N° d'ordre : 01/2007-E/EL

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene

Faculté d'Electronique et d'Informatique

THESE

Présentée pour l'obtention du diplôme de DOCTORAT D'ÉTAT

EN : ELECTRONIQUE
Spécialité : Electronique des Systèmes

Par : Mohamed DEBYECHE

Reconnaissance Automatique de la Parole Appliquée à la Langue Arabe

Soutenue publiquement le 18/02/2007, devant le jury composé de :

Redouane TOUMI	Professeur (USTHB)	Président
Jean Paul HATON	Professeur (Univ. Henry Poincaré, Nancy)	Directeur de thèse
Amrane HOUACINE	Professeur (USTHB)	Codirecteur de thèse
Malika BOUDRAA	Professeur (USTHB)	Examinatrice
Amar DJERADI	Professeur (USTHB)	Examinateur
Latifa HAMAMI	Maître de Conférence (ENP d'Alger)	Examinatrice
Mhania GUERTI	Maître de Conférence (ENP d'Alger)	Examinatrice

Résumé

La reconnaissance automatique de la parole (RAP), processus de transformation permettant le passage du signal acoustique à une suite d'unités phonétiques discrètes, joue un rôle fondamental dans tout système de Dialogue Homme-Machine (DHM). Ce travail est à inscrire dans la problématique générale posée par la reconnaissance automatique de la parole avec comme langue d'intérêt la langue arabe. Nous avons ainsi élaboré des systèmes de reconnaissance émanant d'approches différentes : une approche analytique fondée sur les connaissances gérées par un système expert, une approche statistique basée sur les modèles de Markov cachés (hidden Markov model pour l'anglais) et une approche hybride mixant les modèles de Markov et les réseaux de neurones artificiels.

La première approche mise en œuvre est l'approche analytique fondée sur les connaissances acoustiques et phonétiques extraites de la lecture de spectrogrammes numériques. Ces connaissances sont formalisées à l'aide d'outils propres à l'intelligence artificielle, les règles de production. Pour les besoins de l'élaboration de cette approche, un logiciel d'analyse interactif de la parole que nous avons appelé **APHAK** (**A**cquisition-**PH**onetic-**A**coustic-**K**nowledge) a été développé.

La deuxième approche de reconnaissance mise en œuvre est l'approche statistique basée sur les modèles de Markov cachés (MMCs). Ces modèles gèrent les distorsions temporelles du signal acoustique en s'appuyant sur des densités de probabilités pour modéliser les distorsions en fréquence. Plusieurs modifications sont introduites sur ces modèles afin d'augmenter leur capacité de discrimination. Ces modifications ont concerné la résolution acoustique (nature du vecteur acoustique d'entrée), l'utilisation de dictionnaire spécifique à chaque composante du vecteur d'entrée, le type de modèles (modèles indépendants et dépendants du contexte) et enfin la modélisation à dictionnaires multiples où chaque modèle possède son propre dictionnaire.

Nous avons également mis en œuvre l'approche modèle de Markov caché multibandes. Le modèle multibandes est un modèle parallèle de reconnaissance permettant de traiter de manière indépendante l'information partielle du signal de parole. Ce modèle a été étudié en environnements calme et bruité.

Un des problèmes inhérents à la structure des modèles de Markov cachés discrets est lié à la perte d'information sur le signal acoustique original engendrée par la procédure de quantification vectorielle (QV). Afin de palier à ce problème, une approche originale de QV que nous avons appelé quantification vectorielle distribuée (QVD) a été mise en œuvre. Celle-ci permet une distribution optimale des composantes du dictionnaire sur les états du modèle et réalise une unification entre la micro structure acoustique du signal de parole et la macro structure phonétique lors du processus d'estimation des paramètres des modèles. Cette nouvelle technique de QV a été implémentée en deux variantes. La première variante utilise l'algorithme des K-moyennes afin d'optimiser le processus de QV alors que la seconde exploite, pour le même objectif, la capacité propre de classification des réseaux de neurones.

Afin de réaliser les expériences de reconnaissance comparatives (approche experte vs approche markovienne) et de valider les différentes améliorations proposées, un corpus d'apprentissage et de test prononcé par des locuteurs natifs algériens a été établi. Ces expériences ont porté sur l'identification des phonèmes spécifiques à la langue arabe à savoir les consonnes glottales, pharyngales et emphatiques. Le choix de ces consonnes est motivé par le fait qu'elles sont reconnues unanimement comme responsables des limites des systèmes de reconnaissance dédiés à la langue arabe.

Les résultats de reconnaissance obtenus indiquent que l'approche statistique basée sur les modèles de Markov cachés est plus performante que l'approche analytique fondée sur les connaissances. Aussi, la QV distribuée en particulier dans sa variante combinant les réseaux de neurones artificiels entraînés sur le critère de maximisation de l'information mutuelle améliore la performance des modèles en terme de taux de reconnaissance et de vitesse de décodage.

Remerciements

Mes premiers remerciements vont à Monsieur Jean Paul HATON, Professeur à l'université Henry Poincaré de Nancy, pour son soutien de toujours et l'intérêt constant avec lequel il a suivi mon travail. Qu'il me soit permis de lui témoigner ici ma profonde gratitude.

Je tiens également à remercier Monsieur Amrane Houacine, Professeur à la Faculté d'Electronique et d'Informatique, qui m'a accordé sa confiance et accepter de co-diriger ce travail. Qu'il trouve ici l'expression de toute ma gratitude pour tous les conseils prodigués.

Je tiens à remercier Monsieur Redouane TOUMI, Professeur à la Faculté d'Electronique et d'Informatique, pour avoir accepter la présidence de ce jury.

Je remercie également Madame Latifa Hamami, Maître de Conférence à l'Ecole Nationale Polytechnique, Madame Malika Boudraa, Professeur à la Faculté d'Electronique et d'Informatique, Mademoiselle Mhania Guerti, Maître de Conférence à l'Ecole Nationale Polytechnique et Monsieur Amar Djeradi, Directeur du laboratoire LCPTS et Professeur à la faculté d'Electronique et d'Informatique, pour avoir accepter de faire partie de ce jury.

A tous les membres du laboratoire LORIA que j'ai eu le plaisir de côtoyer lors de mes différents stages de recherche, je leur adresse mes amitiés les plus sincères.

Qu'il me soit permis enfin d'associer dans une même pensée amicale tous ceux qui ont contribué à la réalisation de ce travail. Qu'ils trouvent tous ici l'expression de ma très vive sympathie.

Tables des Matières

Liste des Figures

Liste des Tableaux

Listes des Abréviations

DHM	Dialogue Homme-Machine
RAP	Reconnaissance Automatique de la Parole
TALN	Traitement Automatique du Langage Naturel
DTW	Dynamic Time Warping
HMM	Hidden Markov Model
MMC	Modèle de Markov Caché
RAL	Reconnaissance Automatique du Locuteur
TF	Transformée de Fourier
TFR	Transformée de Fourier Rapide
LPC	Linear Predictf Coding
AR	Auto-Régressif
MFCC	Mel Frequency Cepstral Coefficient
TFR^{-1}	Transformée de Fourier Rapide Inverse
LP	Linear Predict
PLP	Perceptual Linear Prediction
DCT	Discrete Cosinus Transform
RASTA	RelAtive SpecTrAl.
RNA	Réseau de Neurone Artificiel
MLP	Multi Layer Perceptron
TDNN	Time Delay Neural Network
IA	Intelligence Artificielle
MAP	Maximum A Posteriori
MS-TDNN	Multi States TDNN
RNN	Recurrent Neural Networks
PMC	Perceptron Multi Couches
GMM	Gaussian Mixture Model
EM	Expectation Maximization
MLE	Maximum Likelihood Estimate
QV	Quantification Vectorielle
QVC	Quantification Vectorielle Conventionnelle
QVD	Quantification Vectorielle Distribuée

Introduction Générale

1.1 Introduction

La parole est le moyen de communication le plus naturel entre les humains. Avec le développement de la technologie de l'information et l'utilisation massive de l'ordinateur, le Dialogue Homme-Machine (DHM) utilisant la parole comme moyen de communication fait l'objet d'un intérêt accru de la part aussi bien de la communauté scientifique que de la communauté industrielle. La reconnaissance automatique de la parole (RAP), composante principale du système de DHM, constitue un thème de recherche central dans le domaine plus vaste celui du Traitement Automatique du Langage Naturel (TALN). Les applications de la RAP sont aussi nombreuses que diversifiées, elles existent là où la parole peut remplacer une interface existante pour communiquer avec la machine. Elles peuvent être grossièrement regroupées en quatre catégories :

- **Les applications multimédia :** création de lettres, rapports et autres documents par l'intermédiaire de la parole. Plusieurs produits de dictée vocale sont sur le marché. Nous pouvons citer le logiciel ViaVoice d'IBM ou encore WINSAPI de Microsoft. L'interaction vocale dans les logiciels pédagogiques (ex : outils d'apprentissage des langues), etc.

- **Les applications industrielles :** Commander avec la parole des équipements industriels telles que les machines (robots), contrôle mains libres des équipements de voiture tels que la radio, le conditionnement d'air, le moteur, le téléphone sans fil (ex : voice dialing), etc.

- **Les applications médicales :** aides aux personnes handicapées. Nous pouvons citer les cabines téléphoniques dotées d'un système de reconnaissance pour la composition d'un numéro d'appel pour les handicapés moteurs. Dans le cas des malentendants, les prothèses auditives basées sur la reconnaissance de la parole.

- **Les applications téléphoniques :** l'automatisation des transactions téléphoniques (ex : les opérations bancaires). Le self service téléphonique pour l'accès à des services d'informations (ex : consulter un bulletin météorologique), etc.

1.2 Facteurs de complexité liés à la reconnaissance

Le but de la reconnaissance automatique de la parole est de développer des techniques et des systèmes permettant aux ordinateurs d'accepter la parole comme entrée. La reconnaissance de la parole peut donc être vue comme une opération de transformation du signal de parole en texte en utilisant l'information contenue dans le signal et la connaissance a priori du domaine. En termes plus simples les personnes souhaitent que leur voix, signal de parole comme illustré par la **Fig. I.1**, soit transcrite en texte.

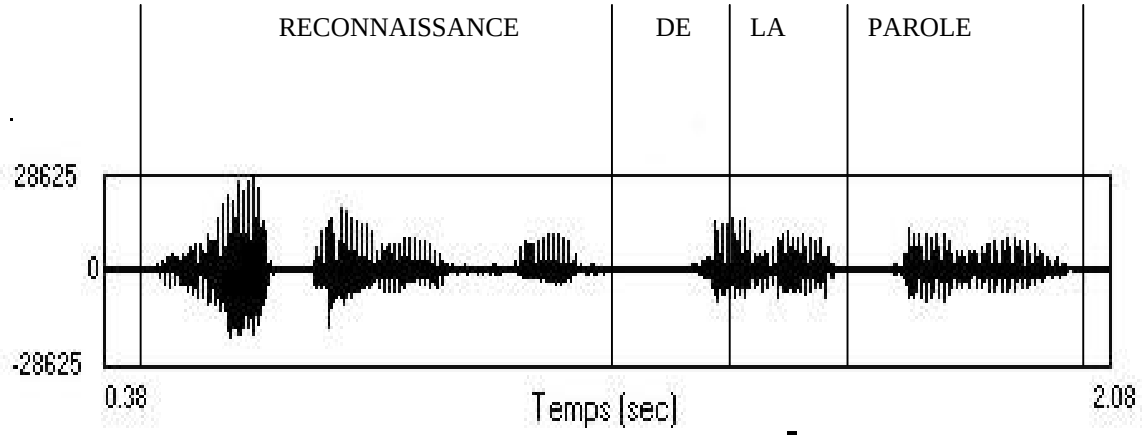


Fig. 1.1 – Exemple de représentation temporelle du signal de parole

Le signal de parole présente plusieurs propriétés qui lui sont spécifiques et qui posent un véritable défi aux chercheurs s'efforçant de comprendre les processus sous-jacents à son traitement automatique et en particulier à sa reconnaissance et cela depuis les années 50. En premier lieu la parole est un phénomène séquentiel étalé dans le temps par conséquent les processus mis en œuvre dans la reconnaissance sont assujetties à une contrainte temporelle externe : l'ordre dans lequel les sons de la parole aboutissent à l'oreille. En deuxième lieu, la parole est continue. Le caractère continu, ininterrompu de la parole soulève un problème majeur qui est celui du passage entre continu et discret, c'est-à-dire la mise en correspondance entre un signal d'entrée continu et des représentations lexicales discrètes.

1.3 Classification des systèmes de reconnaissance

La difficulté de la tâche de reconnaissance fait que les systèmes développés existants sur le marché peuvent être classés en fonction du mode de fonctionnement, du mode d'élocution, de la taille du vocabulaire à reconnaître, de la syntaxe du vocabulaire et de la robustesse.

Le mode de fonctionnement

Les systèmes de reconnaissance peuvent être classés en trois catégories selon leurs mode de fonctionnement. On distingue les systèmes dépendants du locuteur ou systèmes mono locuteur, les systèmes multi-locuteurs (pouvant être utilisés par plusieurs locuteurs) et les systèmes indépendant du locuteur. Dans le cas mono locuteur, le système de reconnaissance a été configuré pour un locuteur spécifique. Ce type de système conduit généralement à de bons résultats de reconnaissance mais pour chaque nouveau locuteur une nouvelle session de configuration est requise. La plupart des logiciels de reconnaissance de parole se trouvant sur le marché actuellement sont basés sur ce type de fonctionnement. Les systèmes multi-locuteurs fonctionnent bien pour certains groupes de locuteurs pour lesquels le système a été configuré. Le passage d'un locuteur à un autre du même groupe se fait sans phase d'adaptation. Les systèmes indépendant du locuteur sont configurés une seule fois afin de modéliser une variété de voix (différents locuteurs). En raison de la grande variabilité induite, les systèmes indépendants du locuteurs sont moins précis que les systèmes mono locuteur ou multi-locuteurs. Ils sont donc les plus difficiles à mettre en œuvre.

Le type d'élocution

Le type d'élocution caractérise la façon dont on peut parler au système. Il existe quatre type d'élocution distincts. Le type mots isolés, le locuteur doit prononcer chaque mot isolément ou dans le cas d'une phrase chaque mot doit être séparé du reste par une pause distincte. Le type mots connectés, dans ce cas chaque mot est clairement articulé sans pause volontaire pour les séparer. Le type parole continue lue, dans ce cas il n'existe pas de pauses entre les mots et les mots ne sont pas toujours clairement articulés. C'est le discours usuel, si ce n'est que les textes sont lus. Le type parole continue spontanée, c'est le discours usuel sans aucune contrainte. Les systèmes de reconnaissance de type mots isolés peuvent atteindre des taux de reconnaissance relativement bien de nos jours. Ces types de systèmes sont généralement destinés à des commandes de machines nécessitant un vocabulaire limité ou à des compositions d'appels téléphoniques dans des cabines dédiées aux personnes handicapées. Les systèmes mots connectés sont des cas particuliers comme la reconnaissance de nombres quelconques. Les systèmes parole continue sont plus difficile à mettre en œuvre. Il n'existe pas encore de systèmes avec des performances comparables à celle des humains. La complexité de la tache devient davantage ardue pour le cas des systèmes parole continue spontanée, la parole spontanée est souvent agrammatical et mal structurée.

La taille du vocabulaire

Le vocabulaire est l'ensemble des mots que le système est capable de reconnaître. Il est évident que la performance du système diminue quant la taille du vocabulaire augmente. Pour des systèmes large vocabulaire il est impossible d'utiliser les modèles de mots parce que cela nécessite non seulement une large base de données mais aussi un nombres de paramètres prohibitifs . Dans ce cas le choix d'unités de reconnaissance plus compact et en nombre limité (phonèmes, syllabes) est nécessaire.

La syntaxe

La syntaxe spécifie les contraintes imposées sur les suites de mots prononcés. Elle peut être inexistante (tout mot peut suivre n'importe quel autre mot), ou bien contraignante (après un mot donné seuls certains autres sont autorisés). Dans beaucoup de cas la tâche du système de reconnaissance se trouve simplifiée par l'utilisation de la syntaxe. Des modèles de types N-grammes (tout mot peut être suivi par une séquence de N-1 autres avec une probabilité fixée) sont utilisées. En général des modèles de grammaire du type bi-grammes (N=2) et tri-grammes (N=3) sont communément utilisés en reconnaissance de la parole.

La robustesse

La robustesse caractérise la capacité du système à fonctionner dans des conditions difficiles. De nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance ont été identifiées :

- Les bruits d'environnement tels que bruits additifs stationnaires ou non stationnaires (par exemple, bruits des moyens de transports ou bruits des systèmes industriels).
- Utilisation de différents microphones et différentes caractéristiques (fonctions de transfert) du système d'acquisition du signal (filtres), conduisant généralement à du bruit de convolution.
- Acoustique déformée et bruits (additifs) corrélés avec le signal de parole utile (par exemple, distorsions non linéaires et réverbérations).
- Bande passante fréquentielle limitée (par exemple dans le cas des lignes téléphoniques).

- Elocution inhabituelle ou altérée, comprenant entre autres : l'effet Lombard, (qui désigne toutes les modifications, souvent inaudibles, du signal acoustique lors de l'élocution en milieu bruité), le stress physique ou émotionnel, une vitesse d'élocution inhabituelle, ainsi que les bruits de lèvres ou de respiration.

I.4 Approches de reconnaissance

La reconnaissance automatique de la parole a été appréhendée jusqu'à présent selon deux types d'approches fondamentalement différentes. Une première approche dite analytique fondée sur des connaissances d'ordre acoustiques, phonétiques et phonologiques. Ces connaissances sont formalisées à l'aide d'outils propres à l'intelligence artificielle comme les systèmes experts, les systèmes à base de frames, etc. Une deuxième approche dite globale basée sur les principes de la reconnaissance des formes. Dans cette approche, on tente de modéliser au mieux des unités représentatives du signal de parole. Deux types de modélisation sont utilisées: la modélisation déterministe qui a conduit aux premiers systèmes de reconnaissance basés sur les principes de la programmation dynamique utilisant différentes variantes de l'algorithme DTW (Dynamic Time Warping, pour l'anglais) et la modélisation statistique basée sur le modèles de Markov cachés (MMC). Cette modélisation par MMC est à l'origine des avancées majeures réalisées actuellement en RAP.

1.5 Objet de la thèse

La langue anglaise a capté toutes les avancées majeures réalisées dans le domaine de la reconnaissance automatique de la parole par conséquent ces propriétés acoustiques, phonétiques et linguistiques pertinentes sont maintenant bien établies. De même, des langues telles que le français, l'espagnol, le japonais ou le mandarin peuvent être à un degré moins considérées comme des langues assez dotées. Actuellement avec l'expansion fulgurante des nouvelles technologies de l'information le développement de systèmes de reconnaissance multilingues devient un objectif de beaucoup de laboratoires de recherche. Il est donc impératif que des études sur la langue arabe soient menées afin de la rendre accessibles à ces nouvelles technologies.
Dans ce cadre, l'objet de cette thèse est la reconnaissance automatique de la parole appliquée à la langue arabe. Pour atteindre cet objectif nous avons eu à traiter plusieurs axes de recherche comprenant l'analyse du signal vocal, l'intelligence artificielle, la reconnaissance des formes. Cette pluridisciplinarité du travail réalisé ne se limite évidement pas aux seuls axes précités, il faut leur adjoindre l'acoustique, la phonétique, la linguistique, la théorie de l'information, les probabilités ou encore l'algorithmique.
Dans cette thèse plusieurs approches de reconnaissance ont été mise en œuvre. Nous avons commencé par l'approche analytique où nous avons mis au point une base de connaissances relative aux consonnes spécifiques à la langue arabe. Ces connaissances ont été formalisées sous forme de règles de production. Le décodage est réalisé grâce à un système expert. L'acquisition des connaissances indispensables à la mise en œuvre de cette approche a nécessité le développement d'outils d'analyse et d'affichage graphiques. Un logiciel interactif d'analyse de la parole que nous avons appelé **APHAK** (**A**cquisition-**PH**onetic-**A**coustic-**K**nowledge) a été, à cet effet, développé.
Ces dernières années les modèles de Markov cachés (MMC) se sont imposés comme l'approche la plus utilisée en reconnaissance automatique de la parole aussi bien pour l'anglais que pour des langues telles que le français, l'espagnol ou le japonais. Dans cet objectif, Il était donc intéressant pour nous de développer cette approche et de déterminer si effectivement l'approche MMC peut complètement intégrer les aspects phonétiques, particulièrement dans le contexte des consonnes emphatiques et arrières arabes, pour finalement les considérer comme des modèles ordinaires. Cette approche "tout automatique" a été opposée à celle analytique fondée sur les connaissances.

