en utilisant 7 bits.

$$N'_{cj} = N_j \quad ; \quad j=0,\dots,3 \quad (B.19)$$

A la réception et en supposant que la transmission est achevée sans perte d'information. Le décodage de ces valeurs permet de restaurer la position réelle :

$$N'_j = N_{cj} \quad ; \quad j=0,\dots,3 \quad (B.20)$$

d. Codage/décodage du gain LTP

Le gain b_j ($j=0,\dots,3$) de prédiction long terme est codé en utilisant 2 bits par l'algorithme suivant :

$$\text{Si } b_j \leq DLB(i) \quad \text{alors } b_{cj} = 0 \quad ; i=0 \quad (B.21)$$

$$\text{Si } DLB(i-1) < b_j \leq DLB(i) \quad \text{alors } b_{cj} = i \quad ; i=1,20 \quad (B.22)$$

$$\text{Si } DLB(i-1) < b_j \quad \text{alors } b_{cj} = 3 \quad ; i=30 \quad (B.23)$$

Où $DLB(i)$, $i=0,\dots,3$ désigne les niveaux de décision du quantificateur, et b_{cj} est la valeur du gain codé les niveaux de décision de quantification sont donnés par le tableau B.3

Tableau B.3 : Tableau de quantification du gain LTP

i	Niveau de décision DLB(i)	Niveau de quantification QLB(i)
0	0.2	0.10
1	0.5	0.35
2	0.8	0.65
3		1.00

La règle de décodage est mise en œuvre par :

$$b'_j = QLB(b_{c_j}) ; \quad j=0,...,30 \quad (B.24)$$

QLB est le niveau de quantification et b'_j est la valeur décodée du gain.

- e. *Filtrage par analyse long terme* : Le signal résiduel court terme $d(k_0+k)$, $k=0,...,159$ est traité par sous trame de 40 échantillons. A partir de chaque sous trame, on calcule le signal résiduel long terme.

$$e(k_j + k) = d(k_j + k) - d''(k_j + k) \left\{ \begin{array}{l} j = 0, \dots, 3 \\ k = 0, \dots, 39 \\ k_j = k_0 + j40 \end{array} \right\} \quad (B.25)$$

Où $d''(k_j+k)$, $k=0,...,39$ est calculé à partir des échantillons du signal résiduel court terme reconstitué ultérieurement

$$d''(k_j + k) = b'_j . d'_j(k_j + k - N_j) \left\{ \begin{array}{l} j = 0, \dots, 3 \\ k = 0, \dots, 39 \\ k_j = k_0 + 40j \end{array} \right\} \quad (B.26)$$

- f. *Filtrage par synthèse long terme*

Le signal résiduel long terme $d(k_0+k)$, $k=0,...,159$ est traité par sous trame de 40 échantillons. A partir de chaque sous trame

$$d'(k_j + k) = e'((k_j + k) + d''(k_j + k) \left\{ \begin{array}{l} j = 0, \dots, 3 \\ k = 0, \dots, 39 \\ k_j = k_0 + 40j \end{array} \right\} \quad (B.27)$$

III.5 Codage RPE

- a. Filtre de pondération : Un filtre RIF est appliqué à chaque sous trame de 5 ms par convolution des 40 échantillons $e(k)$ avec la réponse impulsionnelle $H(i)$ $i=0,...,10$ tableau (3.4)

Tableau B.4 : Réponse impulsionnelle du filtre de pondération

i	5	4(6)	3(7)	2(8)	1(9)	0(10)
$H(i) \times 2^{13}$	8192	5741	2054	0	-374	-134

$$H(\Omega = 0) = 2,779 \quad (B.28)$$

La convulsion classique d'une séquence de 40 échantillons en utilisant 11 prise de la réponse impulsionnelle constitue $40+11-1=50$ échantillons. En revanche l'algorithme

de filtrage utilisé ici constitue les 40 échantillons centraux de l'opération de convolution classique. Pour simplifier la notation chaque sous-segment est représenté par $x(k)$, $k = 0, \dots, 39$.

$$x(k) = \sum_{i=0}^{10} H(i) \times e(k+5-i) \quad \text{avec } k = 0, \dots, 39. \quad (\text{B.29})$$

Note : $e(k+5-i) = 0$ pour $k+5-i < 0$ et $k+5-i > 39$.

b. Décimation adaptatif du taux d'échantillonnage par la grille de sélection RPE :

Le signal filtré x est sous échantillonné par :

$$x_m(i) = x(k_j + m + 3 \times i) ; \quad i=0, \dots, 12 \text{ et } m=0, \dots, 3 \quad (\text{B.30})$$

m est la position de la grille de décimation.

