République Algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de
la recherche scientifique

Centre universitaire Cheikh El-Arbi Tébessi
de Tébessa

Institut des sciences exactes et de technologie

Département de l'électronique

MEMOIRE

Présenté en vue de l'obtention du diplôme de **MAGISTER**

Le 27/ 06 / 2007

DÉVELOPPEMENT D'UN SYSTÈME DE RECONNAISSANCE ROBUSTE DE LA PAROLE

Option :

Contrôle et automatique

Par
Triki Abdallah

RAPPORTEUR :	Mr	M. BEDDA	Professeur	U. ANNABA

Devant le jury

PRÉSIDENT :	Mr	N. GUERFI	Maître de conférence	C.U.TEBESSA
EXAMINATEURS :	Mr	N. DOGHMANE	Professeur	U. ANNABA
	Mr	M.MAAMRI	PHD	C.U.TEBESSA
INVITÉ	Mr	A.GATTAL	C.C	C.U.TEBESSA

Remerciements

Je voudrais tout d'abord exprimer ma profonde reconnaissance au Professeur **Mouldi BEDDA**, mon Directeur de mémoire, qui a dirigé mon travail; Ses conseils et ses commentaires précieux m'ont permis de surmonter mes difficultés et de progresser dans mes études. Qu'il soit remercié pour ses qualités humaines et scientifiques.

Je tiens à remercier tous les membres du jury qui ont bien voulu consacrer une partie de leur temps précieux à examiner ce travail

Je remercie sincèrement, le Maître de Conférence **N. GUERFI** pour l'honneur qu'il me fait en présidant ce jury de mémoire. Je voudrais également remercier les membres examinateurs : le Docteur **M. MAAMRI,** le Professeur **N.DOGHMANE**

J'exprime aussi mes remerciements distingués à **messieurs A.GATTAL et M.A.CORBA**, qui par leurs expériences, m'a beaucoup aidé tout au long de ce mémoire, un gros merci pour tout mes collègues pour leur soutien moral.

Enfin, plus que tous, je voudrais exprimer mes plus profonds remerciements à mes amis et à ma famille, pour leurs encouragements, leur patience et leur amour... Ils sont à l'origine de tout ce que j'ai accompli de bien.

TABLE DES MATIÈRES

Table des matières

Chapitre 5 Etude expérimentale

Liste des figures

Liste des tableaux

Liste des abréviations

RAP :	Reconnaissance Automatique de la Parole
LPC :	Linear Predictive Coding
AR :	Auto-Régressif
EQM :	Erreur Quadratique Moyenne
LPCC :	linear prediction cepstral coefficients
MFCC :	Mel Frequency Cepstral Coefficients
FFT :	Fast Fourier Transform
DCT :	. Discrete Cosinus Transform
DFT :	Discrete Fourier Transform
PLP :	PLP (Perceptual Linear Predictive)
RASTA :	RelActive SpecTrAl processing
DWT :	Discrete Wavelet Transform
WP :	Wavelet Packet
AWP :	Admissible Wavelet Packet
HTK :	Hidden Markov Toolkits
MATLAB :	MAtrix LABoratory
HMM :	Hidden Markov Model
ML :	Maximum Likelihood
MAP :	Maximum A Posteriori
BA :	Base d'Apprentissage
BT :	Base de Test
BACC :	Bark Auditory Cepstral Coefficients
QV :	Quantification Vectorielle
DAP :	Décodage Acoustico-Phonétique
IEEE :	Institute of Electrical and Electronics Engineers

INTRODUCTION GÉNÉRALE

Introduction générale

La reconnaissance automatique de la parole a pour but de permettre à un utilisateur de s'adresser oralement à une machine pour des tâches divers : commande, traduction, dictée ….
La recherche dans le domaine de la communication parlée, notamment celle axée sur le développement de nouvelles technologies de l'information, est une activité en pleine expansion, dont plusieurs disciplines et compétences interagissent dans le but d'améliorer les performances des systèmes de Communication Homme-Machine.

La reconnaissance automatique de la parole (RAP) fait partie intégrante de cette discipline et représente l'un des thèmes essentiels de la communication parlée. Plusieurs projets visent à intégrer la parole dans des interfaces Homme-machine en vue d'aboutir à des systèmes sophistiqués capables de simuler le comportement humain à tous les niveaux.
Grâce aux efforts de recherches menés dans différents domaines, de nombreux systèmes de reconnaissance de la parole sont actuellement proposés. Malgré l'importance des efforts fournis, les performances se dégradent sensiblement lorsque les conditions deviennent réelles comme lors du traitement de la parole spontanée.

Aujourd'hui, l'impact des systèmes de RAP est encore minime dans la vie courante et la commande des ordinateurs ne s'effectue toujours pas par la voix, malgré les promesses de fabricants de logiciel ou de matériel informatique (Microsoft, Apple).
La plupart des applications en reconnaissance de la parole peuvent être regroupées en 4 catégories :

- Commande et contrôle: Contrôler à l'aide de la parole des équipements particuliers (machines, robots….) ou des programmes (ouvrir par exemple des fenêtres ou naviguer sous Windows, aide aux personnes handicapées).
- Accès à des bases de données ou recherches d'informations: Compositions automatiques de numéros de téléphones, serveurs vocaux, réservation d'un vol, guidage automatique (dans une voiture), remplir un questionnaire.
- Dictée vocale : Création de lettres, rapports et autres documents par l'intermédiaire de la parole.

- Transcription automatique de la parole : Indexation de programmes télévision ou radio, sous-titrage et traduction automatiques.

Notre étude s'intègre dans le cadre du développement d'un système de reconnaissance de mots arabes isolés multi-locuteur. Dans des applications réelles, l'analyse acoustique par les méthodes les plus performantes de l'état de l'art reste insuffisante; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer la qualité de l'analyse acoustique, on utilisant des techniques de paramétrisation basées sur des transformées en ondelettes. La paramétrisation MFCC est la plus largement répandue dans les systèmes de reconnaissance de la parole, il est difficile de les concurrencer. En général, quand une nouvelle paramétrisation est proposée, les taux de reconnaissance peuvent être améliorés par rapport aux MFCC, mais seulement dans des conditions de parole bruitée.

Motivé par le désir d'établir des modèles temps-fréquence de la parole, nous avons étudié s'il était possible de concevoir des paramétrisations de la parole qui ont les avantages des MFCC mais pas leurs inconvénients.

Plan du mémoire

Ce mémoire comporte cinq chapitres :

Le premier chapitre donne un aperçu sur la parole, et ses caractéristiques. Il présente par la suite les modes de fonctionnement des systèmes de reconnaissance automatique de la parole et décrit les méthodes de la reconnaissance. Ce chapitre sera l'occasion de présenter les modèles de production et de la perception de la parole.

Le deuxième chapitre présente un état de l'art des méthodes d'analyse du signal utilisées en reconnaissance de la parole.

Le troisième chapitre présente un rappel sur la transformée en ondelettes, ainsi que les méthodes proposées pour l'extraction de paramètres.

Le quatrième chapitre présente les modèles de Markov cachés utilisés en reconnaissance de la parole. Ce type de modélisation Markovienne conduit à résoudre trois types de problèmes, le problème d'évaluation des probabilités d'observation étant donné un modèle, le calcul du chemin optimal et finalement l'estimation du modèle. Le chapitre présente aussi un système de reconnaissance de mots arabes isolés développé sur la base de l'algorithme de Baum-Welch et L'algorithme de Viterbi. A la fin de ce chapitre une description de la plate forme HTK(Hidden Markov Tool Kits), on présentant ces outils et sa structure.

Le cinquième chapitre présente les réalisations, tests et validations de système présenté.

Conclusion générale du travail effectué et les perspectives espérées.

CHAPITRE 1 :

RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Chapitre 1 :

Reconnaissance Automatique de la Parole

1.1 Introduction :

Le problème de la reconnaissance automatique de la parole consiste à extraire l'information lexicale contenue dans un signal parole. La conception d'un système de reconnaissance automatique de la parole est rendue difficile par la complexité du signal parole et la tâche de reconnaissance envisagée.

Depuis plus de deux décennies, des recherches intensives dans ce domaine ont été accomplies par de nombreux laboratoires internationaux. Des différentes approches ont été développées pour réaliser la reconnaissance de parole.

La reconnaissance de mots isolés en vocabulaire limité et pour mono- locuteur est un problème bien maîtrisé grâce à la technique de l'alignement temporel dynamique. Les systèmes actuels de reconnaissance sont fondés sur une approche probabiliste utilisant les modèles de Markov cachés. Elles rendent concevable la reconnaissance de parole continue à grand vocabulaire et indépendamment du locuteur.

Dans la première partie de ce chapitre, nous exposons les notions qui se rattachent à l'étude des organes biologiques de production et de perception de la parole et nous décrivons le signal de parole et nous présenterons les différents problèmes posés lors de son traitement.

Dans la deuxième partie, nous présentons la description du fonctionnement d'un système de reconnaissance automatique de la parole, les modes de reconnaissance, et les différentes méthodes de reconnaissance.

1.2 Le signal de la parole

Le signal de la parole n'est pas un signal ordinaire. Il est le vecteur d'un phénomène complexe : la communication parlée. La reconnaissance de la parole pose de nombreux problèmes aux chercheurs depuis 1950. D'un point de vue mathématiques, il est difficile de modéliser le signal de parole, compte tenu de sa variabilité.

L'étude des mécanismes de phonation permettra donc de déterminer, dans un certaine mesure ce qui est de la parole et ce qui n'en est pas, l'étude des mécanismes d'audition et des propriétés perceptuelles nous permettrons de dire ce qui dans le signal parole, est réellement perçu [B3]

Nous allons ici tenter de mettre en évidence quelques caractéristiques importantes du signal non stationnaire afin de faire ressortir les problèmes posés lors de son traitement. Nous allons tout d'abord exposer les notions qui se rattachent à l'étude des organes biologiques de la production et de la perception de la parole

1.2.1 Production de la parole

1.2.1.1 L'appareil phonatoire

L'appareil phonatoire nous permet de produire des sons très variés dans un espace fréquentiel et énergétique pourtant limité. L'appareil phonatoire humain (paragraphe 1.2.1.2) a été la base de recherches visant à simuler mécaniquement ses capacités, recherches ayant permis, en retour, de mieux comprendre son fonctionnement.

1.2.1.2 L'appareil phonatoire humain [B1] [B2] [B4]

La production de la parole est assurée, chez l'homme, par plusieurs organes successifs. Les poumons sont indispensables dans ce processus puisqu'ils assurent la génération d'un composant incontournable : de l'air sous pression. Cet air, expulsé, traverse alors les cordes vocales qui entrent ou non en action pour produire un voisement. Ce voisement correspond à la fréquence fondamentale qui est le timbre de la voix. Cette fréquence fondamentale étant produite, elle est propagée dans l'ensemble du conduit vocal. Ce conduit est de forme et de volume variable. Plusieurs organes concourent à ces possibles modifications qui permettent de produire des sons différents. Parmi ces organes se trouve la langue, acteur principal des modifications qui peut agir par constriction ou occlusion du conduit vocal. Les dents et les lèvres agissent également par occlusion ou constriction, à des degrés cependant moindres.

Le conduit vocal est, la plupart du temps, constitué d'un seul conduit buccal. La luette et son prolongement vers le palais, le vélum, assurent normalement la fermeture du conduit nasal pendant la production de parole. Le conduit nasal peut, dans certains cas, être connecté au conduit vocal.

Cette connexion permet de générer des sons supplémentaires en modifiant le volume de la caisse de résonance normalement constituée par le seul conduit buccal. Une coupe de l'appareil phonatoire humain est fournie en figure 1.1.

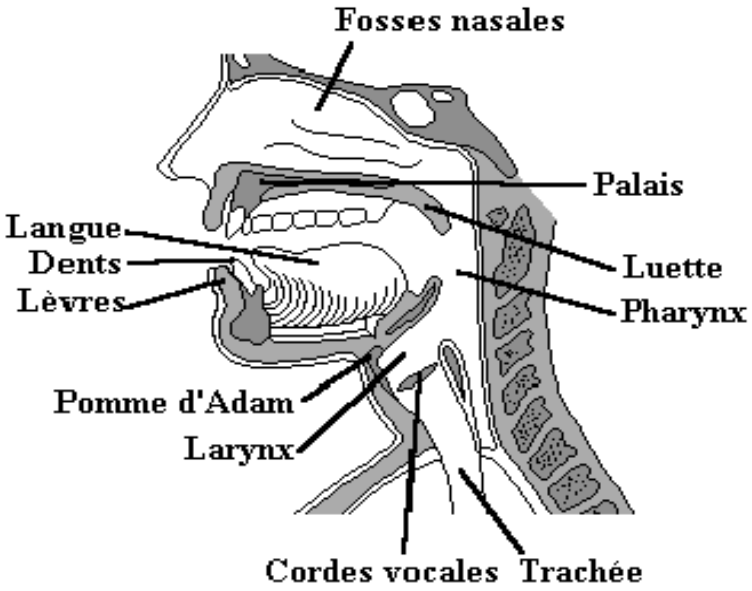


Figure 1.1 : Coupe de l'appareil phonatoire humain

Les différents organes de la parole et leur agencement peuvent servir de base à des modélisations.

1.2.1.3 Modélisation du processus de production de la parole

Le signal de parole est avant tout un son produit par l'homme et qui nécessite [B6]

- une excitation ou une source : vibration des cordes vocales.
- un milieu de propagation : l'air.
- une cavité de résonance : les cavités du conduit vocal.

La figure 1.2 représente le modèle source-filtre qui regroupe les différents éléments cités précédemment.

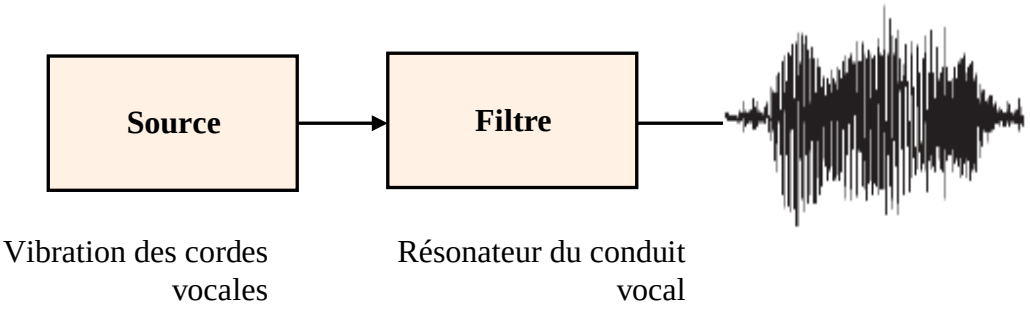


Figure 1.2 : Modèle source-filtre

– Source

La source, qui joue un rôle très important, est composée d'une partie périodique et d'une partie assimilée à un bruit. Les sons produits peuvent donc être De deux types : voisés (quasi-périodiques) ou non voisés. La vibration des cordes vocales se fait à la fréquence fondamentale F_0. Cette fréquence est communément appelée le pitch. La valeur du pitch varie d'un individu à un autre. Elle dépend également du genre : homme (≈ 120Hz), femme (≈ 220Hz), enfant (≈ 330Hz). Le pitch semble à première vue une caractéristique intéressante pour la reconnaissance du locuteur [B7] mais les difficultés d'extraction de F0 ont limité son utilisation. Il peut cependant être combiné à des méthodes de codage traditionnelles [B8].

– Filtre

L'effet du conduit vocal sur la source est modélisé par un filtre. Le filtre permet de transformer la source pour produire des sons différents. Ce filtre diffère selon le son prononcé. La figure 1.3 représente le modèle de production de la parole. Sur cette figure, on peut constater que la production d'un son oral est différente de celle d'un son nasal. La principale différence réside dans les cavités exploitées lors de la production. Une caractéristique importante des filtres est leurs fréquences de résonances. Ces fréquences sont nommées formants. Les voyelles peuvent être discriminées selon la valeur de leurs formants (F1, F2, ...).

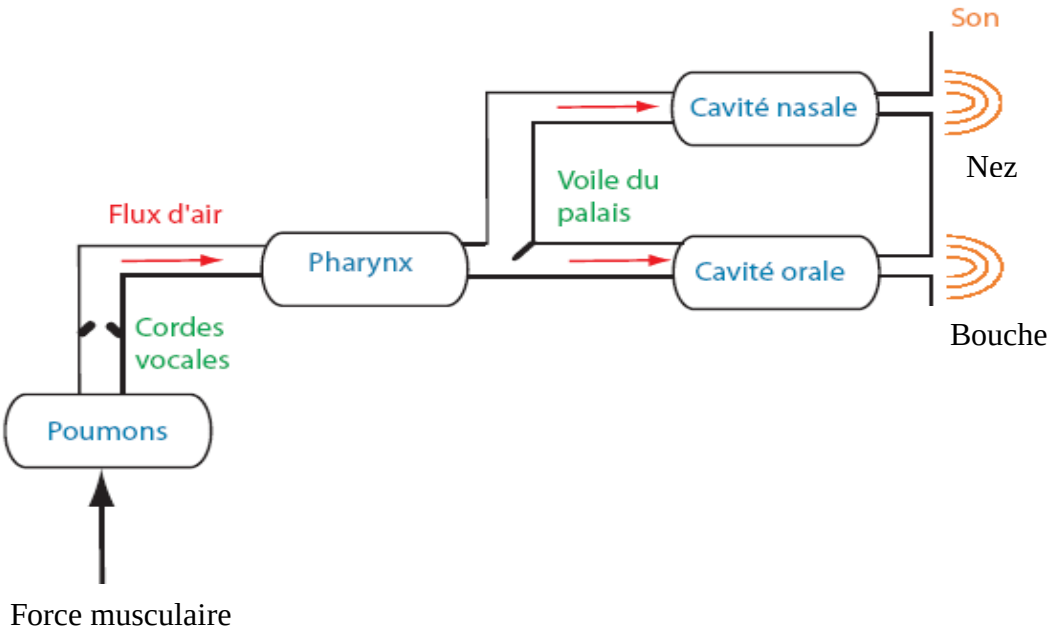


Figure 1.3 : Modèle de production de la parole

1.2.2 Perception de la parole

1.2.2.1 L'appareil auditif [B5]

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur. Nous allons maintenant présenter l'anatomie de l'oreille, organe récepteur de l'information sonore, et les capacités de perception qui caractérisent cet organe lorsqu'il est en parfait état et n'a subi aucune atteinte venue amoindrir ses capacités intrinsèques.

1.2.2.2 L'appareil auditif humain

L'oreille est divisée en trois parties distinctes, cette division se faisant en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau

en charge du traitement. Une description détaillée de l'oreille figure 1.4 permettra au lecteur de mieux appréhender les différents organes la constituant et de mieux visualiser leur répartition..

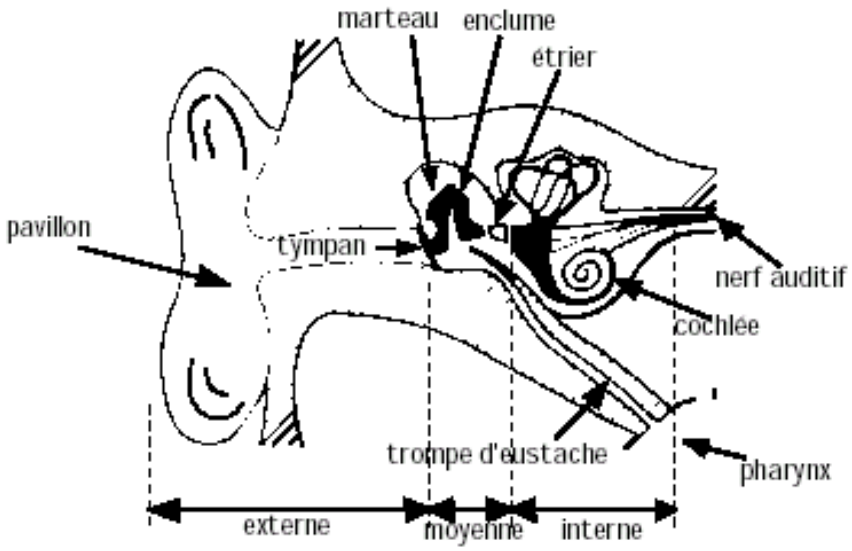


Figure 1.4 : Coupe de l'appareil auditif humain

1.2.2.3 Modélisation du processus de la perception de la parole

La seconde modélisation consiste naturellement à considérer non plus la production mais la perception humaine [B9]. Il s'agit donc de modéliser les caractéristiques de l'oreille humaine [B10]. Ces modèles sont appelés modèles auditifs

Depuis de nombreuses années, des études ont été menées pour essayer de "mimer" l'oreille humaine [B11]. Cependant, la réplique des différents phénomènes n'améliore pas systématiquement les performances des systèmes de reconnaissance [B12]. Les principales étapes de la perception humaine sont représentées sur la figure 1.5.

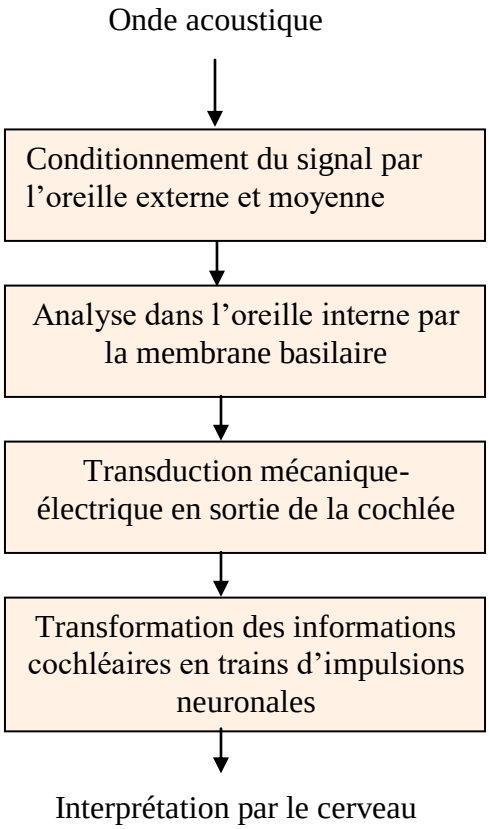


Figure. 1.5 – Modèle de la perception humaine.

Le conditionnement du signal et l'analyse dans l'oreille interne se traduisent dans le domaine de traitement du signal par des opérations de filtrage et la mise en place d'échelle non-linéaire (Mel ou Bark). Ces deux techniques sont largement utilisées dans les méthodes d'extraction de paramètres traditionnels.

1.2.2.4 L'aire d'audition

L'homme est en effet très limité dans ses capacités de perception auditive vis-à-vis d'autres membres du règne animal. Il lui est ainsi impossible de distinguer des sons de plus de 20 kilohertz, les ultrasons, alors que certains animaux qui lui sont familiers peuvent percevoir des sons allant jusqu'à 50 kilohertz. De même lui est-il impossible de distinguer des sons d'une fréquence inférieure à 20-25 hertz, les infrasons. À l'intérieur de cet espace fréquentiel existe un sous-espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie en deçà de laquelle l'homme ne percevra pas un son d'une fréquence appartenant pourtant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition et il est variable en fonction de la fréquence. Inversement, il existe une limite d'énergie maximale.

Cette limite ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être irrémédiablement endommagées. Cette limite s'appelle le seuil de douleur et elle aussi est variable en fonction de la fréquence. Il est intéressant de noter qu'il existe dans l'oreille deux muscles qui permettent à l'homme de débrayer le transfert des vibrations du tympan à la cochlée pour limiter les dégradations qui peuvent survenir dans le cas où un bruit dépassant le seuil de douleur est perçu. L'espace de fréquences et d'énergies ainsi défini (figure 1.6) constitue la zone d'audition à l'intérieur de laquelle l'homme peut recevoir des informations de son environnement. C'est bien sûr à l'intérieur de cet espace que se trouve le champ de la musique qui circonscrit lui-même le champ de la parole.

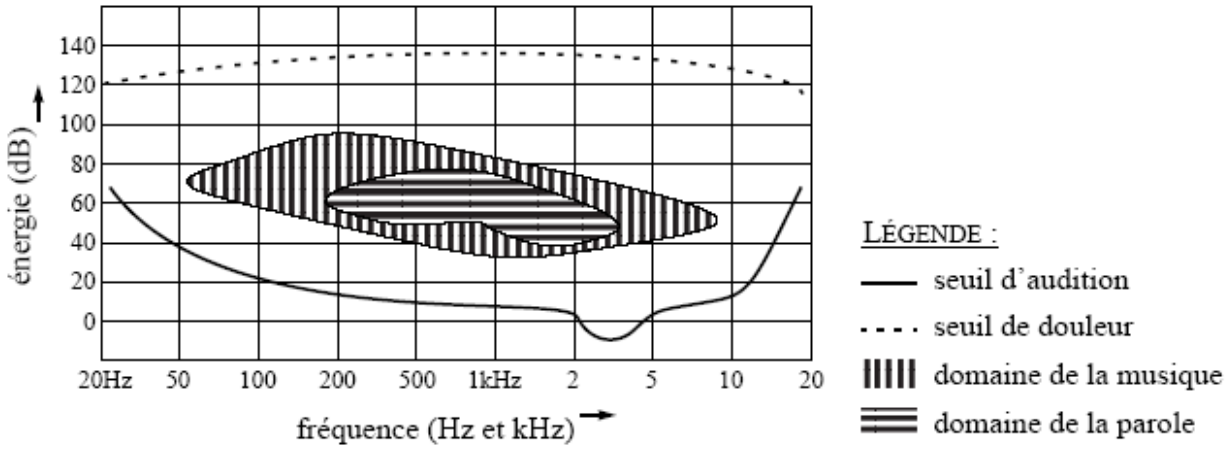


Figure 1.6 : L'aire d'audition

1.2.2.5 L'échelle des Mels

L'échelle des Mels est une échelle biologique. C'est une modélisation de l'oreille humaine. A noter que le cerveau effectue en quelque sorte une reconnaissance vocale complexe avec filtrage des sons. On considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence. L'échelle des Mels modélise assez fidèlement la perception de l'oreille : linéairement jusqu'à 1000 Hz, puis logarithmiquement au dessus. La formule donnant la fréquence en Mel à partir de celle en Hz est :

$$Mel(f) = \frac{1000.\ln(1+\frac{f}{700}}{\ln(1+\frac{1000}{700}} \approx 1127.\ln(1+\frac{f}{700}) \approx 2595.\log(1+\frac{f}{700}) \qquad (1.1)$$

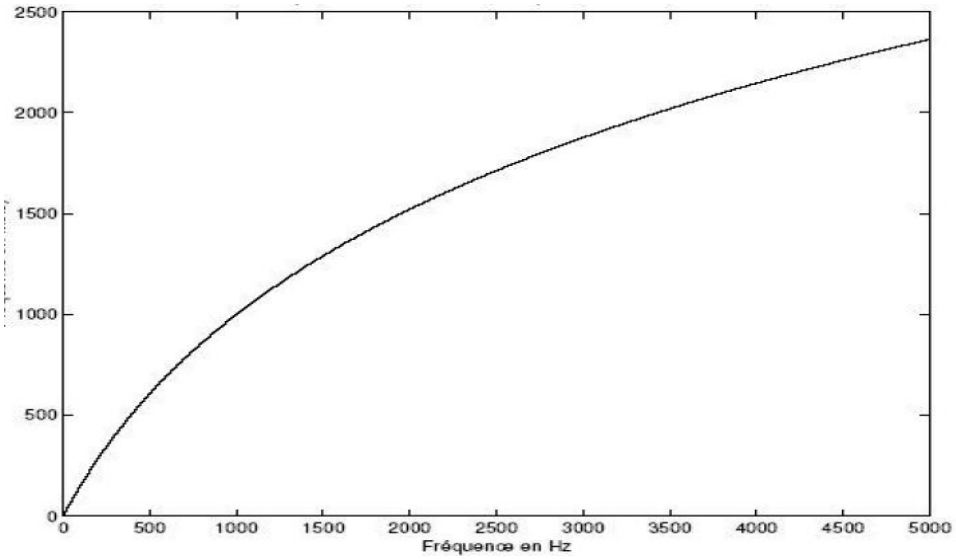


Figure 1.7 : Graphe de conversion de la fréquence de Hertz à Mels

L'échelle des Mels permet donc de modéliser une perception de l'oreille linéairement.