Plusieurs types de modèles de Markov cachés tels que les modèles discrets, continus ou semi-continus **[1]**, **[2]** sont actuellement développés et appliqués avec succès pour la reconnaissance de la parole. Il reste que les modèles discrets sont attractifs de par leur relative complexité algorithmique et leur vitesse de décodage c'est pourquoi plusieurs travaux leurs sont consacrés **[3]**, **[4]**, **[5]**. Plus récemment, dans le contexte de la croissance prodigieuse des applications réseaux, les systèmes de reconnaissance basés sur les modèles discrets qui utilisent la quantification vectorielle (QV) constituent une solution utile, peu coûteuse et prometteuse **[6]**, **[7]**. Nous nous sommes donc intéressés aux modèles de Markov cachés de types discrets en proposant des modifications dans le but d'améliorer leurs discriminations. En effet, à travers cette étude, nous établirons la corrélation entre la performance et la topologie des modèles. Nous montrerons l'apport de l'analyse acoustique multi-variables permettant d'aboutir à une résolution acoustique optimale, l'intérêt de l'approche multi-dictionnaires consistant à allouer aux composantes du vecteur acoustique multi-variables des dictionnaires spécifiques et enfin l'utilité de la modélisation explicite des effets de coarticulation par l'utilisation de modèles dépendants du contexte.

Toujours dans la perspective d'amélioration des performances des modèles nous proposons une nouvelle approche de quantification vectorielle (QV) combinée aux modèles discrets. L'originalité de la QV proposée est de permettre une distribution optimale des composantes du dictionnaire sur les états du modèle et de réaliser une unification entre la micro structure acoustique du signal de parole et la macro structure phonétique lors du processus d'estimation des paramètres des modèles. Cette nouvelle technique de QV a été implémentée en deux variantes. La première variante utilise l'algorithme des K-moyennes afin d'optimiser le processus de QV alors que la seconde exploite, pour le même objectif, la capacité propre de classification des réseaux de neurones artificiels.

1.6 Structure de la thèse

Cette thèse est structurée en six chapitres. Après ce premier chapitre introduction générale, nous décrirons dans le second chapitre, le signal de parole en mettant en évidence ces principales caractéristiques ainsi que les difficultés rencontrées lors de sa modélisation. Nous présenterons ensuite les analyses acoustiques spécifiques effectuées sur le signal de parole avant tout processus de reconnaissance. Enfin, nous traiterons de la reconnaissance de la parole, un état de l'art qui aboutit à une synthèse des différentes approches de reconnaissance rencontrées jusqu'à présent.
Le troisième chapitre sera consacré à la phonétique arabe et à l'approche analytique mise en œuvre. Dans ce chapitre seront décrits : les particularités de la langue arabe, les problèmes liés à sa reconnaissance, le système de reconnaissance analytique fondé sur les connaissances, l'analyse acoustique des consonnes arrières et emphatiques arabes réalisée avec les connaissances déduites. Nous terminerons ce chapitre par le formalisme adopté et les résultats de reconnaissance obtenus.
Le quatrième chapitre sera consacré à l'approche globale mise en œuvre. Nous présenterons les différentes étapes rentrants dans le processus de reconnaissance basé sur les modèles de Markov cachés (MMC), les expériences comparatives de reconnaissance ainsi que les différentes améliorations apportées à l'approche MMC.
Dans le cinquième chapitre nous présenterons l'approche MMC multibandes mise en œuvre ainsi que les expériences réalisées en environnement calme et bruité.
Le sixième chapitre sera lui consacré à la nouvelle approche de quantification vectorielle (QV) proposée à savoir la QV distribuée avec ces deux variantes d'implémentation, les expériences réalisées ainsi que les résultats comparatifs obtenus.
Nous résumerons dans le dernier chapitre, conclusions et perspectives, les différentes méthodes testées, les améliorations obtenues et les perspectives de ce travail.

Signal de Parole et Reconnaissance Automatique

2.1 Introduction

Dans ce chapitre nous y exposerons tout d'abord les mécanismes de production et de perception de la parole chez les humains. Nous mettrons en relief les caractéristiques particulières du signal de parole, responsables de la difficulté de la tâche de reconnaissance. Nous présenterons les méthodes d'analyse les plus adaptées à la RAP. Nous traiterons enfin de la reconnaissance automatique de la parole, un état de l'art qui aboutit à une synthèse des différentes approches de reconnaissance rencontrées jusqu'à présent.

2.2 Le signal de parole

2.2.1 Production et perception de la parole

La **Fig. 2.1** illustre le mécanisme de production et de perception de la parole chez les humains. Le processus de production commence quand un locuteur formule dans son cerveau le message à communiquer (préparation conceptuelle du message), s'ensuit l'étape de conversion du message dans un langage structuré (syntaxe). Le locuteur exécute ensuite une série d'actions volontaires et coordonnées d'un certain nombre de muscles du système articulatoire (produire un souffle, faire vibrer les cordes vocales, modeler et faire résonner les vibrations) permettant de produire la séquence de parole correspondante. Cette séquence se matérialise à la sortie du conduit vocal par un signal acoustique. Une fois le signal acoustique généré et propagé vers l'auditeur, le processus de perception commence. D'abord, l'auditeur traite le signal acoustique à travers la membrane basilaire dans l'aire intérieure de l'oreille, une analyse spectrale du signal entrant est réalisée, un processus neural de transcodage convertit le spectre en signaux d'activité du nerf auditif qui correspond presque exactement à un processus d'extraction de paramètres pertinents du signal. L'activité neurale le long du nerf auditif est alors convertie dans un langage structuré à l'intérieur du cerveau où la perception du message est réalisée.

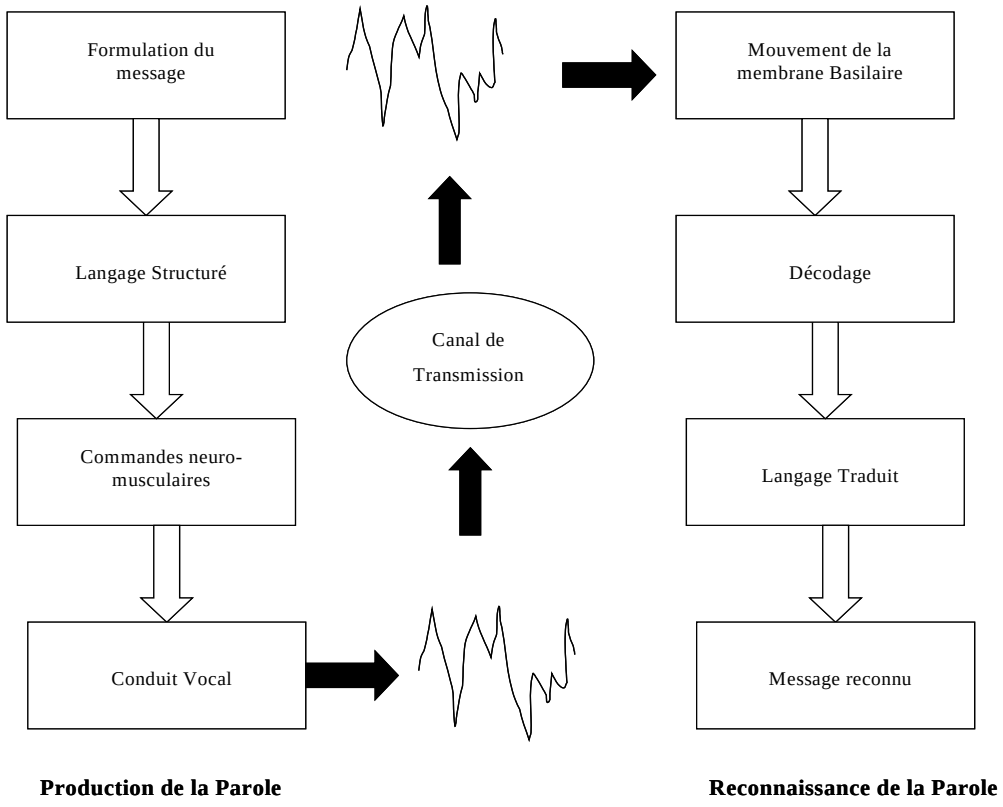


Fig. 2.1 - Mécanisme de production et de perception de la parole chez les humains.

2.2.2 Informations contenues dans le signal de parole

Le signal de parole véhicule des informations de nature très différentes, ceci impose aux systèmes de reconnaissance de n'extraire que l'information nécessaire à son application. L'information sur celui qui a émis le message pour la reconnaissance du locuteur (RAL), l'information sur l'état psychologique de celui qui a émis le message pour la reconnaissance des émotions, dans quelle langue le message a été émis pour la reconnaissance de la langue et enfin l'information linguistique du message émis pour la reconnaissance de la parole (l'objet de cette thèse). Ces différentes informations sont portées par des paramètres tels que le fondamental, les formants, la prosodie, l'enveloppe spectrale, le timbre, les phonèmes, etc.

2.2.3 Caractéristiques du signal de parole

Le signal de parole présente plusieurs caractéristiques spécifiques qui rendent son traitement extrêmement complexe. Parmi ces caractéristiques deux sont fondamentales : la redondance de l'information dans le signal vocal et son extrême variabilité qui se manifeste à différents niveaux.

2.2.3.1 Redondance du signal

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 reconnaissances robustes aux bruits **[8], [9], [10]**.

2.2.3.2 Variabilité du signal

Le signal de parole possède une très grande variabilité. Cette variabilité peut être classée en trois niveaux : le niveau intra-locuteur, le niveau inter-locuteurs et le niveau contextuel.

- **Variabilité intra-locuteur :** cette variabilité identifie les différences dans le signal produit par une même personne. En effet, une même personne ne prononce jamais un mot deux fois de façon identique. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son système d'élocution, son intonation ou sa phraséologie. La variabilité intra-locuteur est cependant beaucoup plus limitée que la variabilité inter-locueturs que nous allons décrire maintenant. Il est en effet possible, malgré les problèmes énoncés ci-avant, de mettre en œuvre des systèmes automatiques d'identification du locuteur, à la manière d'une personne reconnaissant une voix familière. Cette capacité est la preuve qu'une certaine constance existe dans la phase de production de la parole par un même individu.

- **Variabilité inter-locuteurs :** Cette variabilité est un phénomène majeur en reconnaissance de la parole. Comme nous venons de le rappeler, un locuteur reste identifiable par le timbre de sa voix malgré une variabilité qui peut parfois être importante. La contrepartie de cette possibilité d'identification à la voix d'un individu est l'obligation de donner aux différents sons de la parole une définition assez souple pour établir une classification phonétique commune à plusieurs personnes. La cause principale des différences inter-locuteurs est de nature physiologique. La parole est principalement produite comme nous l'avons vu précédemment, grâce aux cordes vocales qui génèrent un son à une fréquence de base, le fondamental. Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par le conduit vocal, cette transformation par convolution permet de générer des sons différents. Or le conduit vocal est de forme et de longueur variable selon les individus et, plus

généralement, selon le genre et l'âge. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes. La variabilité inter-locuteurs trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux. Un exemple de cette variabilité apparaît lorsque vous comparez la voix d'un locuteur du Nord avec celle d'un locuteur du sud de l'Algérie par exemple.

- **Variabilité contextuelle :** Une variabilité aussi importante est la variabilité contextuelle attribuée aux phénomènes dits de coarticulation. On désigne par coarticulation le fait que les mouvements accomplis par les articulateurs dans la production de la parole se chevauchent sur l'axe temporel. Le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis subissent alors l'influence des sons qui les précèdent ou les suivent. Ces effets de coarticulation sont des interférences sur le signal de parole. Ils entraînent l'altération des formes sonores en fonction du contexte. On doit noter d'ailleurs que la prise en compte de cette variabilité améliore la performance des systèmes de reconnaissance.

2.3 Méthodes d'analyse du signal de parole

Cette partie présente les grands principes qui ont conduit aux différentes méthodes d'analyse du signal de parole actuellement utilisées pour la reconnaissance automatique de la parole. Ces méthodes sont les cepstres, le codage par prédiction linéaire et les méthodes basées sur les modèles d'audition. Il faut noter ici qu'avant toute analyse (extraction de paramètres), le signal de parole subit des opérations de mise en forme. La **Fig. 2.2** illustre ces différentes opérations. Le signal est tout d'abord filtré puis échantillonné à une fréquence donnée Fe. Une pré-accentuation est effectuée afin de relever les hautes fréquences, ensuite le signal est fragmenté en trames. Chaque trame est constituée d'un nombre N d'échantillons de parole. En général N est fixé de telle manière que chaque trame corresponde à environ 30ms de parole. Enfin, le fait de traiter un petit morceau de signal amène des problèmes dans le filtrage (effets de bords). Pour éviter cela, des fenêtres de pondération sont utilisées. Ce sont des fonctions que l'on applique à l'ensemble des échantillons prélevés dans la fenêtre du signal original de façon à diminuer les effets de bords. Parmi les fenêtres les plus courantes, nous avons la fenêtre de Hamming. En général, les fenêtres successives se recouvrent et elles doivent avoir une longueur suffisante.

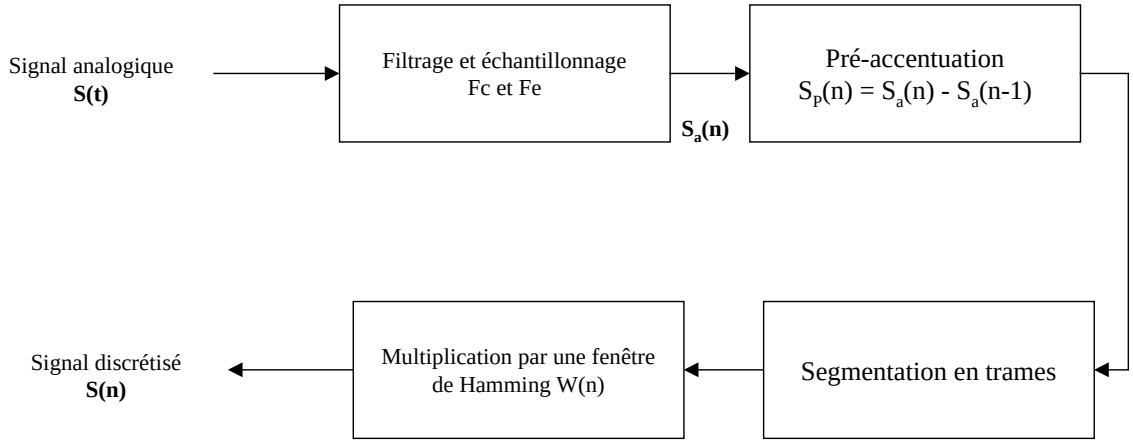


Fig. 2.2 – Mise en forme du signal de parole.

2.3.1 Analyse par transformée de Fourier

L'analyse fréquentielle de la parole se ramène aux opérations de la transformée de Fourier (TF) et n'a d'intérêt que si elle s'applique à une période stable du signal vocal, donc sur une période assez courte. Le spectre à court terme du signal s(n) se calcule à partir d'une fenêtre h(n) qui permet d'isoler une portion du passé récent de s(n) :

$$S(\omega,n)=\sum_{k=-\infty}^{k=+\infty} s(n)\,h(n-k)\exp(-j\omega n) \qquad \textbf{(2.1)}$$

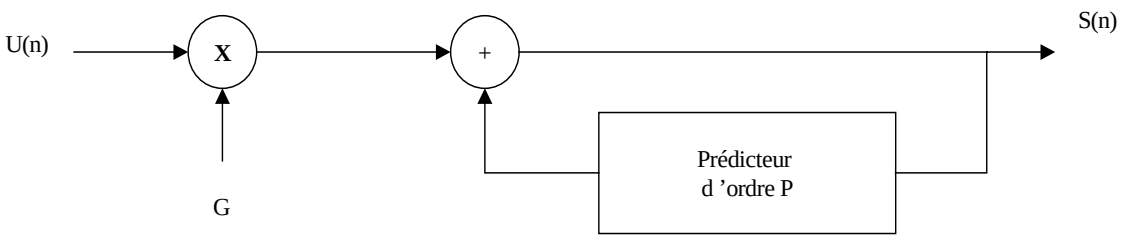
La quantité $|S(\omega,n)|^2$ est le spectre de puissance à court terme. L'implantation algorithmique efficace associée à la TF est la transformée de Fourier rapide (TFR). Elle présente de nombreux avantages en tant que méthode d'analyse fréquentielle. La rapidité de sa mise en œuvre l'a propulsé au rang d'élément incontournable des systèmes de traitement du signal. La TFR permet aussi une représentation fréquentielle du signal aussi fine que l'on souhaite. De plus, pour une étude qualitative de la parole, la TFR est très intéressante parce qu'elle permet une représentation par spectrogramme (évolution du spectre dans le temps) de qualité. Mais, après la naissance de la notion de représentation temps-fréquence, des études théoriques ont permis de mettre à jour quelques désavantages de la TFR qui sont impossibles à éliminer et qui constituent ainsi les limites de l'exploitation de la TF **[11].** Malheureusement les limites théoriques relatives aux représentations temps-fréquence ne sont pas les seuls problèmes de la TF. Le défaut majeur de la TF pour l'étude de la parole vient de l'inévitable intermodulation source/conduit présente dans le spectre qui ne permet pas de connaître précisément la hauteur du fondamental. Cette intermodulation est due à la convolution qui est réalisée par le conduit vocal sur la fréquence fondamentale produite par les cordes vocales. La déconvolution ne pouvant pas être réalisée par une simple transformée, il a donc fallu développer une technique particulière capable de la réaliser pour fournir ces deux informations utiles à l'analyse de la parole. L'étude des représentations temps-fréquence et les limites de la TF ont donc poussé à créer des méthodes de traitements de signal plus adaptées à la parole.