Selon la solution explicite du critère de l'erreur quadratique moyenne, la sous séquence optimum x_m est sélectionnée en maximisant l'énergie :

$$E_M = \max_m \sum_{i=0}^{12} x_m(i) ; \quad m=0, \dots, 3 \quad (\text{B.31})$$

La position optimum de la grille est codé en tant que M_c par 2bits.

c. Quantification APCM de la séquence RPE :

La sous séquence $x_m(i)$ (séquence RPE) est quantifiée par la méthode APCM (Adaptive Pulse Code Modulation). Pour chaque séquence RPE constituée d'un ensemble de 13 échantillons $x_m(i)$, le maximum de la valeur absolue $|x_m(i)|$ est sélectionné, ensuite quantifié par une fonction logarithmique utilisant 6 bits (tableau B.5).

Tableau B.5 : Quantification des maximum de la séquence RPE

xmax	x'max	xmaxc	xmax	x'max	xmaxc
0 .. 31	31	0	2048 .. 2303	2303	32
32 .. 63	63	1	2304 .. 2559	2559	33
64 .. 95	95	2	2560 .. 2815	2815	34
96 .. 127	127	3	2816 .. 3071	3071	35
128 .. 159	159	4	3072 .. 3327	3327	36
160 .. 191	191	5	3328 .. 3583	3583	37
192 .. 223	223	6	3584 .. 3839	3839	38
224 .. 255	255	7	3840 .. 4095	4095	39
256 .. 287	287	8	4096 .. 4607	4607	40
288 .. 319	319	9	4608 .. 5119	5119	41
320 .. 351	351	10	5120 .. 5631	5631	42
352 .. 383	383	11	5632 .. 6143	6143	43
384 .. 415	415	12	6144 .. 6655	6655	44
416 .. 447	447	13	6656 .. 7167	7167	45
448 .. 479	479	14	7168 .. 7679	7679	46
480 .. 511	511	15	7680 .. 8191	8191	47
512 .. 575	575	16	8192 .. 9215	9215	48
576 .. 639	639	17	9216 .. 10239	10239	49
640 .. 703	703	18	10240 .. 11263	11263	50
704 .. 767	767	19	11264 .. 12287	12287	51
768 .. 831	831	20	12288 .. 13311	13311	52
832 .. 895	895	21	13312 .. 14335	14335	53
896 .. 959	959	22	14336 .. 15359	15359	54
960 .. 1023	1023	23	15360 .. 16383	16383	55
1024 .. 1151	1151	24	16384 .. 18431	18431	56
1152 .. 1279	1279	25	18432 .. 20479	20479	57
1280 .. 1407	1407	26	20480 .. 22527	22527	58
1408 .. 1535	1535	27	22528 .. 24575	24575	59
1536 .. 1663	1663	28	24576 .. 26623	26623	60
1664 .. 1791	1791	29	26624 .. 28671	28671	61
1792 .. 1919	1919	30	28672 .. 30719	30719	62
1920 .. 2047	2047	31	30720 .. 32767	32767	63

Finalement, la normalisation est effectuée en divisant les 13 échantillons par x'_{max}

$$x'(i) = \frac{x_M(i)}{x'_{max}} ; \quad i=0,\dots,1 \quad (\text{B.32})$$

$x'(i)$ sont quantifiés par 3 bits (tableau B.6)

Tableau B.6 : Quantification des échantillons RPE normalisés

$x' \cdot 2^{15}$ (Interval-limits)	$\underline{x}_M \cdot 2^{15}$	x_{Mc} (Channel)
-32768 ... -24577	-28672	0 = 000
-24576 ... -16385	-20480	1 = 001
-16384 ... -8193	-12288	2 = 010
-8192 ... -1	-4096	3 = 011
0 ... 8191	4096	4 = 100
8192 ... 16383	12288	5 = 101
16384 ... 24575	20480	6 = 110
24576 ... 32767	28672	7 = 111

- d. *Quantification inverse APCM* : Les $x_{Mc}(i)$ sont décodés en $x'_M(i)$ et normalisés par les valeurs décodées x'_{maxC} donnant ainsi la sous séquence décodée $x'_M(i)$.