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

1.2.3 Complexité du signal parole

La complexité de signal parole provient de la combinaison de plusieurs facteurs (la redondance de signal acoustique, la grande variabilité, les effets de coarticulation en parole continue …..) rendent la tâche de reconnaissance de la parole difficile à réaliser Nous allons maintenant voir quelque caractéristiques de ce signal afin de faire ressortir les problèmes posés lors de son traitement.

- Redondance de signal parole :

Le signal de parole est extrêmement redondant. Cette grande redondance lui confère une robustesse à certains types de bruits. De nombreuses recherches sont menées afin de rendre les systèmes de reconnaissance robustes aux bruits, mais les performances humaines sont encore loin d'être atteintes

- Variabilité du signal

Le signal de parole possède une très grande variabilité. Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution peut varier, la durée du signal est alors modifiée. Toute altération de l'appareil phonatoire peut modifier la qualité de l'émission (exemple : rhume, fatigue…). De plus, la diction évolue dans le temps. La voix est modifiée au cours des étapes de la vie d'un être humain (enfance, adolescence, âge adulte…). La variabilité interlocuteur est encore plus accentuée. La hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. Un exemple pertinent de cette variabilité apparaît lorsque nous comparons la voix d'un locuteur originaire du Nord avec celle d'un locuteur originaire du sud de l'Algérie. Enfin, la parole est un moyen de communication où de nombreux éléments entrent en jeu, tels que le lieu, l'émotion du locuteur, la relation qui s'établit entre les locuteurs (stressante ou amicale). Ces facteurs influencent la forme et le contenu du message. L'acoustique du lieu (milieu protégé ou environnement bruité), la qualité du microphone, les bruits de bouche, les hésitations, les mots hors vocabulaire sont autant d'interférences supplémentaires sur le signal de parole.

- Les effets de coarticulation :

La production parfaite d'un son suppose un positionnement précis des organes phonatoire. Le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis subissent alors l'influence de ceux les précèdent ou les suivent.

L'effet de coarticulation est un facteur de variabilité supplémentaire. La prise en compte de ce facteur améliore sensiblement les performances de système de reconnaissance.

- Continuité :

Tout discours peut être retranscrit par des mots, qui peuvent à leur tour être décrit comme une suite de symboles élémentaires appelés phonèmes. Cela laisse supposer que la parole est un processus séquentiel, au cours duquel des unités indépendantes se succèdent. Malheureusement, les phonéticiens eux-mêmes ont des difficultés à identifier individuellement ces unités discrètes dans le signal. La parole est en réalité un flux continu, et n'existe pas de pause entre les mots qui pourrait faciliter leur localisation automatique pour les systèmes de reconnaissance.

1.3 Système de reconnaissance automatique de la parole

1.3.1 Description du fonctionnement

La démarche suivie en reconnaissance automatique de la parole consiste à opérer selon le schéma général de la figure 1.8.

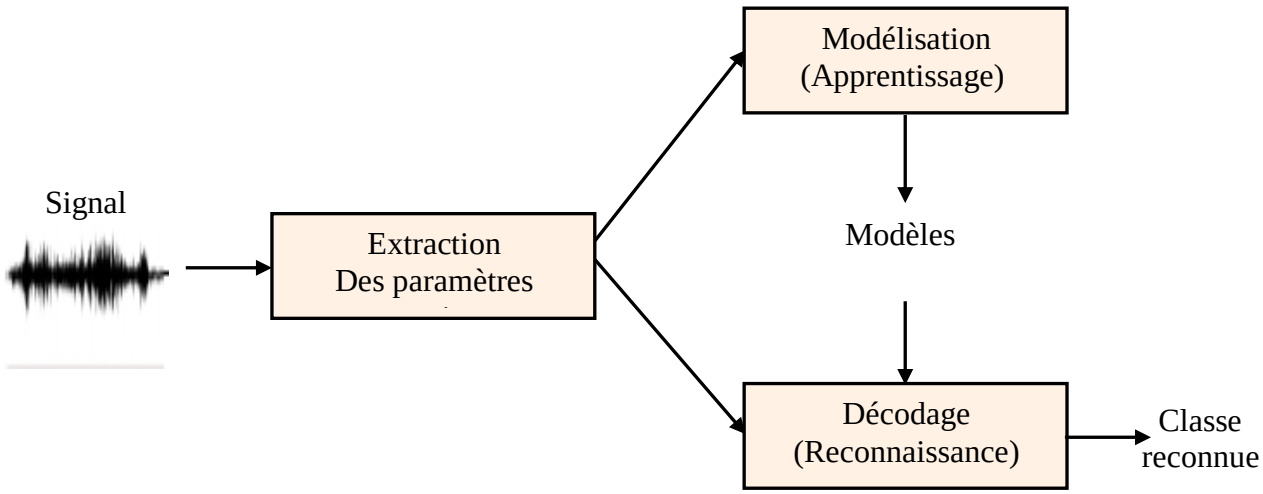


Figure 1.8 : Schéma général d'un système de RAP

Le système comporte 3 étapes principales : l'analyse acoustique, l'apprentissage et la reconnaissance L'analyse acoustique traite le signal et en extrait les vecteurs acoustiques qui seront utilisés pour les phases suivantes. A partir d'un corpus de données étiquetées, l'apprentissage consiste à apprendre les motifs vocaux. On construit ici des modèles acoustiques pour chaque unité acoustique (mots, phonème,……..). Ceci se fait le plus souvent via des modèles de Markov cachés (HMM pour « Hidden Markov Model «. En anglais). Cette étape franchie, il est nécessaire de mettre en correspondance une suite d'éléments acoustiques avec une forme lexicale. C'est ici qu'intervient la modélisation du langage. Elle permet d'obtenir une information a priori sur le positionnement d'un mot dans le signal à reconnaître par différentes techniques de modélisation La phase de reconnaissance reconnaît le signal test en utilisant des modèles qui viennent de la phase d'apprentissage.

1.3.2 Le mode d'élocution

Le mode d'élocution caractérise la façon dont on peut parler au système. Il existe quatre modes d'élocution distincts :

- Le mode '' mots isolés'' c'est le mode utilisé dans notre sujet pour la reconnaissance des chiffres arabes. Ce mode est purement acoustique. , ne nécessite aucune notion de contexte, sémantique, ou de syntaxe, c'est à ne dire aucun modèle de langage. Il permet de reconnaître des mots clairement séparés par des pauses.

  La reconnaissance automatique de mots isolés permet de réaliser des interfaces homme-machine à commandes vocales, afin de piloter tout outil électronique, en prononçant de simples mots clés. Ceci est très utile dans le cas de systèmes portables miniaturisés, sur lesquels il est impossible d'implanter une interface graphique complète à cause de leur taille réduite.
- L'identification de '' mots-clés connectés'' est purement acoustique et ne fait pas intervenir de modèles de langage.

  Le système reconnaît des séquences de quelques mots sans pause volontaire pour les séparer (exemple : reconnaissance de chiffres connectés ou de nombres quelconques….)
- Le mode '' reconnaissance de parole continue '' est plus sophistiqué .Son but est d'associer, à la partie acoustique, une partie modèle de langage.
- Le mode '' reconnaissance de parole spontanée ''

1.3.3 Les modes de fonctionnement d'un système de reconnaissance

Un système de reconnaissance peut être utilisé sous plusieurs modes :

- **Dépendant du locuteur (mono-locuteur)**

  Dans ce cas particulier, le système de reconnaissance est configuré pour un locuteur spécifique. C'est le cas de la plupart des systèmes de reconnaissance de parole disponibles sur le marché. Les principaux systèmes de dictée vocale actuels possèdent une phase d'apprentissage recommandée avant toute utilisation (voire même une adaptation continue des paramètres au cours de l'utilisation du logiciel) afin d'effectuer une adaptation des paramètres à la voix de l'utilisateur.

- **Pluri-locuteur (ou multi-locuteur)**

  Le système de reconnaissance est élaboré pour un groupe restreint de personnes. Le passage d'un locuteur à un autre du même groupe se fait sans adaptation.

- Indépendant du locuteur

Tout locuteur peut utiliser le système de reconnaissance.

1.3.4 Les différents méthodes de reconnaissance de la parole

On distingue usuellement en reconnaissance de la parole l'approche analytique et l'approche globale.

- Approche globale

Cette méthode considère le plus souvent le mot comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans ce type de méthode on compare globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique ou des modèles de Markov cachés (HMM : Hidden Markov Model). Cette méthode a pour avantage d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes.

Ce type de méthode est utilisé dans les systèmes suivants :

- Reconnaissance de mots isolés.
- Reconnaissance d'unités enchaînées.
- Reconnaissance de parole dictée avec pauses entre les mots.

- Approche analytique

Cette méthode fait intervenir un modèle phonétique du langage. Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabe, phonème, …….etc.).

Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance dans cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification de ces unités en utilisant des méthodes de reconnaissance des formes (classification statistique, réseau de neurones, etc.) ou des méthodes d'intelligence artificielle (systèmes experts par exemple). Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation (délétions, insertions, substitutions, recouvrements) et d'étiquetage phonétique. C'est pourquoi le DAP (Décodage Acoustico-Phonétique) est fondamental dans une telle approche.

- **Formalisation statistique de l'approche**

Les systèmes actuels de reconnaissance sont fondés sur une approche probabiliste [B6]. Ils consistent généralement en deux parties fondamentales que sont le module de modélisation acoustique et celui de modélisation du langage.

La formule générale, dans le cadre d'un système entièrement probabiliste, s'exprime sous la forme d'une équation bayésienne. Le but du système est de trouver l'hypothèse $\tilde{M}$ qui maximise pour toutes les séquences de mots M possibles et pour une observation acoustique O, l'équation suivante :

$$\tilde{M} = \arg\max_{M} P(M/O) = \arg\max_{M}\left(\frac{P(O/M).P(M)}{P(O)}\right) \quad (1.2a)$$

$$= \arg\max_{M}\left(P(O/M).P(M)\right) \quad (1.2b)$$

Dans cette équation, nous pouvons identifier plusieurs facteurs :

- $P(O)$: est la probabilité de l'observation acoustique O. Celle-ci est constante pour toutes les séquences de mot M , d'où la formule finale de l'équation précédente.
- $P(O/M)$ est la probabilité de l'observation acoustique O connaissant une séquence de mots M .
- $P(M)$ est la probabilité *a priori* de la séquence de mots M , sans aucune notion d'acoustique, dans le langage considéré.

Les unités acoustiques modélisées peuvent être des mots comme dans l'approche globale, ou des unités plus courtes telles que le phonème comme dans l'approche analytique.

La modélisation markovienne est la plus utilisée. Son application à la reconnaissance de la parole continue a été rendue possible par l'augmentation continue de la puissance des ordinateurs et de la taille des bases de données disponibles.

1.4 Conclusion :

Ce chapitre qui porte un aperçu sur la reconnaissance automatique de la parole, a permis de dégager les caractéristiques du signal en vue de leur utilisation en reconnaissance vocale.

Divers modes de fonctionnement ont été évoqué dans ce chapitre tel que le mode mono-locuteur et le mode multi-locuteur. Cette étude a permis de retenir la méthode de reconnaissance globale pour la reconnaissance de mots arabes isolés. Nous pensons que la conception d'un système de reconnaissance automatique de la parole doit être effectuée de façon à tenir compte à la fois la charge de calcul de l'algorithme de reconnaissance et de la complexité des signaux parole. Dans cet esprit, nous avons présenté l'approche statistique successible de résoudre de façon plus ou moins notre problème de reconnaissance.

CHAPITRE 2 :

ANALYSE DU SIGNAL PAROLE

Chapitre 2

Analyse du Signal Parole

2.1 Introduction :

Le signal acoustique présente, dans le domaine temporel, une redondance qui rend indispensable un traitement préalable à toute tentative de reconnaissance. .

L'analyse de la parole permet la mise en forme du signal mais aussi l'extraction de paramètres nécessaires pour les prochaines étapes telles que la reconnaissance. Un des objectifs de cette analyse est d'obtenir une représentation compacte et informative du signal. Le signal de parole est non stationnaire mais il peut être considéré comme localement stationnaire. L'analyse du signal de parole se fait pendant ces périodes stationnaires dont la durée varie de 10 à 30 ms. Cette durée correspond aussi `à la durée de stabilité du modèle de production. L'analyse de la parole (figure 2.1) consiste à effectuer des prétraitements, nécessaires pour la mise en forme du signal, tels que le découpage en fenêtres. La seconde étape de l'analyse est l'extraction de paramètres qui est l'objet de ce mémoire.

Un système d'analyse du signal parole se décompose en deux blocs figure 2.1

- Bloc Prétraitement.
- Bloc d'extraction de paramètres.

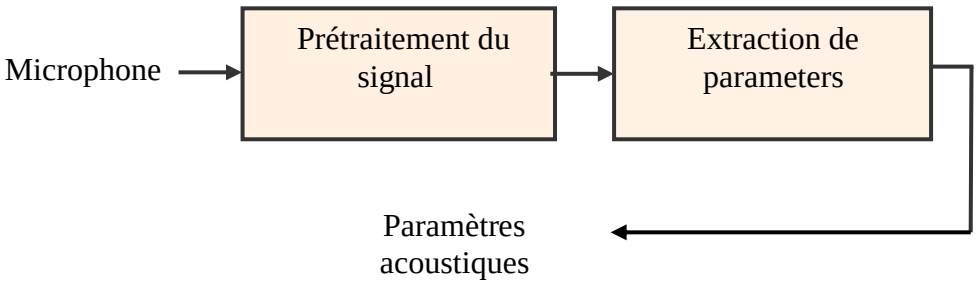


Figure 2.1 : schéma général d'un traitement acoustique

2.2 Prétraitements du signal :

Avant tout calcul, il est nécessaire de mettre en forme le signal de parole. Pour cela, quelques opérations sont effectuées avant tout traitement. La figure2.2 illustre l'ensemble de ses opérations. Le signal est tout d'abord filtré puis échantillonné. Une préaccentuation est effectuée. Puis, le signal est segmenté en trames

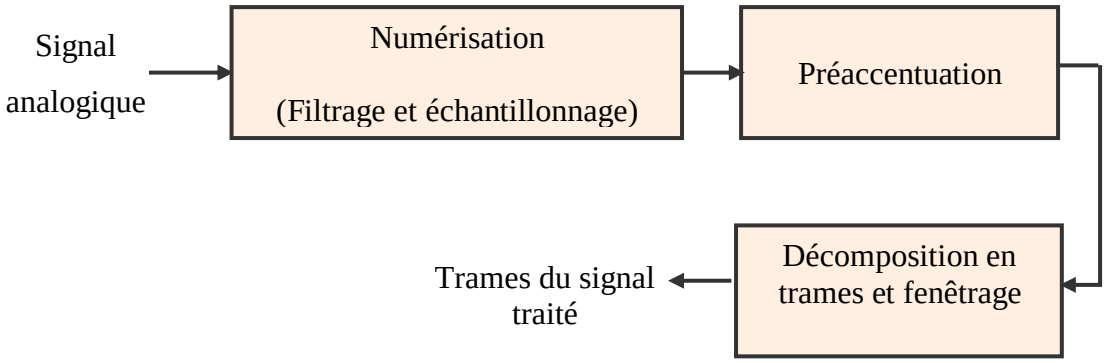


Figure 2.2 : Bloc de prétraitements

D'autres prétraitements, ayant pour but d'augmenter la robustesse, sont parfois mis en oeuvre comme par exemple la normalisation des signaux ou bien la soustraction spectrale qui a pour effet d'éliminer les bruits additifs.

2.2.1 Numérisation :

Pour être utilisable par ordinateur, un signal doit être tout d'abord numérisé. Cette opération tend à transformer un phénomène analogique, le signal sonore dans notre cas, en une suite d'éléments discrets. La numérisation sonore repose sur deux paramètres :

- La quantification
- La fréquence d'échantillonnage

La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore chaque pas d'échantillonnage. Le choix de la fréquence d'échantillonnage est aussi déterminé pour la définition de la bande passante représentée dans le signal numérisé. L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50 Hz - 8 kHz], la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon [B14] ; mais elle peut varier en fonction du domaine d'application ou des besoins ou contraintes matériels. Pour les applications de type téléphonique, cette fréquence descend à 8 kHz.

2.2.2 Préaccentuation du signal :

L'étape de préaccentuation (ou pré-emphase) consiste à accentuer les hautes fréquences. On fait généralement appel à un filtre de la forme :

$$H(z) = 1 - \alpha . z^{-1} \qquad 0 \le \alpha \le 1 \tag{2.1}$$

α : est le paramètre de préaccentuation.

La valeur typique de α est 0.97 [B16].

L'intérêt de préaccentuation est d'aplatir le spectre de signal de parole et de filtrer la composante continue de façon à le placer dans des conditions « optimales » vis-à-vis des traitements ultérieurs, notamment dans le calcul d'un modèle autorégressif [B 15].

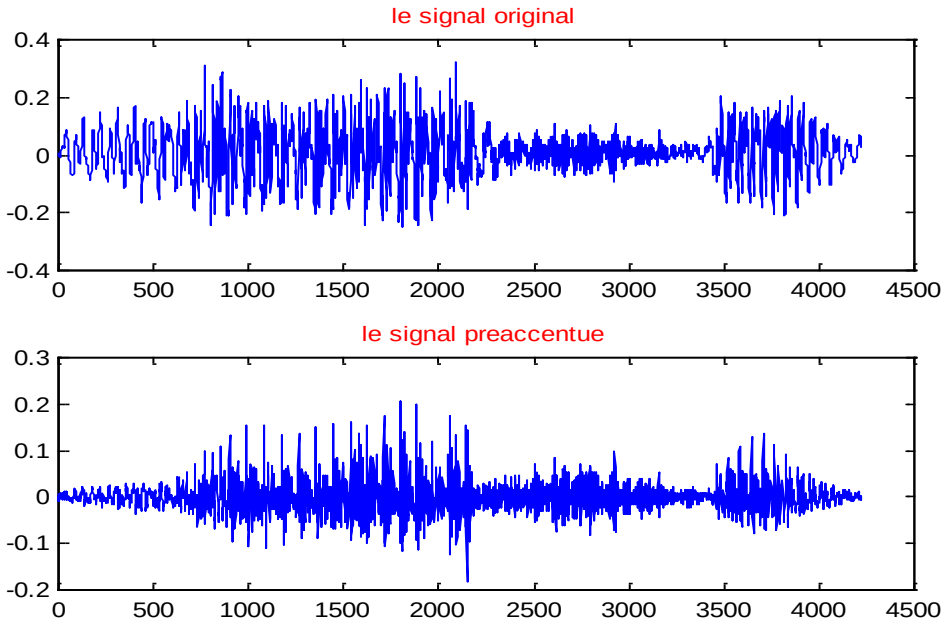


Figure 2.3. : Le signal de parole (chiffre "1"),

2.2.3 Décomposition en trames et fenêtrage :

La parole est un son complexe dans son ensemble, mais si on considère des intervalles de temps très réduits (environ 30 ms), le signal vocale est alors presque stationnaire. C'est alors pour cela que l'on considère le signal vocal comme un signal quasi-stationnaire, en première approximation. On pourra utiliser des fonctions de fenêtrage pour améliorer les résultats.

Une fois la formule de préaccentuation du signal effectuée, nous découpons le signal en fenêtres. Le signal est découpé en tranches de 2^n échantillons appelées trames ou encore fenêtres qui ont la particularité de se recouvrir de moitié dans l'objectif d'avoir un meilleur traitement pour FFT (Fast Fourier Transform).

La longueur N d'une tranche est choisie de façon à avoir des tranches dont la durée est de l'ordre de 30ms. En conséquence, pour un mot de durée de 0.5s (comme par exemple la liste des chiffres : ouahad, ithnani, . . ., cela donne entre 30 et 40 tranches à traiter .Ensuite une fenêtre de pondération est appliquée à chaque trame ceci dans l'objectif d'harmoniser les échantillons pour permettre un meilleur traitement pour l'algorithme FFT. En effet la FFT ne donne pas de bons résultats quand une pente trop importante est détectée dans une partie du signal. La fenêtre de pondération a pour objectif de minimiser les erreurs produites par FFT. Le choix se porte généralement sur les fenêtres de Hamming [B17]

La figure 2.4 illustre la fenêtre de pondération Hamming :

$$w_n = \begin{cases} 0.54 - 0.46.\cos\left(2.\pi.\dfrac{n}{N-1}\right) & n = 0,..., N-1. \\ 0 & \text{ailleurs} \end{cases} \qquad (2.2)$$

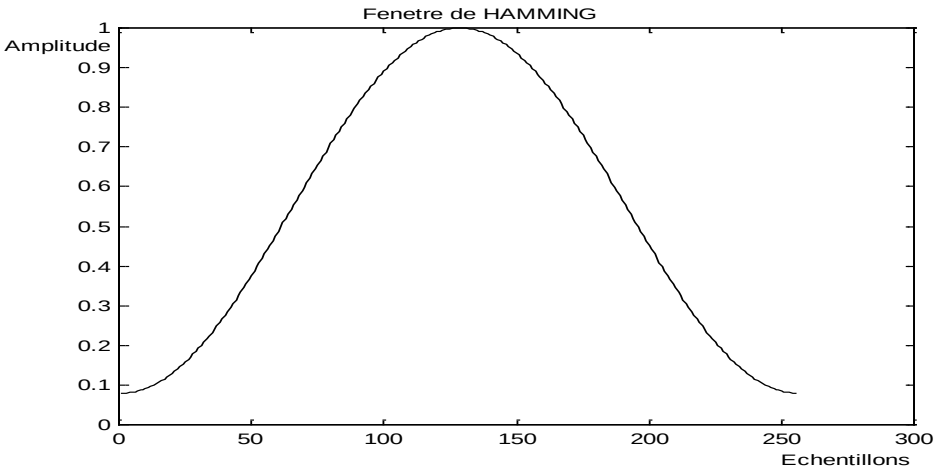


Figure 2.4 : Fonction de pondération de Hamming sur 256 points

.

2.3 Extraction de paramètres :

L'extraction de paramètres est l'objet principal de l'analyse de la parole et c'est le passage obligé de toutes les applications en traitement de la parole

2.3.1 Energie de signal :

L'énergie de signal est un indice qui peut par exemple contribuer à la détection du début et fin de la parole et de voisement d'un segment de la parole. L'énergie totale E (fenêtre) est calculée directement dans le domaine temporel sur une trame de signal comme :

$$\text{E (fenêtre)} \quad = \sum_{n=0}^{N-1} s^2(n) \quad , \tag{2.3}$$

avec s(n) : l'échantillon à l'instant n

2.3.2 Coefficients d'auto-corrélation :

Les coefficients d'auto corrélation $\{r_{ss}(k)\}$ de signal s sont utilisés dans le cadre de la modélisation autorégressive. L'estimation de ces coefficients est calculée par :

$$r_{ss}(k) = E[s(n).s(n-k)] = \sum_{n=k}^{N-1} s(n).s(n-k) \qquad avec\ 0 \le k \le N-1 \tag{2.4}$$

D'autres paramètres peuvent être calculés dans le domaine temporel, comme la fréquence fondamentale, taux de passage par zéro…, mais ils ne sont pas utilisés dans notre système d'expérimentation.

2.3.3 Représentation prédictive :

C'est une méthode d'analyse du signal qui a été largement utilisée dans les systèmes de reconnaissance de la parole. Elle fournit une bonne représentation du signal vocal et permet de déduire les paramètres de bases tels que la fonction de transfert du conduit vocale.

Cette méthode connue sous le sigle LPC (Linear Predictive Coding) est fondé sur les connaissances de la production de la parole et suppose que le modèle de production est linéaire[B14]. Le modèle se décompose en une source active et un conduit passif, on modélise l'onde vocale comme la sortie d'un filtre dont la fonction de transfert est :

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{A(z)} \tag{2.5}$$

avec $$A(z) = a_0 + a_1.z^{-1} + a_2.z^{-2} + ... + a_p.z^{-p} \quad (2.6)$$

où $a_0 = 1.$ P est l'ordre du modèle.

Le principe de base de codage par prédiction linéaire est qu'un échantillon du signal peut être approximé par une combinaison linéaire des P échantillons précédents (modélisation AR.)