2.3.2 Codage prédictif linéaire

Le codage prédictif linéaire (LPC, Linear Predictif Coding) est une technique de codage et de représentation de la parole **[12]**. Elle s'appuie principalement sur l'idée que le système phonatoire peut être modéliser par un filtre linéaire. Ce filtre est excité par un train d'impulsions pour les sons voisés et aléatoires pour les sons non voisés. Il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents (équation **2.2**)

$$S(n)=-\sum_{k=}^{P} a_k\,S(n-k)+GU(n) \qquad \textbf{(2.2)}$$

U(n) étant l'entrée du filtre prédicteur suivant :



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

$$\hat{S}(n) = -\sum_{k=1}^{P} a_k \, S(n\_k) \qquad \textbf{(2.3)}$$

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

$$e(n) = S(n) - S(n) = S(n) + \sum_{k=1}^{P} a_k \, S(n-k) \qquad \textbf{(2.4)}$$

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

$$E_P = \sum e^2(n) = \sum_{n_1}^{n_2} \left(S(n) + \sum_{k=1}^{P} a_k \, S(n-k) \right)^2 \qquad \textbf{(2.5)}$$

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

$$\frac{\partial E_P}{\partial a_i} = 0 \qquad \text{i = 1, 2, ……, P} \qquad \textbf{(2.6)}$$

Le calcul de cette relation (équation **2.6**) conduit aux équations suivantes:

$$\sum_{k=1}^{P} a_k \, \Phi_{ik} = -\Phi_{i0} \qquad \text{, i = 1, 2, ……., P} \qquad \textbf{(2.7)}$$

$$\Phi_{ik} = \sum_{n=n_1}^{n_2} S(n-i) S(n-k) \qquad \textbf{(2.8)}$$

Ces équations normales (équation **2.7**), dites de Yule Walker, constituent un système linéaire de P équations à P inconnues. La résolution de ce système permettra d'obtenir les coefficients a_k du filtre. Parmi les méthodes de minimisation de l'énergie résiduelle de prédiction donc de résolution du système, on trouve principalement la méthode d'autocorrélation et la méthode de covariance.

- **Méthode d'autocorrélation :** L'énergie résiduelle de prédiction Ep, définie dans l'équation **2.5**, est minimisée sur une durée infinie :

$$E_P = \sum_{n=-\infty}^{+\infty} e^2(n) \qquad \textbf{(2.9)}$$

La fonction d'autocorrélation est définie par :

$$R(i) = \sum_{n=-\infty}^{+\infty} S(n) S(n-i) \qquad \textbf{(2.10)}$$

$$R(i) = R(-i)$$

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

longueur finie correspondant à N échantillons. La sommation infinie de l'équations **2.9** se ramène donc à une somme finie, soit :

$$E_P = \sum_{n=0}^{N-1+P} e^2(n) \qquad \textbf{(2.11)}$$

L'équation **2.7** devient alors :

$$\Phi_{ik} = \sum_{n=0}^{+\infty} S(n-i)S(n-k) = \sum_{n=-\infty}^{+\infty} S(n)S(n+i-k) \qquad \textbf{(2.12)}$$

Dans ce cas Φ_{ik} n'est autre que la fonction d'autocorrélation évaluée pour *(i-k)*, soit :

$$\Phi_{ik} = R(i-k) \qquad \textbf{(2.13)}$$

Le système donné par l'équation **2.7** s'écrira alors sous la forme suivante :

$$\sum_{k=1}^{P} a_k R(i-k) = -R(i)\,, \quad \text{i = 1, 2, ……., P} \qquad \textbf{(2.14)}$$

La méthode d'autocorrélation est largement utilisée parce que d'une part elle conduit à un système d'équations d'une structure particulière. La matrice des valeurs d'autocorrélation est une matrice Toeplitz. En effet, elle est symétrique et tous les éléments sur chaque diagonale sont identiques, ce qui facilite la résolution. D'autre part, elle assure la stabilité du modèle auto-régressif trouvé. Différents algorithmes permettent en effet la résolution du système (équation **2.14**) par une récursion sur l'ordre de prédiction. Parmi ces algorithmes nous pouvons citer l'algorithme de Levinson-Durbin et l'algorithme de Leroux-Gueguen **[12], [13], [14]**.

- **Méthode de covariance :** Dans cette méthode l'énergie de prédiction Ep est minimisée sur une durée finie :

$$E_P = \sum_{n=M_1}^{M_2} e^2(n) \qquad \textbf{(2.15)}$$

$$E_P = \sum_{n=0}^{N-1} e^2(n)$$

La relation (équation **2.8**) devient alors :

$$\Phi_{ik} = \sum_{n=0}^{N-1} S(n-i)S(n-k) \qquad \textbf{(2.16)}$$

Φ_{ik} définie dans l'équation II.15 n'est rien d'autre que la fonction de covariance $\sigma(i,k)$. Ainsi le système de l'équation **2.7** s'écrira alors :

$$\sum_{k=1}^{P} a_k \sigma(i,k) = -\sigma(i,0) \ , \qquad \text{i = 1, 2, ……., P} \qquad \textbf{(2.17)}$$

Dans ce cas la matrice (P x P) des valeurs de covariance est symétrique mais non de Toeplitz. Dans cette méthode, les propriétés statistiques de l'estimateur sont meilleures que dans le cas de la méthode d'autocorrélation, mais la matrice a une structure moins simple que celle de Toelplitz ce qui rend la résolution des équations de Yule-Walker un peu plus coûteuse. Un autre inconvénient de cette méthode est qu'elle ne garantit pas la stabilité du modèle auto-régressif trouvé. Parmi les algorithmes de résolution nous pouvons citer l'algorithme de Gram-Schmit et l'algorithme de Morf **[12], [14], [15].**

2.3.3 Analyses cepstrales

L'intermodulation source/conduit observée sur le spectre calculé par TFR rend son utilisation pour la mesure des formants Fi et de la fréquence fondamentale F_0 très difficile. Le lissage cepstrale est une méthode qui vise à séparer la contribution du conduit et de la source d'excitation par déconvolution. Pour cela on fait l'hypothèse que le signal vocal s(n) est produit par un signal excitateur g(n) (source glottique) traversant un système linéaire passif de réponse impulsionnelle b(n) (conduit). Avec ces hypothèses on peut écrire :

$$s(n) = b(n) \otimes b(n) \qquad \textbf{(2.18)}$$

Pour déconvoluer plus aisément s(n) il suffit de transposer le problème par homomorphisme dans un espace où l'opérateur $\otimes$ (convolution) correspond à un opérateur + (addition).

En pratique cette transposition par homomorphisme est réalisée par les étapes schématisées sur la **Fig. 2.3** suivante.

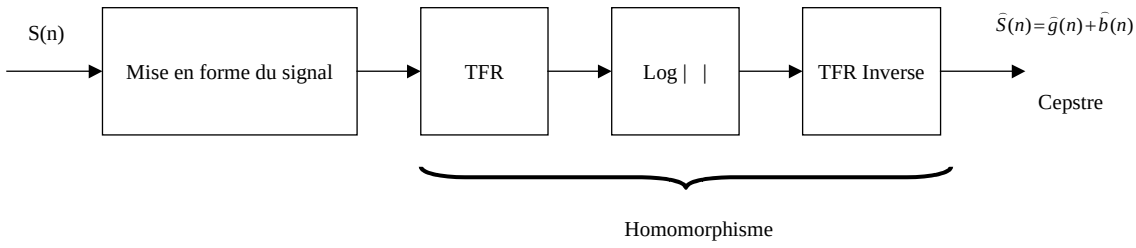


Fig. 2.3 - Analyse homomorphique de la parole.

Où $\hat{S}(n)$ sont les coefficients cepstraux approchés prenant leurs valeurs dans un domaine pseudo-temporel réel appelé domaine quéfrentiel. La structure de la parole et les hypothèses sur la source d'excitation et du conduit vocal permettent de dire que :

- $\hat{g}(n)$ se réduit théoriquement à une séquence d'impulsions de période n_0 (n_0 correspond à la fréquence fondamentale F_0).

- $\hat{b}(n)$ décroît rapidement (en 1/n) avec n et devient négligeable pour $n > n_0$.

Dans ces conditions, on peut admettre que la contribution du conduit est localisée dans les basses quéfrences ($n < n_0$) et que la séquence d'impulsions reflète la contribution de la source. Cette méthode de calcul des cepstres est élémentaire**[15],** il existe d'autres méthodes itératives effectuant un lissage, ce qui permet d'obtenir des cepstres de meilleure qualité.

2.3.3.1 Calcul des cesptres à partir des coefficients LPC

Le calcul des cespstres par la méthode décrite précédemment (analyse homomorphique) est généralement moins utilisé du fait de la charge de calcul importante associée au calcul de la TFR et de la TFR inverse. On lui préfère une méthode paramétrique. Cette méthode paramétrique permet de déterminer les coefficients cepstraux des signaux de parole à partir des coefficient LPC. C'est ainsi que le signal de parole est considéré comme engendré par un filtre auto-régressif (AR)[1] dont il faut déterminer les coefficients a_i en utilisant des méthodes classiques de prédiction linéaire comme la méthode d'autocorrélation par exemple. Le calcul des cepstres est alors basée sur une procédure récursive liant les coefficients cepstraux (C_m) et les coefficients de prédiction (a_i). Cette procédure récursive est traduite par les équations suivantes :

$$\ln\left[\frac{1}{A_p(z)}\right] = \sum_{n=1}^{\infty} C_q(n)\, z^{-n} \qquad \textbf{(2.19)}$$

$C_0 = \ln\sigma^2$, σ^2 est le gain du modèle LPC **(2.20)**

$$C_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) C_k\, a_{m-k} \quad , \text{ pour } 1 \le m \le P \qquad \textbf{(2.21)}$$

$$C_m = \sum_{k=1}^{m-k}\left(\frac{k}{m}\right) C_k\, a_{m-k} \quad , \qquad \text{pour } m > P \qquad \textbf{(2.22)}$$

2.3.3.2 MFCC (Mel-scale Frequency Cepstral Coefficients)

Les coefficients MFCC sont une extension des coefficients cepstraux par le passage de l'échelle fréquentielle linéaire à une échelle fréquentielle non linéaire proche de l'audition humaine. Cette échelle non linéaire est l'échelle perceptive Mel. Celle-ci est plus exactement linéaire pour les basses fréquences (inférieures à 1000Hz) et logarithmique pour les hautes fréquences. C'est ainsi que des filtres répartis linéairement en basses fréquences et logarithmiquement en hautes fréquences (**Fig. 2.4**) sont utilisés afin de capturer les caractéristiques phonétiques importantes du signal de parole. Ces filtres possèdent la caractéristique suivante : plus la fréquence est élevée, plus la bande passante est large ce qui permet une meilleure résolution temporelle des hautes fréquences.

[1] Il existe un autre modèle de production, le modèle ARMA(Auto Regressive Moving Average). Seulement, l'estimation de ces modèles étant plus délicate que celle des modèles AR, on préfère utiliser les modèles auto-régressifs dans la plupart des applications.

L'échelle Mel peut être définie par la relation suivante entre la fréquence en Hertz et sa correspondante en mels :

$$Mel(f) = x \log_{10}\left(1 + \frac{f_{Hertz}}{y}\right) \qquad \textbf{(2.23)}$$

Plusieurs valeurs sont utilisées pour x et y par exemple x=1000/log(2) et y=1000. De nos jours, les valeurs les plus couramment utilisées sont x=2595 et y=700.

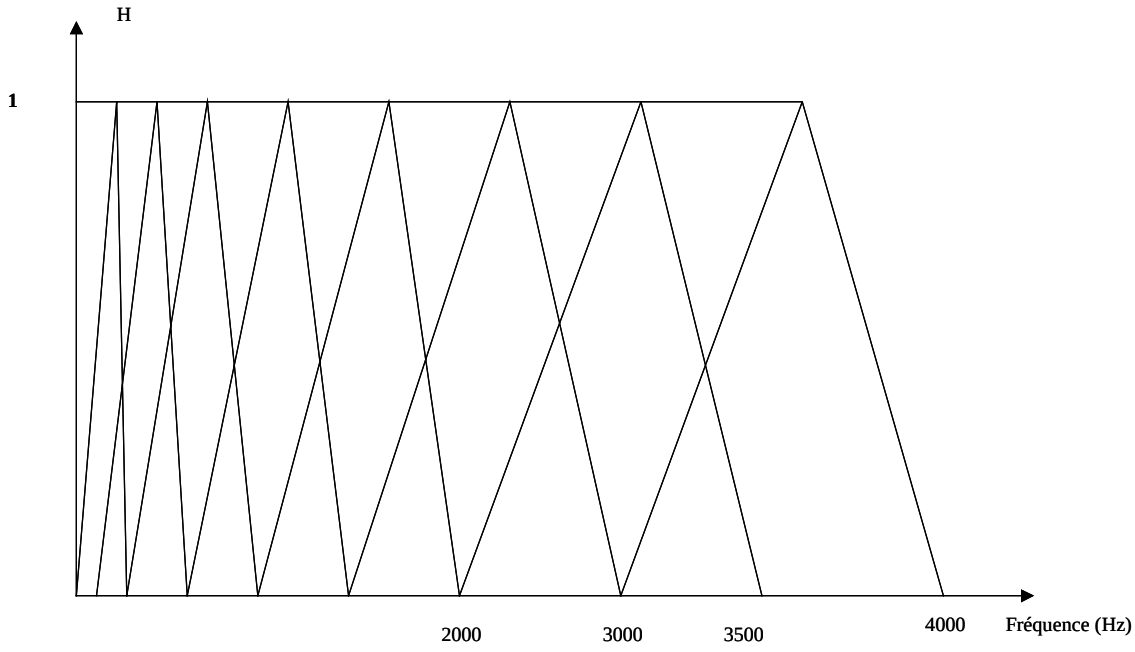


Fig. 2.4 – Bancs de filtres Mel.

Les coefficients MFCC sont très utilisés en RAP du fait des bons résultats qu'ils ont permis d'obtenir. La majorité des systèmes de reconnaissance actuels utilisent ces coefficients pour représenter le signal de parole. D'ailleurs nous avons nous aussi utilisé ces coefficients dans notre travail. De ce fait le calcul détaillé de ces coefficients sera donné dans le chapitre III. Pour l'heure nous nous contentons de la description succincte suivante de la procédure de calcul.

- Calcul du spectre de puissance P(f) par transformée de Fourier rapide (TFR)
- L'échelle fréquentielle linéaire du spectre P(f) est transformée en une l'échelle non linéaire Mel aboutissant au spectre P(Mel). Ce spectre est passé à travers le banc de filtres triangulaires, les sorties des filtres sont alors récupérées.
- Les valeurs logarithmiques des sorties de chaque filtre sont transformées dans le domaine temporel par Transformée de Fourier Inverse (TFR^{-1}) donnant ainsi les coefficients MFCC.

2.3.4 Analyse PLP (Perceptual Linear Prediction)

La méthode PLP est une méthode inspirée du principe de prédiction linéaire. Elle combine ce principe à une représentation du signal qui suit l'échelle humaine d'audition **[16]**, **[17]**. L'estimation par prédiction linéaire (LP) des coefficients du modèle auto-régressif (tout pôle) est très utilisée en reconnaissance de la parole. La méthode LP identifie uniformément le spectre sur toutes les fréquences de la bande d'analyse. Or cette propriété est loin d'être vérifiée pour l'oreille humaine, car il a été établi que celle-ci est plus sensible aux fréquences situées au milieu de la bande d'analyse du spectre. Ainsi, il est possible que certains détails importants du spectre ne soient pas pris en compte lors de l'analyse LP. L'analyse PLP permet de résoudre ce problème. Son but est d'estimer les paramètres d'un filtre auto-régressif tout pôle, modélisant au mieux le spectre auditif. Le processus de calcul des coefficients PLP peut être décrit par la **Fig. 2.5** suivante.

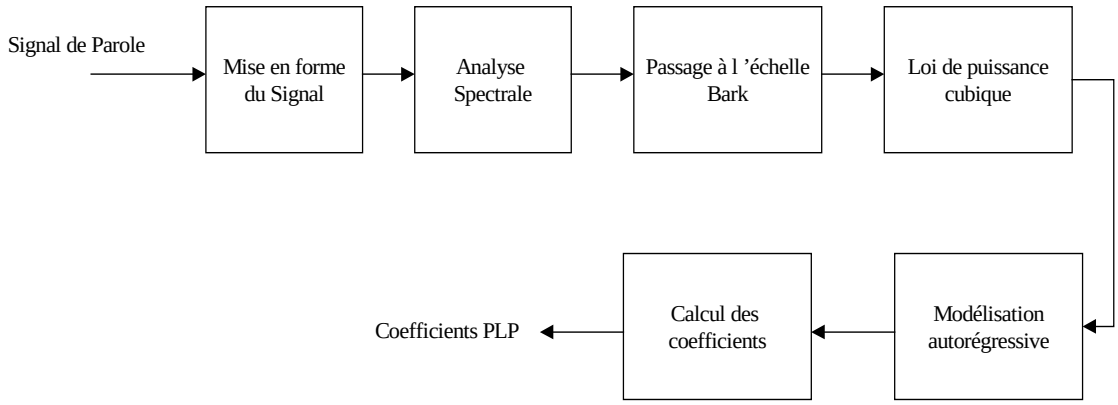


Fig. 2.5 – Processus de calcul des coefficients PLP

Après une mise en forme du signal de parole, le spectre de puissance P(ω) est calculé. Ensuite, un passage de l'échelle de fréquence usuelle à l'échelle de Bark **(**équation **2.24)** est effectué.

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi}\right)^2 + 1\right)^{0.5}\right) \qquad \textbf{(2.24)}$$

ω représentant la fréquence angulaire exprimée en rd/s et Ω la fréquence de Bark.

Ce passage à l'échelle Bark, permet d'approximer de manière grossière ce que nous savons de la forme des filtres auditifs. Celle-ci est approximativement constante le long de l'échelle de Bark. Le spectre de puissance dans l'échelle de Bark est convolué avec le spectre de puissance de la courbe de bande critique en utilisant l'équation **2.25** suivante.

$$\psi(\Omega)=\begin{cases}0 & \Omega<-1.3\\ 10^{2.5(\Omega+0.5)} & -1.3\le\Omega\le-0.5\\ 1 \quad pour & -0.5\le\Omega\le2.5\\ 10^{-1.0(\Omega-0.5)} & 0.5\le\Omega\le2.5\\ 0 & \Omega>2.5\end{cases} \qquad \textbf{(2.25)}$$

Cette courbe de masquage est une approximation de la courbe de masquage asymétrique de Schroeder **[18]**.

On essaye ensuite d'approximer la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert *E(ω).* Le spectre de puissance est multiplié par cette fonction de transfert.

$$E(\Omega)=E(\omega).\Theta(\Omega) \qquad \textbf{(2.26)}$$

$$\Theta(\Omega_t)=\sum_{\Omega=-1.3}^{\Omega=2.3}P(\Omega-\Omega_t).\Psi(\Omega) \qquad \textbf{(2.27)}$$

La non-linéarité entre l'intensité d'un son et sa force de perception par l'oreille est ensuite approximée par une loi de puissance :

$$\Phi(\Omega)=E(\Omega)^{0.33} \qquad \textbf{(2.28)}$$

L'étape finale consiste en une modélisation auto-régressive classique du spectre du modèle auditif tout pôle, en calculant les coefficients auto-régressifs du filtre. L'analyse PLP est très similaire à l'analyse MFCC. La différence est que l'analyse PLP utilise l'échelle Bark au lieu de l'échelle Mel et un modèle auto-régressif tout pole au lieu de la transformée en cosinus discrète (DCT) pour le calcul des coefficients.

Cette méthode PLP a été par la suite améliorée pour résister à certaines conditions de bruit. C'est ainsi que l'analyse RASTA-PLP a été développée, RASTA étant l'acronyme de RelAtive SpecTrAl. La méthode PLP, dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsion spectrales linéaires, Hermansky **[19],** propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP. La mise en œuvre de ce filtrage (RASTA) permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication.

2.4 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole a pour but d'extraire l'information linguistique contenue dans le signal vocal elle peut donc être interprétée comme une tâche de transformation du signal acoustique en texte. Deux approches principales complètement différentes ont été utilisées ces quinze dernières années pour réaliser cette tache, une approche analytique fondée sur les connaissances utilisant les techniques de l'intelligence artificielle et une approche globale basée sur les techniques de la reconnaissance des formes. Dans ce qui suit, nous allons détailler ces deux approches en précisant les différentes étapes qui interviennent dans leurs processus respectifs.