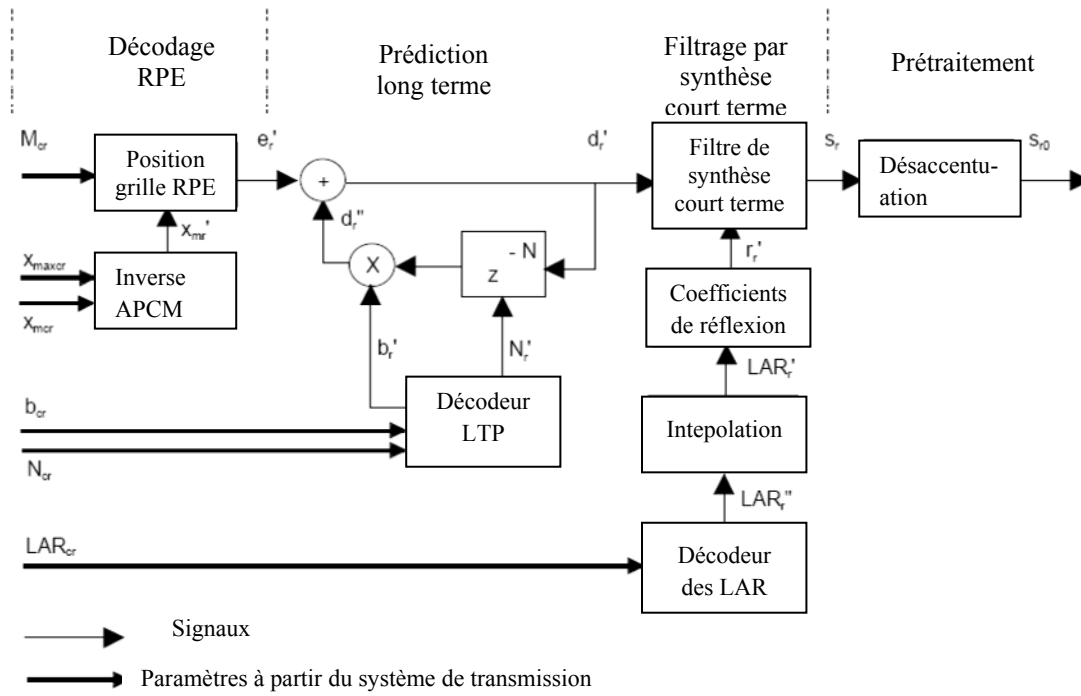
- e. *Positionnement de la grille RPE* : La sous séquence quantifiée est sur échantillonnée par un rapport de 3 en ajoutant des zéros en fonction de la position de grille M_c .

III.5 Décodeur

La plupart des blocs du décodeur ont déjà été décrit dans la partie précédente. Les seuls blocs ajoutés dans le décodeur sont :

- Le filtre de synthèse court terme
- Le filtre de désaccentuation

Le schéma complet du décodeur est illustré dans la figure 3.4, les variables et les paramètres du décodeur se caractérisent par l'indice « r » pour désigner les valeurs reçues des valeurs du codeur.



- a. *Filtrage par synthèse à court terme* : Les coefficients du filtre de synthèse à court terme (voir figure 3.5) sont reconstitués en appliquant la procédure identique à celle du codeur. Le filtre de synthèse à court terme est mis en œuvre en fonction de la structure en treillis représenté dans la figure 3.5.

$$s_{r(0)}(k) = d'_r(k) \quad (\text{B.33})$$

$$s_{r(i)}(k) = s_{r(i-1)}(k) + r'_{r(9-i)} \times v_{8-i}(k-1) \quad i = 0, \dots, 8 \quad (\text{B.34})$$

$$v_{9-i}(k) = v_{8-i}(k-1) + s_{r(i)}(k) \times r'_{r(9-i)} \quad i = 0, \dots, 8 \quad (\text{B.35})$$