Le système est excité, soit par un train d'impulsions pour générer un son voisé, soit par une séquence aléatoire pour produire un son non voisé figure 2.5. Ainsi, les paramètres du modèle sont l'indicateur du voisement, la fréquence fondamentale (pitch), le gain G et les coefficients (a_k) du filtre numérique.

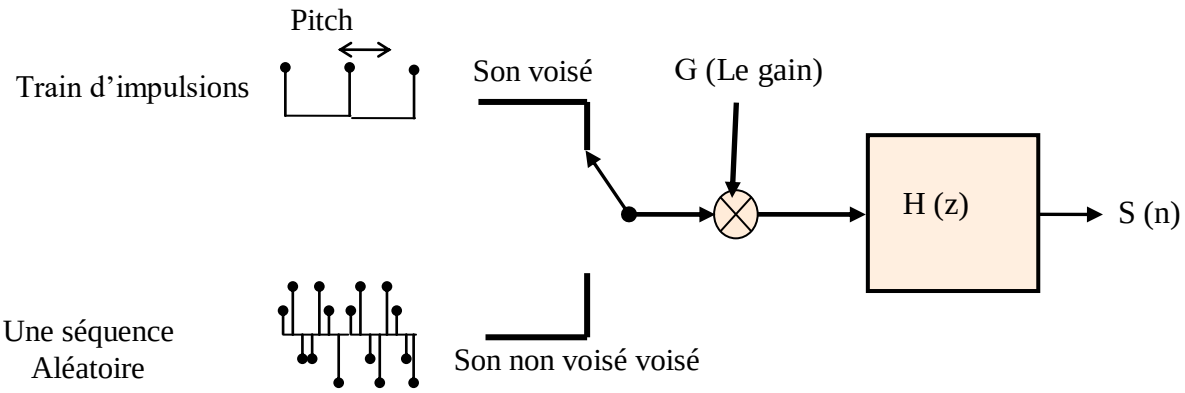


Figure 2.5 : Modèle de production de la parole utilisée pour l'extraction des paramètres

l'équation aux différences caractérisant le système et reliant l'entrée u (n) à la sortie S (n) est donnée par :

$$s(n) = -\sum_{k=1}^{p} a_k.s(n-k) + G.u(n) \quad (2.7)$$

Un prédicteur linéaire est défini par les coefficients $a(k)$ comme étant un système dont la sortie est :

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k.s(n-k) \quad (2.8)$$

où $\hat{s}(n)$ est la partie prédite

les coefficients a_k sont les estimés des coefficients de prédiction

L'erreur de prédiction e (n) est définie par :

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k.s(n-k) \quad (2.9)$$

où les constantes a_k sont les coefficients de prédiction, P l'ordre du prédicteur, et e(n) l'erreur de prédiction..

Le problème consiste à ajuster les coefficients de combinaison pour que l'énergie de l'erreur résiduelle e (n) (l'erreur quadratique moyenne), dans un intervalle de temps, soit minimale. On distingue deux méthodes suivant la manière de définir cet intervalle [B1] :

- Méthode de covariance : l'énergie résiduelle est calculée sur l'ensemble des échantillons de l'erreur en régime établi
- Méthode d'auto-corrélation : l'intervalle de temps est infini et on considère le signal nul hors de l'intervalle $[0, N-1]$où N est le nombre des points considérés. Cependant, le signal est supposé stationnaire sur tout l'intervalle du temps

- **La méthode d'auto-corréction**

Dans le système de reconnaissance de la parole, la méthode d'auto-corrélation est presque exclusivement utilisée en raison de son efficacité et la stabilité inhérente. La méthode d'auto-corrélation toujours produit un filtre de prédiction dont les zéros de polynôme $A(z)$ se trouvent à l'intérieur du cercle unitaire dans le plan Z [B18] [B20.] le schéma bloc de cette méthode est illustré par la figure 2.6.

l'erreur quadratique moyenne de prédiction linéaire définie par :

$$EQM = E\left[e^2(n)\right] = E\left[\left(s(n) + \sum_{k=1}^{p} a_k.s(n-k)\right)^2\right] \tag{2.10}$$

la minimisation de l'erreur quadratique moyenne revient à annuler son gradient par rapport aux coefficients a_k conduit aux équations suivantes :

$$\frac{\partial E\left[e(n)^2\right]}{\partial.a_j} = 2E\left[\left(s(n) + \sum_{k=1}^{p} a_k.s(n-k)\right).s(n-j)\right] = 0 \qquad j = 1....p \tag{2.11}$$

on notant : $r_{ss}(k) = E\left[s(n).s(n-k)\right]$ l'auto-corrélation de s (n), avec $r_{ss}(k) = r_{ss}(-k)$

on obtient : $$\sum_{k=1}^{p} r_{ss}(k-j).a_k = -r_{ss}(j). \qquad j = 1.....p \tag{2.12}$$

Ces dernières équations sont appelées «les équations normales », dite de Yule-Walker. qui peuvent être écrites sous la forme matricielle suivante

$$\underline{R}.\underline{a} = \underline{r} \tag{2.13}$$

Avec : $\underline{r} = \left[-r_{ss}(1) \quad -r_{ss}(2) \quad . \quad . \quad -r_{ss}(p-1)\right]^T$, $\underline{a} = \left[\, a_1 \quad a_2 \quad . \quad . \quad a_p \,\right]^T$

$$\underline{R} = \begin{bmatrix} r_{ss}(0) & r_{ss}(1) & .. & .. & r_{ss}(p-1) \\ r_{ss}(1) & r_{ss}(0) & .. & .. & r_{ss}(p-2) \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ r_{ss}(p-1) & r_{ss}(p-2) & ... & .. & r_{ss}(0) \end{bmatrix}.$$

où : $\underline{R}$. : matrice d'auto corrélation est une matrice de toeplitz.

$\underline{r}$: Vecteur d'auto corrélation.

$\underline{a}$: Vecteur paramètres.

L'énergie résiduelle minimum est donnée par l'expression suivante :

$$\sigma^2 = r_{ss}(0) + \sum_{k=1}^{p} a_k r_{ss}(k) \tag{2.14}$$

est l'ensemble peut s'écrire

$$\left[\begin{array}{c|ccc} r_{xx}(0) & r_{xx}(1) & \cdots & r_{xx}(P) \\ \hline r_{xx}(1) & r_{xx}(0) & \cdots & r_{xx}(P-1) \\ \vdots & \vdots & \ddots & \vdots \\ r_{xx}(P) & r_{xx}(P-1) & \cdots & r_{xx}(0) \end{array}\right] \left[\begin{array}{c} 1 \\ a_1 \\ \vdots \\ a_P \end{array}\right] = \left[\begin{array}{c} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{2.15}$$

La solution de cette équation détermine les coefficients de prédiction

Un algorithme rapide pour résoudre l'équation 2.15 et calculer les coefficients a_i été développé par Levinson en 1947 et modifié par Durbin en 1960.

Les applications de la prédiction linéaire en analyse de la parole sont nombreuses et anciennes. Une transformation directe des coefficients auto régressifs vers les coefficients cepstraux est possible.

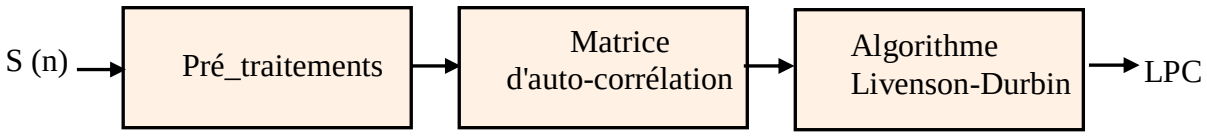


Figure 2.6 : Extraction des coefficients LPC par la méthode d'auto corrélation

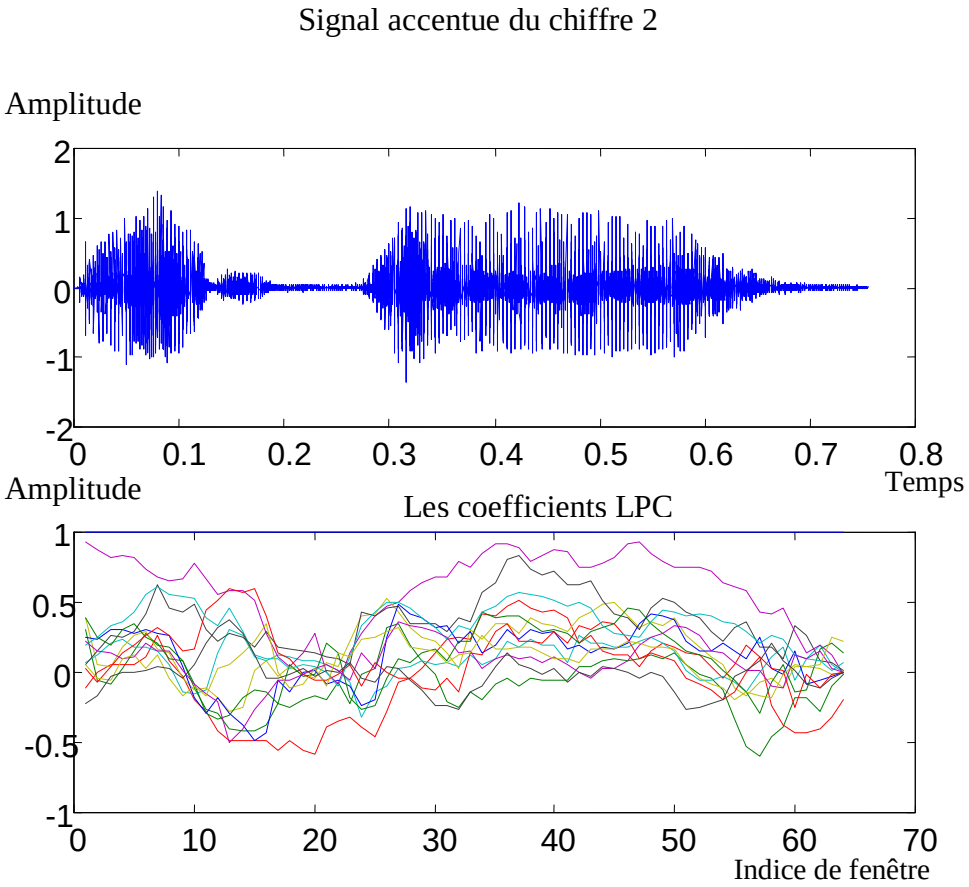


Figure 2.7 **: Coefficients LPC du chiffre 2**

2.3.4 Représentation cepstrale

Le modèle de parole suppose que le signal vocal s_n résulte de la convolution entre la source (signal excitateur g_n) et le conduit vocal (filtre passif de réponse impulsionnelle h_n) par l'équation temporelle :

$$s_n = g_n * h_n \tag{2.16}$$

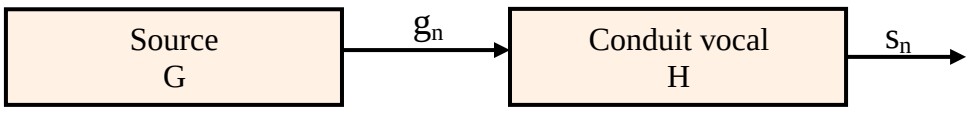


Figure 2.8 : Signal vocal

Dans le domaine spectral, la convolution devient un produit

$$S_K = G_K . H_K \tag{2.17}$$

où $\{S_K\}$, $\{G_K\}$, $\{H_K\}$sont les spectres respectifs de $\{s_n\},\{g_n\},\{h_n\}$

pour obtenir le cepstre réel, les étapes suivantes sont nécessaires (figure 2.9)

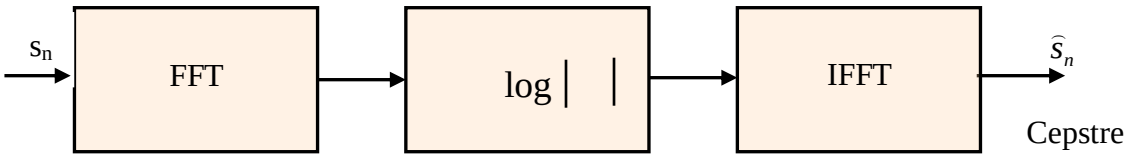


Figure 2.9 : Calcul de cepstre par analyse cepstrale

Le cepstre, qui est la somme des contributions de la source et du conduit. Les coefficients d'indice faible contiennent la quasi-totalité de l'énergie correspondant au conduit, alors que la fréquence fondamentale du locuteur fait apparaître un pic a une "quefrence" d'indice plus élevé. Un filtrage cepstral passe-bas permet de ne conserver que la contribution du conduit.

Le cepstre possède autres propriétés intéressantes qui en font une représentation efficace en reconnaissance de la parole. Parmi ces propriétés, le logarithme qui permet par un procédé de filtrage l'élimination des effets convolutifs dans le domaine temporel. En effet, grâce à la fonction logarithmique les bruits deviennent additifs. Une simple soustraction permet l'annulation de ces bruits. On parle alors de soustraction cepstrale.

Le cepstre est la base de plusieurs méthodes de codage comme par exemple le codage MFCC (Mel Frequency Cepstral Coding).et le codage LPCC

2.3.5 Représentation LPCC

Cette analyse permet de déterminer les coefficients cepstraux à partir des coefficients de la prédiction linéaire d'une modélisation AR du signal parole (voir figure 2.10) par LPCC (linéaire prédiction cepstral coefficients).

La méthode de calcul de cepstre est basée sur une relation de récurrence liant les coefficients cepstraux et les coefficients $a(i)$ [B1].

$$c_n = \begin{cases} \log(\sigma^2) \ , & n = 0 \\ -a_1 & n = 1 \\ -a_n - \sum_{i=1}^{n-1} \left(\frac{n-i}{n} \right) . c_{n-i} . a_i \ , & 2 \le n \le p \\ -\sum_{i=1}^{n-1} \left(\frac{n-i}{n} \right) . c_{n-i} . a_i & n \rangle p \end{cases} \quad \text{avec} \quad n = 1....Q \tag{2.18}$$

Où σ^2 l'énergie résiduelle minimale. *n et Q* sont respectivement les dimensions des vecteurs LPC et LPCC., ensuite, un liftrage est effectué pour augmenter la robustesse des coefficients cepstraux .Ce liftrage consiste en une multiplication par une fenêtre (voir formule 2.19) de poids des coefficients cepstraux augmentant l'amplitude des coefficients pour être moins sensibles au canal de transmission et au locuteur.

$$w_n = 1 + \frac{Q}{2}.\sin\left[\ \frac{\pi.n}{Q}\ \right], \qquad n = 1....Q \tag{2.19}$$

Q : nombre des coefficients

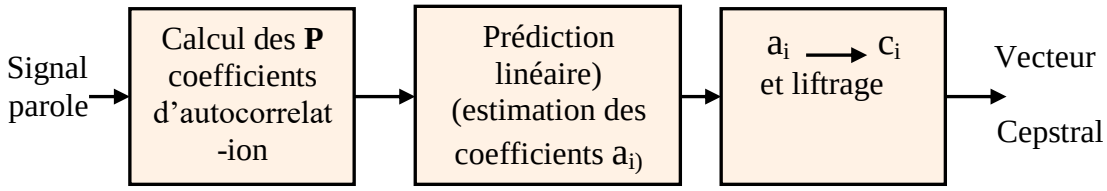


Figure 2.10. : Processus de calcul des coefficients LPCC

2.3.6 Les coefficients MFCC (Mel-scale Frequency Cepstral Coefficients)

Le codage MFCC (Mel Frequency Cepstral Coefficients) est sûrement la technique de codage la plus utilisée en reconnaissance de la parole

La motivation d'extraction de coefficients MFCC est qu'ils s'intègrent deux notions importantes. La première est la notion de bancs de filtres qui modélisent la membrane basilaire. Ces bancs de filtres sont d´employés non pas sur une échelle en Hertz mais sur une échelle non linéaire : l'échelle Mel. Cette échelle est issue de connaissances sur la perception humaine. La résolution perceptive des fréquences diffère selon que l'on écoute des sons de basses ou hautes fréquences. La deuxième est la décorrélation des coefficients produite par la projection DCT. (DCT : Discrète Cosinus Transform)

Pour transformer une fréquence linéaire en une fréquence Mel correspondant aux bandes critiques" du système auditif [B16], on peut utiliser la formule de transformation suivante :

$$Mel(f) = 2595.\log(1 + \frac{f}{100}) \tag{2.20}$$

où f est la fréquence en Hz, $Mel(f)$ est la fréquence mel-échelle de f . Il existe plusieurs approximations pour la fonction de correspondance entre la fréquence mel et la fréquence réelle mais la formule ci-dessus est la plus utilisée.

Le processus de calcul de coefficients MFCC est détaillé par le schéma de la figure 2.11 :

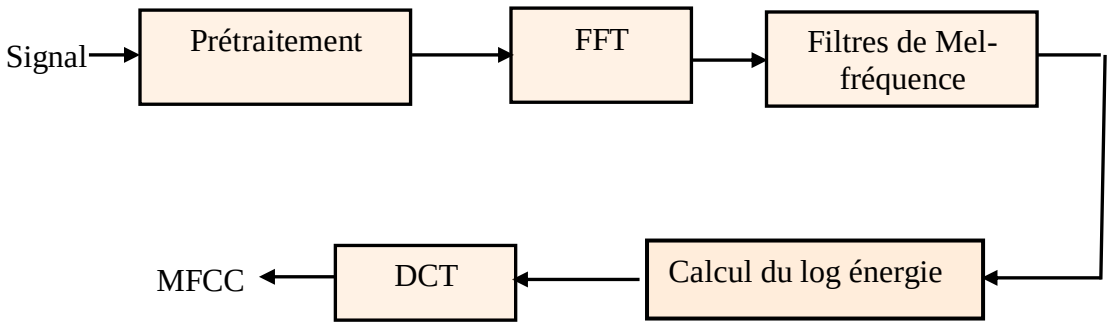


Figure 2.11 : Calcul de coefficients MFCC.

−FFT (Fast Fourier Transform).

Cette étape consiste à prendre chaque fenêtre et à appliquer la transformée de Fourier, on convertit ainsi chaque trame du domaine temporel en domaine fréquentiel. La DFT est définie sur l'ensemble de N échantillons $\{s_n\}$:

$$S_n = \sum_{k=0}^{N-1} s_k . e^{-2.j.\pi\, n.k / N}, \qquad n = 0,1,....., N-1 \tag{2.21}$$

La transformée de Fourier rapide (Fast Fourier Transform - FFT) a été introduite en 1960. C'est un algorithme très puissant pour l'implantation de la DFT.

- Banc de filtres Mel.

Parce que l'étendue des fréquences présentes dans le spectre est encore très large, donc beaucoup de données à traiter, on a recours au banc de filtres dans l'échelle de Mel. On relie ainsi le système de reconnaissance vocale au fonctionnement de l'oreille humaine. Il s'agit de filtres passes bandes centrées linéairement dans le domaine fréquentiel de Mel et de largeur telle qu'ils divisent l'espace des fréquences de manière égale dans le domaine de Mel et qu'ils se recouvrent chacun par moitié. Les filtres sont donc disposés logarithmiquement dans l'échelle des fréquences usuelles (Hz), on a ainsi beaucoup de filtres pour les basses fréquences

alors que les hautes fréquences sont disposées plus largement. Cette répartition est illustrée ci-dessous (figure 2.12) :

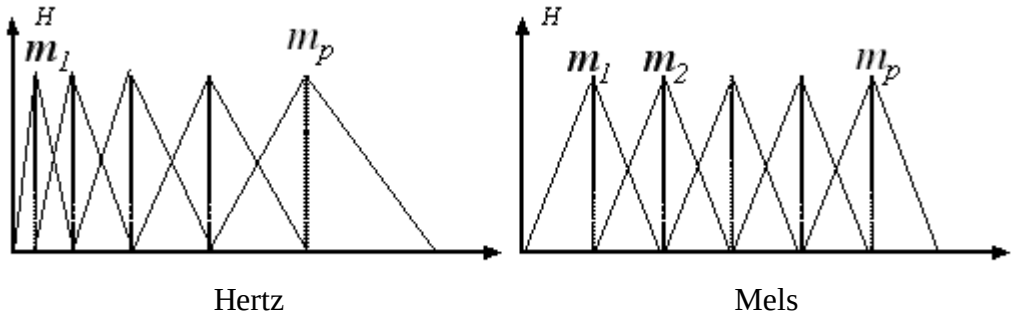


Figure 2.12 : Répartition des filtres triangulaires sur les échelles fréquentielle et Mel

- **Calcul du log énergie**

L'étape suivante consiste à calculer le logarithme des énergies E_i sortantes d'un banc de filtres de Mel. On note que le traitement du logarithme est bien exécuté par l'oreille. De plus, le logarithme exécute la compression dynamique, création d'une extraction de formes moins sensitive à la variation dynamique. [B16]

- Calcul du cepstre Mel fréquence

C'est l'étape finale, où on transforme les données dans l'échelle de Mel (fréquentielle donc) vers l'échelle des temps. Le résultat de cette étape sera les MFCC proprement dit. Il suffit d'effectuer l'inverse de la transformée de Fourier. Dans la pratique, on effectue une transformée en Cosinus Discrète (DCT), ce qui revient au même puisque la transformée en Cosinus donne la partie réelle de la transformée de Fourier, or ici on a que des réels. Les d premiers coefficients cepstraux C_n peuvent être calculés directement à partir du logarithme des énergies E_i sortant d'un banc de F filtres par la formule suivante :

$$C_n = \sum_{k=1}^{F} \log(E_i).\cos\left[\frac{n.\pi}{F}.\left(k-\frac{1}{2}\right)\right], \qquad n = .1,....., d \tag{2.22}$$

Il s'agit d'une représentation compacte; à partir d'une vingtaine de filtres en échelle de Mel, une dizaine de coefficients cepstraux sont généralement considérés comme suffisants pour les expériences de reconnaissance de la parole.

Le coefficient C_0 qui est la somme des énergies n'est pas utilisé, il est éventuellement remplacé par le logarithme de l'énergie totale E_0 calculée dans le domaine temporel et normalisée. .Certains chercheurs calculent des paramètres BACC (*Bark Auditory Cepstral Coefficients*) en échelle Bark; mais les différences entre les deux échelles sont peu importantes.

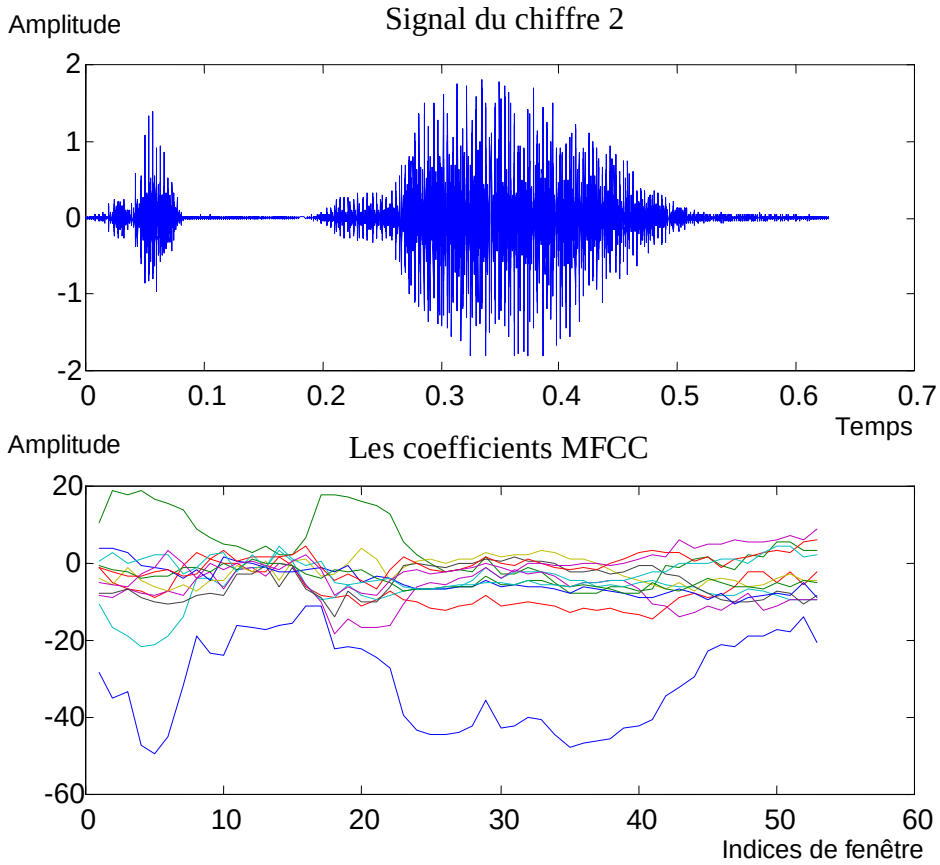


Figure 2.13 : Calcul de coefficients MFCC du chiffre **2**

2.3.7 Coefficients différentiels

Le but final de l'extraction des paramètres est de modéliser la parole, un phénomène très variable. Par exemple, même si elle a de l'importance, la simple valeur de l'énergie n'est pas suffisante pour donner toute l'information portée par ce paramètre. Il est donc souvent nécessaire de recourir à des informations sur l'évolution dans le temps de ces paramètres. Pour cela, les dérivées première et seconde sont calculées pour représenter la variation ainsi que l'accélération de chacun des paramètres. Même si la robustesse de la représentation obtenue est accrue, cela implique aussi de multiplier par 3 l'espace de représentation.

Soit C_n le coefficient cepstral d'indice k de la trame i, alors le coefficient différentiel ΔC_n correspondant est calculé sur $2.n._{\Delta}+1$ trames d'analyse par :

$$C_n(t) = \frac{\sum_{-n_\Delta}^{n_\Delta} i.C_n(t+i)}{\sum_{-n\Delta}^{n_\Delta} i^2} \qquad (2.23)$$

ces coefficients, appelés aussi coefficients delta, estiment la pente de la régression linéaire d'un coefficient à un instant donné. La dérivée de l'énergie ΔE est calculée de la même façon. Leur utilisation améliore sensiblement les performances des systèmes markoviens de reconnaissance.