2.4.1 Approche analytique fondée sur les connaissances

L'approche analytique tente de détecter et d'identifier les unités linguistiques discrètes du message vocal. Le structure générale adoptée par les systèmes de reconnaissance analytique de la parole continue peut être schématisée par la **Fig. 2.6** suivante :

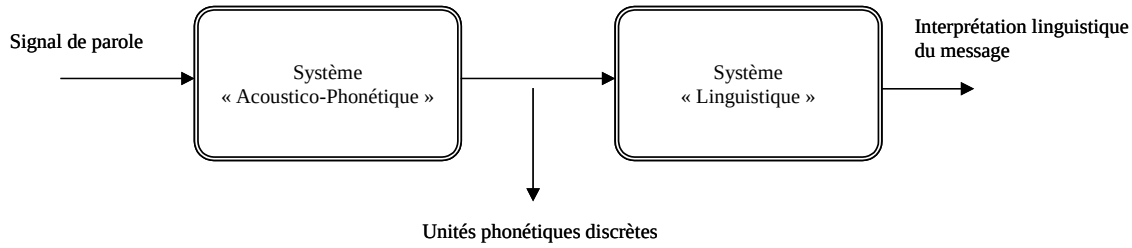


Fig. 2.6 - Structure générale d'un système de reconnaissance analytique.

Dans cette structure, on distingue deux systèmes :

- Le système acoustico-phonétique qui, à partir du signal acoustique, génère des unités phonétiques discrètes. Les unités les plus utilisées sont les phonèmes. Un phonème est un élément sonore d'un langage donné, déterminé par les rapports qu'il entretient avec les autres sons de ce langage. Il en existe une trentaine exactement 34 en langue arabe. Cette notion de phonème est assez importante en reconnaissance vocale.
- Le système linguistique qui utilise en entrée la suite d'unités phonétiques discrètes pour donner en sortie l'interprétation linguistique du message vocal. Ce décodage linguistique est réalisé grâce à diverses informations morphologiques, grammaticales, lexicales, etc.

Dans cette structure, l'étape qui conditionne en grande partie la performance du système de reconnaissance analytique fondée sur les connaissances est l'étape de décodage acoustico-phonétique. Cette étape est menée généralement par un système "intelligent" qui raisonne à partir d'un ensemble de connaissances acoustiques, phonétiques et phonologiques. Les systèmes experts, outils de l'intelligence artificielle (I.A), permettent en effet d'intégrer ces différentes sources de connaissances. Ils tentent de reproduire la compétence d'experts humains dans le domaine. L'expertise humaine en compréhension de la parole est évidente, elle semble tellement instinctive que l'introspection ne fournit guère d'informations. Nous ne savons pas vraiment sur quels éléments acoustiques porte le processus de reconnaissance. Par contre des experts peuvent transcrire phonétiquement des sonagrammes avec des taux de reconnaissance appréciables, il

semble donc naturel de modéliser et d'utiliser les connaissances et stratégies de phonéticiens experts en lecture de sonagrammes. Ces systèmes ont constitué dans un passé récent une solution aux problèmes posés par le décodage acoustico-phonétique. Dans un tel système le signal de parole est d'abord segmenté en classes phonétiques discrètes, ces classes sont par la suite étiquetées en unités phonétiques discrètes (par exemple les phonèmes) en utilisant des connaissances traduites par des règles de production gérées par un système expert.

Un système expert comprend généralement trois parties bien distinctes :

- La base de faits : constituée par des informations relatives à chaque segment vocal.
- La base de connaissances : formée par un ensemble de règles qui modélisent l'ensemble des connaissances nécessaires à l'expert.
- La structure de contrôle : choisit les règles à appliquer pour déterminer la nature exacte du segment. Elle utilise le contenu de la base de faits relatif à ce segment.

Grâce à sa modularité, un tel système est relativement facile à modifier et à étendre, notamment en rajoutant des règles. On peut expérimenter diverses stratégies de contrôle sans avoir à changer tout le reste du système. Parmi les différents systèmes experts en DAP réalisés, on citera **[20], [21], [22]**.

2.4.2 Approche globale basée sur la modélisation des formes

Dans l'approche globale basée sur la modélisation des formes, on tente de modéliser au mieux des unités représentatives du signal de parole. Il existe principalement deux types de modélisation des propriétés d'un signal donné : la modélisation déterministe et la modélisation statistique. Nous allons détailler dans ce qui suit les principales techniques de reconnaissance des formes basées sur ces deux types de modélisation : le DTW pour la modélisation déterministe et les modèles de Markov cachés pour la modélisation statistique. Nous parlerons ensuite de l'utilisation des réseaux de neurones artificiels (RNA) dans la reconnaissance en particulier leur association aux modèles de Markov cachés qui a donné naissance à l'approche hybride de reconnaissance.

2.4.2.1 Dynamic Time Warping (DTW)

L'alignement temporel, plus connu sous l'acronyme de DTW, Dynamic Time Warping, est une méthode fondée sur un principe de comparaison du signal à reconnaître avec des formes de références. Le DTW est une des premières méthodes de reconnaissance des formes à être appliquée pour la reconnaissance automatique de la parole. Son principe est basé sur l'enregistrement des formes de références pendant une phase d'apprentissage ensuite, lors de la phase de reconnaissance on compare le signal prononcé à ces formes de références. Le signal prononcé est identifié en fonction de sa proximité avec une des formes de références stockées.

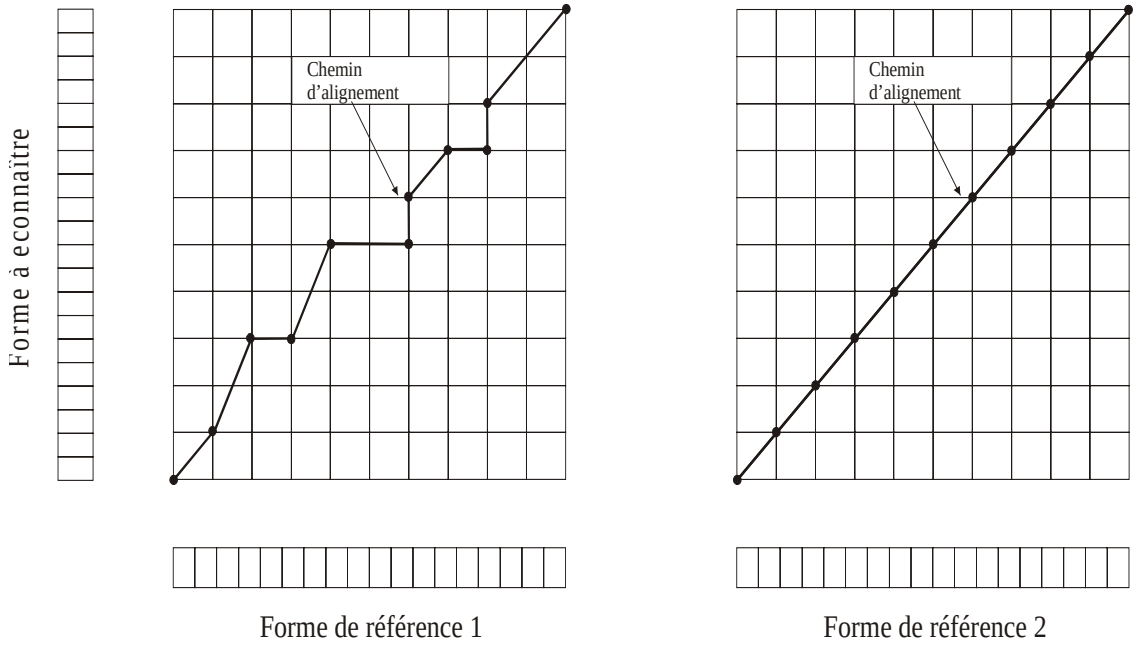


Fig. 2.7 – Alignement temporel entre une forme à reconnaître et des formes de références

Comme le montre le schéma de la **Fig. 2.7**, la forme choisie sera celle pour laquelle le chemin de mise en correspondance (chemin d'alignement) est le plus court, cette taille minimale marquant le peu de différences entre la forme à reconnaître et la forme de référence.
Un point important de l'alignement temporel est la définition de la fonction de recalage qui permet de calculer, selon certaines contraintes, la distance entre la forme à reconnaître et la forme de référence. La forme à reconnaître est mise en correspondance dans le plan temporel par l'algorithme d'alignement qui essaie de trouver le plus cours chemin dans le graphe ainsi constitué. Cette fonction de mise en correspondance définit une valeur pour chaque arc de graphe, ces valeurs favorisant l'axe médian qui correspond à une parfaite mise en relation de la forme à reconnaître comme le montre la **Fig. 2.7.**
Le calcul de la fonction d qui représente la distance entre deux points successifs du graphe n'est cependant pas suffisant pour calculer la longueur totale du chemin parcouru dans le graphe. Une fonction supplémentaire, G, calcule une longueur totale qui permettra, après le calcul de cette longueur des chemins de toutes les formes de base de références, de savoir à quel mot du vocabulaire préenregistré correspond la forme à classer. Le calcul de cette fonction G répond au même principe que le principe général énoncé par Bellman pour la programmation dynamique : toute sous-partie du chemin optimal est lui-même un chemin optimal. Des exemples de fonctions d et G de calcul de distance pourront être trouvé dans **[23]** ou **[24].**
Cette méthode de reconnaissance des formes est, initialement, bien adaptée à la reconnaissance de mots isolés mais des extensions ont été développées pour permettre son application à la reconnaissance de mots connectés. Nous pouvons citer dans ce sens les algorithmes de programmation dynamique suivants : le One-Pass (OP) présenté dans **[25],** le Two-Levels (TL) **[26]** et le Level-Building (LB) **[27], [28]**.
La technique de reconnaissance de mots connectés à deux niveaux (Two Levels) fut la première application pratique du DTW aux mots connectés. Son principe est de trouver pour toutes les portions du message les formes de référence optimales puis de trouver la meilleure manière de les connecter. Deux processus d'alignement sont nécessaires, un au niveau des unités de bases, l'autre au niveau du message.

L'algorithme LB (Level Building) fut développé comme une amélioration du TL ; il conduit surtout à un coût de calcul moins élevé. Le LB consiste à rechercher d'abord le meilleur mot pour le début de la phrase sur toutes les trames. Il détermine ensuite les meilleurs seconds mots à partir des meilleurs premiers mots. Ce processus est répété jusqu'aux meilleurs $L^{ièmes}$ mots. L doit être fixé au préalable, ce qui est une contrainte puisqu'on ne connaît pas le nombre de mots de la phrase. Toutefois, la solution optimale au rang L restera optimale pour les rangs supérieurs. Il suffit donc de choisir L suffisamment grand ce qui, en contrepartie, fait perdre beaucoup d'intérêt à la méthode. Les décodages TL et LB nécessitent des aller-retour sur la suite des vecteurs acoustiques à décoder ; ils ne peuvent donc être implantés en temps réel. La complexité du TL $O(T^2)$, avec T la longueur en trames du message à décoder, est élevée. Celle du LB est plus raisonnable O(LT), L étant le nombre maximal d'unité présente dans le message. Mais dans les deux cas, la modélisation des niveaux interdit des structures de recherche en boucle.
Le One-pass n'a aucun des défauts précédents : le décodage est obtenu après un simple parcours du message, sa complexité est en O(T) et des structures de recherche en boucle peuvent être envisagées. Le OP traite toutes les références en parallèle et progresse de manière synchrone trame par trame. Son principe est de réintroduire les sous-chemins optimaux atteignant la fin d'un modèle vers le début de tous les autres modèles. Cet algorithme est à la base des moteurs de reconnaissance de la parole actuels.
Les principaux problèmes inhérents à ce type de reconnaissance sont la taille de la base des formes de référence et la fonction de calcul des distances. Des méthodes complémentaires ont été développées pour tenter de réduire la taille de la base des formes de références par sélection optimale des formes à conserver **[29].** Ces méthodes reposent surtout sur une exploration statistique de la base des formes de références et permettent d'obtenir une caractérisation des différents ensembles la constituant, ces ensembles correspondant aux différents symboles référencés dans la base. Une des techniques qu'il est possible d'employer pour ce faire est, par exemple, la méthode des plus proches voisins (k-ppv).

2.4.2.2 Modélisation statistique

2.4.2.2.1 De l'intérêt des modèles statistiques

Depuis maintenant plusieurs années, la reconnaissance automatique, se fait principalement à l'aide d'une modélisation statistique de segments de parole. Dans la chaîne de traitement d'un système de reconnaissance de la parole, la première étape consiste à extraire des caractéristiques pertinentes du signal. Cette phase de paramétrisation du signal permet ainsi d'obtenir une représentation plus compacte de l'information contenue dans le signal. Pour s'affranchir des problèmes de non-stationnarités du signal, on a souvent recours à une analyse spectrale ou cepstrale à l'aide d'une fenêtre glissante à intervalles de temps réguliers, typiquement toutes les 10ms. On représente la portion de signal considérée par un vecteur acoustique. Les systèmes de reconnaissance n'utilisent donc pas le signal lui-même mais plutôt la suite de vecteurs acoustiques, considérée comme la réalisation d'un processus aléatoire X .
Les avantages d'une telle conceptualisation sont multiples. Le principal avantage réside dans le fait de pouvoir exprimer intrinsèquement la variabilité de la parole par un modèle mathématique du processus aléatoire, ce qui n'est pas possible avec une approche de type appariement des formes où, pour modéliser la variabilité, on a recours à des formes de références multiples ou à des distances stochastiques comme la distance de Mahalanobis. Un modèle stochastique sera donc capable de rendre compte des variabilités intra et inter locuteurs. De plus, en supposant que l'on dispose d'un modèle du processus aléatoire, il est donc possible de calculer la vraisemblance d'une observation (i.e. d'une réalisation du processus aléatoire) pour un modèle donné, ce qui permet alors d'exprimer de manière simple les problèmes de reconnaissance de la parole. Un autre avantage de ce type d'approche réside dans la possibilité d'avoir un modèle paramétrique du

processus aléatoire. Une forme peut donc être définie par un nombre restreint de paramètres plutôt que par une ou plusieurs formes de référence. L'ensemble de ces avantages justifie la prépondérance de l'approche statistique dans la mise en œuvre des systèmes de reconnaissance de la parole actuels.

2.4.2.2.2 Approche statistique en reconnaissance de la parole

En reconnaissance statistique de la parole, on cherche à trouver le message prononcé connaissant l'observation, comme illustré par la **Fig. 2.8**. Un locuteur prononce une séquence de mots w^* qui donne lieu à une réalisation acoustique $X = x$. Le décodeur cherche la séquence de mots $\hat{w}$ qui approxime le mieux, selon un critère donné, w^* connaissant l'observation.

Fig. 2.8 – Schéma générique d'un système de reconnaissance basé sur l'approche statistique.

Dans le cadre de l'approche statistique, le critère le plus approprié est naturellement celui du Maximum A Posteriori (MAP). On cherche donc la séquence $\hat{w}$ pour laquelle la probabilité a posteriori de la séquence w connaissant l'observation $X = x$ est maximale, ce qui se traduit mathématiquement par :

$$\hat{w} = \arg\max_{w} P(W = w \mid X = x) \qquad \textbf{(2.29)}$$

$$= \arg\max_{w} P(X = x \mid W = w)\ P(W = w)$$

La deuxième formulation faisant apparaître deux termes distincts, le premier, étant appelé score acoustique tandis que le second correspond au score linguistique. En effet, le premier terme représente la probabilité de l'observation (i.e. de la suite de vecteurs acoustiques) pour une séquence de mots donnée tandis que le deuxième terme est la probabilité a priori de la séquence de mots. Les modèles statistiques de segments de parole rendent possible le calcul du score acoustique comme nous le verrons par la suite dans le cadre de la modélisation par modèles de Markov cachés. Le score linguistique est lui donné par le modèle de langage.

2.4.2.2.3 Les Modèles de Markov cachés

Les modèles de Markov cachés, appelés aussi HMM, l'abréviation anglaise de Hidden Markov Model, sont les modèles les plus utilisés pour la reconnaissance de la parole. Ils ont été appliqués

avec succès à la reconnaissance des phonèmes **[30],** à la reconnaissance des mots isolés **[31],** à la reconnaissance de mots enchaînés **[32]** et à la reconnaissance de la parole continue large vocabulaire **[33]**. L'utilisation des MMC pour la reconnaissance de la parole trouve ses fondements dans l'existence d'un formalisme statistique adéquat conduisant au développement d'algorithmes puissants permettant leurs implémentations et dans la structure même d'un MMC qui n'est en fait qu'un modèle approprié simulant le conduit vocal **[34].** Dans ce qui suit seront donnés une brève présentation des modèles de Markov cachés ainsi que les fondements mathématiques à l'origine de leur utilisation pour la reconnaissance automatique de la parole. Plus de détails sur l'approche markovienne de reconnaissance de la parole seront donnés dans le chapitre 4.

2.4.2.2.3.1 Formalisme

Les modèles MMC tentent de modéliser les unités représentatives de la parole par des modèles statistiques. Ils supposent que la suite de vecteurs acoustiques $X = x_1, x_2, ...x_T$ représentatifs du signal de parole est stationnaire par morceaux, ce qui signifie que, par morceau, les vecteurs acoustiques suivent la même loi de probabilité. On associe donc au processus X un processus caché Y où Y_t est une indicatrice de la loi correspondant à X_t. Pour modéliser l'évolution temporelle de la parole, la loi du processus Y est donnée par une chaîne de Markov homogène, généralement d'ordre 1. On représente habituellement le processus Y sous forme d'un automate stochastique comme illustré par la **Fig. 2.9,** une densité de probabilité étant associée à chacun des états de l'automate. L'automate étant capable, après apprentissage, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle Formellement, un modèle de Markov caché est défini par le nombre d'états N de l'automate et de l'ensemble des paramètres λ suivants :

$$\lambda = \left\lfloor A=(a_{ij}\ ;\ i, j=1,....., N),\ B=b_j(.),\ \Pi=(\pi_i\ ,\ i=1,....., N) \right\rfloor$$

où π_i est la distribution des probabilités initiales des états, a_{ij} la probabilité de transiter de l'état i à l'état j, soit $a_{ij} = P(q_t = j \,|\, q_{t-1} = i)$ et $b_j(.)$ la fonction densité de probabilité associée à l'état j. Les fonctions densités de probabilités associées aux états déterminent le type de modèle c'est ainsi qu'on parle de MMC discrets, continus, semi-continus, etc.

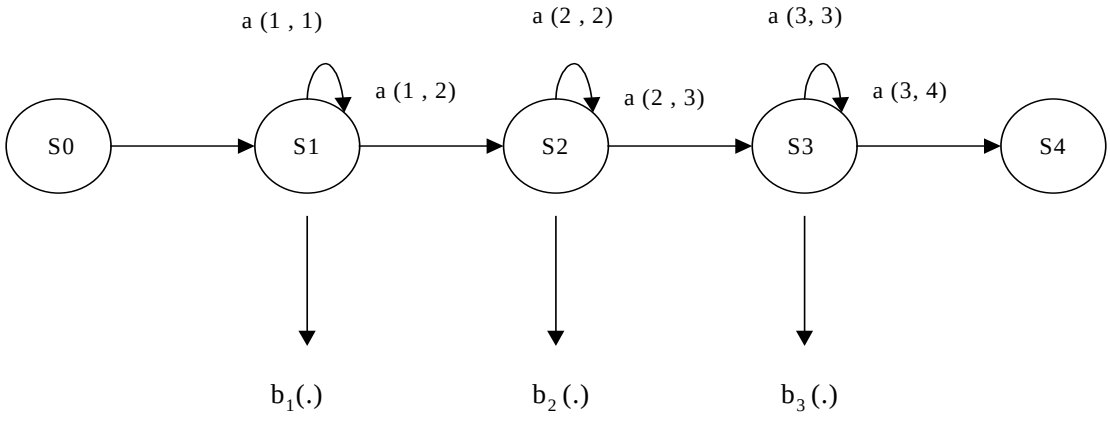


Fig. 2.9 – Un exemple de modèle de Markov caché.