$$s'_r(k) = s_{r(8)}(k) \quad (\text{B.36})$$

$$v_0(k) = s_{r(8)}(k) \quad (\text{B.37})$$

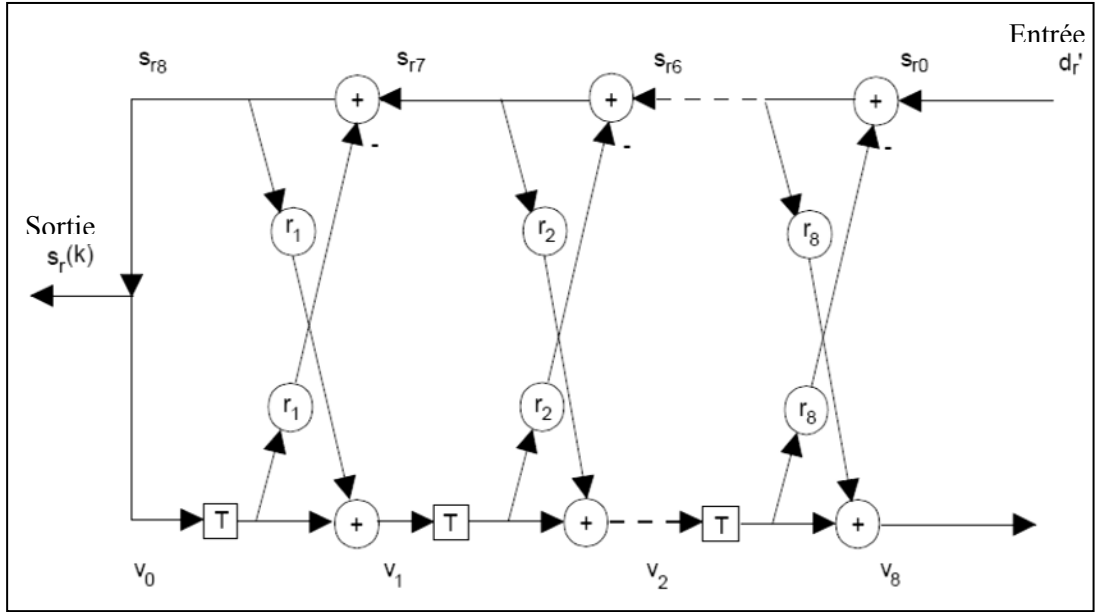


Figure B.5 : Filtre de synthèse court terme.

- b. Prétraitement :** La sortie du filtre de synthèse $s(k)$ est introduite dans le filtre IIR de désaccentuation conduisant au signal de sortie.

$$s_{r0}(k) = s_r(k) + \beta \times s_{r0}(k-1) \quad (\text{B.38})$$

Avec $\beta = 28180 \times 2^{-15}$

ANNEXE C

Annexe C : Système de RAL sous HTK

La plate-forme HTK (*Hidden Markov Model Toolkit* ou "boîte à outils de modèles de Markov cachés") a été développée à l'Université de Cambridge. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole et du locuteur continue à base de HMM [Young, Jansen *et al.*, 1995].

Les modèles peuvent représenter des mots ou des locuteurs (selon l'application), et leur topologie peut être librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, peuvent être discrètes ou continues (GMM). Les modèles sont initialisés avec l'algorithme de Viterbi, puis réestimés par l'algorithme optimal de Baum-Welch. Le décodage est réalisé par l'algorithme de Viterbi, sous la contrainte d'un réseau syntaxique défini par l'utilisateur, et le résultat est enfin évalué par alignement dynamique avec la chaîne phonétique ou lexicale de référence.

L'ensemble de ces outils est écrit en langage C, et la documentation détaille leur utilisation et les principes de leur implémentation, ce qui permet d'intégrer de manière efficace les modifications souhaitées dans le système de reconnaissance.

Dans cette annexe nous présentons les étapes de conception d'un système de reconnaissance de locuteur en utilisant l'outil HTK. Les étapes de la reconnaissance d'un locuteur par HTK sont :

C.1. Représentation acoustique du signal

Dans cette phase, une paramétrisation des fichiers son de la base de données d'apprentissage est effectuée (Figure A.1). Dans notre cas nous avons utilisé des fichiers .wav de la base de données ARADIGIT.