Des coefficients différentiels du second ordre peuvent aussi contribuer à l'amélioration du système. Ces coefficients $\Delta.\Delta C_n$ et $\Delta\Delta E$ sont calculés par régression linéaire des coefficients delta sur $n_{\Delta\Delta}$ trames (typiquement $n_{\Delta\Delta}$ =1 ou 2).

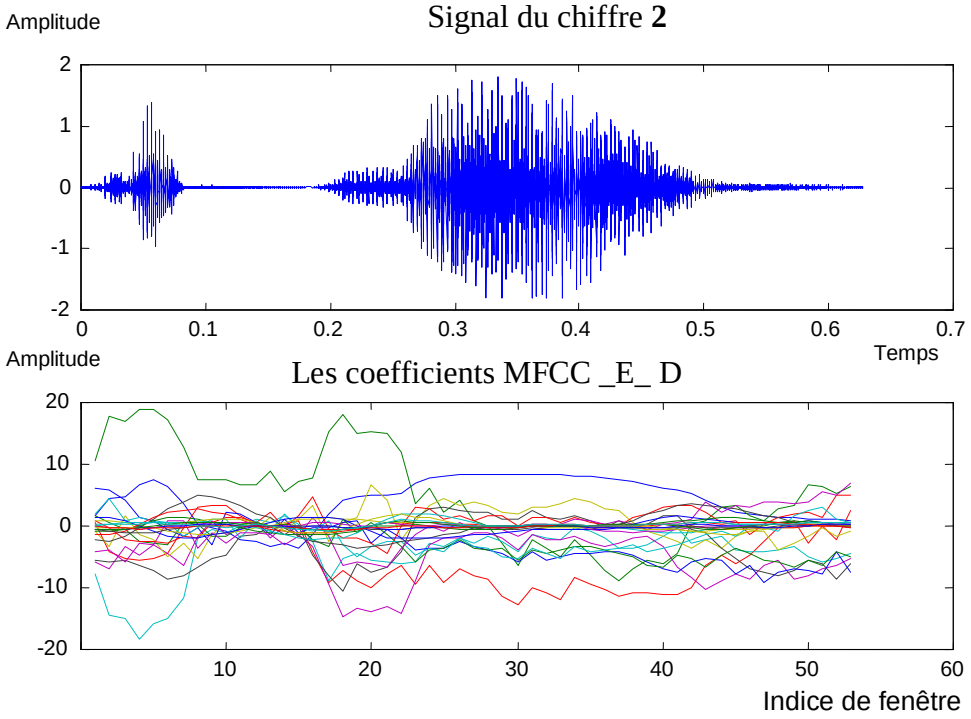


Figure 2.14 : Calcul des coefficients MFCC_E_D du chiffre **2**

3.2.8 Autres paramétrisations du signal

Nous n'énumérerons pas tous les types de paramètres employés dans le domaine de la recherche en parole car il y en a énormément et ce n'est pas le propos de notre mémoire. Pourtant, il est à noter que d'autres approches plus proches de l'audition humaine. Le codage PLP (Perceptual Linear Predictive) [B21] est un exemple d'intégration de connaissances sur la perception humaine pour la conception d'extracteurs de paramètres efficaces, le RASTA-PLP [B22], version approfondie de PLP. Cette liste ne se veut pas exhaustive mais permet d'avoir un aperçu des différents paramètres qu'il est possible.

2.4 Conclusion

Les paramètres traditionnels (LPC, LPCC, MFCC) sont issus de connaissances sur la production et/ou la perception humaine. Dans le cadre du modèle LPC, on cherche à modéliser le conduit vocal alors que le codage MFCC cherche à modéliser l'oreille.

Le choix des paramètres est important car il conditionne toute la méthodologie mise en oeuvre pour la reconnaissance. Il est irréversible et ne peut pas donc être remis en cause au cours du système. . Nous avons choisi dans cette phase pour la conception de notre système les paramètres les plus couramment utilisés qui sont les LPC, MFCC et les LPCC comme paramètres de référence dans l'étude comparative avec les paramètres provenant de la transformée en ondelettes.

CHAPITRE 3 :

PARAMÉTRISATION DU SIGNAL PAROLE A PARTIR DE REPRÉSENTATIONS EN ONDELETTES

Chapitre 3 :

Paramétrisation du Signal Parole à Partir de Représentations en Ondelettes

3.1 Introduction

La transformée en ondelettes d'un signal permet de représenter le signal sur un espace bidimensionnel appelé le plan temps-échelle, fournissant sur le signal des informations conjointes en temps et en fréquence. Le pavage du plan temps-fréquence induit par cette transformée a pour particularité de permettre une résolution temporelle fine aux hautes fréquences et une résolution fréquentielle fine aux bases fréquences. Cette propriété permet souvent une analyse intéressante du signal mais reste rigide. La décomposition en paquets d'ondellettes est une extension de la transformée en ondelettes discrète permettant de choisir le pavage du plan temps- fréquence selon le signal traité et selon un critère répondant aux contraintes de l'application.

Nous allons présenter dans ce chapitre un rappel sur les ondelettes et l'algorithme de décomposition en ondelettes discrètes, et en paquets d'ondelettes, puis, nous présentons les différentes techniques de paramétrisation utilisées à partir de représentations en ondelettes
Nous ne développerons pas ici l'aspect théorique de cette transformée mais nous nous intéressons au côté applicatif.

3.2 Transformée en ondelettes

Dans ce rappel, nous nous sommes limités aux notions nécessaires pour notre application

3.2.1 Transformée en ondelettes continue

La transformée (décomposition) en ondelettes continue $W(a,b)$ d'une fonction $f(t)$ suivant l'ondelette $\Psi(t)$ s'écrit :

$$W(a,b) = \int_{-\infty}^{+\infty} f(t).\Psi^{*}{}_{a,b}(t)dt \qquad (3.1a)$$

$$= \left\langle f, \Psi_{a,b} \right\rangle \qquad (3.1b)$$

$\left\langle f, \Psi_{a,b} \right\rangle$: Produit scalaire de f par $\Psi_{a,b}$

avec : $$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right) \quad (3.2)$$

$\Psi_{a,b}(t)$: Ondelettes filles

a : Facteur d'échelle (ou de délation), relié à la notion de fréquence

b : Le décalage (translation), relie à la notion de position temporelle

Plus a est grand, plus l'ondelette est dilatée. Par conséquent les grandes valeurs de a seront logiquement associées aux bases fréquences, les plus petites aux hautes fréquences.
La décomposition d'un signal en ondelettes est ainsi paramétrable : les ondelettes filles sont, à l'aide d'une fonction d'échelle, réglables en dilatation et en translation de l'ondelette mère.
La formule 3.1b peut être interprétée comme une projection du signal sur une famille de fonctions analysantes $\Psi_{a,b}$ construite à partir d'une fonction mère Ψ suivant l'équation 3.2.

Cette ondelette mère doit posséder certaines caractéristiques dont les plus importantes sont :

- $\Psi(t)$: doit être localisée (c.-à-d. s'annuler rapidement en dehors d'un intervalle donné).
- $\Psi(t)$: doit être de moyenne nulle afin de rendre compte la forme de signal à analysé.
- $\Psi(t)$: doit être oscillante afin de s'adapter aux différentes formes de signal.
- $\Psi(t)$: doit être régulière que possible.

pour plus de détails voir [B32].

La transformation est en théorie infiniment redondante puisque l'ondelette est translatée de manière continue ; cependant il existe des méthodes pour diminuer cette redondance : l'une de ces méthodes consiste en l'emploi de la transformée en ondelettes discrète.

3.2.2 Transformée en ondelettes discrète

Contrairement à la transformée continue, dans laquelle l'ondelette est dilatée et translatée de manière continue, la transformée en ondelettes discrète translate et dilate l'ondelette selon des valeurs discrètes [B28].

Ces coefficients a et b seront discrétisés de la manière suivante :

$a = a_0^n \;\; et \;\; b = m.b_0.a_0^n$ Avec $a_0 \rangle 1 \;\; et \;\; b_0 \rangle 0$ fixés et appartenant à Z.

$$W(a,b) = \frac{1}{\sqrt{a_0^n}} \int_{-\infty}^{+\infty} f(t).\Psi^{\bullet}\left(\frac{1}{a_0^n}.t - m.b_0\right).dt \quad (3.3)$$

En pratique, pour réduire la redondance, on choisit a_0 *et* b_0 de telle sorte que les fonctions $\Psi_{a,b}$ forment une base orthonormée. En général on prend :

$a_0 = 2$ *et* $b_0 = 1$: on parle de la transformée dyadique

3.2.3 Algorithme de décomposition (algorithme de Mallat)

L'algorithme pyramidal de Mallat réalise la décomposition d'un signal en une somme d'ondelettes ainsi que sa reconstitution, suivant un processus itératif rapide . La transformée en ondelttes a la même longueur que le signal.

Le signal décomposé est constitué à chaque étape de deux parties figure 3.1 :

- la partie approximation : signal d'approximation
- la partie détail : signal de détail.

Les échantillons des signaux de détails sont appelés coefficients d'ondelettes (ou de détails) et les échantillons du signal d'approximation sont appelés coefficients de fonction d'échelle (ou d'approximation).

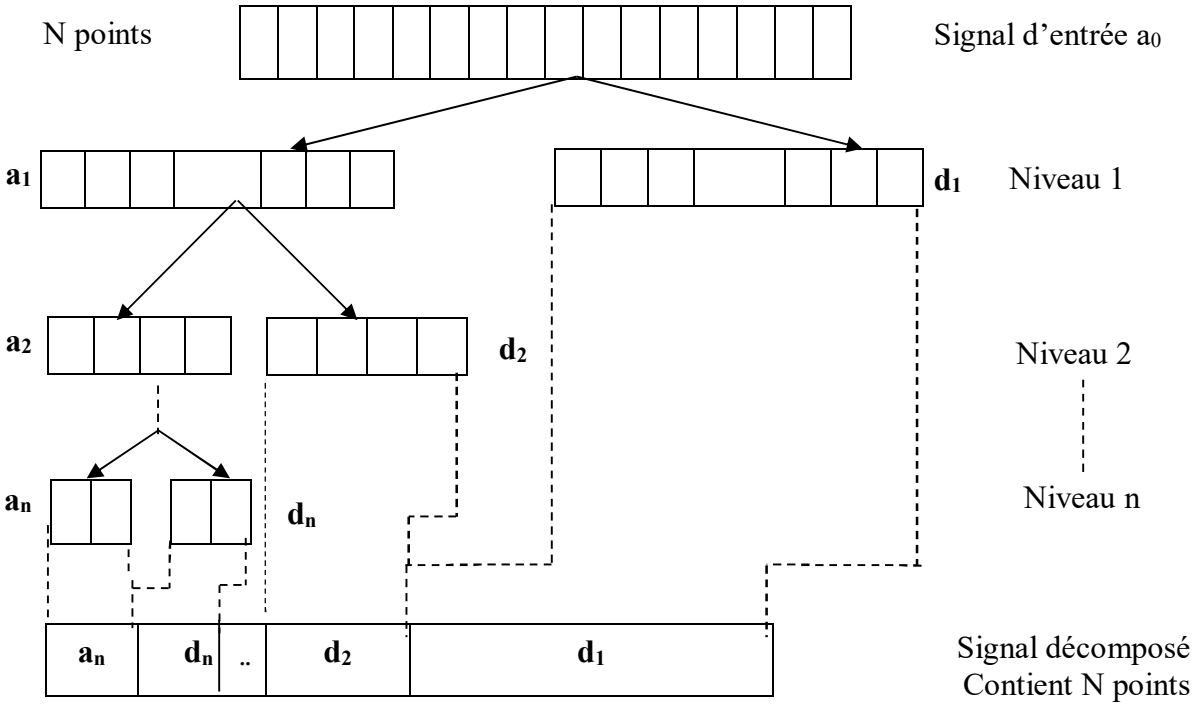


Figure 3.1. : Schéma de la décomposition

Stéphane Mallat [B23], [B29] a montré qu'une telle décomposition peut être obtenue par des filtrages passe-bas et passe-haut du signal temporel. Après chaque filtrage le signal est sous échantillonné d'un facteur de deux. Ce processus de décomposition est itéré Sur les résultats du filtrage passe-bas jusqu'à l'obtention du nombre de bandes de fréquence désiré. Au final, le signal est décomposé en coefficients d'approximation et en coefficients de détail. Les coefficients d'approximation correspondent à des moyennes locales du signal. Les coefficients de détail, appelés coefficients d'ondelettes, représentent les différences entre deux moyennes locales successives, c'est à dire entre deux approximations successives du signal, l'ensemble constituant une pyramide de filtres figure 3.2.

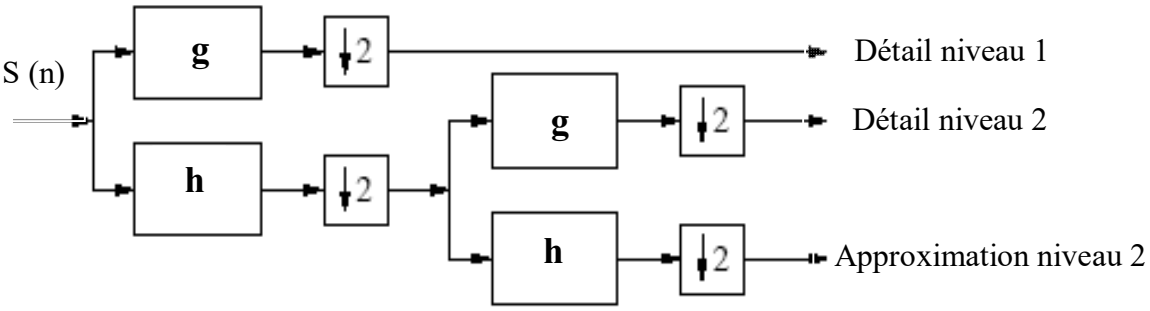


Figure 3.2. : Schéma de l'algorithme de la décomposition à l'aide de bancs de filtres (deux niveaux)

$s(n)$: correspondant au signal d'origine au niveau 0

g : Filtre passe-haut h : Filtre passe-bas

Les filtres h et g dépendent des ondelettes utilisées. De plus, le lecteur trouvera des informations sur le calcul et la synthèse de ces filtres dans [B32], [B33].

Les coefficients du niveau $j+1$ sont donnés par les équations

$$d_{j+1}(n) = \sum_k g(2.n-k).a_j(k) \tag{3.4}$$

$$a_{j+1}(n) = \sum_k h(2.n-k).a_j(k) \tag{3.5}$$

avec : $a_0(k) = s(k)$

d_{j+1} : Coefficients d'ondelettes au niveau (à l'échelle) $j+1$

a_{j+1} : Coefficients d'approximation au niveau $j+1$

La résolution en sortie de chaque paire (ou banc) de filtres étant deux fois inférieure à la résolution d'entée, on parle d'analyse multi- résolution dyadique.

3.2.4 Paquets d'ondelettes

Les paquets d'ondelettes constituent une généralisation de l'analyse multi-résolution [B23] comme pour la décomposition en ondelettes, il s'agit toujours de décomposer le signal au moyen d'un filtre passe-bas et d'un filtre passe-haut complémentaires. La différence réside dans le fait que les différents signaux de détails vont également faire l'objet à chaque itération d'une décomposition selon le même principe. La figure 3.3 illustre la procédure en question.

Une cellule de cette décomposition qui engendre deux autres cellules sur le niveau immédiatement inférieur s'appelle ' *père*', tandis que les cellules engendrées s'appellent cellules '*enfants*'

On appelle "coefficients de Paquets d'ondelettes " l'intégralité des coefficients obtenus (approximations et détails). Ces coefficients calculés à partir de niveau j et de sous espace p sont donnés par les équations suivantes :

$$w_{j+1}^{2.p}(n) = \sum_k h(2.n-k).w_j^p(k) \tag{3.6}$$

$$w_{j+1}^{2p+1}(n) = \sum_k g(2.n-k).w_j^p(k) \tag{3.7}$$

avec : $w_0^0(n) = s(n)$

La figure 3.3 représente l'algorithme pyramidal étendu permettant d'obtenir ces coefficients.

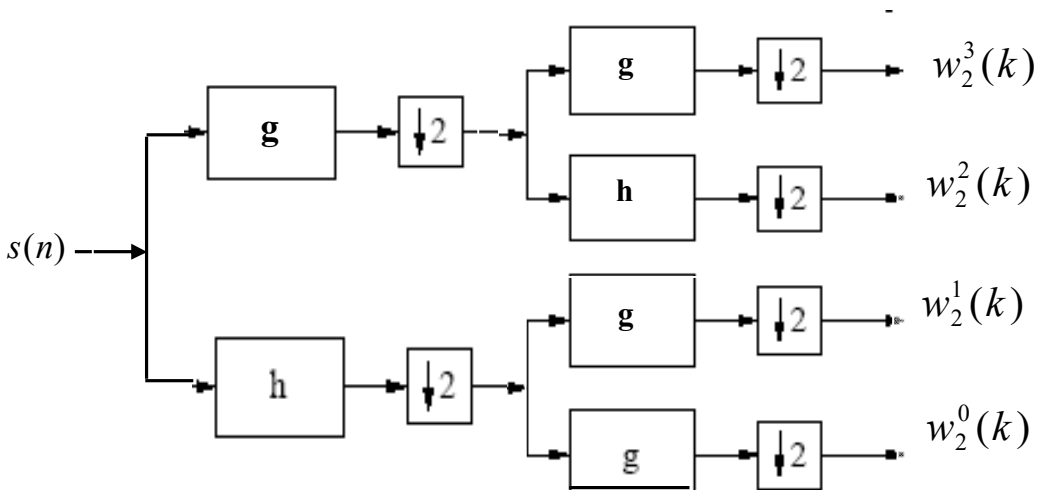


Figure 3.3 : Schéma de l'algorithme de la décomposition en Paquets d'ondelettes d'un signal (Deux niveaux)

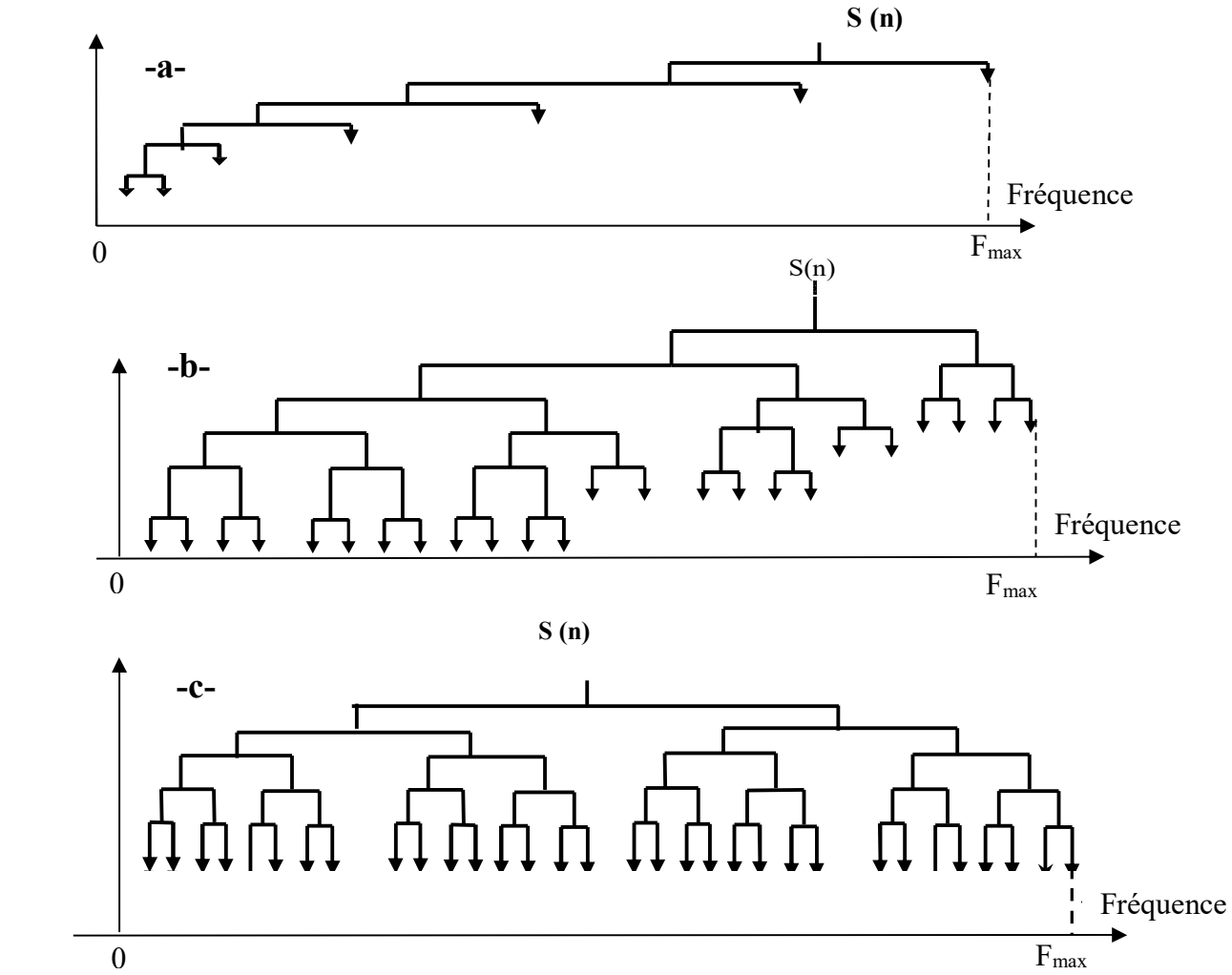
La décomposition totale d'un signal en paquets d'ondelettes fournit une représentation redondante. Chaque niveau de la décomposition contient toute l'information du signal .Ainsi, seules certaines combinaisons de paquets, constituent des représentations non redondantes du signal. Le choix de ces combinaisons se fait selon le signal traité et selon un critère répondant aux contraintes de l'application. Les espaces correspondants sont définis à partir de la notion d'arbre admissible AWP (Admissible Wavlet Packet) [B23]

3.3 Ondelettes en reconnaissance de la parole

L'analyse par ondelettes peut être interprétée comme une analyse de Fourier à fenêtre glissante. En effet, la particularité des ondelettes c'est que la durée de leur support temporel varie selon la bande de fréquences a analysé. Par conséquent, elle permet d'observer un phénomène basse fréquence sur une longue durée et réciproquement, d'observer un phénomène haute fréquence sur une courte durée. [B13].

La motivation d'utiliser les ondelettes dans la paramétrisation du signal parole est qu'elles mesurent les variations temporelles des composantes spectrales et qui permettent donc d'extraire les caractéristiques temporelles et fréquentielles du signal. De plus, la décomposition multi-bandes, qu'effectue la transformée en ondelettes dyadique, se rapproche fortement de ce que fait l'oreille humaine, c'est à dire qu'elle ressemble à une échelle logarithmique. Par rapport aux coefficients MFCC, les ondelettes ont une résolution temps fréquence différente, permettent une représentation plus compacte du signal, possèdent un ensemble plus riche de fonctions de base et sont plus robustes à la non stationnarité du signal et à ses distorsions.

La plupart des techniques qui emploient l'analyse d'ondelettes dans la paramétrisation du signal parole tente à remplacer l'analyse de Fourier classique par l'analyse par ondelettes, ou de faire une approximation de l'échelle Mel par la transformée en paquets d'ondelettes. Plusieurs structures de décomposition sont proposées figure 3.4.



Avec :

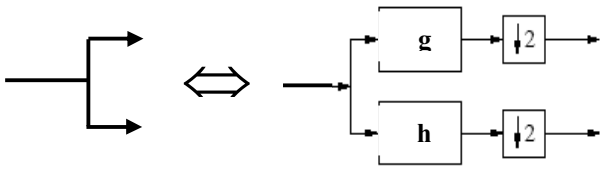


Figure 3.4 : Différentes décompositions d'ondelettes **a** : Structure dyadique, **b :** Structure de paquets d'ondelettes admissibles [B24], **c :** pleine décomposition de paquets d'ondelettes.

La structure binaire de paquets d'ondelettes admissibles (AWP : Admissible Wavelet Packet) présentée par Farooq et Datta [B24], donne une décomposition en bandes de fréquences similaires à celle de l'échelle de Mel. Elle permet de représenter le signal avec une certaine redondance. Cette structure est illustrée par la figure 3.4b.

3.4 Extraction de Paramètres

Après avoir réalisé la décomposition et le choix de l'arbre de décomposition (voir figure 3.4), dans chaque bande de fréquence, nous calculons à partir des coefficients d'ondelettes, différents paramètres par échantillon (normalisés par le nombre d'échantillons).
Les paramétrisations réalisées sont constituées d'un seul coefficient dans chaque bande de fréquence .

- . **Logarithme de l'énergie par échantillon : E**

$$p_j = 20\log\left(\frac{1}{N_j}\sum_{k=1}^{N_j} d_j^2(k)\right) \qquad j = 1,2,.....,J \tag{3.8}$$

où $d_j(k)$ dénote le coefficient d'ondelettes à la position k et au niveau j, N_j le nombre de coefficients au niveau j, J nombre des niveaux de décomposition et p_j le vecteur paramètres au niveau j

Nous employons non seulement les coefficients de détail mais également les coefficients d'approximation

$$g_J = 20\log\left(\frac{1}{N_J}\sum_{k=1}^{N_J} a_J^2(k)\right) \tag{3.9}$$

où $a_J(k)$ dénote le coefficient d'approximation à la position k et au niveau J.

- **Logarithme de l'amplitude moyenne : A-M**

$$p_j = 20\log\left(\frac{1}{N_j}.\sum_{k=1}^{N_j}\left|d_j(k)\right|\right) \tag{3.10}$$

- **Paramètres hiérarchiques**

Les paramètres hiérarchiques correspondent au calcul de paramètres au centre de la fenêtre d'analyse en utilisant le même nombre de coefficients quelque soit la bande [B35].