Nous avons vu précédemment que l'approche statistique de la reconnaissance de la parole reposait sur le calcul de l'équation **2.29** en particulier la vraisemblance $P(X = x \mid W = w)$ où w est une suite de modèles d'unités acoustiques. Dans le cas des modèles de Markov, le calcul de la vraisemblance est donné par

$$P(X = x \mid W = w) = \sum_{Q} P(X = x \mid W = w, Y = y) P(Y = y \mid W = w) \qquad \textbf{(2.30)}$$

$$P(X = x \mid W = w) = \sum_{Q} \pi_{y1} b_{y1}(x_1) \prod_{t=2}^{T} a(y_{t-1}, y_t) b_{y1}(x_t) \qquad \textbf{(2.31)}$$

Où *Q* représente toutes les séquences d'états de longueur T possibles. Deux algorithmes permettent de calculer cette probabilité selon que l'on considère toutes les séquences d'états possibles autorisées (algorithme Forward Backward) ou que l'on considère la séquence d'états la plus probable (algorithme de Viterbi). Nous allons nous limiter ici à une description succincte du décodage par l'algorithme de Viterbi.

2.4.2.2.3.2 Décodage par l'algorithme de Viterbi

Le calcul du score acoustique est donné par l'équation **2.30**. L'algorithme de Viterbi **[35]**, **[36]** permet d'approximer cette équation par

$$P(X = x \mid W = w) \propto \max_{Q} P(X = x \mid W = w, Y = y) P(Y = y \mid W = w) \qquad \textbf{(2.32)}$$

La segmentation optimale étant donnée par le terme pour lequel le maximum est atteint. Le calcul se fait de manière itérative sur le logarithme de la vraisemblance. En effet, le produit sur le temps de l'équation **2.31** devient alors une somme et on peut maximiser l'ensemble par un ensemble de maximisations locales. Si on définit $\Phi_i(t)$ comme étant le log-vraisemblance le long du meilleur chemin finissant dans l'état i à l'instant t, on montre alors aisément que :

$$\Phi_j(t+1) = \log b_j(x_{t+1}) + \max_{i} (\Phi_i(t) + \log a_{ij}) \qquad \textbf{(2.33)}$$

où la maximisation se fait sur l'ensemble des états prédécesseurs possibles pour l'état j. Cette récursion permet de calculer l'approximation de l'équation **2.32**, le score final étant donné par $\max_i (\Phi i(T) + \log a_{i,F})$, $a_{i,F}$ étant la probabilité de sortir du modèle à partir de l'état i.
La théorie des modèles de Markov cachés en reconnaissance de la parole sera reprise plus en détail dans le chapitre 4 étant donné qu'elle est l'approche principale utilisée tout au long de ce travail.

2.4.2.3 Approche connexionniste

Après avoir décrit dans la section précédente la modélisation probabiliste basée sur les modèles de Markov cachés nous introduisons maintenant une nouvelle forme de modélisation celle, basée sur les réseaux de neurones artificiels (RNA). Les réseaux de neurones artificiels présentent en effet

plusieurs aspects du connexionnisme, qui voulait s'inspirer du cerveau humain dans ses qualités de distribution et du parallélisme du traitement de l'information, et ses capacités d'apprentissage. Certains réseaux sont le résultat d'ailleurs de la recherche de structure et d'interprétation des codages internes de l'information. Les RNA peuvent être utilisés comme des fonctions discriminantes non paramétriques dans une tache de classification, comme des modèles de régression ou comme des approximateurs universels. Dans le domaine de la reconnaissance automatique de la parole, les réseaux de neurones sont utilisés comme reconnaisseurs de formes pour des vocabulaires limités. En reconnaissance de la parole continue ils sont plutôt combinés avec d'autres types de modélisation en particulier la modélisation MMC. Cette combinaison aboutit à la conception de systèmes de reconnaissance dits hybrides. Il existe, bien entendu, plusieurs types de réseaux de neurones utilisés dans le traitement de la parole **[37], [38]**. Nous allons présenter dans ce qui suit les réseaux les plus utiles pour la reconnaissance automatique de la parole.

2.4.2.3.1 Perceptrons multi-couches

Les perceptrons multi-couches (MLP Multi Layer Perceptron, pour l 'anglais) sont les réseaux les plus courants et les plus utilisés. La **Fig. 2.10** représente un réseau de neurones perceptron à trois couches.

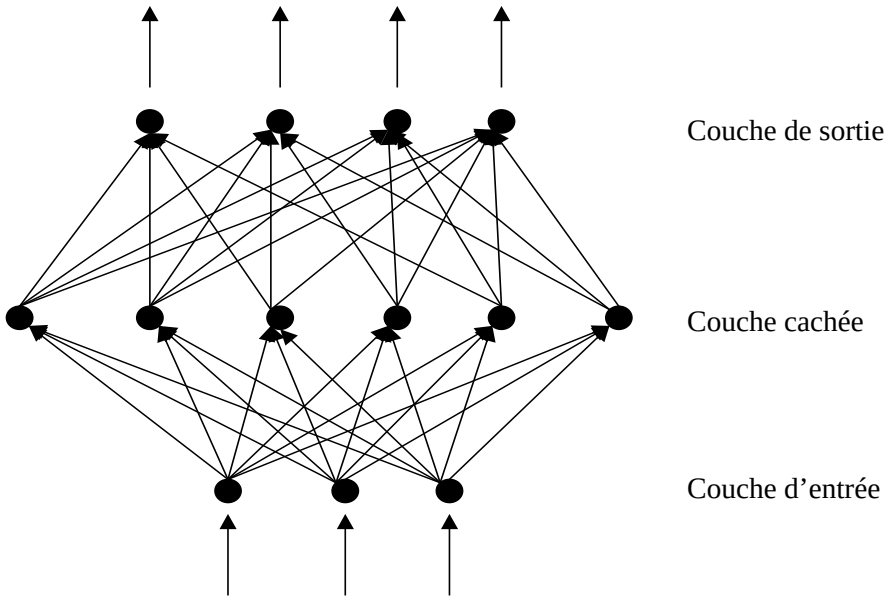


Fig. 2.10 - Perceptron multi couches avec une couche cachée

Les perceptrons multi-couches sont des réseaux à propagation directe (sans cycle), ils sont généralement constitués d'unités élémentaires interconnectées par des liens avec des poids variables. Le réseau de la **Fig. 2.10** est constitué de trois couches : une couche d'entrée, une couche cachée et une couche de sortie. Chaque couche est constituée d'unités élémentaires appelées neurones. Le neurone calcule une somme pondérée des composantes du vecteur d'entrée et ajoute un biais à cette somme. Ensuite une fonction non linéaire est appliquée au résultat. L'architecture du neurone peut donc être résumée simplement sur la **Fig. 2.11**. Plusieurs types de fonctions non-linéaires peuvent être définis :

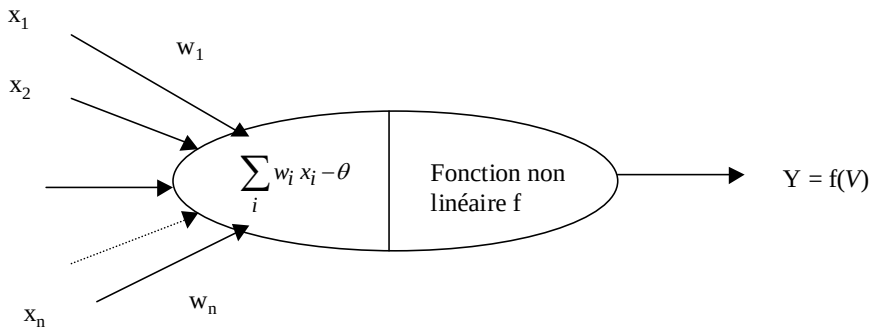


Fig. 2.11 – Neurone artificiel.

- fonction softmax :

$$f(V_j) = \frac{\exp(V_j)}{\sum_{i=1}^{n} \exp(V_i)} \qquad \textbf{(2.34)}$$

- fonction signe :

$$f(V) = \begin{cases} +1 & si\ V \rangle 0 \\ -1 & si\ V \langle 0 \end{cases} \qquad \textbf{(2.35)}$$

- fonction sigmoïde :

$$f(V) = \frac{1}{1 + \exp(-V)} \qquad \textbf{(2.36)}$$

La fonction la plus utilisée est la fonction sigmoïde. Les différents termes de la somme pondérée des nœuds du réseau constituent les paramètres (ou poids) de celui-ci. Ils sont estimés lors d'un processus d'entraînement dit supervisé. Un entraînement est dit supervisé lorsque les entrées sont présentées aux réseaux les unes après les autres avec la sortie imposée correspondant à chaque entrée. Le problème de l'entraînement d'un réseau revient à trouver un ensemble de paramètres (poids) W qui minimise une fonction d'erreur E. Il existe plusieurs critères d'erreur qui peuvent être utilisés. Notons par T l'ensemble des données d'entraînement, les sorties réelles y et les sorties désirées ε. Les critères les plus connus sont :

- le critère des moindres carrés :

$$E = \frac{1}{2} \sum_{t \in T} \sum_{l=1}^{L} \left(y_l^{(t)} - \varepsilon_l^{(t)} \right)^2 \qquad \textbf{(2.37)}$$

- Le critère entropique

$$E = \frac{1}{2} \sum_{t \in T} \sum_{l=1}^{L} \varepsilon_l^{(t)} \ln(y_l^{(t)}) \qquad \textbf{(2.38)}$$

Les techniques du type descente du gradient sont les plus adaptées pour estimer ces paramètres (les poids du réseau w). Elles sont basées sur le calcul du gradient de l'erreur par rapport aux paramètres du réseau et à l'ajustement de ces paramètres jusqu'à ce qu'un minimum de l'erreur soit atteint.
La méthode de rétro-propagation (back-propagation, pour l'anglais) **[38], [39]**, minimisant un des critères d'erreur (erreur quadratique ou entropie) entre les sorties réelles et désirées pour chacun des vecteurs d'entrées, est utilisée pour modifier la valeur des poids. La fonction est une fonction non linéaire des poids avec plusieurs minima possibles. Le gradient de l'erreur est employé pour ajuster les poids de l'élément de sortie et est propagé pour modifier les poids des couches cachées (d'où le nom de rétro-propagation de la couche de sortie vers la couche cachée).
L'ajustement des paramètres se fait généralement selon la formule :

$$W^{T+1} = W^T - \eta \frac{\delta E}{\delta w_{ij}}\bigg|_{W^T} \qquad \textbf{(2.39)}$$

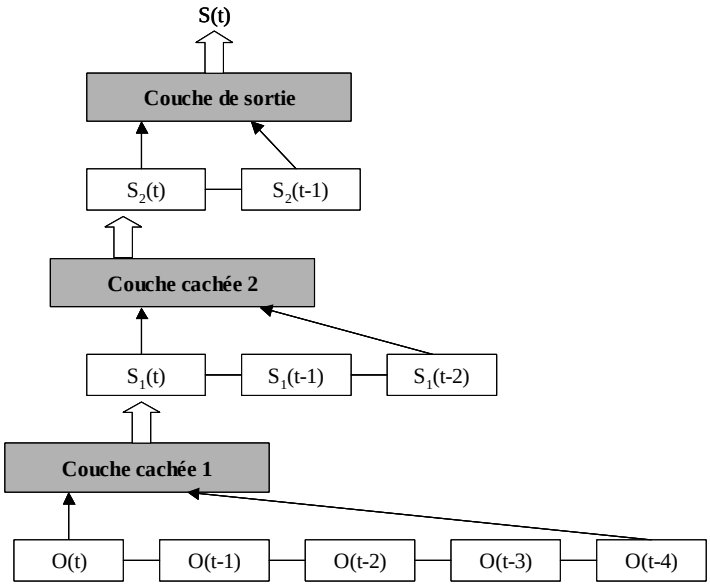
Où η est appelé taux d'apprentissage et doit être assez faible pour garantir la convergence du processus et assez grand pour éviter une convergence trop lente. Cet algorithme est le plus utilisé de nos jours pour l'entraînement des MLPs.

L'inconvénient majeur des réseaux MLPs en reconnaissance de la parole est qu'ils ne sont pas adaptés pour prendre en compte une des principales caractéristiques du signal de parole, sa dimension temporelle. Des solutions sont proposées pour la prise en compte de cet aspect dynamique par l'utilisation d'autres types de réseaux comme par exemple les réseaux TDNN (Time Delay Neural Network) et les réseaux récurrents que nous allons présenter dans ce qui suit.

2.4.2.3.2 Les réseaux à délais

La parole est essentiellement un processus dynamique qui évolue dans le temps. Comme les modèles neuromimétiques de base (par exemple, les MLPs) ne sont pas adaptés au traitement de cette dimension temporelle, diverses solutions sont proposées. Les réseaux de neurones à délais peuvent être qualifiés de réseaux dynamiques étant donné que l'objectif recherché à travers cette architecture à retard est la prise en compte de l'aspect dynamique de la parole. Ce type de réseaux appelés TDNN emploie la structure de base MLP et incorpore des retards pour prendre en compte le contexte temporel de l'unité élémentaire basse vers l'unité élémentaire haute. La **Fig. 2.12** décrit la structure opératoire d'un réseau TDNN à deux couches cachées. Le nombre de retard de chaque couche est déterminé par l'application recherchée.
La décision de reconnaissance est faite en accord avec les activations des cellules de la couche de sortie cumulées sur une fenêtre d'analyse fixe. Ces réseaux sont entraînés en utilisant une version légèrement modifiée de l'algorithme de rétro propagation.
Les réseaux TDNN ont été en premier proposés et utilisés par A. Waibel **[40]** pour réaliser la classification de trois consonnes plosives [b], [d] et [g]. L'architecture du réseau utilisé dans ce travail est une architecture à deux couches cachées. La couche d'entrée possède 16 cellules (correspondant aux 16 coefficients spectraux du vecteur acoustique), chacune avec deux retards.

Les deux couches cachées comportent 8 et 3 cellules chacune avec respectivement 4 et 8 retards, la couche de sortie comporte elle trois cellules. Pour la reconnaissance en mode dépendant du locuteur un taux de reconnaissance global de 98.5% est rapporté.



O(t) : Vecteur d'entrée à l'instant t
Sj(t) : la sortie de la couche cachée j à l'instant t
S(t) : la sortie du réseau à l'instant t

Fig. 2.12- Structure opératoire d'un réseau TDNN.

Il faut noter que des variantes TDNN ont été aussi proposées pour la reconnaissance des mots telles que les réseaux TDNN à plusieurs états (MS-TDNN : multi states TDNN). Dans ce cas des MS-TDNN au lieu de cumuler directement au niveau des trames les scores de vraisemblance dans la couche de sortie il est inclue une couche d'état pour réaliser une programmation dynamique sur ses scores. Chaque phonème est représenté par un état particulier et possède une variable de durée. La reconnaissance des mots est faite en trouvant la séquence d'états optimale.
Il faut aussi remarquer que les TDNN ne sont pas des réseaux dynamiques réels dans le sens qu'ils emploient une représentation spatiale du temps. Comme le nombre de retards est fixé, le réseau peut avoir des problèmes dans le cas de la reconnaissance des phonèmes de durées très variables. Nous savons que dans une séquence de parole, le contexte est important et aide à la classification. Ainsi pour la reconnaissance des phonèmes l'idée est donc d'augmenter le contexte mais en gardant un nombre de coefficients restreint **[41]**. Les poids sont donc partagés et chaque vecteur du contexte est traité par les mêmes fonctions. Ceci permet également de factoriser une partie des calculs.

2.4.2.3.3 Les réseaux récurrents

Contrairement aux réseaux de neurones à propagation directe, les RN récurrents ou RNN (Recurrent Neural Networks) ont des connexions en boucles. Ceci permet de garder et de modéliser

le contexte des formes d'une séquence, en limitant le nombre d'entrées, contrairement aux réseaux à délais (TDNN) qui mettent les poids en commun sur les formes consécutives pour prendre en compte le contexte. La **Fig. 2.13** illustre l'architecture d'un MLP récurrent.

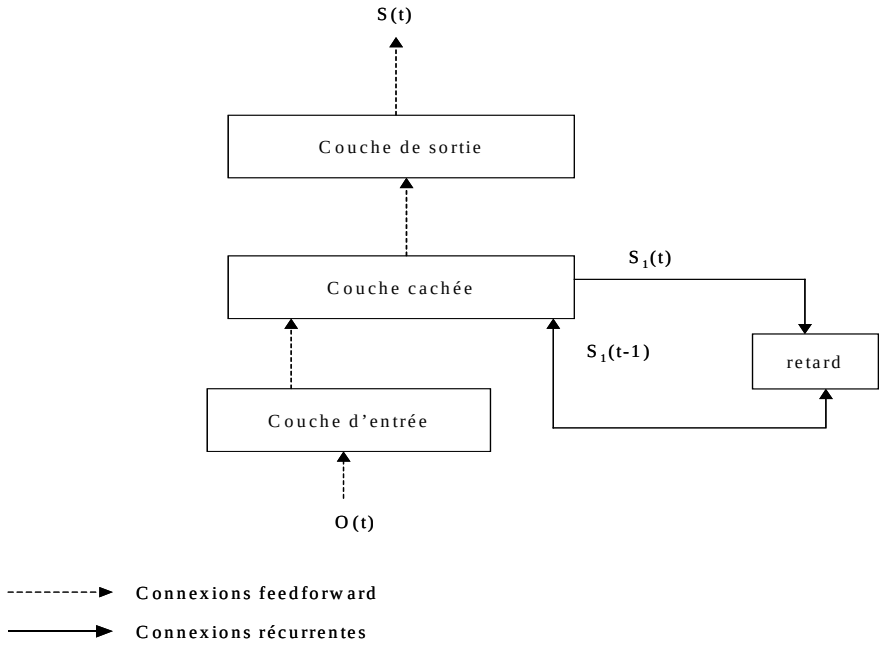


Fig. 2.13 - Architecture d'un réseau PMC récurrent à trois couches.

Avec les connexions récurrentes l'état du réseau dépend non seulement de l'entrée actuelle mais aussi de l'entrée passée. Par conséquent, les RNN sont capables de modéliser des relations séquentielles complexes. Les RNN ont été utilisés dans la prédiction des séries temporelles, l'inférence grammaticale**,** identification et contrôle des systèmes dynamique, etc. Pour la reconnaissance de la parole leur utilisation n'est pas aussi développée que celle des réseaux TDNN **[42], [43]**. Ceci est principalement dû à des problèmes majeurs non encore résolus. Le premier problème est relatif à la difficulté de déterminer des sorties appropriées pour entraîner un RNR. Ces sorties doivent être spécifiées à chaque instant et sont nécessaires pour avoir les moyennes temporelles propres. Aussi, pour s'adapter simultanément aux indices statiques et dynamiques de la parole un réseau de grande taille devient essentiel afin de réaliser une convergence raisonnable. Comme résultat pour une application large vocabulaire en particulier le temps nécessaire à l'entraînement est juste très long pour être accepter.
Cho *et al*. **[44]** ont adopté une structure RNN hiérarchique pour la reconnaissance des syllabes isolées. Dans cette méthode chaque syllabe est segmentée en deux composantes phonémiques modélisée par l'utilisation de différents réseaux RNN. Ceci réduit considérablement la taille du réseau, plusieurs syllabes utilisent les mêmes composantes phonémiques. Les résultats rapportés en mode indépendant du locuteur sont un taux de 73.5% pour 54 syllabes étudiées.
Ces réseaux ont été beaucoup utilisés pour modéliser le signal de parole avec son contexte. A cause des boucles de connexions, l'apprentissage par rétro propagation doit être adapté pour rétro propager l'erreur à travers la séquence. Plusieurs algorithmes ont été proposés, par exemple celui appelé BPS **[45]**.