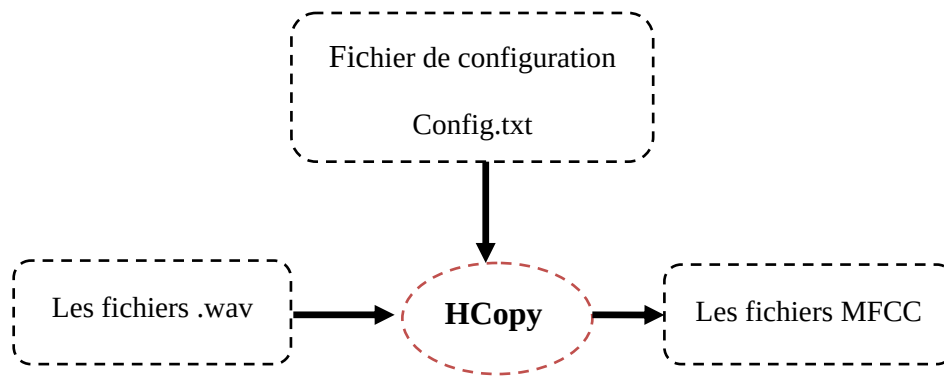


Figure C.1 : Représentation acoustique du signal.

On commence par définir le modèle de la langue, appelé aussi grammaire (Figure C.2), ensuite, on construit le dictionnaire des locuteurs (figure C.3).

```

/* gramamaire*/

$WORD = LOC001 | LOC002 | LOC003 | LOC004 | LOC005 | LOC006 | LOC007 |
LOC008 | LOC009 | LOC010 | LOC011 | LOC012 | LOC013 | LOC014 | LOC015 |
LOC016 | LOC017 | LOC018 | LOC019 | LOC020 | LOC021 | LOC022 | LOC023 |
LOC024 | LOC025 | LOC026 | LOC027 | LOC028 | LOC029 | LOC030| locuteur001 |
locuteur002 | locuteur003 | locuteur004 | locuteur005 | locuteur006 | locuteur007 |
locuteur008 | locuteur009 | locuteur010 | locuteur011 | locuteur012 | locuteur013 |
locuteur014 | locuteur015 | locuteur016 | locuteur017 | locuteur018 | locuteur019 |
locuteur020 | locuteur021 | locuteur022 | locuteur023 | locuteur024 | locuteur025 |
locuteur026 | locuteur027 | locuteur028 | locuteur029 | locuteur030;

($WORD)
  
```

Figure C.2 : Grammaire des locuteurs de la base ARADIGIT.

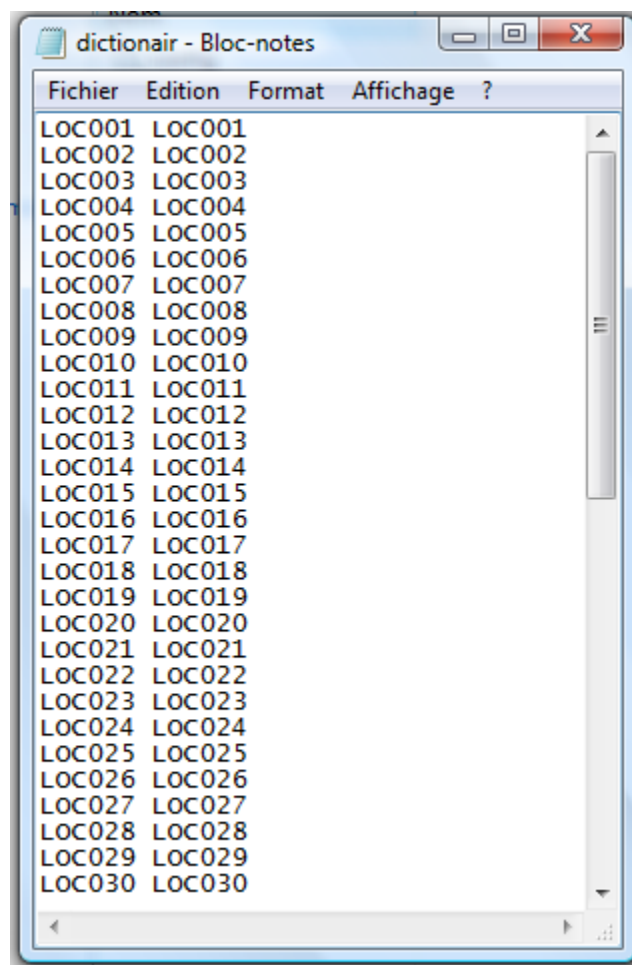


Figure C.3 : Dictionnaire des locuteurs de la base ARADIGIT.