:

- **Logarithme de l'énergie hiérarchique par échantillon : E_H**

$$p_j = 20\log\left(\frac{1}{N_j}\sum_{k=\left(N_j-N_J\right)/2}^{\left(N_j+N_J\right)/2} d_j^2(k)\right) \tag{3.11-a}$$

➢ **Logarithme de l'amplitude moyenne hiérarchique : A-M_H :**

$$p_j = 20\log\left(\frac{1}{N_j}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2}\left|d_j(k)\right|\right) \qquad (3.11\text{-b})$$

- **Logarithme de l'énergie Teager par échantillon : T_E :**

Nous utilisons ici l'opérateur TEO (Teager Energy Operator) introduit par Kaiser [B30]. Cet opérateur permet d'obtenir des paramètres robustes :

$$p_j = 20\log\left(\frac{1}{N_j}\sum_{k=1}^{N_j}\left|\left(d_j^2(k) - d_j(k-1).d_j(k+1)\right)\right|\right) \qquad (3.12)$$

- **Seuillage des coefficients**

Les coefficients d'ondelettes correspondent aux détails de signal. Lorsqu'un détail est faible, il peut être ignoré sans que cela affecte les données de manière visible. Le seuillage des coefficients d'ondelettes est donc un bon moyen d'éliminer les détails les plus faibles, que l'on considère comme du bruit

$$p_j = \left\{ d_j(k) \right\} \qquad (3.13\text{ a})$$

avec $p_j = \left| d_j(k) \right| \rangle \lambda$ λ : Le seuil (3.13 b)

Les deux coefficients d'ondelettes les plus importants en valeur absolue sont utilisés comme paramètres, les autres étant mis à zéro [B27].

- **Coefficients cepstraux multi résolution**

Les paramètres proposés issus de la transformée en ondelettes sont fondés sur le principe des coefficients MFCC, après la décomposition en sous bandes, la transformée en cosinus discrète est appliquée aux logarithmes des énergies de sous bandes [B31]. Quelques coefficients de la DCT choisi parmi les premiers sont utilisés comme paramètres acoustiques (figure 3.5)

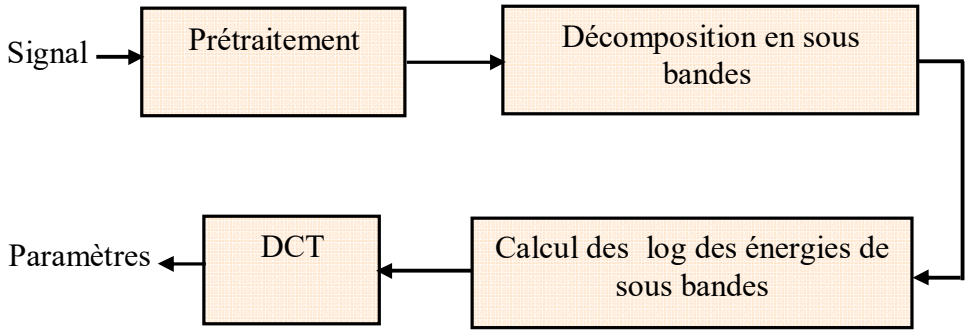


Figure 3.5 : Processus de calcul des coefficients fondé sur le principe des coefficients MFCC

Une autre technique proposée basée sur la projection des énergies d'un banc de filtre Mel sur une base d'ondelette (figure 3.6). Cette technique permet de calculer les coefficients cepstraux à différentes résolutions de sous bandes [B51] [B35].

.

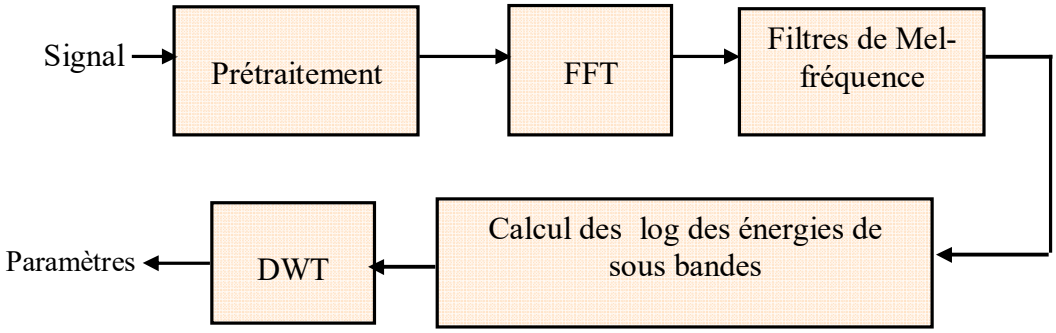


Figure 3.6 : Processus de calcul des coefficients fondé sur la projection des énergies d'un banc de filtre Mel sur une base d'ondelette.

3.5 CONCLUSION

Nous avons parcouru dans ce chapitre les techniques utilisées pour extraire les paramètres acoustiques à partir de la représentation en ondelettes. Malgré le succès des méthodes traditionnelles (LPC, LPCC, MFCC,…), on peut espérer que la mise en oeuvre de techniques de paramétristion proposées permettra d'améliorer les performances de notre système de reconnaissance.

CHAPITRE 4 :

MODÈLES DE MARKOV CACHÉS

Chapitre 4

Modèles de Markov Cachés

4.1 Introduction

Le problème de la reconnaissance de parole peut être formulé dans ces termes : Comment modéliser au mieux des unités représentatives du signal de parole ? Il existe en fait deux types de modélisation possible des propriétés d'un signal donné :

- Les modèles déterministes, qui exploitent les propriétés intrinsèques du signal,
- Les modèles statistiques, qui caractérisent les propriétés statistiques du signal.

Dans ce travail, nous avons opté pour des modèles statistiques : les modèles de Markov cachés, appelés aussi HMM, l'abréviation anglaise HMM sera utilisée couramment pour parler des modèles de Markov cachés. Les HMM se sont imposés comme une technique prédominante en reconnaissance de la parole ces dernières années. Nous allons présenter dans ce qui suit les bases nécessaires à l'utilisation de ce type de modèle pour la reconnaissance automatique de la parole. Ces modèles se sont avérés les mieux adaptés aux problèmes de la reconnaissance de la parole et la quasi-totalité des outils de reconnaissance de la parole disponibles actuellement sur le marché sont basés sur cette technologie.

4.2. Présentation des Modèles de Markov Cachés

4.2.1. Définition

Un modèle de Markov est un automate probabiliste d'états finis constitué de N états. Un processus aléatoire se déplace d'état en état à chaque instant, et on note q_t le numéro de l'état atteint par le processus à l'instant t. L'état réel q_t du processus n'est pas directement observable, on dit qu'il est "caché" mais le processus émet après chaque changement d'état un symbole discret o_t qui appartient à un alphabet fini de n_V symboles $V = \{ v_k \}_{1 \le k \le n_V}$.

Dans le cas d'un processus markovien du premier ordre, la probabilité de passer de l'état *i* à l'état *j* à l'instant *t* et d'émettre le symbole v_k ne dépend ni du temps, ni des états aux instants

précédents. Un modèle de Markov caché ou HMM est alors défini par l'ensemble des paramètres (S, A, B, π) [B38]:

- $S = \{s_i\}_{1 \le i \le N}$ l'ensemble des N états.
- $A = \{a_{ij}\}_{1 \le i, j \le N}$ l'ensemble des probabilités de transition entre les états *i* et *j*:

$$a_{ij} = p(q_t = j / q_{t-1} = i) \qquad \text{pour un HMM d'ordre 1} \tag{4.1}$$

avec $a_{ij} \ge 0$ et $\sum_{j=1}^{N} a_{ij} = 1$

- $V = \{v_k\}_{1 \le k \le n_V}$ l'ensemble des n_V symboles observables,
- $B = \{b_j(k)\}_{1 \le j \le N, 1 \le k \le n_V}$ l'ensemble des probabilités d'émission du symbole v_k lors de l'arrivée dans l'état *j*:

$$b_j(k) = p(o_t = v_k / q_t = j) \tag{4.2}$$

avec $b_j(k) \ge 0 \quad et \quad \sum_{k=1}^{n_V} b_j(k) = 1 \quad \forall\, j$

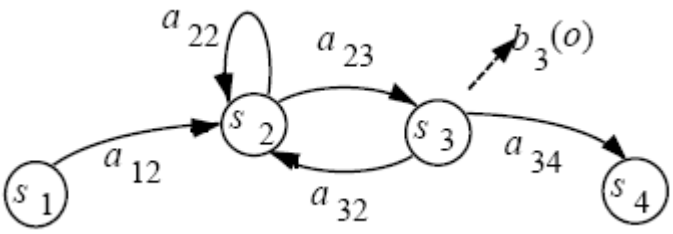
- $\pi = \{\pi_i\}_{1 \le i \le N}$ l'ensemble des probabilités initiales des états :

$$\pi_i = p(s_i = q_1) \tag{4.3}$$

avec $\forall\, i \quad \pi_i \ge 0, \quad \sum_{i=1}^{N} \pi_i = 1$

on appelle $\lambda = (A, B, \pi)$ les paramètres complets du modèle HMM

La figure 3.1 donne l'exemple d'un modèle autorisant le bouclage sur le deuxième état et des oscillations entre le deuxième et le troisième état; les transitions non représentées par un arc sont supposées de probabilité nulle. Enfin, l'émission d'une observation O par le troisième état suit la probabilité $b_3(o)$).



La figure 4.1 : Exemple de modèle de Markov.

4.2.2. Types de Modèles de Markov Cachés

Il existe deux types principaux de modèles de Markov cachés, le modèle ergodique et le modèle gauche droite [B39]

- **Le modèle ergodique**

C'est un modèle sans contrainte où toutes les transitions d'un état vers un autre sont possibles comme le montre la figure 3.2. On peut remarquer que la connaissance de π est primordiale pour le choix de l'état de départ

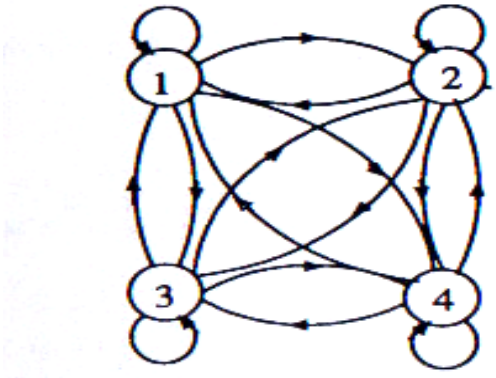


Figure 4.2 : Le modèle ergodique

- **Le modèle gauche-droite**

C'est un modèle contenant des contraintes sur les transitions (existence des transitions $(i, j) / a(i, j) = 0$. Il existe deux types de ce modèle qui sont le modèle parallèle et le modèle séquentiel comme le montre la figure 4.3. On peut remarquer que le modèle gauche-droite couvre souvent la plupart des applications et de ce fait, il est largement suffisant en parole.

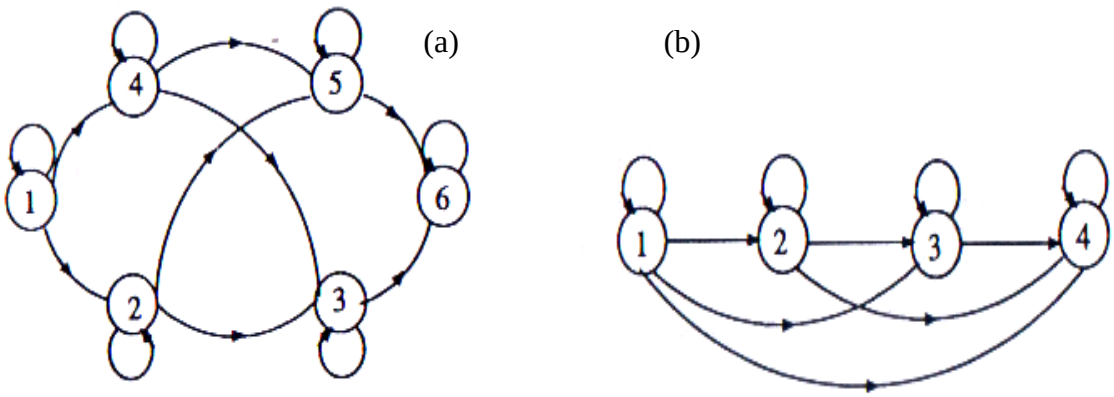


Figure 4.3 : Le modèle gauche-droite : (a) le modèle séquentiel, (b) le modèle parallèle.

4.2.3 Procédure de génération des observations

La procédure de génération des observations de longueur T par HMM est une procédure itérative en quatre états :

- La première étape est consacrée au choix de l'état initial :

  $t=1$, choisir l'état initial $q_1 = s_i$ avec une probabilité π_i

- La deuxième étape est relative à l'observation dans l'état sélectionné :

  Choisir l'observation $o_t = v_k$ avec une probabilité $b_i(k)$

- La troisième étape est relative à la sélection de l'étape suivante :

  $q_{t+1} = s_j$ Avec une probabilité a_{ij}

- La quatrième étape concerne le changement effectif d'état :

  $t \leftarrow t+1$

  Si $t \langle T$ **alors** retourner à l'étape 2 **Sinon** fin

  Fin si

4.2.4. Les problèmes à résoudre [B40]

La réalisation par la machine λ d'un processus markovien de durée *T* est décrite par:

- $Q = (q_1 .....q_T)$ un chemin *a priori* caché parmi les *N* états.
- $O = (o_1 .....o_T)$ une suite d'observations appartenant à l'alphabet de n_V symboles.

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

- **Problème 1:** " Evaluer une observation selon un modèle"
  Le problème de l'estimation des probabilités peut être énoncé de la façon suivante étant donné un modèle de Markov $\lambda = (A, B, \pi)$ et la séquence d'observation $O = (o_1 .....o_T)$ comment calculer la probabilité que la suite des observations ait été émise par un modèle, lorsque plusieurs modèles existent, cette évaluation permet le choix du modèle le plus probable.
- **Problème 2** Problème de décodage"
  Etant donné la séquence d'observations $O = (o_1 .....o_T)$ et le modèle $\lambda = (A, B, \pi)$ comment choisir la séquence d'états cachés la plus probable d'un modèle ayant produit les observations.
- **Problème 3 :** " L'apprentissage des paramètres d'un modèle "
  A partir d'un modèle donné *a priori* 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 simplifié de la reconnaissance de mots isolés. Les modèles de mots émettent des observations du signal de parole. L'évaluation de la probabilité d'émission d'un mot par différents modèles permet alors l'identification d'un mot inconnu. Le chemin optimal parmi les états cachés sert pour l'initialisation des modèles avant leur apprentissage, mais la procédure est surtout utile pour le décodage de la parole continue.

4.2.5. Solution des trois problèmes des HMM

4.2.5.1 Evaluation de la probabilité d'observation

Il existe plusieurs manières d'évaluer la probabilité d'observation. Nous décrivons les deux principales dans la suite.

- Evaluation directe

La probabilité que la suite d'observations $O=(o_1.....o_T)$ soit émise par la machine λ, est égale à la somme des probabilités conjointes de l'observation O et du chemin Q pour l'ensemble E_Q de tous les chemins possibles de longueur T:

$$p(O/\lambda)=\sum_Q p(O,Q/\lambda)=\sum_Q p(O/Q,\lambda).p(Q/\lambda)\ , \tag{4.4}$$

or, on à les relations :

$$p(Q/\lambda)=\pi_{q1}.a_{q1q2}.a_{q2q3}.....a_{qT-1qT}\quad , \tag{4.5}$$

$$p(O/Q,\lambda)=b_{q1}(o_1).b_{q2}(o_2).....b_{qT}(o_{T)}\quad , \tag{4.6}$$

d'où $$p(O/\lambda)=\sum_Q \pi_{q1}.b_{q1}(o1).a_{q1q2}.b_{q2}(o2).a_{q2q3}.....a_{qT-1qT}.b_{qT}(o_T)\quad , \tag{4.7}$$

Cependant, cette formule possède une complexité, bien trop importante. Aussi, l'algorithme "*Forward-Backward* " a été conçu pour diminuer la complexité de calcul.

- Evaluation par les fonctions *Forward-Backward*

Dans cette approche, on considère que l'observation peut se faire en deux temps, d'abord, émission du début de l'observation $O=(o_1.....o_t)$ en aboutissant à l'état q_i au temps t, puis, émission de la fin d'observation $O=(o_t.....o_T)$ sachant que l'on part de l'état q_i au temps t .dans ce cas l'évaluation de l'observation est égale à :

$$p(O/\lambda)=\sum_{qi}\alpha(t,q_i).\beta(t,q_i) \tag{4.8}$$

où $\alpha(t,q_i)$ est la probabilité d'émettre le début $O=(o_1.....o_t)$ et d'aboutir à q_i à l'instant t, et $\beta(q_i,t)$ est la probabilité d'émettre la fin $O=(o_t.....o_T)$ sachant que l'on part de q_i à l'instant t. Le calcul de α se fait avec t croissant tandis que le calcul de β se fait avec t décroissant, d'où les expressions *Forward-Backward.*

on notera par la suite, $\alpha(t,q_i)$ par $\alpha_t(i)$ et $\beta(q_i,t)$ par $\beta_t(i)$.

L'algorithme *Forward*

L'algorithme **Forward** délivre deux informations : $p(O/\lambda)$ et $\alpha_t(i)$ Où

$$\alpha_t(i) = p(o_1, o_2, o_3, ..., o_t, q_i = s_i / \lambda) \quad (4.9)$$

.

l'algorithme *Forward* se détermine ainsi :

- **Initialisation**

$$\alpha_1(i) = \pi_i . b_i(o_1) \qquad 1 \le i \le N \ ; \quad (4.10\ a)$$

- **Induction**

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i).a_{ij}].b_j(o_{t+1}) \qquad 1 \le t \le T-1 \ , \ 1 \le j \le N \ ; \quad (4.10\ b)$$

- **Terminaison**

$$p(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i) \ . \quad (4.10\ c)$$

La procédure *Forward* détermine $p(O/\lambda)$ en utilisant exactement la même formule (4.7) développe précédemment, mais de manière inductive, ce qui diminue fortement le temps et le nombre de calculs.

L'algorithme *Backward*

$$\beta_t(i) = p(o_{t+1}, o_{t+2}, o_{t+3}, ..., o_T /, q_T = s_i, \lambda) \ . \quad (4.11\ a)$$

on déduit $\beta_t(i)$ de $\beta_{t+1}(i)$ de manière inductive comme suit :

- **Initialisation**

$$\beta_T(i) = 1 \qquad 1 \le i \le N \quad ; \quad (4.11\ b)$$

- **Induction**

$$\beta_t(i) = [\sum_{j=1}^{N} \beta_{t+1}(j).a_{ij}].b_j(o_{t+1}) \qquad T-1 \ge t \ge 1 \ , \ 1 \le i \le N \ ; \quad (4.11\ c)$$

L'algorithme Backward est conçu de manière similaire à l'algorithme Forward. . Il fonctionne également de manière inductive

La probabilité d'observation est obtenue en prenant les valeurs de α et de β à un instant t quelconque

$$p(O/\lambda) = \sum_{qi} \alpha_t(i).\beta_t(i) \quad , \tag{4.12}$$

Les variables *Forward* et *Backward* sont utilisés conjointement lors de l'apprentissage des modèles. Ce problème de décodage sera précisément étudié dans les parties suivantes.

4.2.5.2 Calcul de chemin optimal

La procédure d'estimation *Forward* ou *Backward* fournit la probabilité d'émission des observations cumulées 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* cherche la séquence d'états cachés la plus probable et calcule la probabilité d'émission le long de ce chemin. La probabilité ainsi estimée néglige les chemins moins probables, et une reconnaissance à partir de cette probabilité est sous optimale par rapport à l'estimation *Forward* ou *Backward*. Mais la procédure fournit une segmentation du signal qui peut être exploitée pour l'apprentissage initial des modèles ainsi que pour le décodage de la parole continue.

Algorithme de viterbi : [B41] [B42]

Pour trouver le meilleur chemin $Q = \{ q_1, q_2, ...q_T \}$ pour une suite d'observations $O = \{o_1, o_2, ...o_T\}$, on définit la fonction

$$\delta_i(t) = \max_{q1.....qt-1} p\ (q_1 q_2 .....q_t = i, o_1 o_2 .....o_t / \lambda) \tag{4.13}$$

$\delta_t(i)$ Est la probabilité du meilleur chemin amenant à l'état s_i à l'instant t, en étant guidé par les t premières observations. En gardant trace, lors du calcul, de la suite d'états qui donne le meilleur chemin amenant à l'état s_i à l'instant t.

L'algorithme de viterbi se construit ainsi :

- **Initialisation**

$$\delta_1(i) = \pi_i b_i(o_1) \qquad 1 \le i \le N \tag{4.14 a}$$

$$\psi_1(i) = 0 \tag{4.14 b}$$

- **Recursion**

$$\delta_t(j) = \underset{1 \le i \le N}{Max}\left[\delta_{t-1}(i).a_{ij}\right] b_j(o_t) \qquad 2 \le t \le T \;,\; 1 \le j \le N \tag{4.14 c}$$

$$\psi_t(j) = \underset{1 \le i \le N}{Arg\max}[\delta_{t-1}(i)a_{ij}] \qquad 2 \le t \le T \;,\; 1 \le J \le N \tag{4.14 d}$$

- **Terminaison**

$$p^* = \underset{1 \le i \le N}{Max}[\delta_T(i)] \tag{4.14 e}$$

$$q_T^* = \underset{1 \le i \le N}{Arg\max}[\delta_T(i)] \tag{4.14 f}$$

- **Chemin**

$$q_t^* = \Psi_{t+1}^*(q_{t+1}^*) \qquad \text{t =T-1, T-2,…….,1} \tag{4.14g}$$

4.2.5.3 L'apprentissage des paramètres d'un modèle

Le troisième problème consiste à trouver une méthode pour ajuster les paramètres du modèle $\lambda = (\pi, A, B)$ afin de maximiser la probabilité d'une séquence donnée. Il s'agit d'une estimation sur le critère du maximum de vraisemblance (*Maximum Likelihood Estimation* ou MLE).

- Maximum de vraisemblance

L'estimation par maximum de vraisemblance (*Maximum Likelihood Estimation* ou *MLE*) consiste à choisir les paramètres du modèle afin de rendre maximale la probabilité d'émission des observations *O* par le modèle :

$$\lambda^* = \arg\max_{\lambda} p(O/\lambda) \tag{4.15}$$

Une résolution analytique directe n'est pas possible, mais les formules de *Baum-Wel* permettent une ré-estimation itérative des paramètres du modèle en appliquant ce critère. A la suite de la ré-estimation des paramètres du modèle λ^*_n, le nouveau modèle λ^*_{n+1} vérifie:

$$p(O/\lambda^*_{n+1}) \rangle\ p(O/\lambda^*_n) \tag{4.16}$$

La convergence vers un optimum local est démontrée, mais les valeurs initiales des paramètres π *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.

- Algorithme de Baum -Welch

L'algorithme de Baum-Welch est un algorithme à apprentissage. Son but étant la maximisation de la vraisemblance d'un modèle, celui-ci modifie substantiellement les paramètres du modèle étudie afin d'augmenter sa vraisemblance. L'algorithme réalise son optimisation en ré-estimant les différents paramètres (A, B et π), suivant la (ou les) séquence(s) observée(s).

Pour décrire comment ré-estimer les paramètres du HMM, on définit la probabilité $\xi_t(i, j)$ qui représente la probabilité d'être à l'état i au temps t et de faire une transition à l'état j au temps $t+1$ étant donnée la séquence d'observation O et le modèle λ

$$\xi_t(i, j) = p(q_t = s_i, q_{t+1} = s_j \ / O, \lambda) \tag{4.17}$$

ce qui se réécrit :

$$\xi_t(i, j) = \frac{p(q_t = s_i \,,\, q_{t+1} = s_j \,, O / \lambda)}{p(O / \lambda)} \tag{4.18}$$

par définition des fonctions *forward-backward,* on en déduit *:*

$$\xi_t(i, j) = \frac{\alpha_i(t).a_{ij}.b_j(o_{t+1}).\beta_{t+1}(j)}{p(O / \lambda)} \tag{4.19 a}$$

$$= \frac{\alpha_t(i).a_{ij}.b_j(o_{t+1}).\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i).a_{ij}.b_j(o_{t+1}).\beta_{t+1}(j)} \tag{4.19 b}$$

dans laquelle, le numérateur est $p(q_t = S_i, q_{t+1} = S_j / O, \lambda)$ et la division par $p(O/\lambda)$ donne une mesure de probabilité.

on définit aussi la quantité :

$$\gamma_t(i) = p(q_t = s_i / O, \lambda) \tag{4.20}$$

soit $$\gamma_t(i) = \sum_{j=1}^{N} p(q_t = s_i \ , \ q_{t+1} = s_j \ / \ O, \lambda) \tag{4.21}$$

on déduit la relation

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{4.22}$$

où $\sum_t \xi_t(i, j)$ donne l'information sur le nombre de transitions à partir de l'état s_i à l'état s_j,

ou $\sum_t \gamma_t(i)$ donne l'information sur le nombre de fois possible d'être dans l'état s_i

des formules de ré-estimation convenable pour π , A, B peuvent être

π_i^* = Nombre fois possible d'être dans l'état s_i à l'instant t ;

a_{ij}^* = (nombre de transitions possibles de s_i à s_j / nombre de transitions possibles

à partir de s_i);

$b_j^*(k)$ = (nombre de fois possible d'être en s_j en observant v_k / nombre de fois possible

d'être en s_j).

soient : $$\pi_i^* = \gamma_1(i) \tag{4.23}$$

$$a_{ij}^* = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{4.24}$$

$$b_j^*(k) = \frac{\sum_{t=1, o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{4.25}$$

Il est important de remarquer que la structure du modèle est conservée au cours de la ré-estimation; les transitions initialement interdites entre deux états le restent:

$$a_{ij} = 0 \Rightarrow a_{ij}^* = 0 \tag{4.26}$$

Un important aspect de la procédure de ré-estimation est que les contraintes stochastiques sont automatiquement vérifiées à chaque itération.

$$\sum_{i=1}^{N} \pi_i^* = 1 \tag{4.27}$$

$$\sum_{j=1}^{N} a_{ij}^* = 1 \qquad 1 \le i \le N \tag{4.28}$$

$$\sum_{k=1}^{M} b_j^*(k) = 1 \tag{4.29}$$

l'algorithme de Baum-Welch se construit ainsi :

- Fixer les valeurs initiales :

  a_{ij}^0 , $b_j^0(k)$, π_i^o Pour $1 \le i, j \le N$, $1 \le k \le n_V$

- Calculer à l'aide des fonctions *Forward-Backward* :

  $\xi_t(i, j)$, $\gamma_t(i)$, $1 \le i, j \le N$ $1 \le t \le T-1$ et λ^* en utilisant les formules de ré_estimation

- Recommencer en deuxième étape jusqu'à certain point limite.