2.4.2.4 L'approche hybride RNA/MMC

L'inconvénient majeur des RNAs appliqués seuls à la reconnaissance automatique de la parole réside dans leur relative inadéquation à traiter des signaux séquentiels. Comme nous l'avions vu dans la section précédente, malgré le développement de nouvelles architectures de réseaux plus complexes **[43], [45]** afin de dépasser cette limitation par la prise en compte de l'aspect dynamique du signal de parole, les RNA n'arrivent pas encore à surpasser les modèles de Markov cachés. Cela est dû principalement à leur incapacité de modéliser les dépendances à longs termes, ce que fait par contre très bien un MMC par l'intermédiaire de ces contraintes topologiques (traitant les contraintes phonologiques, lexicales et syntaxiques). Les réseaux de neurones seuls ne constituent donc pas, pour l'instant, une solution aux problèmes de la reconnaissance de la parole. C'est pourquoi de nouveaux types de systèmes mixant des techniques connexionnistes avec d'autres symboliques ou stochastiques en particulier les MMCs ont été développés, ces modèles sont couramment appelés modèles hybrides. Nous allons décrire dans ce qui suit les trois principaux formalismes de combinaison des RNA et des MMC conduisant chacun à une catégorie d'approche hybride. L'approche hybride où les RNA sont utilisés comme estimateurs des probabilités *a posteriori* des états des MMC. L'approche hybride dans laquelle les RNA sont utilisés comme quantificateurs vectoriels pour les modèles de Markov cachés discrets et l'approche hybride d'optimisation globale où les RNA constituent un module de pré-traitement dont les sorties forment le vecteur d'observation des fonctions de densité de probabilités des modèles.

2.4.2.4.1 Les RNA comme estimateurs statistiques

Bourlard et *al* **[46], [47], [48]** ont proposé cette catégorie d'approche hybride RNA/MMC pour la reconnaissance de la parole continue dans laquelle les réseaux de neurones de type PMC sont entraînés afin d'estimer les probabilités a posteriori (et non les vraisemblances comme dans le cas des distributions gaussiennes) des états des MMC. L'objectif poursuivi est de maximiser les probabilités a posteriori d'un modèle donné MMC (modèle gauche-droite) M_j étant donné une séquence d'observations acoustiques *X*.

Si l'on suppose que le modèle M_j possède N états $(S_1,.....,S_N)$, et que la séquence d'observations acoustiques $X=(x_1,.....,x_T)$ est de longueur T, on peut développer la probabilité a posteriori selon :

$$P(M_j \mid X) = \sum_{q_1 q_{21} \ldots\ldots q_T} P(q_1 \ldots\ldots q_T, M_j \mid X) \qquad \textbf{(2.40)}$$

$$P(M_j \mid X) = \sum_{q_1 q_{21} \ldots\ldots q_T} P(q_1 \ldots\ldots q_T, \mid X)\, P(M_j \mid q_1 \ldots\ldots q_T, X) \qquad \textbf{(2.41)}$$

$$P(M_j \mid X) = \sum_{q_1 q_{21} \ldots\ldots q_T} P(q_1 \ldots\ldots q_T \mid X)\, P(M_j \mid q_1 \ldots\ldots q_T) \qquad \textbf{(2.42)}$$

acoustique probabilité a priori

Dans cette dernière expression une hypothèse est faite sur la quantité $P(M_j \mid q_1,.....,q_T)$, elle est considérée comme étant indépendante de la séquence d'observations acoustiques X. Elle dépend uniquement des choix faits dans la définition des modèles. Elle peut, par conséquent, être calculée séparément.

En appliquant les propriétés des probabilités jointes, l'équation **2.42** peut être développée selon la relation :

$$P(M_j \mid X) = \sum_{q_1 \ldots\ldots q_T} P(q_1 \mid X)P(q_{\acute{e}} \mid X, q_1).....P(q_T \mid X, q_1, ..., q_{T-1})P(M_j \mid q_1, ..., q_T) \qquad \textbf{(2.43)}$$

$$= \sum_{q_1 \ldots\ldots q_T} \left[\prod_{t=1}^{T} P(q_t \mid X, q_1, ..., q_{t-1}) \right] P(M_j \mid q_1, ..., q_T)$$

En faisant encore l'hypothèse habituelle que les modèles de Markov cachés sont d'ordre 1 et qu'à l'instant t la dépendance sur *X* est limitée à une fenêtre de taille k centrée sur l'observation x_t (un contexte acoustique)

$$X_{t-k}^{t+k} = \left\{ x_{t-k}, ..., x_t, ..., x_{t+k} \right\} \qquad \textbf{(2.44)}$$

L'équation précédente est dans ce cas approximée par l'équation suivante :

$$P(M_j \mid X) \approx \sum_{q_1 \ldots\ldots q_T} \left[\prod_{t=1}^{T} P(q_t \mid x_{t-k}, ..., x_{t+k}, q_{t-1}) \right] P(M_j \mid q_1, ..., q_T) \qquad \textbf{(2.45)}$$

Les probabilités $P(q_t \mid x_{t-k}, ..., x_{t+k}, q_{t-1})$, appelées probabilités de transition conditionnelles, jouent maintenant le rôle de probabilités d'émission. Elles peuvent être estimées par un réseau de neurones de type PMC comme représenté à la **Fig. 2.14**. La couche de sortie du réseau possède un nombre de cellules égale au nombre d'état du modèle et chaque cellule produit donc la probabilité a posteriori de l'état correspondant.

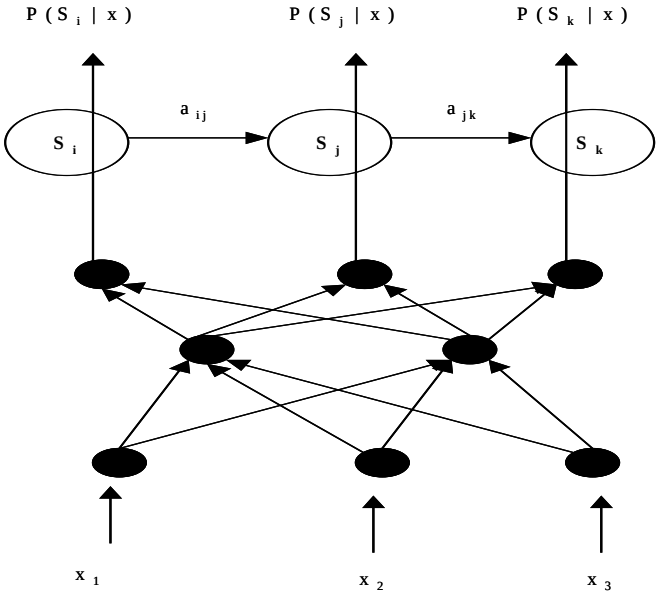


Fig. 2.14 - Exemple d'un PMC avec entrée contextuelle et générant des probabilités a posteriori des états d'un modèle MMC gauche-droite.

Les avantages principaux des réseaux de neurones pour estimer les probabilités d'observation sont les suivantes :

- Les réseaux de neurones n'imposent pas d'hypothèses sur la forme des distributions (gaussienne ou multi-gaussiennes) statistiques associées à chaque état du modèle. En effet, il a été montré en théorie et en pratique que l'apprentissage du réseau de neurones permettait d'estimer des distributions statistiques de n'importe quelle forme.
- Les réseaux de neurones sont soumis à un apprentissage discriminant (une de leurs propriétés majeures). Cela permet de diminuer le manque de pouvoir de discrimination des MMC, au moins localement.
- L'utilisation de l'information temporelle est plus aisée avec les réseaux de neurones : il est facile de fournir plusieurs vecteurs acoustiques à l'entrée du réseau. Une information contextuelle est donc prise en compte dans les probabilités estimées et la corrélation entre des fenêtres successives n'est pas négligée.

L'apprentissage des réseaux de neurones se fait généralement par l'algorithme de Viterbi des séquences d'observations sur les MMC. On obtient alors une annotation pour chaque observation.
Une autre annotation, proposée par Yan et al. **[49]**, consiste à utiliser les fonctions avant et arrière (forward-backward) pour le calcul des probabilités a posteriori des états, étant donné la séquence d'observations. Cette annotation fournit non plus simplement une classe, mais la distribution complète des probabilités a posteriori. Cette information beaucoup plus riche permet d'obtenir de meilleurs résultats. En effet, dans le cas de la RAP, la transition entre phonèmes est relativement ambiguë à l'échelle d'une observation.
Diverses architectures hybrides utilisants des réseaux de neurones pour estimer les probabilités d'observation ont été proposés. Le **Tableau 2.2** présente les principales architectures.

Modèle	Description	Performance
Modèle de référence **[47], [48], [50]**.	Un réseau PMC est entraîné avec l'algorithme BP sur un contexte acoustique pour estimer les probabilités a posteriori des états. Le critère du maximum a posteriori a été considéré.	DARPA Ressource Management (RM) «février 91» testé pour une tache de reconnaissance de la parole continue, un vocabulaire de 991 mots, 5.8% de taux d'erreur (WER) rapporté.
Le réseau MLP est remplacé par un réseau RNN **[51], [52]**.	Des réseaux dynamiques pour estimer les probabilités a posteriori	Le système, nommé ABBOT, a été testé sur ARPA Wall Street Journal. Un taux d'erreur (WER) de 8.8% a été rapporté.
Modélisation explicite du contexte **[53]**.	Le contexte acoustique est explicitement considéré. Le critère utilisé est celui de la vraisemblance des observations étant donné le modèle et le contexte.	DARPA RM, SI, a été testé sur une tache de reconnaissance de parole continue, un taux d'erreur (WER) de 6.3% (CI-hybride) a été rapporté.
Sorties continues **[49]**.	Le mode d'entraînement du réseau est basé sur les variables avant et arrière (Forward-backward) pour calculer les probabilités a posteriori.	Tache de reconnaissance sur des chiffres enchaînés collectés sur le canal téléphonique. Un taux d'erreur de 4.9% a été rapporté.

Tableau 2.1 : Les principales architectectures hybrides basées sur les réseaux de neurones pour estimer les probabilités a posteriori des états des MMC.

2.4.2.4.2 Les RNA comme quantificateurs vectoriels

La quantification vectorielle est souvent mise en cause dans les systèmes de reconnaissance basés sur les modèles de Markov cachés discrets. Pour l'améliorer, plusieurs chercheurs ont tenté d'appliquer les réseaux de neurones afin de générer les dictionnaires. La quantification vectorielle est soumise dans ce cas à un apprentissage par un algorithme de type LVQ (Learning Vector Quantization) **[54]**, **[55]**, **[56]**. Cette méthode constitue une réelle alternative aux méthodes standard de quantification.

D'autres variantes de quantification vectorielle neurale sont proposées. Parmi lesquelles, nous pouvons citer celle présentée par H.P Huter **[57]** où la quantification vectorielle est effectuée par un réseau de neurones de type PMC soumis à l'apprentissage de classes de phonèmes. Dans les résultats présentés, ce système est moins performant que le système MMC discret de départ. Cela peut s'expliquer par le fait que la taille du vocabulaire de quantification vectorielle faite par le réseau de neurones est réduite de 64 à 20 symboles. On perd alors beaucoup d'informations et les erreurs de classification du PMC sont plus difficiles à prendre en compte.

Diverses autres architectures hybrides basées sur une quantification vectorielle connexionniste des vecteurs acoustiques pour les modèles de Markov discrets ont été proposées, plus de détails concernant ses architectures peuvent être consultés dans les références **[58]**, **[59]**, **[60].**

2.4.2.4.3 Approche d'optimisation globale

Dans le modèle hybride présenté par Y. Bengio **[61]**, les réseaux de neurones servent à transformer (non linéairement) l'espace des observations acoustiques en entrée des gaussiennes.
Le système est construit de la sorte en amont du MMC. Deux RN en parallèle font une classification des phonèmes, intégrant un certain contexte. Un troisième RN est initialisé en composantes principales des sorties des deux RN précédents, pour diminuer le nombre de dimensions. La sortie de ce troisième réseau est modélisée par des gaussiennes (diagonales puisque les sorties sont décorrélées). Les MMC, avec leurs probabilités de transition et les paramètres des gaussiennes, sont alors appris. La particularité intéressante de ce système est l'optimisation globale conjointe des MMC et des réseaux de neurones. En effet, la fonction de coût (MLE ou MMIE) est dérivée à travers les observations (gaussiennes) jusqu'aux sorties des RN, qui sont optimisés par rétro-propagation.
D'autres systèmes ont été construits sur des variantes de ce principe d'optimisation globale conjointe des MMC et des RN. Nous pouvons citer le système présenté dans **[62]** où des réseaux de neurones de type PMC sont utilisés comme extracteurs de paramètres complémentaires pour un MMC entraîné sur le critère du maximum de l'information mutuelle. Le système présenté dans **[63]** représente un cadre plus général de l'approche d'optimisation globale conjointe. Dans ce dernier une extension du rôle des RN est réalisée, les réseaux sont utilisés aussi bien pour l'extraction des paramètres que pour l'estimation des modèles MMC.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les mécanismes de production et de perception de la parole chez les humains ainsi que les caractéristiques du signal vocal responsables en grande partie des difficultés de la tâche de reconnaissance. Par la suite, nous avons décrit les différentes approches de reconnaissance avec le processus de traitement relatif à chacune d'entre elles. C'est ainsi que nous avons présenté les méthodes d'analyse du signal de parole les mieux adaptées et les plus utilisées, les stratégies de reconnaissance adoptées dans l'approche analytique fondée sur les connaissances, dans l'approche globale DTW, dans l'approche statistique Markovienne et enfin dans l'approche hybride mixant les modèles de Markov cachés et les réseaux de neurones artificiels.

Phonétique Arabe et Reconnaissance Analytique Fondée sur les Connaissances

3.1 Introduction

La reconnaissance analytique appelée aussi reconnaissance phonétique est basée sur l'identification d'éléments phonétiques du continuum vocal. Elle tente, à partir d'informations de types acoustiques, phonétiques et phonologiques, d'interpréter le message émis. De ce fait, la conception de systèmes de reconnaissance fondés sur cette approche nécessite préalablement une étude phonétique de la langue qu'on voudrait reconnaître. Dans ce sens, nous avons jugé plus intéressant d'associer dans ce même chapitre l'étude phonétique de la langue arabe et notre contribution dans l'approche analytique. C'est ainsi que les particularités phonétiques de la langue arabe et la problématique de sa reconnaissance sont d'abord présentées. Ensuite, sont présentées, notre contribution dans la mise en œuvre de cette approche à savoir les connaissances acquises sur les consonnes arrières et emphatiques formalisées sous forme de règles de production et gérées par un système expert outil de l'intelligence artificielle (I.A).

3.2 Eléments de phonétique arabe

La phonétique a occupé une position de choix chez les anciens linguistes arabes particulièrement Sibaway et son maître Al-Khalil Ibn Ahmed. Sibaway, grammairien du 8ème siècle, donne dans son livre une description large et précise des sons de l'arabe parlée de l'époque. Bien qu'elle ne soit pas la langue du quotidien des peuples l'arabe standard, version moderne de l'arabe classique, est la langue de la science, de l'enseignement, de la littérature, etc. C'est la langue commune à tous les arabophones et à près d'un milliard de musulmans. Le système phonétique de l'arabe standard comporte 34 phonèmes. Le **Tableau 3.1** donne pour chaque phonème /harf/ y figurant le nom, la transcription A.P.I, le symbole arabe et le code informatique attribué. L'originalité de la phonétique arabe se fonde, pour une grande partie, sur la pertinence de la durée dans le système vocalique, sur la présence du trait de gémination, des consonnes emphatiques et des consonnes arrières (pharyngales et glottales).

- Le système vocalique de l'arabe standard se compose de trois voyelles brèves /a/, /u/ , /i / et leurs correspondantes longues /aa/, /uu/, /ii/. La durée d'une voyelle longue est approximativement égale à 1.5 à 2 fois la durée moyenne de sa correspondante brève. Le paramètre durée est donc très important dans la langue arabe tant au niveau sémantique qu'au niveau grammatical. Il caractérise les voyelles mais également les consonnes géminées.

- La gémination est le dédoublement de deux consonnes identiques en une géminée. La gémination en arabe a une fonction différenciative morphologique et sémantique. Toutes les consonnes arabes peuvent être géminées. Au plan spectral, la gémination se traduit par un allongement de la durée **[64]** et un renforcement des caractéristiques acoustiques.

- Les consonnes arrières sont des consonnes spécifiques à la langue arabe. Elles se caractérisent par un lieu d'articulation se trouvant à l'arrière du conduit vocal. Ces consonnes sont difficiles à étudier et leur investigation par les moyens classiques des phonéticiens est problématique à cause de leurs points et modes d'articulations qui se trouvent dans la région laryngale et pharyngale qui sont des régions difficilement accessibles. Dans la langue arabe, il existe quatre consonnes arrières, 2 glottales / ?/, /h/ la première classée comme plosive et la seconde comme fricative et 2 pharyngales /$\hbar$/ , /ε/ classées respectivement fricative sourde et sonnante **[65]**.

- L'emphase est dans le cas des langues sémitiques un trait phonétique caractérisant certaines consonnes. Il existe dans la langue arabe 4 consonnes emphatiques, 2 plosives /$\underline{d}$/ , /$\underline{t}$/ et 2 fricatives /$\underline{s}$/ , /$\underline{\partial}$/ . Dans ses travaux sur la nature des consonnes emphatiques, Bonnot **[66]** trouve que ces consonnes possèdent deux lieux d'articulation, l'un antérieur et l'autre postérieur mais que la corrélation d'emphase doit être limitée à six consonnes (/$\underline{t}$/ -/t/ , /$\underline{d}$/ - /d/ et /$\underline{s}$/ - /s/). En outre, sur le plan spectral, l'emphase se traduit par un renforcement des propriétés acoustiques et une influence particulière sur les formants des voyelles adjacentes.

Phonème Code Choisi	**Nom**	**Transcription A.P.I[1]**	**Symbole Arabe**	**Phonème Code Informatique**	**Nom**	**Transcription A.P.I[1]**	**Symbole Arabe**
s.	**saad**	$\underline{s}$	ص	t	ta	t	ت
d.	**daad**	$\underline{d}$	ض	s	seen	s	س
t.	**tta**	$\underline{t}$	ط	j	jeem	z	ج
D.	**tha**	$\underline{\partial}$	ظ	X	kha	χ	خ
E	**ain**	ε	ع	d	daal	d	د
h	**ha**	h	ـه	T	thal	θ	ت
H	**hha**	$\hbar$	ح	r	ra	r	ر
A	**hamza**	?	ء	Z	za	Z	ز
G	gin	γ	غ	c	sheen	ʃ	ش
f	fa	f	ف	D	dhal	∂	د
q	qaaf	q	ق				
k	kaaf	k		a	fatha	a	ـَ
l	lam	l	ل	u	damma	u	ـُ
m	mim	m	م	i	kasra	i	ـِ
n	noun	n	ث	aa	Fatha longue	a:	
w	waw	w	و	uu	Damma longue	u:	
y	yaa	y	ي	ii	Kasra longue	i:	
b	baa	b	ب				

Tableau 3.1: Les phonèmes de l'arabe standard

[1] Alphabet Phonétique International

3.3 Problématique de reconnaissance

Bien que la langue arabe ait fait l'objet de plusieurs travaux ayant trait à l'aspect phonétique et linguistique depuis maintenant plusieurs années **[67]**, **[68], [69]**, **[70]**, il reste que dans le domaine de la technologie des langues en particulier celui de la reconnaissance automatique, elle est toujours considérée comme une langue peu dotée comparativement à des langues telles que l'anglais, le français, l'espagnol ou le japonais. Le problème de la reconnaissance de la langue arabe reste donc posé malgré les travaux de qualité menés ces dernières années **[71]**, **[72], [73]**, **[74]** qui sont malheureusement insuffisants comparativement aux nombreux travaux réalisés pour les autres langues. Nous avons choisi en ce qui nous concerne d'apporter notre contribution au problème de la reconnaissance de la langue arabe standard en s'intéressant à l'identification de phonèmes complexes et spécifiques tels que les consonnes emphatiques et arrières car, de par leurs structures articulatoires et acoustiques, ces phonèmes sont unanimement reconnus comme responsables des limites des systèmes de reconnaissance dédiés à la langue arabe.