Les coefficients MFCC sont extraits des fichiers wav et sur des fenêtres de 25ms grâce à l'outil **Hcopy** en se servant du fichier de configuration **config** comme paramètre d'entrée .

Le fichier de configuration **config** (Figure C.4) peut définir les paramètres indispensables pour la phase d'analyse acoustique. Le choix s'est porté sur les 12 premiers coefficients MFCC. Pour chaque coefficient, on attribue une dérivée première (12 dérivées premières au total) ainsi qu'une dérivée seconde (12 dérivées secondes) pour prendre en compte la dynamique du signal. En somme on obtient un vecteur acoustique de 36 coefficients correspondant à chaque trame du signal.

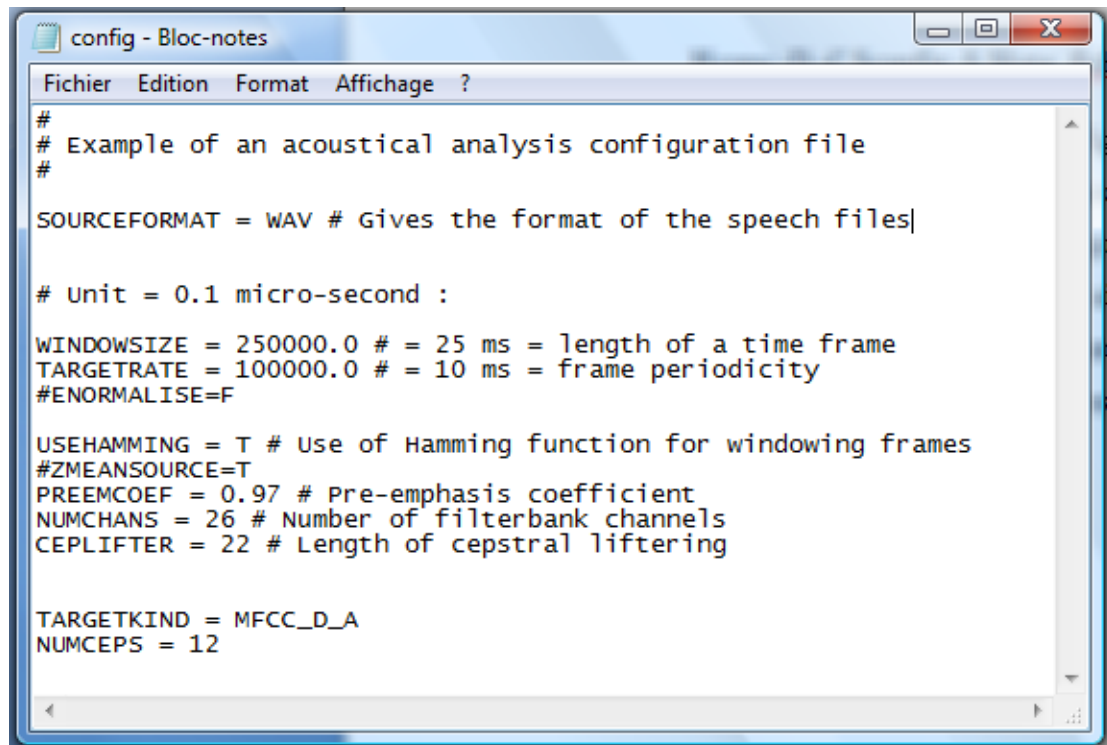


Figure C.4 : Fichier de configuration pour la phase de l'analyse acoustique.

Le résultat de cette phase est un ensemble de fichiers .mfcc qui constituent une représentation vectorielle pour modéliser les locuteurs.

Commande HTK sous perl : le calcul des paramètres MFCCs peut être effectué avec la commande perl 'system' comme suit :

```
system("HCOPY -C $fconfig_mfcc $fichier_liste_chemins ");
```

tel que :

- **\$fich_config_mfcc :** est le fichier de configuration des paramètres acoustiques.
- **\$list_liste_chemins :** est le fichier qui contient la liste des fichiers parole et des coefficients MFCC.

C.2 Apprentissage

L'apprentissage des modèles revient à faire apprendre à partir des données acoustiques déjà préparées, le vecteur moyen, la matrice de covariance et la matrice de probabilité de transition entre états, pour chaque modèle HMM.