4.2.6 Densité d'observation continue dans les modèles de Markov cachés

Jusqu'à présent nous n'avons considéré que le cas où les observations prennent des valeurs dans un alphabet fini discret et nous pouvons donc utiliser une loi de probabilité discrète dans chaque état de modèle .une telle approche ne soit pas compatible avec les observations qui sont des signaux continus.

Les vecteurs de coefficients calculés sur une trame de signal sont des points dans un espace multidimensionnel continu. Il est possible de modéliser ces observations par des modèles de Markov à émissions discrètes au moyen de la quantification vectorielle (QV) [B44]. Cependant, la QV introduit des distorsions. Il est donc préférable d'utiliser des modèles de Markov cachés avec des densités d'observations continus. [B43]

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 probabilités 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 est fréquente:

$$b_j(O) = \sum_{k=1}^{G} C_{jk}.N(O, \mu_{jk}, \sigma_{jk}) \tag{4.30}$$

où μ_{jk} *et* σ_{jk} sont respectivement le vecteur de moyenne et la matrice de covariance, G est le nombre de gaussiennes C_{jk} le coefficient de pondération de la k^{eme} Gaussienne en état j et O est le vecteur qui doit être modélisé.

avec $$\sum_{k=1}^{G} C_{jk} = 1 \tag{4.31}$$

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

$$N(o, \mu, \sigma) = \frac{1}{\sqrt{(2.\pi)^d.\det(\sigma)}}.\exp\left(\left[(\frac{-1}{2}.(O-\mu)^T.\sigma^{-1}.(O-\mu))\right]\right) \tag{3.32}$$

Dans le cas d'une distribution mono-gaussienne, les formules de ré-estimation de la moyenne, matrice de covariance de la densité de probabilité sont données par les équations suivantes :

$$\mu_j^* = \frac{\sum_{t=1}^{T} \gamma_t(j).o_t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{4.33}$$

$$\sigma_j^* = \frac{\sum_{t=1}^{T} \gamma_t(j).(o_t - \mu_j).(o_t - \mu_j)^T}{\sum_{t=1}^{T} \gamma_t(j)} \tag{4.34}$$

ces formules peuvent être généralisées au cas de multi-gaussiennes (G gaussiennes)

$$\mu_{jk}^* = \frac{\sum_{t=1}^{T} \gamma_t(j,k).o_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{4.35}$$

$$\sigma^*_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k).(o_t - \mu_{jk})(o_t - \mu_{jk})^T}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{4.36}$$

$$C^*_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M} \gamma_t(j,k)} \tag{4.37}$$

où l'exposant T dénote le vecteur transposé et $\gamma_t(j,k)$ est la probabilité d'être dans l'état j à t avec la k^ième composante de la mixture calculée pour o_t par la formule.

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right].\left[\frac{c_{jk} N(o_t, \mu_{jk}, \sigma_{jk})}{\sum_{m=1}^{M} c_{jm} N(o_t, \mu_{jm}, \sigma_{jm})}\right] \tag{4.38}$$

Le terme $\gamma_t(j,k)$ généralisé de $\gamma_t(j)$ de l'équation (4.23) dans le cas de simple gaussienne, ou la densité discrète. Les formules de réestimation pour a_{ij} est identique de celui utilisé pour la densité d'observation discret (4.25).

4. 2. 7 Implémentation

Seuls les principes généraux des algorithmes ont été présentés. Leur mise en oeuvre nécessite quelques précisions. Ainsi, les probabilités calculées sont souvent trop faibles pour être manipulées directement, et il est préférable de transposer les calculs dans le domaine logarithmique ou d'introduire un facteur d'échelle. De plus, afin de permettre une estimation plus robuste des paramètres, les modèles sont généralement appris à partir de plusieurs séquences d'observation plutôt qu'une seule.

4.2.7.1 Facteur d'échelle [B42]

L'algorithme de Baum-Welch, contient à la fois des produits et des sommes dans les quantités qu'il calcule. En effet, cet algorithme d'apprentissage nécessite le calcul des probabilités partielles Forward $\alpha_t(i)$ et Backward $\beta_t(i)$ pour $1 \le t \le T$, $1 \le i \le N$

D'après les formules récursives de calcul de α et de β, il est clair que lorsque T croit, α et β tendent vers 0. Pour un grand nombre d'observations, la fonction de probabilité aura ainsi une valeur trop petite pour être représentée dans un ordinateur. La solution consiste à introduire un facteur d'échelle qu'on multiplie par la probabilité $\alpha_t(i)$ pour permettre à cette dernière de rester dans une échelle convenable. Une opération similaire est alors faisable pour le calcul de $\beta_t(i)$ .Il suffit d'effacer toute trace de ce coefficient à la fin du calcul pour retrouver les valeurs réelles. Le facteur d'échelle utilisée est :

$$c_t = \frac{1}{\sum_{i=1}^{N} \alpha_t(i)} \tag{4.39}$$

ainsi pour une t fixée, on calcule d'abord:

$$\alpha_t(i) = \sum_{j=1}^{N} \hat{\alpha}_{t-1}(j).a_{ij}.b_j(o_t) \tag{4.40}$$

l'ensemble des coefficients $\hat{\alpha}_t(i)$ est donné par :

$$\hat{\alpha}_t(i) = \frac{\sum_{j=1}^{N} \hat{\alpha}_{t-1}(j).a_{ij}.b_j(o_t)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \hat{\alpha}_{t-1}(j).a_{ij}.b_j(o_t)} \tag{4.41}$$

par induction, on peut écrire :

$$\hat{\alpha}_{t-1}(j) = \left[\prod_{\tau=1}^{t-1} c_\tau\right] \alpha_{t-1}(j) \tag{4.42}$$

$$\hat{\alpha}_t(i) = \frac{\sum_{j=1}^{N} \alpha_{t-1}(j).\left[\prod_{\tau=1}^{t-1} c_\tau\right].a_{ij}.b_j(o_t)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{t-1}(j)..\left[\prod_{\tau=1}^{t-1} c_\tau\right].a_{ij}.b_j(o_t)} = \frac{\alpha_t(i)}{\sum_{i=1}^{N} \alpha_t(i)} \tag{4.43}$$

ensuite, on va calculer les termes bêtas, on utilise les mêmes coefficients utilisés pour les alphas, nous aurons :

$$\hat{\beta}_t(i) = c_t.\beta_t(i) \tag{4.44}$$

en fonction des nouvelles variables, l'équation de réestimation a_{ij} devient :

$$a^*_{ij} = \frac{\sum_{t=1}^{T-1} \hat{\alpha}_t(i).a_{ij}.b_j(o_t).\hat{\beta}_{t+1}(j)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \hat{\alpha}_t(i).a_{ij}.b_j.(o_{t+1})\hat{\beta}_{t+1}(j)} \tag{4.45}$$

mais chaque terme $\hat{\alpha}_t(i)$ peut s'écrire :

$$\hat{\alpha}_t(i) = \left[\prod_{s=1}^{t} c_s\right].\alpha_t(i) = C_t.\alpha_t(i) \tag{4.46}$$

comparativement :

$$\hat{\beta}_{t+1}(j) = \left[\prod_{s=t+1}^{T} c_s\right].\beta_{t+1}(j) = D_{t+1}.\beta_{t+1}(j) \tag{4.47}$$

donc l'expression (3.46.) devient :

$$a^*_{ij} = \frac{\sum_{t=1}^{T-1} C_t.\alpha_t(i).a_{ij}.b_j(o_{t+1}).\beta_{t+1}(j).D_{t+1}}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} C_t.\alpha_t.(i).a_{ij}.b_j(o_{t+1}).D_{t+1}.\beta_{t+1}(j)} \tag{4.48}$$

finalement :

$$C_t.D_{t+1} = \left[\prod_{s=1}^{t} c_s\right]..\left[\prod_{s=t+1}^{T} c_s\right] = \prod_{s=1}^{T} c_s = C_T \tag{4.49}$$

Dans ce cas $C_t.D_{t+1}$ peut disparaître de l'équation (4.49) laissant exactement l'équation de re-estimation.

Il est claire que la procédure décrite ci-dessus sera applicable directement pour le cas de la matrice *B* en introduisant seulement des modifications analogues à celles faites pour la re-estimation de la matrice A, et sera implémentée chaque fois que le calcul des termes alphas et bêta est requis ou bien une fois le programme détecte un débordement lors du calcul de ces termes.

Enfin, le seul changement à opérer après application de la procédure de normalisation, est dans le calcul de la probabilité $p(O/\lambda)$, on ne peut en effet sommer $\hat{\alpha}_t(i)$ puisque ceux-ci ont été changés, cependant on peut utiliser la propriété :

$$\prod_{t=1}^{T} c_t . \sum_{i=1}^{N} \alpha_T(i) = C_T . \sum_{i=1}^{N} \alpha_T(i) = 1 \tag{4.50}$$

alors : $$\prod_{t=1}^{T} c_t . p(O/\lambda) = 1 \tag{4.51}$$

ceci donne finalement :

$$p(O/\lambda) = \frac{1}{\prod_{t=1}^{T} c_t} \tag{4.52}$$

Notons par ailleurs, que lors de l'utilisation de l'algorithme de Viterbi pour donner la probabilité maximum, aucune normalisation numérique n'est faite si on transpose les calculs dans le domaine logarithmique.

L'algorithme de viterbi s'écrit ainsi :

$$\delta_t(i) = \underset{q_1 q_2 \ldots\ldots q_t}{Max} \{\log p[q_1 . q_2 \ldots\ldots q_t, o_1 . o_2 \ldots\ldots o_t / \lambda]\} \tag{4.53 a}$$

l'étape initiale est :

$$\delta_1(i) = \log(\pi_i) + \log(b_i(o_1)) \tag{4.53 b}$$

l'étape récursive est donnée par :

$$\delta_t(j) = \underset{1 \le i \le N}{Max} [\delta_{t-1}(i) + \log(a_{ij})] + \log[b_j(o_t)] \tag{4.53 c}$$

et l'étape finale :

$$\log(p^*) = \underset{1 \le i \le N}{Max} [\delta_T(i)] \tag{4.53 d}$$

4.2.7.2 Séquences multiples [B42]

Pour avoir une estimation robuste, il est nécessaire qu'un nombre important d'observations soit disponible. Les formules de ré-estimation sont légèrement modifiées, sans que cela change les propriétés de convergence de l'algorithme.

Considérons un ensemble de *R* suites d'observation $\left\{O^{(r)}\right\}_{1\le r\le R}$

$$O=(o^{(1)}, o^{(2)}, o^{(3)},\ldots\ldots\ldots, o^{(R)}) \tag{4.54}$$

où $O^{(r)}=(o_1^r, o_2^r, o_3^r,\ldots\ldots\ldots, o_{Tr}^r)$ est la r^{ime} séquence d'observation

Nous supposons que les séquences sont indépendantes, et notre but est d'ajuster les paramètres du modèle λ afin de maximiser :

$$p(O/\lambda)=\prod_{r=1}^{R} p\left(o^{(k)}/\lambda\right)=\prod_{r=1}^{R} p_r \tag{4.55}$$

Les variables *Forward* et *Backward* doivent être calculées indépendamment pour chaque séquence d'observations, ce qui permet d'exprimer le compteur de l'équation (4.20 a) pour chaque séquence comme:

$$\xi_t^r(i,j)=\frac{\alpha_i^r(t).a_{ij}.b_j(o_{t+1}^r).\beta_{t+1}^r(j)}{p(O^r/\lambda)} \tag{4.56}$$

$$\gamma_t^r(i)=\sum_{j=1}^{N}\xi_t^r(i,j) \tag{4.57}$$

Il faut utiliser dans le décompte toutes les observations de toutes les séquences. , nous obtiendrons les formules finales de ré-estimation comme suit :

$$\hat{\pi}_i=\frac{1}{R}.\sum_{r=1}^{R}\gamma_1^r(i) \tag{4.58}$$

$$\hat{a}_{ij}=\frac{\sum_{r=1}^{R}\left(\sum_{t=1}^{Tr-1}\xi_t^r(i,j)\right)}{\sum_{r=1}^{R}\left(\sum_{t=1}^{Tr-1}\gamma_t^r(i)\right)} \tag{4.59}$$

$$\hat{b}_j(k)=\frac{\sum_{r=1}^{R}\left(\sum_{t=1,o_t^r=vk}^{Tr}\gamma_t^r(j)\right)}{\sum_{r=1}^{R}\left(\sum_{t=1}^{Tr}\gamma_t^r(j)\right)} \tag{4.60}$$

Rien n'impose que toutes les séquences utilisées proviennent du même locuteur, ce qui donne la possibilité de réaliser avec la même méthode l'apprentissage de modèles dépendants ou non du locuteur.

4.3 Application des HMM à la reconnaissance des mots isolés

Dans le but de réaliser une reconnaissance de mots isolés, Chaque mot m du vocabulaire V est modélisé par un HMM. Les observations émises lors de l'arrivée dans un état correspondent aux trames acoustiques. Ces trames sont usuellement représentées par des vecteurs de paramètres continus. Donc il est préférable de modéliser les probabilités d'émission par des densités de probabilité continues (multi-gaussienne).

4.3.1 Modèles de mots

Les modèles de Markov utilisés pour représenter la parole sont des modèles "gauche-droite séquentiels ", ces derniers sont bien adaptée au problème de la reconnaissance de la parole. L'avantage d'utiliser tels modèles réside dans le fait qu'ils introduisent des contraintes temporelles fortes sur les chaînes de Markov, ce qui est bien adapté à la reconnaissance de la parole, en particulier la reconnaissance de mots isolés. Leurs probabilités de transition vérifient:

$$pour \quad i \rangle j \quad \Rightarrow a_{ij} = 0 \tag{4.61}$$

et

$$\pi_i = \begin{cases} 1 & pour\ i = 1 \\ 0 & pour\ i \neq 1 \end{cases} \tag{4.62}$$

Des modèles gauche-droite particuliers autorisant le bouclage à l'état courant, le passage à l'état suivant ou le saut d'un état ont été introduits par R. Bakis pour représenter des mots Figure 4.4

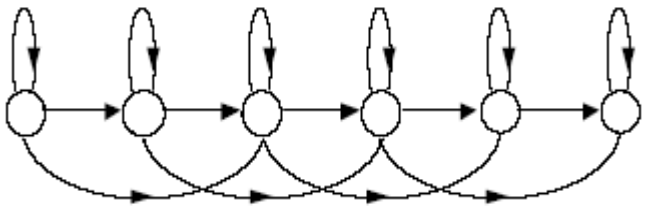


Figure 4.4 : modèle de Bakis 5 états

4.3.2 Initialisation

Comme dans tout algorithme de maximisation, La convergence des modèles vers un maximum le plus proche possible du maximum global au cours de leur ré-estimation nécessite une bonne initialisation des probabilités d'émission. Ici une façon simple d'opérer consiste à découper la suite des T vecteurs, correspondant au mot à apprendre, en N portions égales. Ce qui consiste à supposer que les N états sont d'égales durées. On estime alors, dans chaque portion, les moyennes et les variances respectives de chaque composante. Cela donne les N vecteurs moyens initiaux $\mu_{1:N}$ et les N matrices initiales $\sigma_{1:N}$.

Lorsque la densité de probabilité d'un état est représentée par une multi-gaussienne, une procédure de k-mean pour k=G est utilisée pour fournit les moyennes des G gaussiennes de la loi. En ce qui concerne la matrice de transition, on choisit, comme valeur initiale, une matrice de dimension $N \times N$, avec

$$a_{ij} = 0 \quad pour \quad i \rangle j \qquad \text{(HMM gauche-droite)} \tag{4.63.}$$

et $$a_{ij} = \frac{1}{L} \tag{4.64}$$

L : nombres des éléments non nuls dans la ligne i de la matrice A.

4.3.3 Apprentissage

Pour chaque mot m du vocabulaire, on construit un modèle de Markov noté λ^m , pour ce faire on doit estimer les paramètres (A, c, μ, σ) afin de maximiser la vraisemblance des observations pour le mot m. On fait appel à l'algorithmes itératif de ré-estimeation de Baum-Welch les étapes de cette procédure sont illustrées dans la figure 4.5.

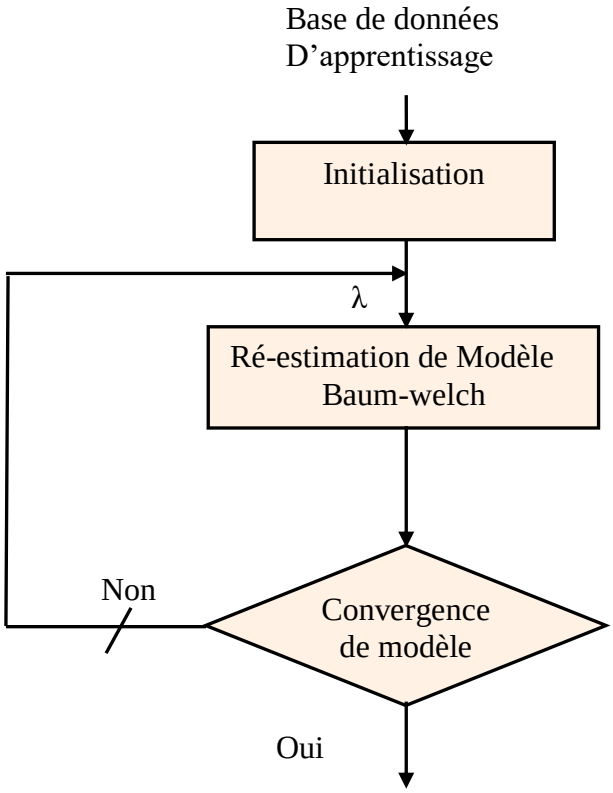


Figure 4.5 : Processus de ré-estimation des paramètres du modèle HMM

4.3.4 Reconnaissance :

Pour chaque mot à reconnaître, on calcule la probabilité $p(O/\lambda^v)$ pour$1 \prec v \leq V$ (V=nombre des modèles) par l'intermédiaire de l'algorithme de Viterbi [B42]; le mot reconnu sera celui pour lequel la probabilité est maximale figure 4.6

$$v^* = Arg \max_{1 \leq v \leq V} \left[prob\left(O/\lambda^v\right) \right] \tag{4.65}$$

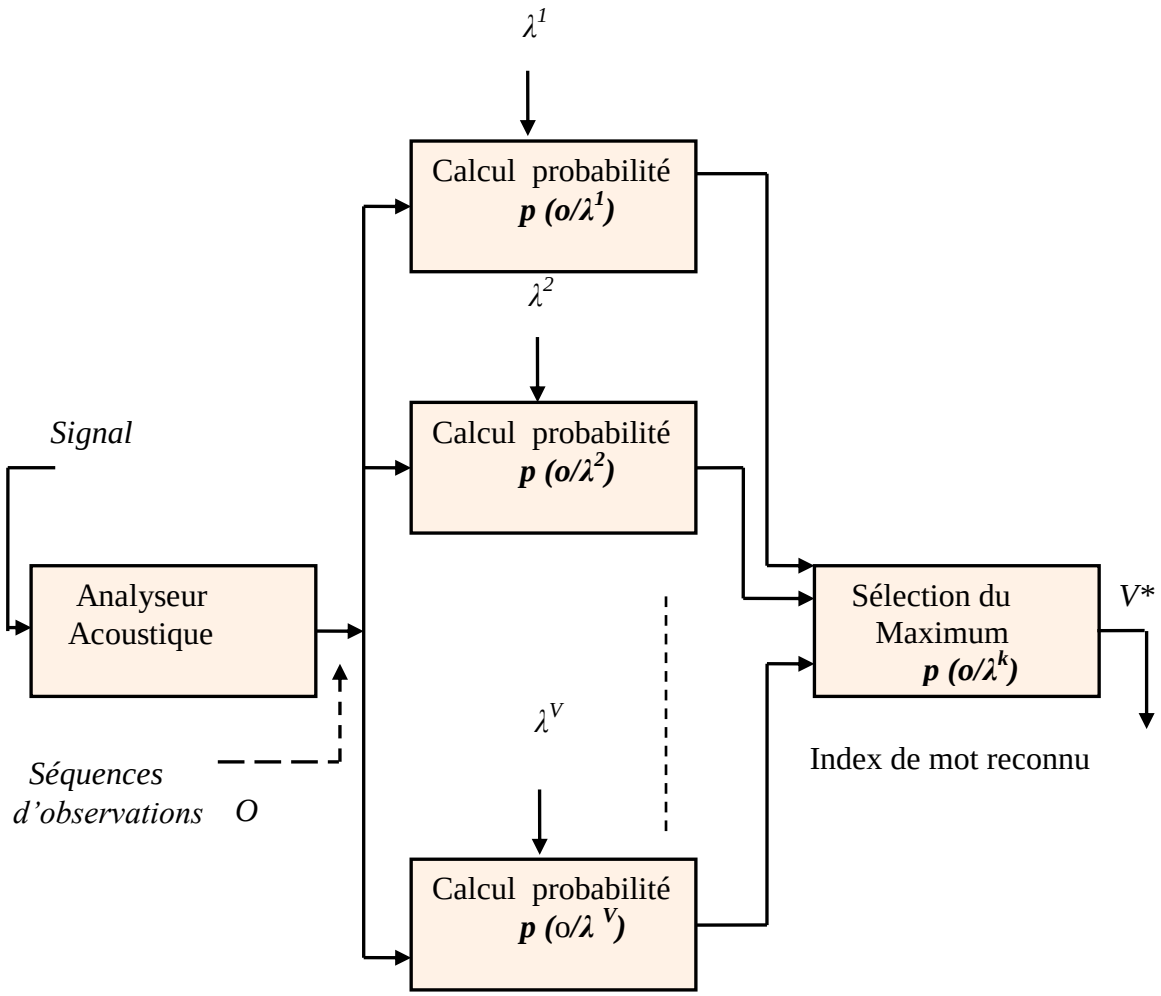


Figure 4.6 : Système de reconnaissance basé sur les HMM

4.4 Plate-forme 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 par S.J. Young et son équipe. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole continue à base de modèles de Markov cachés [B49].

HTK est remarquable par la très grande liberté de choix laissée tout au long de la construction du système de reconnaissance. Les modèles peuvent représenter des mots ou tout type d'unité sub-lexicale, et leur topologie est librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des multi-gaussiennes. Les modèles sont initialisés avec l'algorithme de Viterbi, puis ré-estimés 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. De plus, HTK est un système largement répandu dans le monde de la recherche; en 1992, ses concepteurs revendiquaient déjà plus d'une centaine d'utilisateurs.

4.4.1 Présentation d' HTK

HTK dans sa version 3.1 est structuré en 19 librairies utilisées par 19 outils de base (Tableau 4.1). Les outils manipulent des fichiers de différents types: signaux, étiquettes, paramètres, description des modèles, définition de réseaux. Les formats des fichiers de signaux et d'étiquettes des bases de données les plus répandues sont reconnus. Les autres fichiers sont dans un format particulier à HTK, décrit dans le manuel de référence [B49]. En particulier, les modèles et les réseaux sont définis dans des fichiers texte, ce qui facilite leur création et leur modification par l'utilisateur. Les options d'utilisation des outils sont transmises en argument sur la ligne de commande. Il est donc facile d'automatiser les processus d'apprentissage et de décodage avec des scripts écrits dans le langage de commande du système d'exploitation.

Librairies	Outils de base
HShell interface avec le système d'exploitation	**HCopy** Calcul des paramètres de signal.
Hmath Procédures mathématiques	**HInit** Initialisation d'un modèle.
Hg P Procédures de traitement de signal	**HRest** Ré-estimationd'un modèle.
HDBase Stockage en mémoire des paramètres	**HERest** Ré-estimation des modèles enchaînés
HSpiO Gestion des fichiers des données.	**HVite** Décodage en parole continue.
HLabel Gestion des fichiers d'étiquettes.	**HResults** Résultats du décodage.
HModel Gestion des fichiers des modèles.	**HList** Affichage des fichiers des données.
HParse lecture du réseau syntaxique.	**HLstats** Calcul de statistiques sur les étiquettes.
HGraf affichage graphique.	**HSlab** Affichage du signal et des étiquettes.
HLed édition des fichiers d'étiquettes	**HLabNet** Génération automatique d'un réseau.
HHled Edition des modèles	**HComp** Initialisation globale.
Haudio Permet de l'accès à l'entrée audio	……
…….	…….

Tableau 4.1 : Librairies et outils de base de HTK

4.4.2 Utilisation d'HTK

Les principaux outils de base de HTK s'enchaînent naturellement pour réaliser les différentes étapes d'un système de reconnaissance: calcul des paramètres du signal, apprentissage des modèles, et expériences de reconnaissance (Figure 4.7).