3.4 Approche analytique

3.4.1 Motivations

L'approche analytique fondée sur les connaissances tente de détecter et d'identifier les unités linguistiques discrètes (les phonèmes dans notre cas) du message vocal. Le décodage acoustico-phonétique est mené dans ce cas par un système "intelligent" qui raisonne à partir d'un ensemble de connaissances acoustiques, phonétiques et phonologiques. Les systèmes experts, outils de l'intelligence artificielle (I.A), permettent en effet d'intégrer ces différentes sources de connaissances. Ils tentent de reproduire la compétence d'experts humains dans le domaine. L'expertise humaine en compréhension de la parole est évidente, elle semble tellement instinctive que l'introspection ne fournit guère d'informations. Nous ne savons pas vraiment sur quels éléments acoustiques porte le processus de reconnaissance. Par contre des experts peuvent transcrire phonétiquement des sonagrammes avec des taux de reconnaissance appréciables, il semble donc naturel de modéliser et d'utiliser les connaissances et stratégies de phonéticiens experts en lecture de spectrogrammes. La **Fig. 3.1** illustre un exemple de spectrogramme numérique à bande étroite.

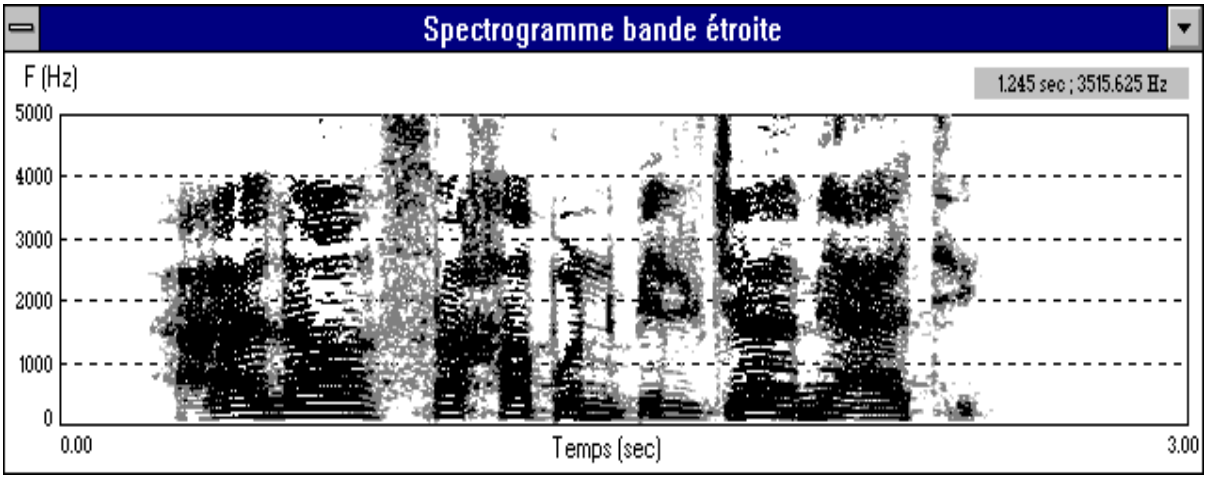


Fig. 3.1 – Exemple de spectrogramme numérique.

3.4.2 Structures acoustiques

Nous présentons dans cette section l'analyse acoustique réalisée qui nous a permis d'établir la structure acoustique de chacune de ces consonnes étudiées et de dégager les connaissances nécessaires à leur identification. Le signal acoustique est riche en informations phonétiques. Ces informations peuvent le plus souvent être extraits de la représentation spectrographique du signal. C'est ainsi qu'à partir de l'analyse d'un nombre important de spectrogrammes de ces consonnes prononcées par différents locuteurs dans différents contextes phonétiques, nous avons relevé les caractéristiques acoustiques de ces consonnes ainsi que la variation de ces caractéristiques en fonction de l'environnement phonétique. Ceci a été possible grâce au logiciel **APHAK** (**A**cquisition **PH**onetic **A**coustic **K**nowledge) **[75]** mis au point. APHAK est un logiciel interactif d'analyse de la parole possédant différentes fonctions telles que l'acquisition-restitution du signal, la visualisation du signal ainsi que ses différents traitements, la manipulation de corpus, le calcul de spectrogrammes et d'autres fonctions permettant une étude très fine de la parole. Ce logiciel est décrit en **annexe A**..

3.4.2.1 Corpus d'analyse.

Nous avons fait prononcer les consonnes d'étude dans des mots (éventuellement sans signification) en tenant compte de tous les contextes dans lesquels ces consonnes peuvent être prononcées. Les enregistrements ont eu lieu dans une salle non bruyante selon des conditions semblables pour tous les locuteurs. Les locuteurs au nombre de vingt, tous Algériens, ont répété le corpus cinq fois. Ce corpus est donné en **annexe B**.

3.4.2.2 Les consonnes arrières

Phonème /H/ - est une fricative pharyngale rétrécie sourde de durée relative 100-160ms. Toutefois elle peut être voisée en position intervocalique. Sur le spectrogramme, elle apparaît comme un bruit d'intensité plus forte que pour la consonne /h/. Suivie de la voyelle /a/ ou de la voyelle /aa/, le bruit de friction est concentré autour de 3500-4500Hz et il inexistant en dessous de 800-1000Hz. Suivie de la voyelle /u/ ou de la voyelle /uu/, la concentration du bruit est entre 3200 et 4200Hz avec une limite inférieure autour de 600-1300Hz. Suivie de la voyelle /i/ ou de la voyelle /ii/, la concentration du bruit est autour de 3200-4500Hz et la limite inférieure autour de 1800-2200Hz. Cette consonne affecte les formants vocaliques. Dans un contexte gauche **/H/**, les voyelle /a/ et /aa/ possèdent un formant F2 plat et un formant F3 montant, les voyelles /u/ et /uu/ le formant F2 descendant et le formant F3 montant alors que pour les voyelles /i/ et /ii/ le formant F2 est montant et le formant F3 est plat.

Phonème /E/ - Traditionnellement cette consonne est classée comme fricative voisée. Al-ani **[67]**, dans son étude acoustique et physiologique des consonnes pharyngales de l'arabe, trouve que l'allophone le plus commun est plosive sourde plutôt que fricative voisée. Pour notre part, la classe retenue est celle des sonnantes. En effet **/E/** est apparue généralement sur le spectrogramme comme un bruit voisé avec une nette structure pseudo-formantique. Cette structure est fonction du contexte. Une étude statistique afin de déduire les structures formantiques principales a été réalisée, les résultats de cette étude sont résumés dans le **Tableau 3.2**. Cette consonne se distingue aussi par son formant F1 élevé.

Contexte-droit	Formants	Moyenne	Intervalle de confiance
/a/ ou /aa/	F1	767	740-790
	F2	1483	1440-1520
	F3	2442	2400-2480
/u/ ou /uu/	F1	465	440-480
	F2	1255	1200-1290
	F3	2186	2100-2220
/i/ ou /ii/	F1	466	445-490
	F2	1800	1740-1850
	F3	2645	2600-2700
Non vocalique	F1	650	600-696
	F2	1630	1550-1720
	F3	2450	2400-2496

Tableau 3.2: Structure pseudo-formantique du phonème **/E/.**

Phonème /A/ - Cette consonne apparaît principalement sur le spectrogramme comme un intervalle de silence de durée variable suivi occasionnellement par un burst (barre d'explosion). Le burst est dans certains cas suivi par un silence de durée 15-30ms dans d'autres par un bruit faible. La présence du burst, ses caractéristiques sont fonction du contexte immédiat de la consonne. Suivie de la voyelle /a/ ou de la voyelle /aa/, le burst est souvent présent avec une durée de 15-30ms, son énergie est concentrée autour de 4000-5000Hz. Suivie de la voyelle /i/ ou de la voyelle /ii/, la concentration de l'énergie du burst est plus élevée en fréquence que pour les voyelles /a/ et /aa/. Suivie de la voyelle /u/ ou /uu/, le burst est très souvent absent, **/A/** attaque directement la voyelle. Entre deux voyelles, **/A/** peut apparaître comme un lien qui connecte les formants F1, F2 et F3 des voyelles précédente et suivante. En position finale, **/A/** est en variation libre. Elle apparaît comme un burst suivi ou non d'un bruit faible, ce burst est précédé par un intervalle de silence de durée 80-120ms. La consonne **/A/** n'affecte pas les formants vocaliques dans le sens d'une transition.

Phonème /h/ - est une fricative glottale sourde de durée relative 100-150ms, toutefois elle peut être voisée en position intervocalique. Sur le spectrogramme, elle apparaît comme un bruit avec une relative structure formantique. La structure formantique, la zone de concentration du bruit varient en fonction de l'environnement phonétique. La structure formantique de la fricative **/h/** dont le contexte droit immédiat est la voyelle /a/ ou la voyelle /aa/ est F1 : 500-650Hz, F2 : 1450-1550Hz. Avec le contexte droit immédiat la voyelle /u/ ou la voyelle /uu/ F1 : 280-350Hz, F2 : 200-2800Hz enfin suivie de la voyelle /i/ ou la voyelle /ii/ la structure est F1 : 320-400Hz et F2 : 900-1100Hz.
Pour le bruit de friction, quand la consonne **/h/** est suivie par la voyelle /a / ou la voyelle /aa/ , la concentration du bruit est autour de 3000-4000hz. Suivie de la voyelle /i/ ou de la voyelle /ii/, le bruit est concentré entre 3200-4200Hz, il devient très léger voir inexistant en dessous de 2000Hz. Suivie de la voyelle /u/ ou de la voyelle /uu/, le bruit de friction est nettement plus léger et est concentré dans les basses fréquences autour de 1000Hz. Les formants vocaliques sont à peine affectés par la consonne **/h/.** C'est un phonème instable, influencé par le voisinage vocalique, il est difficile de lui trouver un modèle.

3.4.2.3 Les consonnes emphatiques

Phonème /s./ - possède une structure acoustique spécifique d'une consonne fricative sourde de durée relative 100-170ms. Le bruit de friction est concentré dans la bande 4000-5000Hz, il est inexistant au-dessous de 2800-3000Hz. Ce phonème affecte les formants des voyelles adjacentes. Dans le contexte gauche immédiat /**s./,** les voyelles /a/ et /aa/ possèdent un formant F1 autour de 700Hz avec une transition plate et le formant F2 autour de 1200-1250Hz avec une transition montante, les voyelles /u/ et /uu/ possèdent le formant F2 descendant et le formant F1 plat alors que les voyelles /i/ et /ii/ possèdent les formants F1 et F2 tous les deux montants.
Suivie de la voyelle /a/ ou de la voyelle /aa/, le bruit de friction est concentré autour de 4000-4300Hz et il inexistant en dessous de 3000Hz. Suivie de la voyelle /u/ ou de la voyelle /uu/, la concentration du bruit est entre 4200 et 4600Hz avec une limite inférieure autour de 3200Hz. Suivie de la voyelle /i/ ou de la voyelle /ii/, la concentration du bruit est autour de 40000-4500Hz et la limite inférieure autour de 3000-3300Hz.

Phonème /D./ - apparaît comme une consonne fricative sourde avec une durée relative comprise entre 90-120ms. Cette consonne apparaît aussi comme un phonème possédant une pseudo-structure formantique F1 : 250-275Hz, F2 : 900-1000Hz et F3 : 2300-2350Hz. Dans le contexte précédent immédiat /**D./,** les voyelle /a/ et /aa/ possèdent un formant F1 localisé autour de 600Hz avec une transition plate et le formant F2 localisé autour de 1100-1200Hz avec aussi une transition plate, les voyelles /u/ et /uu/ possèdent le formant F1 descendant et le formant F2 montant alors que les voyelles /i/ et /ii/ possèdent les formants F1 et F2 tous deux montants.

Phonème /t./ - Cette consonne apparaît principalement sur le spectrogramme comme un intervalle de silence de durée relative 100-180ms suivi par un burst de durée relative 20-30ms. L'énergie du burst est concentrée autour de 1500-2400Hz quand elle est suivie par la voyelle /a/ ou la voyelle /aa/. Quand elle est suivie par la voyelle /u/ ou la voyelle /uu/, la concentration de l'énergie du burst est dans la bande 1500-2200Hz. Dans le cas des voyelles /i/ et /ii/, la concentration du bruit est dans la bande 1700-2400Hz. Dans un contexte précèdent immédiat /**t./,** les voyelles /a/ et /aa/ possèdent des formants F1 et F2 plats. Les voyelles /u/ et /uu/ ont un formant F1 plat et le formant F2 descendant. Les voyelles /i/ et /ii/ ont un formant F1 descendant et F2 montant.

Phonème /d./ - est une plosive sourde avec une durée relative de 80-100ms. C'est un phonème très difficile à prononcer et on le confond presque souvent avec le phonème **/D./**. Quand ce phonème est bien articulé, avec le contexte droit immédiat la voyelle /a/ ou la voyelle /aa/, la concentration du burst est dans la bande 3500-4000Hz, avec le contexte droit immédiat la voyelle /i/ ou la voyelle /ii/, cette concentration est autour de 4000-5000Hz. Les transitions formantiques des voyelles adjacentes /a/ et /aa / sont F1 et F2 plats, celles des voyelles /u/ et /uu/ sont F1 descendant et F2 montant alors que celles des voyelles /i/ et /ii/ sont montantes pour les deux formants F1 et F2.

3.4.3 Le système de reconnaissance

Le système de reconnaissance **[21]** dont le synoptique est donné par la **Fig. 3.2** permet de faire la reconnaissance analytique des phonèmes de l'arabe en parole continue. Ce système de décodage

proprement dit est un système expert à base de règles de production. Il fonctionne selon les étapes suivantes:

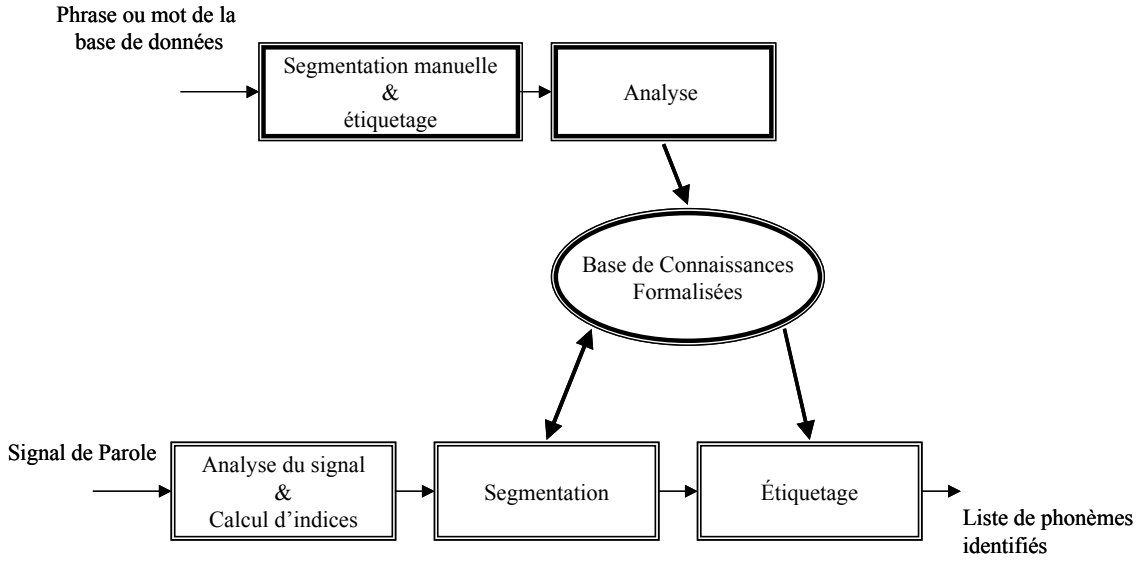


Fig. 3.2 - Schéma du système de reconnaissance analytique.

3.4.3.1 L'analyse du signal et le calcul d'indices

L'extraction des indices phonétiques pertinents est une étape très importante dans le processus de reconnaissance. Nous avons développé une procédure pour chacun des indices phonétiques suivants : la durée du segment, le degré de voisement, la position et les caractéristiques de la barre d'explosion (Burst), les valeurs de formants, les transitions formantiques, la limite du bruit de friction.

- **La barre d'explosion** (burst) : Les consonnes plosives se caractérisent par une obstruction suivie d'une explosion d'énergie de durée brève. Cette explosion d'énergie se traduit sur le spectrogramme par une barre verticale suivie d'un bruit de friction. La localisation de cette barre d'explosion repose sur un critère énergétique. En effet après avoir délimiter la plosive par ces prélèvements de début et de fin avec une marge de sécurité, on procède au calcul de l'énergie totale en décibel dans chacune des huit bandes de fréquences en Hertz réparties comme suite : [0-1000], [1000-2000], [2000-3000], ……., [7000-8000] et ceci pour chaque prélèvement situé entre le début et la fin de la plosive, en utilisant les valeurs obtenues par un algorithme de TFR sur le signal temporel échantillonné à 16KHz. Il apparaît suffisant, pour détecter le burst, de rechercher la position du maximum de l'énergie dans chaque bande de fréquences. Seulement, la grande variabilité des valeurs et la présence éventuelle d'un second burst impose une interpolation des points par un algorithme de la fonction SPLINE **[76]** et de travailler sur les valeurs de la dérivée première. Un numéro de prélèvement est candidat à une position du burst si son énergie dans la bande est supérieure à un certain seuil calculé en fonction de l'énergie de la partie silencieuse de la plosive et celle de la voyelle suivante. Pour accepter un numéro de prélèvement comme position finale du burst, il doit occurer modulo une marge d'un prélèvement un nombre acceptable de fois

sur l'ensemble des huit bandes de fréquences afin qu'il corresponde bien à sa visibilité sur le spectrogramme.

- **Fréquence du burst :** on calcule le spectre par LPC du prélèvement détecté préalablement comme étant le burst. La position en fréquence du maximum de l'énergie correspond à la fréquence du burst.

- **Le pitch :** Le pitch ou fréquence fondamentale est un paramètre important aussi bien pour la synthèse de la parole, le codage ou la reconnaissance automatique de la parole. L'oreille est en effet très sensible à ces variations qui constituent la prosodie. Son estimation est réalisée par la méthode de l'autocorrélation modifiée introduite par Dubnowsky **[77].** L'algorithme de calcul est résumé dans ce qui suit :

  - Lecture d'une trame d'échantillons du signal de parole (la fenêtre doit être choisie de façon à diminuer son influence sur la fonction d'autocorrélation, une fenêtre idéale doit contenir deux à trois périodes du signal de parole). Sa longueur est de cinq à vingt millisecondes pour des valeurs du fondamental élevées et de vingt à cinquante millisecondes pour des faibles valeurs.