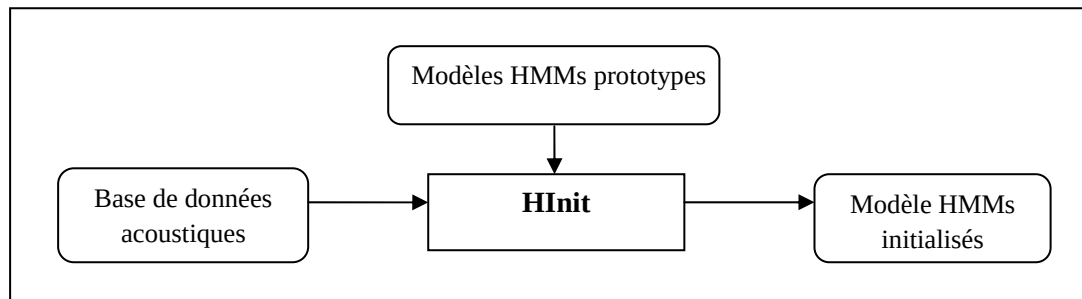


Figure C.6 : Initialisation des modèles prototypes par l'outil HInit.

Commande HTK sous perl

```
system("HInit -S fichiers_acoustiques.txt prototype");
```

Avec:

-fichiers_acoustiques.txt : est un fichier texte contient la liste des fichiers MFCCs de la base d'apprentissage.

-prototype : est le fichier prototype définie manuellement dans l'étape précédente.

C.2.3 Entraînement des modèles HMMs

Après l'initialisation des modèles acoustiques, l'étape suivante consiste à les entrainer en utilisant la base acoustique et la base phonétique.

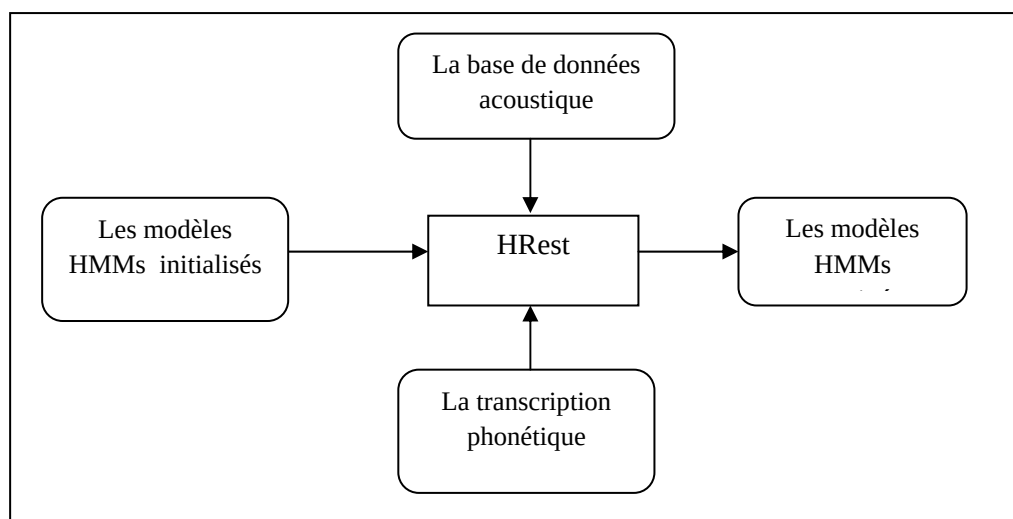


Figure C.7: Entraînement des modèles HMMs par l'outil HRest.

Commande HTK sous perl

```
system("$HRest -M $rep_hmms_entraines -S $liste_donnee_apprentissage  
$rep_hmms_initiales\\$_");
```

Avec:

`liste_donnee_apprentissage` : c'est une liste qui contient tous les noms de modèles HMMs qui vont être entraînés lors de cette étape.

`rep_hmms_inities`: est le répertoire contenant l'ensemble des modèles HMMs déjà initialisés dans l'étape précédente.

C.3 Reconnaissance

La reconnaissance se fait avec l'algorithme de Viterbi, implémenté sous HTK avec l'outil **HVite**, cet outil nécessite l'utilisation de différentes connaissances : les modèles HMMs (déjà entraînés), le réseau syntaxique des différents chemins à suivre, le modèle du langage et le dictionnaire.