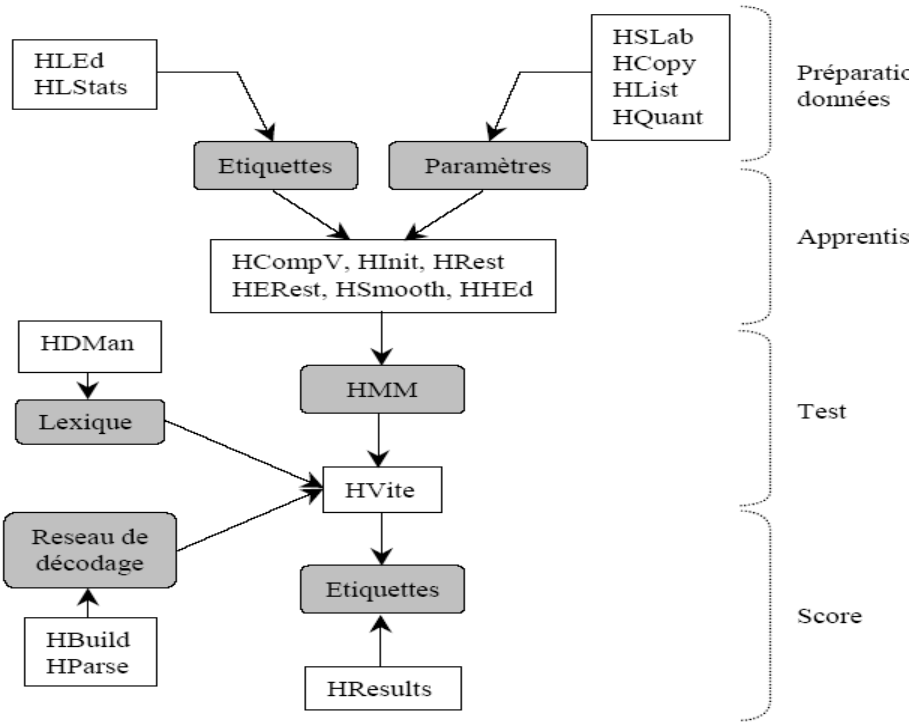


Figure 4.7 : Structure d'un système de reconnaissance avec HTK

4.4.3 Prétraitement des données

Avant l'apprentissage des modèles, il est nécessaire de préparer les données d'apprentissage en calculant les paramètres du signal et en étiquetant les phrases d'apprentissage. La représentation du signal est obtenue avec l'outil HCopy, qui produit en particulier des coefficients LPCC ou MFCC ainsi que l'énergie. Les coefficients différentiels du premier et du second ordre peuvent être calculés ultérieurement lors de la lecture des fichiers de paramètres, ce qui économise leur stockage en mémoire de masse. Les phrases d'apprentissage doivent être toutes étiquetées en fonction des unités acoustiques modélisées. Les bases de données de parole sont parfois fournies avec un étiquetage phonétique qui ne correspond pas exactement aux unités acoustiques modélisées. L'éditeur HLEd permet alors de modifier les étiquettes, par exemple pour regrouper plusieurs phonèmes différents dans une seule classe ou pour fusionner deux segments adjacents.

4.4.4 Topologie des modèles

Pour chaque unité acoustique, il faut définir un modèle prototype contenant la topologie choisie, à savoir le nombre d'états du modèle, les états entre lesquels les transitions sont possibles, le type de loi de probabilité associée à chaque état. L'état initial et l'état final ont la particularité de ne pas émettre d'observation, mais de servir uniquement à la connexion des modèles en parole continue. Il n'est pas nécessaire que tous les modèles utilisent le même prototype. Les probabilités d'émission sont associées aux états et sont décrites par une combinaison linéaire de gaussiennes multi-variables, caractérisées par leur moyenne et leur matrice de covariance dans l'espace des paramètres. La matrice de covariance est théoriquement symétrique, mais peut être choisie diagonale si l'on suppose l'indépendance entre les composantes des vecteurs de paramètres, et les vecteurs de paramètres peuvent être séparés en flux de données indépendants.

4.4.5 Apprentissage

Pour chacune des machines modélisant une unité acoustique, l'outil HInit initialise les probabilités d'émission des états du modèle au moyen de la procédure itérative des "k moyennes segmentales" basée sur l'algorithme de Viterbi. Cette phase nécessite l'étiquetage fin des phrases d'apprentissage utilisées, car il faut extraire tous les segments correspondant à l'unité modélisée. Cette fastidieuse segmentation manuelle peut cependant être limitée à une fraction de l'ensemble d'apprentissage, de manière à disposer de quelques représentants pour l'initialisation de chaque modèle acoustique. L'estimation des paramètres d'un modèle est affinée avec HRest, qui applique l'algorithme optimal de Baum-Welch jusqu'à la convergence et ré-estime les probabilités d'émission et de transition. Dans une phase suivante, il est possible d'appliquer plusieurs itérations de l'outil HERest, qui ré-estime simultanément l'ensemble des modèles sur de la parole continue non segmentée. Les modèles obtenus peuvent être améliorés, en augmentant par exemple le nombre de gaussiennes servant à estimer la probabilité d'émission d'une observation dans un état. Le choix du nombre optimal de gaussiennes est un problème délicat, généralement guidé par des heuristiques. Une commande de l'éditeur de modèles HHEd réalise l'augmentation du nombre de gaussiennes modélisant une densité de probabilité. Les modèles doivent être ensuite ré-estimés par HRest ou HERest. HTK offre aussi des facilités pour travailler avec des modèles contextuels, dépendants des contextes phonétiques gauches ou droits.

4.4.6 Reconnaissance

Le module de décodage de la parole continue, HVite, utilise l'algorithme de Viterbi pour trouver la séquence d'états la plus probable correspondant aux paramètres observés dans un modèle composite, et en déduire les unités acoustiques correspondantes. Le modèle composite autorise la succession des modèles acoustiques en fonction d'une syntaxe choisie par le concepteur du système. Il est possible d'écrire la syntaxe aux niveaux phonétique ou lexical, les mots du lexique pouvant être définis eux-mêmes par la concaténation d'unités sub-lexicales. Parallèlement à cette syntaxe, HVite tient compte d'un modèle de langage de type bigramme, estimé sur les étiquettes des phrases d'apprentissage par l'outil HLStats. Le résultat du décodage est comparé aux étiquettes de référence par un alignement dynamique réalisé par HResults, afin de compter les étiquettes identifiées, omises, substituées par une autre, et insérées, et de calculer le taux de reconnaissance.

4.5 Conclusion

Les modèles de Markov cachés, présentés dans ce chapitre sont des techniques largement utilisées en reconnaissance de formes, et sont les plus utilisés en reconnaissance de la parole. Ils bénéficient d'algorithmes d'entraînement et de décodage performants (algorithmes de Viterbi, Baum-Welch).

Toutefois, les hypothèses nécessaires à la mise en oeuvre de ces algorithmes peuvent pénaliser les performances de ces modèles.

Les principales hypothèses les plus contraignantes sont :

- □□□Entraînement non discriminant (maximisation de la vraisemblance au lieu de probabilités a posteriori).
- Forme des densités de probabilité fixées (multi-gaussiennes ou discrète).
- □□□□□□□□□Les composantes des vecteurs acoustiques sont supposées non corrélées.
- La séquence des états est un processus de Markov du premier ordre.

CHAPITRE 5 :

ETUDE EXPÉRIMENTALE

Chapitre 5

Etude expérimentale

5.1 Introduction

Le but de ce chapitre est d'exposer les résultats expérimentaux des deux types d'ensembles des paramètres utilisés (paramètres usuels et paramètres provenant de la transformée en ondelettes), concernant la reconnaissance des chiffres arabes isolés (de 0 à 9) Nous allons comparer différentes variantes d'analyse afin d'observer leurs performances. Et enfin nous complèterons cette étude par une analyse de l'influence du bruit sur les deux types d'ensembles des paramètres.

5.2 Préparation des données

5.2.1 Description de la base de données

La base de données utilisée est conçue au département d'électronique de l'université d'Annaba. Les occurrences sont enregistrées dans des conditions moyennes avec une carte son d'un PC ordinaire. Cette base contient 92 locuteurs : 46 locuteurs de sexe masculin et 46 locuteurs de sexe féminin. Chaque locuteur a prononcé 10 fois chaque chiffre arabe (de 0 à 9) d'une manière isolée, 7 occurrences ont été utilisées pour les besoins d'apprentissage et les trois autres pour la phase des tests.

Nous avons choisi 50 locuteurs (25 hommes, 25 femmes). A partir de cette nouvelle base, nous avons construit trois sous- bases pour l'évaluation du système de reconnaissance,

- **BA** : construite pour la phase d'apprentissage contenant 40 locuteurs (20 femmes et 20 hommes).
- **BT1 :** construite pour la phase de test contenant 30 locuteurs (15 femmes et 15 hommes)). Cette sous-base contient deux catégories de locuteurs, la première contient 20 locuteurs qui ont contribué à la phase d'apprentissage, la deuxième contient 10 locuteurs qui n'ont pas servi pour l'apprentissage, mais pour le test seulement.
- **BT2 :** cette base de test est construite pour l'étude de la robustesse des paramètres au bruit. Les données de cette base sont les données de BT1 auxquelles nous avons ajouté de bruit avec un niveau d'énergie. Pour mesurer cela, le rapport signal sur bruit (SNR : Signal to Noise Ratio) est défini comme le rapport de la puissance du signal sur celle du bruit :

$$SNR = 10.\log_{10} \frac{P_S}{P_b} \quad (dB) \qquad (5.1)$$

P_S : Puissance de signal

P_b : Puissance de bruit

Il faut noter que notre système est conçu pour un mode multi-locuteurs, mais on introduit sans aucune adaptation des locuteurs qui n'ont pas contribué à la phase d'apprentissage pour tester l'efficacité de notre système.

5.2.2 Analyse acoustique

Les données sont prélevées à une fréquence d'échantillonnage de 11025Hz et numérisées à la résolution de 16 bits. Après l'échantillonnage, on fait appel à un filtre de préaccentuation de la forme $H(z) = 1 - 0.97.Z^{-1}$ pour relever les hautes fréquences, puis une analyse sur une fenêtre glissante de longueur de 256 points sur laquelle le signal est supposé quasi-stationnaire, pour un pas de recouvrement entre deux trames successives égal à 1/ 2 et enfin une pondération par fenêtre de hamming est effectuée afin de réduire l'effet de bord.

Notre système a été entraîné et testé par deux ensembles de paramètres

- Paramètres usuels :

  Nous avons utilisé les coefficients LPC, LPCC et les MFCC ajoutés aux coefficients de leurs variations du 1^{er} et 2^{eme} ordre avec l'introduction du coefficient énergie.
- Nouveaux paramètres provenant de la transformée en ondelettes

5.2.3 Caractéristiques du système de base

Les choix qui ont été faits pour le système de base sont les suivants:

- Unités acoustiques : 10 mots (10 chiffres)
- Topologie des modèles : modèle gauche droite
- Probabilité d'émission modélisée par une combinaison linéaire de gaussienne .Tous les modèles ont la même topologie, et les probabilités d'émission de tous les états sont représentées par un nombre identique de gaussiennes.

5.2.4 Nombre de paramètres du système

Le système de reconnaissance doit tenir compte du nombre de paramètres. Une augmentation du nombre de paramètres du système permet une modélisation acoustique plus fine, à condition que le modèle soit pertinent et qu'une quantité suffisante de données d'apprentissage soit disponible.

Pour notre système, nous avons:

- V : Nombres de modèles (10 modèles +1 pour HTK)
- Nombres d'états par modèle $(N \geq 3)$
- G : Nombres de gaussiennes par état $(G \geq 1)$
- d moyennes et d variances par gaussiennes (d : dimension de vecteur paramètres).

5.3 Résultats et discussion

5.3.1 Expériences avec les coefficients usuels

Le but de ces expériences est de construire un système de référence, qui sera utilisé par la suite dans l'étude comparative avec les coefficients provenant de la transformée en ondelettes.

5.3.1.1 Choix des paramètres acoustiques de référence

Les coefficients LPC, LPCC ou MFCC ainsi que l'énergie E sont les coefficients les plus utilisés dans les systèmes de reconnaissance. L'usage complémentaire des coefficients différentiels du premier et du second ordre s'est généralisé depuis une dizaine d'années dans les systèmes de reconnaissance à base de HMM. Le choix d'une représentation du signal parmi celles présentées est en définitive guidé par les résultats expérimentaux dans la tâche considérée. Nous présentons quelques études comparatives entre ces différents coefficients afin de justifier un des choix.

Bien que nous puissions effectuer ces tests avec n'importe quelle structure de modèle continu de Markov Caché, nous choisissons ici, le modèle dont la structure comprend 5 états et 3 mixtures par état car d'après nos expérimentations, cette structure du modèle se situe dans la région optimale quand nous changeons les paramètres du modèle.

A. Expérience 1 : Influence des paramètres acoustiques

Dans le cadre de ces expériences, nous avons calculé pour chaque trame :

- LPC : 12 coefficients LPC
- LPCC : 12 coefficients LPCC
- MFCC : 12 coefficients MFCC.

Le tableau 5.1 donne les taux de reconnaissance pour chaque type de paramètres.

Mot à reconnaître	**Paramètres**		
	LPC	**LPCC**	**MFCC**
0	82,41	86.67	95.56
1	89,17	98.89	98.89
2	84,66	92.22	94.44
3	77,11	84.44	90
4	84,32	95.56	96.67
5	93,47	100	100
6	94,00	96.67	98.89
7	81,11	84.44	94.44
8	84,44	94.44	92.22
9	80,22	96.67	97.78
Taux moyen	**85.09**	**93**	**95.89**
Taux moyen Réf : HTK	**84.03**	**96.67**	**97.12**

Tableau 5.1 : Influence des coefficients d'analyse sur le taux de reconnaissance

Résultats et discussion :

Les résultats obtenus montrent :

- L'utilité d'un prétraitement par banc de filtres, d'une échelle fréquentielle non linéaire et la représentation cepstrale.
- Les coefficients MFCC sont plus performants que les coefficients LPCC

A cet effet, nous choisirons les coefficients MFCC et nous les considérons comme le noyau initial pour tester l'efficacité des autres caractéristiques.

B. Expérience 2 : Influence des coefficients différentiels et de l'énergie

Le système utilise les coefficients MFCC et l'énergie, ainsi que les coefficients différentiels de ces paramètres. Le logarithme de l'énergie de la tranche est ajouté aux 12 coefficients cepstraux pour former un vecteur de 13 coefficients, les coefficients différentiels du premier et deuxième ordre sont utilisés avec les coefficients dits « statiques ». Pour ces expériences, nous avons calculé pour chaque tranche :

- MFCC : 12 coefficients MFCC
- MFCC_E : 13 coefficients (12 coefficients MFCC + 1 coefficient de l'énergie E).
- MFCC_E_D : 26 coefficients (12 coefficients MFCC + 1 coefficient E +la première dérivé de (12 coefficients MFCC + 1 coefficient E).
- MFCC_E_D_A : 39 coefficients (12 coefficients MFCC + 1 coefficient E+ la première dérivé de (12 coefficients MFCC +1 coefficient de l'énergie E) + la deuxième dérivé de (12 coefficients MFCC + 1 coefficient E)).

- Paramètres de modèle HMM : 5 états, 3 mixtures par densité de probabilité pour chaque état. Le tableau 5.2 donne les taux de reconnaissance obtenus.

Mot à reconnaître	**Paramètres**			
	MFCC	**MFCC_E**	**MFCC_E_D**	**MFCC_E_D_A**
0	95.56	97.78	98.89	98.89
1	98.89	98.89	100	100
2	94.44	94.44	100	100
3	90.00	93.33	98.89	97.78
4	96.67	97.78	97.78	95.56
5	100	100	100	100
6	98.89	98.89	100	100
7	94.44	87.78	96.67	95.56
8	92.22	92.22	100	100
9	97.78	100	97.78	96.67
Taux moyen	**95.89**	**96.11**	**99.00**	**98.44**
Taux moyen Réf : HTK	**97.12**	**98.50**	**99.61**	**99.72**

Tableau 5.2 : Influence des coefficients différentiels et l'énergie sur le taux de reconnaissance.

- **Résultats et discussion**

D'après les résultats obtenus, on remarque que :

▪ L'introduction du coefficient d'énergie n'apporte pas des améliorations significatives, mais des autres expériences montrent son efficacité lorsque il est combiné avec leurs coefficients différentiels.

▪ .L'influence des coefficients différentiels du premier ordre (MFCC_E_D) est majeure, le taux de reconnaissance s'améliore d'une façon remarquable (de **3.11 %**) par rapport aux coefficients statiques (MFCC), la dérivée seconde n'apporte pas des améliorations significatives par rapport aux coefficients différentiels du premier ordre (MFCC_E_D) dans le cadre de ces expériences; son intérêt a été mis en évidence lors des expériences de reconnaissance de mots en milieu bruité [B48],

Il faut noter que l'utilisation des coefficients différentiels multiple l'espace de représentation. Cela va augmenter le temps de calcul et l'espace mémoire (voir tableau 5.3).

Modèle : N = 5 G = 3			
MFCC	**MFCC_E**	**MFCC_E**_D	**MFCC_E**_D_A
195 KO	225 KO	835 KO	1.79 MO

Tableau 5.3 Taille de fichier globale contenant le modèle pour différents paramètres

5.3.1.2 Choix des paramètres de modèle HMM

A. Expérience 1 : Influence du Nombre de Mixtures

Pour valider le choix du nombre de mixtures G prototype, nous avons réalisé l'apprentissage de modèles à G mixtures (G =1, 3,5 et 7). Le nombre d'états N a été fixé à 5. Le tableau 5.4 présente les résultats obtenus.

Mot à reconnaître	Nombre des mixtures			
	G=1	G=3	G=5	G=7
0	96.67	98.89	100	100
1	98.89	100	100	100
2	100	100	100	100
3	98.89	98.89	98.89	100
4	98.89	97.78	100	100
5	100	100	100	100
6	98.89	100	100	98.89
7	93.33	96.67	98.89	98.89
8	100	100	100	100
9	96.67	97.78	100	98.89
Taux moyen	**98.22**	**99.00**	**99.78**	**99.67**
Taux moyen Réf : HTK	**99.39**	**99.61**	**99.89**	**99.80**

Tableau 5.4 : Influence de nombre de mixture sur le taux de reconnaissance

- Résultats et discussion

D'après les résultats obtenus, nous remarquons que plus le nombre de mixtures augmente, plus le taux de reconnaissance s'améliore. Les meilleurs résultats sont obtenus pour G = 5 gaussiennes par état. Au delà de 3 gaussiennes, les performances du système ne sont pas significativement supérieures, tandis que la convergence de l'apprentissage est plus longue à atteindre et l'espace mémoire est augmenté.

Il faut noter aussi qu'un nombre trop élevé de gaussiennes par rapport à la quantité de données disponibles conduit à un mauvais apprentissage, ce qui explique la diminution du taux de reconnaissance moyen à partir de 7 gaussiennes par état.

On peut conclure que le nombre de mixtures **G =3** est suffisant pour modéliser les chiffres arabes isolés.

Expérience 2 : Influence du Nombre des états

Pour valider le choix du nombre d'états du modèle prototype, nous avons réalisé l'apprentissage de modèles à N états (N = 3 : 7). Le nombre de gaussiennes G par densité de probabilité a été fixé à 3 Le tableau 5.5 présente les résultats obtenus.

Mot à reconnaître	**Nombre des états**				
	N=3	**N=4**	**N=5**	**N=6**	**N=7**
0	98.83	100	98.89	100	100
1	98.89	98.89	100	100	100
2	100	100	100	100	100
3	97.78	96.67	98.89	98.89	98.89
4	95.56	97.78	97.78	98.89	100
5	100	100	100	100	100
6	100	100	100	100	100
7	96.67	96.67	96.67	97.78	97.78
8	100	100	100	100	100
9	96.67	96.67	97.78	97.78	97.78
Taux moyen	**98.44**	**98.67**	**99.00**	**99.33**	**99.67**
Taux moyen Réf : HTK	**99.08**	**98.89**	**99.61**	**99.50**	**99.45**

Tableau 5.5 : Influence de nombre d'états sur le taux de reconnaissance

- **Résultats et discussion**

La comparaison des résultats obtenus donne l'avantage au modèle à 5 états puisque au delà de 5 états, le taux de reconnaissance moyen n'est pas significativement supérieur ainsi que le temps de calcul et l'espace mémoire sont augmentés. Il s'agit cependant d'un choix global, concernant les 10 modèles simultanément. Il est possible, en fonction des résultats obtenus en observant l'adéquation de différents nombres d'états avec un chiffre, de choisir un nombre d'états plus adapté pour chaque chiffre. Le tableau 5.6 donne les résultats obtenus pour ce choix.

Chiffres	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	9
Nombre des états	4	5	3	5	7	3	3	6	3	5

Tableau 5.6 : Influence de nombre d'états sur le taux de reconnaissance

A la fin de ces expériences, il faut noter que : La différence entre les résultats donnés par HTK et notre système est due à la bonne initialisation utilisée par la plate forme HTK.

5.3.2 Expériences avec les coefficients provenant de la transformée en ondelettes

Plusieurs techniques de paramétrisation sont proposées, le but de ces expériences est de déterminer l'apport des représentations en ondelettes sur le taux de reconnaissance.

Le modèle acoustique utilisé : chaque HMM à 5 états, pour chaque état la probabilité d'émission est modélisée par un mélange de 3 Gaussiennes.

5.3.2.1 Expérience 1 : Influence du niveau de décomposition

Après des expériences préliminaires, nous avons choisi d'utiliser l' ondelette suivante : daubechies à 1 moment nul (db-1) et les énergies de sous bandes comme paramètres acoustique. Différents niveaux de décomposition (nombre de bandes de fréquence) sont testés.

Le tableau 5.7 présente les résultats obtenus.

Mot à reconnaître	Niveau de décomposition (nombre de paramètres)			
	3	**5**	**7**	**8**
0	71.11	77.78	80.00	83.33
1	81.11	93.33	93.33	93.33
2	83.33	88.89	87.78	95
3	85.56	87.78	91.11	91.67
4	78.89	77.78	77.78	71.67
5	98.89	100	98.89	96.67
6	96.67	95.56	97.78	98.33
7	77.78	71.11	71.11	76.67
8	84.44	77.78	82.22	90
9	93.33	92.22	92.22	85
Taux moyen	**85.11**	**86.22**	**87.22**	**88.17**

Tableau 5.7 : Influence de niveau de décomposition sur le taux de reconnaissance.

- **Résultats et discussion**

D'après les résultats obtenus, Nous observons que les meilleurs résultats sont obtenus pour 8 niveaux de décomposition. Les faibles taux de reconnaissance pour les petits niveaux de décomposition peuvent être facilement expliqués, cela est du à la division limitée de bandes de fréquence. La résolution pour les basses fréquences est trop brute et ceci cause une perte d'information discriminante ente les chiffres. Quand nous augmentons le nombre de niveaux de décomposition, nous augmentons aussi la résolution des basses fréquences. Cependant, après un niveau élevé de décomposition, on augmente la résolution des très basses fréquences qui ne contient pas nécessairement l'information utile pour la reconnaissance.

5.3.2.2 Expérience 2 : Choix de la fonction d'ondelette

Dans cette expérience nous employons 5 niveaux de décompositions et les énergies de sous bandes comme paramètres acoustiques et nous analysons l'effet de la fonction d'ondelette sur les résultats de reconnaissance. De diverses ondelettes (daubechies, coiflets et symlets) sont employées.

Mot à reconnaître	**Fonction d'ondelette**				
	db-1	**db-2**	**db-4**	**coiflet-1**	**symlet-4**
0	77.78	93.33	95.56	90.00	94.44
1	93.33	96.67	97.78	97.78	96.67
2	88.89	98.33	92.22	92.22	92.22
3	87.78	88.33	90.00	85.56	90.00
4	77.78	76.67	80.00	80.00	87.78
5	100	100	100	100	100
6	95.56	100	96.67	95.56	95.56
7	71.11	76.67	91.11	84.44	80.00
8	77.78	91.67	84.44	84.44	85.56
9	92.22	90.00	93.33	97.78	97.78
Taux moyen	**86.22**	**91.17**	**92.11**	**90.78**	**92.00**

Tableau 5.8 : Influence de la fonction d'ondelette sur le taux de reconnaissance.

La comparaison des résultats obtenus donne l'avantage db-4 Il s'agit cependant d'un choix préliminaire puisque il y a d'autres fonctions qui ne sont pas testées au cours de ces expériences.

En conclusion la fonction d'ondelette affecte également le taux de reconnaissance. Pour le choix de la fonction ondelette qui s'adapte mieux, il faut faire d'autres expériences avec des différentes fonctions d'ondelettes.

5.3.2.3 Expérience 3 : Effet des coefficients d'approximation et de détails

Le but de cette expérience est de déterminer les coefficients informatifs dans la décomposition. Après la décomposition, nous employons seulement un sous-ensemble de coefficients. Le premier ensemble d'expériences a été réalisé en supprimant certains coefficients de détails pour différents niveaux. Dans le deuxième ensemble nous excluons les coefficients d'approximation pour différents niveaux de décomposition

- **Expérience 3-1 : Exclusion des coefficients d'approximation**

Le signal d'approximation est la composante basse fréquence dans la décomposition. Au fur et à mesure que le niveau de la décomposition augmente ce signal correspond à la composante continue du signal parole. Dans cette expérience, nous évaluons l'effet de ce signal pour différents niveaux de décomposition.

Niveaux de decomposition	**3**	**4**	**6**	**8**
Taux moyen avec coefficients d'approximation	**85.56**	86.22	87.11	**87.56**
Taux moyen sans coefficients d'approximation	**85.11**	86.22	87.22	**88.17**

Tableau 5.9 : Taux de reconnaissance pour l'expérience 3-2

- **Résultats et discussion :**

Quand le signal d'approximation est éliminé de la paramétrisation, le taux de reconnaissance diminue pour les petits niveaux de décomposition (voir le tableau 5.9). Cependant si le niveau de décomposition est élevé, l élimination du signal d'approximation, augmente légèrement le taux de reconnaissance.

Aux niveaux élevés le signal d'approximation a moins d'informations. Dans ces cas, son élimination est équivalente à l'élimination de la composante continue du signal qui est en effet une normalisation de la base de données.

- **Expérience 3-2 : Exclusion des coefficients de détails**

Coefficients exclus	-	**d1**	**d1,d2**	**d1,d2,d3**
db-4, 3 niveaux	**90.00**	**80.33**	-	-
db-4, 4 niveaux	**91.89**	90.56	**85.00**	-
db-4, 5 niveaux	**92.11**	91.78	90.33	**80.89**

Tableau 5.10 : Taux de reconnaissance moyen pour l'expérience 3-1.

- **Résultats et discussion :**

Les résultats sont récapitulés dans le tableau 5.10. Le taux de reconnaissance diminue si nous éliminons des coefficients de détails.

La conclusion de cette expérience est que les coefficients de détails contiennent l'information spécifique de détails. La division de fréquence de l'analyse par ondelette dyadique est trop brute pour les régions à haute fréquence.

L'analyse par ondelette dyadique divise la bande de fréquence en deux régions à chaque échelle, seule la région de basse fréquence est à nouveau décomposée. La région de haute fréquence est laissée de côté à chaque pas. Les coefficients de détails recouvrent une partie importante du spectre.

Pour améliorer l'analyse, on devrait utiliser la décomposition en paquets d'ondelettes qui permet une division plus fine dans les régions des hautes fréquences.