La commande **HVite** s'implémente de la façon suivante :

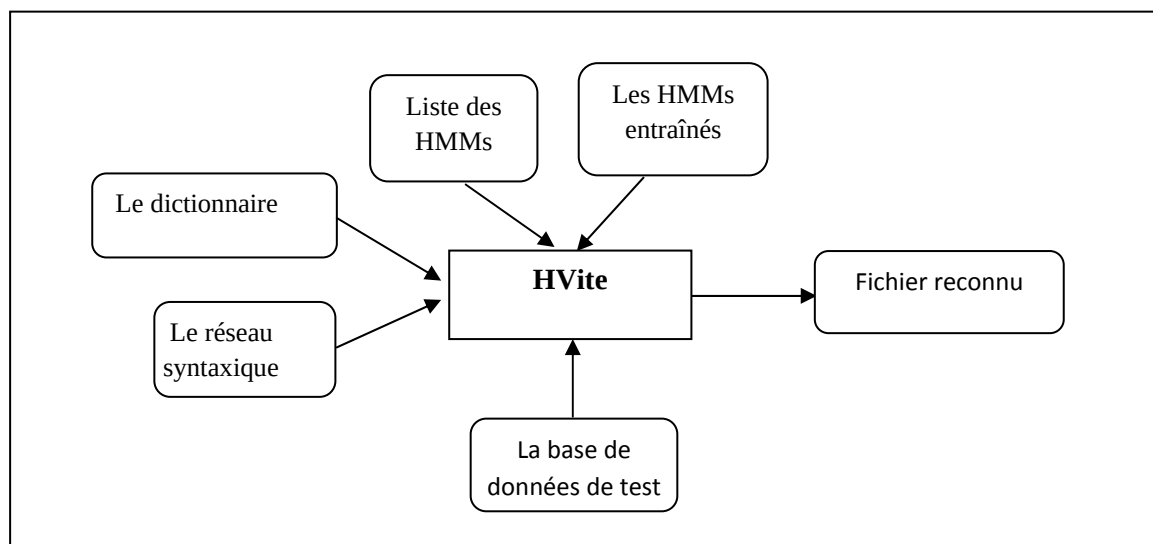


Figure C.8 : Le reconnaissance par l'outil HVite.

Commande HTK sous perl

```
system("$HVite -H hmms_entraînés -S list_test.txt -i fichiers_reconnus
-w réseau_syntaxique dictionnaire liste_hmms) ;
```

Avec :

-fichiers_reconnus : le fichier résultats qui contient le décodage des fichiers de test.

C.4. Evaluation des résultats

L'évaluation des résultats revient à comparer les résultats obtenus après l'étape de décodage avec des résultats de référence. L'outil **HResults** permet de réaliser cette

comparaison par alignement dynamique entre le décodage obtenu et la source de référence.

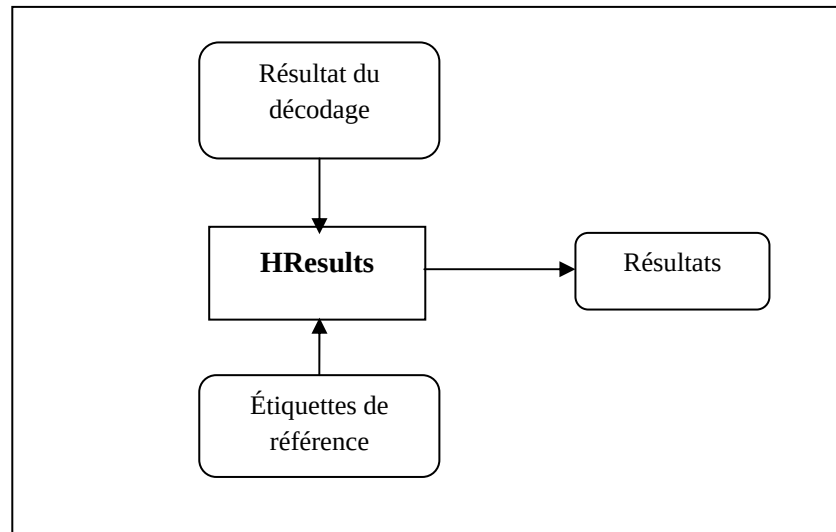


Figure 19: Evaluation des résultats par l’outil HResults.

Commande HTK sous perl

```
HRESULTS -p -I $fich_ref $liste_hmm $reco> $resultats
```