5.3.2.4 Expérience 4 : influence de la résolution fréquentielle

But : La transformation par ondelette dyadique rapporte une division dyadique de la bande de fréquence tels que les basses fréquences sont représentées avec une haute résolution et les hautes fréquences avec une basse résolution (voir la figure 3.4-a). De plus, la sensibilité fréquentielle du système auditif humain suit un échelle logarithmique de fréquence (c.-à-d., Mel). L'utilisation de paquets d'ondelettes peut réaliser une décomposition avec une résolution réglable de fréquence. Il est donc possible de rapprocher l'échelle de Mel en utilisant les paquets d'ondelettes. Dans cette expérience nous utilisons la décomposition en paquets d'ondelettes pour réaliser différentes résolutions de fréquence.

Lors de notre expérimentation, nous employons les ondelettes db-2 et les énergies de sous bandes comme paramètre acoustique.

Trois différentes structures arborescentes sont employées pour la décomposition suivant les indications du schéma 3.4. L'arbre dyadique (le schéma 3.4-a) correspond à la transformation par ondelettes dyadique classique. Ensuite, nous utilisons l'arbre avec une division similaire à l'échelle de Mel AWP (Admissible Wavelet Packet) [B24] (le schéma 3.4-b). Enfin, nous employons également une pleine décomposition de paquet d'ondelettes qui rapporte la même résolution pour toutes les fréquences (voir le schéma 3.4-c).

Structure arborescente	**Taux de reconnaissance**	**Nombre de paramètres**
dyadique	**90.00**	**6**
Approximation de l'échelle Mel	86.89	24
Full décomposition	85.47	32

Tableau 5.11 : taux de reconnaissance en utilisant les structures arborescentes

- **Résultats et discussion :**

Les résultats de reconnaissance sont récapitulés dans le tableau 5.11. Pour chaque structure arborescente le nombre de paramètres utilisé est également donné dans la dernière colonne. Bien que la décomposition dyadique ait le moindre nombre de paramètres elle mène aux meilleurs résultats.

D'autres tests ont montré que par la décorrelation de coefficients de la structure arborescente, nous pouvons atteindre des taux de reconnaissance de 95.80 % (voir tableau 5.12). Ceci indique qu'il y a une forte corrélation entre les coefficients de bases fréquences donc les modèles gaussiens à covariance diagonale ne sont pas suffisants pour modéliser ces paramètres.

5.3.2.5 Expérience 5 : Comparaisons de différentes paramétrisations provenant de la transformée en ondelettes

Dans cette expérience, nous comparons différents paramètres en utilisant les différentes techniques de paramétrisations provenant de la transformée en ondelettes comme décrit dans le chapitre 3. Nous employons l'ondelette Daubechies-4 avec 5 niveaux de décomposition. Les paramètres utilisés sont :

- Logarithme de l'énergie par échantillon : **E**.
- Logarithme de l'amplitude moyenne : **A-M**.
- Logarithme de l'énergie hiérarchique par échantillon : **E_H**.
- Logarithme de l'amplitude moyenne hiérarchique : **A-M_H**.
- Logarithme de l'énergie Teager énergie par échantillon : **E_T**

Paramèters	**E**	**E_T**	**E_H**	**A-M**	**A-M_H**
Taux moyen	**92.11**	84.11	86.28	**91.78**	**90.89**

Tableau 5.12 : Taux de reconnaissance moyen pour l'expérience en utilisant différents provenant de la transformée en ondelettes

Résultats : Les résultats sont donnés dans le tableau 5.12. Pour toutes les techniques de paramétrisation et avec 5 niveaux de décomposition et la fonction d'ondelette db-4, le meilleur taux de reconnaissance est obtenu en utilisant les énergies de sous bandes.

5.3.2.6 Expérience 6 : Influence des paramètres dynamiques

Pour étudier la dynamique de nos paramètres, nous avons ajouté à nos paramètres statiques leurs dérivées première (D) et seconde (A) (voir tableau 5.13).

Nous avons choisi comme paramètres statiques les paramètres qui ont donnés les meilleurs résultats auparavant : les énergies de sous bandes avec db-4 et 5 bandes de fréquence.

Pour ces expériences, nous avons calculé pour chaque tranche :

- E : 5 coefficients E (énergies de sous bandes)
- E_ D : 10 coefficients (5 coefficients E + la première dérivée de (5 coefficients E).
- E_ D_A : 15 coefficients (5 coefficients E + la première dérivée de (5 coefficients E) + la deuxième dérivée de (5 coefficients E)).

Paramèters	E	E_ D	E_ D_A
Taux moyen	92.11	**97.44**	97.11

Tableau 5.13 : Influence des coefficients différentiels et l'énergie sur le taux de reconnaissance.

- **Résultats et discussion :**

Les résultats sont donnés dans le tableau 5.13 montre que l'ajout des dérivées a permis d'améliorer le taux de reconnaissance moyen de 5.33 % (avec D) et de 5% (avec A) par rapport aux paramètres statiques. Toutefois, on constate que l'ajout des dérivées secondes n'a pas apporté d'amélioration par rapport à l'ajout des dérivées premières.

Cette expérience montre que les coefficients différentiels permettent de prendre en compte la dynamique du signal dans les systèmes de reconnaissance

5.3.2.7 Expérience 7: expérience avec les paramètres fondés sur le principe des coefficients MFCC.

Le but de ces expériences est de faire une hybridation entre les paramètres MFCC et les paramètres provenant de la transformée en ondelettes afin d'obtenir une paramètrisation qui regroupe les avantages des deux types des paramètres.

- **Expérience 6.1**

Les paramètres sont calculés suivant le schéma de la figure 3.5

Décomposition en sous bandes	Taux moyen
Dyadique	93.00
AWP : Similaire à l'échelle de Mel	95.80

Tableau 5.14 : Taux de reconnaissance en utilisant les paramètres fondés sur le principe des coefficients MFCC

On remarque que le taux de reconnaissance augmente pour les structures AWP **(**Similaire à l'échelle de Mel) et dyadique, cette amélioration est due à la décorrélation des coefficients produite par la projection DCT (DCT : Discrete Cosinus Transform).

- **Expérience 6.2**

Les paramètres sont calculés suivant le schéma de la figure 3.6

Taux de reconnaissance : **85.20**

Le résultat obtenu montre que la transformation par ondelettes n'est pas un bon décorrélateur pour les énergies de sous bandes par rapport à la DCT

5.3.3. Etude de l'influence du bruit « Robustesse des paramètres »

Le pouvoir discriminant est une qualité essentielle de la reconnaissance, il est impératif de montrer la fiabilité des paramètres dans des conditions réelles d'utilisation c'est-à-dire en présence de bruit dans le signal à reconnaître. En effet, les conditions des tests précédents étaient idéales : utilisation d'un micro de qualité et absence de bruit ambiant. C'est pour cette raison que nous avons tout d'abord volontairement ajouté du bruit (bruit gaussien blanc, moyenne nulle) à nos échantillons, dans un second temps, nous avons mis en évidence le taux de reconnaissance.

- Paramètres de référence : MFCC_E_D_A.
- Paramètres provenant de la transformée en ondelettes : E _D_A. avec db-4 et 5niveaux de décomposition.

Noua avons utilisé quatre valeurs pour le RSB : 00, 10, 20 et 40 dB. La valeur maximale du RSB est limitée par les qualités du système d'acquisition du son à une valeur proche de 40 dB. D'autre part la présence d'un RSB négatif n'a pas été envisagée ici, 0 dB représente déjà une dégradation importante pour un système de reconnaissance

Le tableau ci-dessous présente l'influence du bruit sur le taux de reconnaissance

Environnement de test		**MFCC_E_ D _ A**	**E _D_A**
propre		**98.44**	**97.11**
Bruité	**SNR = 40 dB**	**98.08**	**95.78**
	SNR = 20 dB	**97.67**	89.78
	SNR = 10 dB	85.22	77.89
	SNR = 00 dB	37.56	44.56

Tableau 5.15 : Influence de bruit sur le taux de reconnaissance moyen.

Résultats : les résultats sont donnés dans le tableau 5.15.Nous constatons que pour des valeurs élevés de RSB, toutes nos paramètrisations sont performantes, mais lorsque celui-ci diminue, elles vont successivement toutes "décrocher". Nous constatons aussi que l'utilisation des coefficients MFCC permet de gagner quelques décibels pour RSB > 0. Donc les coefficients d'ondelette ne sont pas robustes au bruit. Le taux de reconnaissance diminue de manière significative. Mais une technique simple de seuillage peut être facilement appliquée pour compenser l'effet du bruit. Le tableau 5.16 donne les résultats d'application de cette technique.

Environnement de test		**Paramètres E_A_D**
Propre		**96.67**
Bruité	**SNR = 10 dB**	**93.00**

Tableau 5.16 : Influence de bruit sur le taux de reconnaissance moyen après le seuillage des coefficients d'ondelette.

Malgré qu'on a utilisé une méthode très simple pour déterminer le seuil, le taux de reconnaissance est améliorer de **15.11 %.**

On conclut que la paramétrisation par ondelettes avec seuillage des coefficients semble être plus résistante au bruit

.

5.4 Conclusion

Notre nouvelle paramétrisation donne de meilleurs résultats de reconnaissance de chiffres arabes isolés comparable aux coefficients MFCC plus leurs paramètres dynamiques. Nous avons ainsi un taux moyen de **92.11 %** (**db-4, 5 niveaux**).

L'utilisation des paramètres dynamiques (dérivées premières et secondes) améliore encore plus nos résultats : un gain relatif significatif de plus de **5%** est obtenu par rapport aux paramètres statiques.

Notons que notre paramétrisation utilise un nombre réduit de coefficients par rapport aux coefficients MFCC, ce qui implique la réduction de l'espace mémoire et le temps de calcul.

L'étude de l'influence de bruit sur le taux de reconnaissance nous a permis de conclure que la paramétrisation par ondelettes avec seuillage des coefficients semble être plus résistante au bruit.

En outre, une fusion de différente paramétrisation nous semble intéressante (MFCC + ondelettes).

CONCLUSION ET PERSPECTIVES

Conclusion

Dans ce mémoire nous avons présenté les technologies de base intervenant dans la conception d'un système de reconnaissance de mots isolés arabes : l'analyse acoustique par traitement du signal l'utilisation de corpus d'apprentissage, la modélisation statistique par les Modèles de Markov cachés continus. Nous avons également présenté une étude comparative entre les différentes techniques d'extraction de caractéristiques (LPC, LPCC, MFCC et les coefficients différentiels) et aussi une étude sur le choix des paramètres de modèle HMM (nombre d'états et le nombre de gaussiennes).

Notre travail présenté dans ce mémoire a consisté à proposer et à étudier un nouveau module d'extraction de paramètres basé sur la représentation en ondelettes. Nous avons présenté une étude expérimentale sur le choix de nombre de niveaux, la fonction d'ondelette, la résolution fréquentielle et les coefficients informatifs dans la décomposition. Nous avons également présenté une étude comparative entre les différentes techniques d'extraction de paramètres provenant de la transformée en ondelette.
Nous avons complété cette étude par une analyse de l'influence du bruit sur les méthodes d'analyse.

Après l'analyse des résultats, nous avons pu voir que :

- La paramètrisation proposée présente un niveau de performance tout a fait comparable à l'état de l'art.
- La nouvelle paramétrisation est une approche intéressante pour :
  - L'amélioration des performances des systèmes de reconnaissance en conditions bruitées (problème de robustesse).
  - La réduction du nombre de paramètres de système de reconnaissance.
- La fusion de différente paramétrisation (MFCC + ondelettes) semble intéressante.

En conclusion, les ondelettes se révèlent en effet être de très bonnes bases d'analyse du signal parole, permettant une représentation temps-fréquence plus exacte que celle obtenue par la transformée de Fourier.

Perspectives

Différentes perspectives existantes pour le prolongement de ce travail.

- La plupart des tests ont été effectués sur une tâche de reconnaissance globale de mots isolés, il serait donc intéressant de tester cette paramètrisation sur une tâche de reconnaissance de phonèmes et/ou de parole continue. Cette évaluation passe entre autre par l'étude de la matrice de covariance.
- Augmentation de nombre de tests qui portent sur le changement de paramètres d'analyse par ondelettes, pour affirmer un choix final.
- Faire des tests sur des données bruitées dans de multiples conditions.
- Introduction d'une fonction de coût pour le choix de la structure arborescente qui s'adapte mieux avec la reconnaissance (entropie par exemple).
- Modélisation de ces paramètres par la combinaison des HMM avec les réseaux de neurones (méthode hybride).
- Il est envisageable d'introduire cette paramètrisation dans la plate forme HTK.

Le mot de la fin

Nous espérons que la lecture a été enrichissante, que les fautes d'orthographe et de grammaire n'étaient pas trop nombreuses, que les références bibliographiques ne comportent plus trop de fautes et que les idées développées dans ce mémoire sont intéressantes.

BIBLIOGRAPHIE

Bibliographie

[B1] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, H. Leich. « *Traitement de la parole* ». Collection Electricité, Presse Polytechniques et Universitaires Romandes, France, 1999.

[B2] R. Boite, M.Kunt. « *Traitement de la parole* ». Presse polytechniques romandes. Edition 1987.

[B3] T.dutoit « *un bilan de des développements récents de traitement automatique de la parole* » Faculté polytechnique de Mons

[B4]- L. Messikh. « *Analyse de la parole continue. Traitement en temps réel* ». Mémoire de Magister, Institut d'électronique, Université d'Annaba, Avril. 1995.

[B5] Laurent Buniet. « *Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques* » .Thèse du Doctorat, février 1997.

[B6] L. Rabiner and B.J. Juang. « *Fundamentals of speech recognition* ». Prentice-Hall, 1993.

[B7] H. Ezzaidi. « *Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques* ». PhD thèse, Université du Québec, 2002.

[B8] H. Ezzaidi and J. Rouat. « *Pitch and mfcc dependent gmm models for speaker* » identification. IEEE Canadian Conference on Electrical and Computer Engineering, 2004.

[B9] J. Rouat, R. Pichevar, and S. Loiselle. « *Perceptive, non-linear speech processing and spiking neural networks* ». International Summe School on Neural Nets"E.R. Caianiello", IX

[B10] J.B. Allen. « *How do humans process and recognize speech?* ». IEEE Trans. on Speech and Audio Processing, 2(4):567–577, 1994.

[B11] J. Mariani. « *Reconnaissance de la parole- Traitement automatique du langage parlé* ». Hermés Science Publications, 2002.

[B12] H. Hermansky « *Should recognizers have ears?* » Proc. ESCA Tutorial and ResearchWorkshop

[B13] Thèse, *« Etude de la représentation du signal parole à partir de représentation en ondelettes* », chapitres 1 et 3, p3-39, p83-119. Par Christophe Gérard, décembre 1995.

[B14] M. Kunt, « *Techniques modernes de traitement numérique des signaux* », Presses polytechniques et universitaires Romandes, 1991

[B15] Joeseph W.Picone. « *Signal modeling techniques in speech recognition*. » Proceeding of the IEEE, 81(9): 1214-1247, September 1993

[B16] C.Becchetti, L.R.Ricotti. « *Speech Recognition '. Theory and c++ implementation* ». John Wiley and Sons. 1999.

[B17] F.J. Harris, « *On the use of windows for harmonic analysis with the discrete Fourier transform* », Proc. Of the IEEE, vol. 66, no. 1, pp. 51-83, 1978.

[B18] J-Makhoul. « *Linear prediction: a tutorial revie*».proc IEEE vol, pp 580 –561, Apr 1975

[B19] «*La parole et son traitement automatique*». Par Calliope (nom collectif représentant les 36 auteurs de cet ouvrage). Editions Masson, 1989.

[B20] Barkatov, K.production, « *Perception du signal vocale* ». Rapport technique, France télécoms R&D Lannion, 2002

[B21] H. Hermensky, « *Perceptual linear predictive analysis of speech*, »Journal of the acoustic Soc.Am, vol87, 1990.

[B22], H. Hermensky and N. Morgan, « *Rasta processing of speech* », "IEEE Trans on speech and Audio Processing, vol2, pp 578, 1994.

[B23] S. Mallat. A Wavelet Tour of Signal Processing. Academic Press, 1998.

[B24] O. Farook and S. Datta , « *Mel Filter-Like Admissible Wavelet Packet Structure for Speech Recognition* », IEEE Signal Processing Letters , vol , 8 , 7, July 2001, pp. 196-198

[B25] R. Gemello, D. Albesano, L. Moisa, and R. De Mori., « *Integration of fixed and multiple Resolution analysis in a speech recognition system.* » In ICASSP, 2001.

[B26] E. Erzin, A.E. Cetin, and Y. Yardimci. «*Subband analysis for robust speech recognition in The presence of car noise.* » In ICASSP, 1995.

[B27] Beng T. Tan, Minyue Fu, Andrew Spray, and Phillip Dermody. «*The use of wavelet transforms in phoneme recognition* ». In ICSLP, 1996.

[B28] Jaideva C. Goswami ET Andrew K. Chan. « *Fundamentals of Wavelets: Theory, Algorithms, and Applications* ». Wiley Series in Microwave and Optical Engineering. Wiley Interscience, 1999. ISBN 0-471-19748-3.

[B29] S. Mallat. . « *A Theory for Multiresolution Signal Decomposition*»: IEEE. Transactions on pattern analysis and machine intelligence.Vol, No 70 July

[B30] J.F. Kaiser. . «On a Simple Algorithm to Calculate the Energy of a Signal». In *ICASSP-90*, 1990.

[B31]. *S.* Chang, Y. Kwon, ***S.*** Yang, . «*Speech feature extracted from adaptive wavelet for speech recognition*», Ectronics Letters**,** vol. 34, no. 23, November 1998, pp.2211 2213.

[B32] Tapan K.Sarkar et Magdalena Salazar-Palma, « *Ondelettes et théorie des filtres ' à partir du text (R 309)* ». Traité Mesures et contrôle, Technique de l'ingénieur.

[B33] Béatrice Persquet-popescu et Jean-Christophe Persquet, « *Ondelettes et Applications* » à partir du text (TE 215). Traité télécoms, Technique de l'ingénieur

[B34] *« Le traitement du signal sous matlab* », Hermes Science Puplications, Paris ,2000

[B35] Murat Deviren, « Systèmes *de reconnaissance de la parole revisités : Réseaux Bayesiens dynamiques et nouveaux paradigmes* » Thèse présentée et soutenue publiquement le 20/10/2004

[B36] R. Sarikaya and J. H. L. Hansen. «*High resolution speech feature extraction parameterization for monophone-based stressed speech recognition*». IEEE Signal Processing Letters, 7(7):182_185, July 2000.

[B37] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen. «*Wavelet packet transform features with Application to speaker identification*». In IEEE Nordic Signal Proc. Symp., 1998.

[B38] Olivier Deroo, «*Modèles dépendants du contexte et méthodes de fusion de données à la reconnaissance de la parole par modèles hybrides HMM/MLP*», Thèse de Doctorat de la Faculté Polytechnique de Mons, Laboratoire TCTS Mons, Décembre 1998.

[B39] Abdel Belaid. Yolande Belaid « *reconnaissance des formes méthodes et applications* » interEditions

[B40] -S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. « *An introduction to the application of theory of probabilistic function of a markov process to automatic speech recognition* ». B.S.T.J vol 62, N° 4, pp 1035-1074, Apr 1983.

[B41] « *La parole et son traitement automatique*». Par Calliope (nom collectif représentant les 36 auteurs de cet ouvrage). Editions Masson, 1989.

[B42] -L.R. Rabiner, «*A Tutorial on Hidden Markov models and Selected Application in Speech Recognition*», Proc IEEE, Vol 7 N°2, Feb 89, pp. 257-286.

[B43]. Cherifa Snani *« Conception d'un système de reconnaissance de mots isolés à base de l'approche stochastique en temps réel – Application : commande calculatrice vocale*» Mémoire du magister 2004. U.Annaba

[B44] M. Djemili, *«Reconnaissance de mots isolés arabes par DTW & HMM*». Mémoire de Magister, Institut d'électronique, Université d'Annaba, 2001.

[B45] H.Bourouba, *« Développement d'un système de reconnaissance de la parole isolée à vocabulaire limité par deux approches déterministe et statistique* » Mémoire de Magister, Institut d'Electronique, Université d'Annaba, 2001.

[B46] H. Dahmani. *« La conception d'un système de reconnaissance de mots isolés et enchaînés*», Mémoire de Magister, Institut d'électronique, Université d'Annaba, 2003.

[B47] Mohamed Chetouani «*Codage neuro-prédictif pour l'extraction de caractéristiques de signaux de parole* ». Thèse de Doctorat Soutenue le 14 décembre 2004

[B48] Claude Barras. *« Reconnaissance de la parole continue, adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés* », Thèse de Doctorat Soutenue le 29 mai 1996.

[B49] Steve young *« The htk book* » Microsoft Corporation first published December 1995.

[B50] Dan Mircea Istrate. « *Détection et reconnaissance des sons pour la surveillance médicale* » Thèse de Doctorat Soutenue le 16 décembre 2003

[B51] H.Wassner and G. Chollet. *«New cepstral representation using wavelet analysis and* spectral transformation for robust speech recognition. » In Proc. ICSLP '96, volume 1, pages 260_263, Philadelphia, PA, 1996.

[B52] «*Transformation for robust speech recognition.* » In Proc. ICSLP '96, volume 1, pages 260_263, Philadelphia, PA, 1996.

Résumé

La conception d'un système de reconnaissance nécessite de porter un soin particulier à chacune des étapes (prétraitement, extraction de paramètres, modélisation,……). Au niveau d'extraction de paramètres les coefficients MFCC , LPCC ou LPC ont permis d'obtenir les meilleurs résultats en reconnaissance de mots isolés, mots enchaînés et parole continue dans des conditions de laboratoire ou pour des tâches simples. En revanche, dans des conditions réelles de traitement de la parole (milieu bruité, parole spontanée. . .), les performances de ces systèmes se dégradent considérablement. Ceci nous a motivé d'étudier et de mettre en oeuvre des nouveaux paramètres acoustiques qui peuvent améliorer, soit la robustesse aux bruits, soit le taux de reconnaissance ou d'obtenir les mêmes performances avec un nombre réduit de paramètres.

Les représentations en ondelettes présentent des propriétés intéressantes pour paramétriser le signal de parole. L'objet de notre travail est donc de développer un système de reconnaissance de chiffres arabes isolés (de 0 à 9) multi-locuteur basé sur les modèles de Markov cachés continus et de déterminer l'apport des représentations en ondelettes sur le taux de reconnaissance. Lors de cette tâche, nous avons comparé les différentes techniques de paramétrisation provenant de la transformée en ondelettes avec celles-ci de l'état de l'art.

Les paramètres classiques LPC, MFCC et LPCC sont testés tout d'abord, les nouveaux paramètres sont ensuite proposés et évalués.

Afin de valider nos paramétrisations, nous avons choisi au premier temps la plate forme logiciel HTK, mais nous avons trouvé des difficultés d'intégrer nos sous-programmes concernant les nouveaux paramètres. On a fait recours au matlab qui contient un toolbox pour les ondelettes.

ملخـــــــــــص

إن إنجاز نظام للتعرف علي الكلام يتطلب عناية خاصة بكل مرحلة من مراحل الإنجاز(المعالجة المسبقة ، استخراج الوسائط ، النمذجة ،.....) . علي مستوي استخراج الوسائط ، المعاملات الصوتية **MFCC, LPC, LPCC** سمحت بالحصول علي نتائج جيدة في التعرف علي الكلمات المعزولة ، المتسلسلة والكلام المستمر في الظروف المخبرية أو من أجل مهام بسيطة، في المقابل و في الظروف الحقيقية لمعالجة الكـــــلام (وسط مشوش ، كلام حر....) مزايا هذه الأنظمة تتقهقر بصفة ملحوظة مما شجعنا علي دراسة وتجسيد وسائـــط صوتية جديدة التي تسمح بتحسين الصمود تجاه التشويش، نسبـــــــة التعرف أو الحصول علي نفس المزايـــا بعدد أقل من الوسائـــط

التمثيل بواسطة الموجات المصغرة يملك خصائص مهمة لتوسيط إشارة الكلام .الهدف من هذا العمـــل هو تطوير نظام تعرف علي الكلام خاص بالأرقام العربية المعزولة(من 0 إلي9) متعدد المتحدثـــــين مؤسس علي نمـــــاذج مركوف الخفية و تعيين تأثير التمثيلات بالموجات المصغرة علي نسبـــــة التعرف . خلال هذه الدراسة قمنـــا بإجـراء مقارنة بين مختلف تقنيات التوسيط المستنتجـة من التمثيل بالموجـــات المصغرة مع الوسائط العتيقـــــة. كما قمنا باختبار الوسائط العتيقة أولا بعد ذلك تم اقتراح الوسائـــــط الجديدة ثم تقديرها.

من أجل تمكين الوسائط الجديدة، في البداية تم اختيار علبـة الأدوات **HTK** لغرس البرامـــــــج الخاصة بالوسائـــــــط الجديدة، لكن وجدنا صعوبة في إدراج البرامج الجديدة، فتم اللجوء إلي برنامـــــج ما ـطـــــــلاب الذي يحـــــتوي علي علبة أدوات خاصة بالموجـــات المصغـرة.

Abstract

The design of a recognition system requires carrying a care particular to each step (pre-treatment, extraction of parameters, modelling ...). The level of features extraction .MFCC, LPCC or LPC made it possible to obtain the best results in recognition of isolated words, connected words and continuous speech under of laboratory conditions or for simple tasks. On the other hand, under real conditions of speech processing (disturbed medium, spontaneous speech…); the performances of these systems are degraded considerably. This motivated us to study and implement new acoustic features which can improve, either the robustness with the noises, or the rate of recognition or to obtain the same performances with a reduced number of features.

The wavelets representations present properties interesting for the speech signal parameterisation The aim of our work is thus to develop a recognition system of isolated Arabic digits (from 0 to 9) multi speaker based on the continuous hidden Markov models and to determine the contribution of wavelets representations on the recognition rate. At the time of this task, we compared the various techniques of parameterization coming from the wavelet transform with those from the state of the art.

Traditional features LPC, MFCC and LPCC are tested first of all; the news features are then proposed and evaluated.

In order to validate our parameterizations, we chose at the first time the HTK toolkits, but we found difficulties in integrating our programs concerning the new features. One made recourse to the matlab which contains a wavelets toolbox .