  - Filtrage passe-bas : un choix judicieux de la fréquence de coupure (aux alentours de 900Hz) permet l'élimination des formants d'ordre supérieur à deux.

  - Passage du signal par un dispositif non linéaire (le 3-level center clipping) donné par la **Fig. 3.3** qui rend le calcul de la fonction d'autocorrélation plus simple et le spectre de puissance plus plat pour mieux détecter le pic correspondant au fondamental.

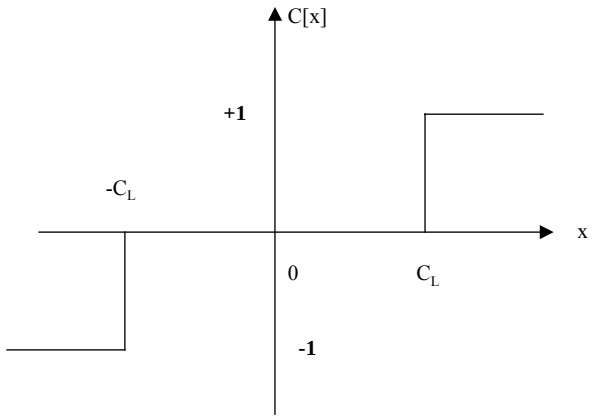


Fig. 3.3 – Graphe de la fonction 3-level center clipping.

Ce qui revient à coder chaque échantillon x(n) sur trois niveaux (1, 0, -1). De ce fait, la fonction d'aucorrélation du signal codé est particulièrement simple à calculer. Le seuil d'écrétage C_L (Center clipped signal) est défini par la relation $C_L = 0.64M$, où M représente le

plus petit des maximums Max1 et Max3 calculés respectivement sur le premier et le dernier tiers de la fenêtre d'analyse.

- Calcul de la fonction d'aucorrélation : une fois le signal codé, on procède au calcul de la fonction d'aucorrélation $R(k)$ puis à la recherche du maximum de cette fonction. La décision de voisement est déterminée en comparant $R(0)$ au plus grand pic $R_{\max}(k)$.

$$\frac{R_{\max}(k)}{R(0)} \leq S \qquad \text{la séquence est non voisée}$$

$$\frac{R_{\max}(k)}{R(0)} \rangle S \qquad \text{la séquence est voisée}$$

Le seuil S est déterminé expérimentalement. La valeur de S retenue est 0.03 (30%).

- Calcul du pitch F_0 par la relation :

$$F_0 = \frac{F_e}{k_{\max}} \qquad \textbf{(3.1)}$$

F_e : fréquence d'échantillonnage
k_{max} : abscisse du maximum de la fonction d'autocorrélation

- **Le degré de voisement :** C'est le rapport entre le nombre de prélèvement voisé et le nombre total des prélèvements d'un segment. Le voisement d'un prélèvement est déterminé lors du calcul de la fréquence fondamentale. Les voyelles ont le plus souvent un degré de voisement égal à un.

- **Energie moyenne :**

$$E_{moy} = \frac{\sum_{j=1}^{Nb\_canaux} E_j(t)}{Nb\_canaux} \qquad \textbf{(3.2)}$$

$Ej(t)$ énergie en fréquence de la trame t.

- **Centre de gravité normalisé (Cdgn) :**

$$Cdgn = \frac{\sum_{j=1}^{Nb\_canaux} j\,(E_j(t) - E_{\min}(t))}{\sum_{j=1}^{Nb\_canaux} (E_j(t) - E_{\min}(t))} \qquad \textbf{(3.3)}$$

avec $E_{\min}(t) = \min(E_1, E_2, ..., E_{Nb\_canaux})$.

- **Suivi de formants :** à partir des coefficients LPC, nous calculons les premiers pôles candidats à des positions de formants. Pour déterminer la suite des valeurs qui correspondent au formant i (i allant de 1 à 3) des contraintes sont imposées : de limitation de la bande (on ne retient que les pôles qui ont une bande passante inférieure à 500Hz), de continuité des valeurs entre des trames successives (critère de continuité) et pour chaque formant des limites de variation sont fixées. Le calcul des fréquences formantiques moyennes est fait uniquement sur la zone stable du segment (nous avons retenu la zone comprise entre le tiers après le début et le tiers avant la fin du segment).

- **Transitions formantiques :** pour chaque formant, sur un intervalle compris entre la fin de la consonne et le tiers de la voyelle contiguë, nous calculons l'ensemble des fréquences formantiques. La courbe formée par l'évolution temporelle de la fréquence d'un formant est approximée par une droite grâce à une procédure de régression linéaire utilisant la méthode des moindres carrées. Le signe du coefficient directeur de la droite permet de déterminer la nature de la transition. Nous avons retenu trois transitions, montante ($a > 0$), descendante ($a < 0$) et plate ($a = 0$).

- **Limites du bruit de friction :** Ce sont des indices visuels pour l'identification des fricatives. La limite inférieure du bruit de friction se calcule en parcourant le spectre de la fricative considérée depuis les basses fréquences jusqu'à atteindre le seuil de visibilité spectrographique. La fréquence du seuil représente la limite inférieure de friction. Pour la limite supérieure, il suffit de parcourir le spectre depuis les hautes fréquences.

3.4.3.2 La segmentation

Elle consiste à segmenter le signal de parole en grandes classes phonétiques. Elle est réalisée en utilisant des algorithmes non contextuels reposants sur des critères énergétiques (**NOVOCA**, **PLOSI**, …..) **[21]**. Nous avons retenu trois grandes classes phonétiques: les voyelles (Vo), les fricatives (Fr) et les plosives (Pl). Le reste des phonèmes de l'arabe standard sont mis dans la classe des sonnantes (So).

- **Noyaux vocaliques : NOVOCA**

  L'algorithme NOVOCA a pour but de localiser tous les noyaux vocaliques contenus dans une phrase et de déterminer la durée vocalique moyenne. Le critère utilisé est l'énergie contenue dans la bande de fréquences [250Hz – 2200Hz]. Cette bande de fréquences a été choisie de manière à défavoriser les sons ayant principalement de l'énergie en très basses fréquences (par exemple les nasales) et ceux qui ont de l'énergie en hautes fréquences (par exemple les fricatives). Les noyaux vocaliques correspondent aux pics de la courbe d'énergie qui vérifient les critères suivants :

  - Le nouveau pic doit atteindre au moins 55% du pic précédent.
  - La vallée de part et d'autre du pic est fonction de la hauteur du pic (plus un pic est important plus la vallée doit être importante)
  - Au moins 55% des échantillons du noyau vocalique doivent être voisés.

  Quand un pic vérifie tous ces critères, on recherche le début et la fin du noyau. On ne décrira que la recherche de la fin du noyau, car la recherche du début du noyau est symétrique.

- A partir du pic, on recherche les points D, F et R qui correspondent respectivement aux bornes de la chute d'énergie et au seuil à partir duquel l'énergie commence à remonter.
- On trace le segment DF et on recherche le point de la courbe d'énergie situé au-dessus de cette droite et qui, de plus, est à la plus grande distance de cette droite. Si le point trouvé se situe entre D+(F-D)/4 et D-(F-D)/4, c'est le marqueur de fin du noyau, sinon c'est (D+F)/2.

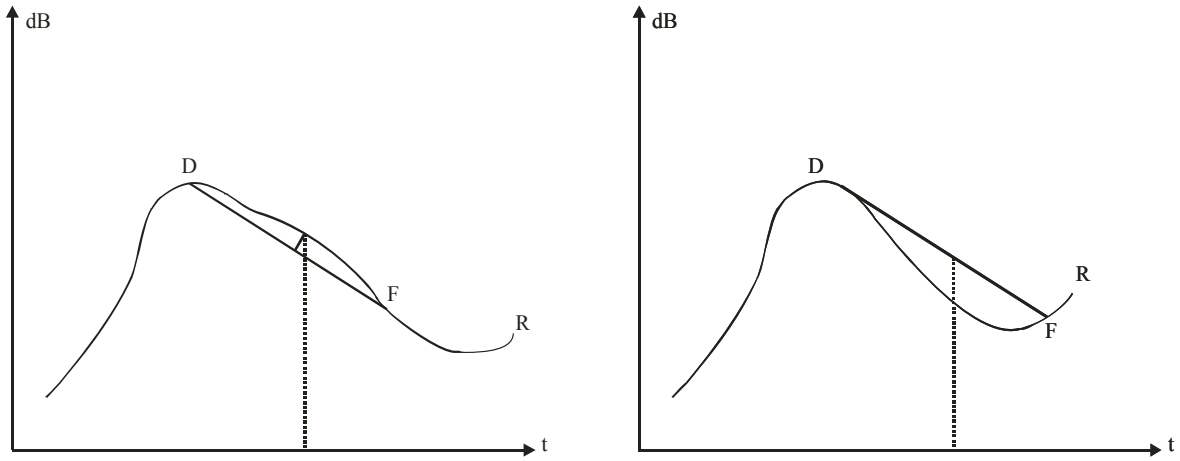


Fig. 3.4 – Calcul de la fin d'un noyau vocalique.

- **Plosives : PLOSI**

  Les plosives sourdes apparaissent sur le spectrogramme comme des colonnes blanches suivies d'une barre d 'explosion. Dans le cas des plosives sonores une vibration des cordes vocales subsiste, elle se matérialise sur le spectrogramme par une barre de voisement. C'est pourquoi, la détection des plosives est basée sur la recherche de segments qui ne présentent pas d'énergie visible au-delà de 650Hz.

- **Fricatives : FRICA**

  Les fricatives possèdent une répartition fréquentielle particulière, beaucoup d'énergie en haute fréquence et pratiquement pas d'énergie visible en basse et moyenne fréquence. Leur détection repose sur le calcul du centre de gravité sur l'énergie visible.

3.4.3.3 L'étiquetage phonétique

Il consiste à identifier les phonèmes en utilisant un système expert à base de règles de production. Ce système consiste en une base de connaissances d'identification de phonèmes sous forme de règles de

production et en un moteur d'inférence. Ce dernier fonctionne en chaînage avant et en chaînage arrière en effectuant une analyse de gauche à droite, segment par segment.

Une règle prend la syntaxe suivante:
R (numéro règle)
CONTEXTE-DROIT (liste de phonèmes)
CONTEXTE-GAUCHE (liste de phonèmes)
SI (prémisses)
PHONEMES (liste de phonèmes)
FIN

Chaque prémisse correspond à un indice caractérisant un segment donné que nous avons déduit lors de la lecture de spectrogrammes de ce segment. La partie contexte gauche et/ou droit est optionnelle dans une règle. Ainsi, la reconnaissance de la consonne **/A/** est fondée sur :

- La présence de la barre d'explosion et sa fréquence;
- Le degré de voisement ;
- Les transitions formantiques des voyelles du contexte droit immédiat de la consonne.

Concernant les consonnes **/h/** et **/E/** nous avons utilisé :
- Le degré de voisement;
- Les valeurs des formants F1 et F2;
- Les transitions formantiques dans les voyelles adjacentes (gauche et droite).

Concernant la consonne **/H/** nous avons utilisé:
- Le degré de voisement;
- La limite inférieure du bruit de friction;
- Les transitions formantiques dans les voyelles adjacentes.

Pour les consonnes **/d./** et **/t./** nous avons utilisé
- La présence de la barre d'explosion et sa fréquence;
- Le degré de voisement;
- Les transitions formantiques des voyelles du contexte droit immédiat de la consonne.

Pour la consonne **/s./** nous avons utilisé:
- Le degré de voisement;
- La limite inférieure du bruit de friction;
- Les transitions formantiques dans les voyelles adjacentes.

Pour la consonne **/D./** nous avons utilisé :
- Le degré de voisement;
- Les valeurs des formants F1, F2 & F3 de la consonne;
- Les transitions formantiques dans les voyelles adjacentes;

- La limite inférieure du bruit de friction.

Pour l'ensemble des consonnes étudiées, nous avons déduit 76 règles de production. Voici par ailleurs deux exemples de règles : la règle sur la barre d'explosion (R3) pour le phonème **/A/** et la règle sur les formants (R53) pour le phonème **/h/**.

R (3)
CONTEXTE-DROIT [a aa]
Si
(Burst-présent &
Burts-fréq-Act 3800-5500)
PHONEMES [A 10]
FIN.

R (53)
CONTEXTE-DROIT [a aa]
Si
(Formant 1_Act 500-650 &
Formant 2_Act 1450-1550)
PHONEMES [h 10]
FIN.

3.4.4 Expériences et résultats

Les connaissances acquises sur les consonnes arrières et les consonnes emphatiques ont été testées sur un corpus de phrases ciblées. Il faut savoir que certaines de ces consonnes se présentent en nombre réduit dans la langue. Par conséquent, nous étions amenés pour l'évaluation de notre système, de construire un corpus de phrases cibles comprenant un nombre suffisant des consonnes d'étude. Ce corpus de test est donné en **annexe C**. Certains critères ont été fixés pour le choix de ces phrases : simplicité de la forme syntaxique, longueur raisonnable et le contenu (faire apparaître les phonèmes d'étude dans les différents contextes où elles peuvent être prononcées). Ces phrases ont été prononcées par 20 locuteurs natifs algériens dans des conditions semblables et calmes.

3.4.4.1 Résultats de la segmentation

L'évaluation de la segmentation consiste à rapprocher les résultats de la segmentation automatique avec la segmentation manuelle effectuée sur les phrases du corpus en utilisant un algorithme de programmation dynamique **[25]**. Pour chaque phonème le nombre d'occurrence dans chaque classe phonétique ainsi que les taux de bonne segmentation sont calculés. L'évaluation est résumée dans le **Tableau 3.3** suivant :

Phonèmes	Pl	Vo	Fr	So	Omis	Taux (%)
/d./	282	6	20	15	17	83
/t./	279	4	5	8	24	87.2
/s./	0	0	265	33	2	88.3
/D./	23	0	174	5	8	83
/A/	302	11	0	89	18	72
/H/	11	0	446	44	19	86
/E/	14	37	5	465	99	75
/h/	0	23	26	252	59	70

Tableau 3.3: Résultats de la segmentation.

Commentaires :

Les remarques suivantes peuvent être déduites des résultats de la segmentation :

- Quand le phonème **/D./** est correctement articulé, il apparaît comme une consonne fricative si non il est segmenté comme une plosive.
- La consonne **/s./** est segmentée comme une fricative. Si le bruit présent possède une structure formantique, **/s./** est segmentée comme une sonnante.
- Quand le phonème **/A/** est correctement articulée, il apparaît comme une consonne plosive. Dans l'environnement intervocalique, **/A/** est segmentée comme une sonnante.
- Le phonème **/H/** est segmentée comme une fricative. Si le bruit présent possède une structure formantique, **/H/** est segmentée comme une sonnante.
- Les phonèmes **/h/** et **/E/** sont considérées comme des consonnes sonnantes avec une structure formantique. Les résultats de la segmentation confirment ce résultat.

3.4.4.2 Résultats de la reconnaissance

L'évaluation de la reconnaissance consiste, en appliquant les règles décrites en **annexe D**, à calculer le taux de reconnaissance phonétique du système par comparaison entre le résultat du décodage et la transcription correcte des phrases obtenues lors de l'étiquetage manuel. Les résultats de la reconnaissance sont résumés dans le **Tableau 3.4**.

Phonèmes	Présents	Reconnus (Taux en %)
/d./	340	79.1
/t./	320	85
/s./	300	87
/D./	210	78.1
/A/	420	70
/H/	520	83
/E/	620	68
/h/	360	69
Global	**3090**	**77.4**

Tableau 3.4: Résultats de la reconnaissance.

Commentaires :

Le taux global de reconnaissance du système est de 77.4%. Les phonèmes /s./, /t./ et /H/ sont les mieux reconnus. Ceci peut s'expliquer par la structure acoustique de ces consonnes qui ne présente pas une très grande variabilité. Les indices limite inférieure de friction et le centre de gravité du spectre sont très discriminants pour les consonnes /s./ et /H/ alors que c'est la barre d'explosion qui est discriminante pour la consonne /t./. Les taux les plus faibles sont obtenus pour les consonnes /A/, /E/ et /h/. Ces consonnes nécessitent donc à notre avis une bonne connaissance du contexte et la recherche d'autres indices plus discriminants.

3. 5 Conclusion

Nous avons présenté dans ce chapitre une étude phonétique de la langue arabe axée sur ces particularités ainsi que, l'approche analytique mise en œuvre pour l'identification en parole continue multi-locuteurs des consonnes arrières et emphatiques. Cette approche utilise des connaissances phonétiques et phonologiques déduites à partir d'une lecture de spectrogrammes numériques de ces consonnes prononcées dans les différents contextes phonétiques possibles. Pour acquérir ces connaissances, un logiciel interactif d'analyse graphique de la parole appelé APHAK a été mis au point. Celui-ci permet en effet grâce à ces différentes fonctions d'analyser très finement la parole. Le moteur de reconnaissance utilisé est un système expert qui gère les connaissances acquises formalisées à l'aide de règles de production (un ensemble de 76 a été déduit). Les tests de reconnaissance réalisés ont porté sur un corpus de phrases. Certains critères ont été fixés dans le choix de ces phrases tels la simplicité de la forme syntaxique, la longueur raisonnable des phrases et le contenu (faire apparaître les consonnes d'étude en nombre suffisant et dans les différents contextes possibles). Les règles utilisées sont pour la plupart contextuelles donc la connaissance du contexte, dans l'application d'une règle, est déterminante. Ces règles utilisent généralement le contexte vocalique, c'est ainsi que pour les tester nous avons eu à réaliser l'identification des voyelles. Cette identification était basée uniquement sur les formants vocaliques F1, F2 et F3 et la durée vocalique. Or, nous sommes persuadés que d'autres paramètres tels que la distance entre formants, le rapport entre formants,…. sont nécessaires pour une identification fiable des voyelles. Ces règles déduites ont permis néanmoins d'atteindre un score global d'identification de l'ordre de 77.4%. Nous pouvons espérer de meilleurs résultats de reconnaissance en affinant encore davantage la technique de segmentation et en introduisant plus de connaissances sur les

autres catégories de sons arabes en particulier sur les voyelles étant donné que la majorité des règles utilisent le contexte vocalique. Cependant, à travers cette étude, nous avons mesuré la difficulté à mettre en œuvre les systèmes à base de règles par le fait que l'expertise est rarement explicite, souvent loin d'être exhaustive et que l'exploitation des connaissances nécessite la gestion d'un grand nombre de seuils empiriques. Tout ceci compromet dans l'état actuel des recherches l'utilisabilité de cette approche.

Reconnaissance par MMC Performances et Paramètres des Modèles

4.1 Introduction

Ce chapitre est consacré à l'approche globale basée sur les Modèles de Markov Cachés (MMC) mise en œuvre. Il nous permettra d'exposer les différents outils théoriques utilisés au cours de la mise en œuvre de cette approche stochastique de reconnaissance de la parole en particulier la théorie de la modélisation Markovienne. Ainsi donc, nous allons présenter dans un premier temps, les algorithmes fondamentaux associés aux modèles de Markov cachés (MMC), ensuite le système de reconnaissance à quantification vectorielle conventionnelle développé. La section suivante présente les expériences comparatives réalisées et les différentes améliorations apportées à cette approche MMC.

4.2 Les modèles de Markov cachés

Les modèles de Markov cachés tentent de modéliser les unités représentatives de la parole par des modèles statistiques. Le formalisme des MMCs a permis des progrès importants en reconnaissance automatique de la parole ces dernières années. C'est ainsi que plusieurs types de modèles ont été développés et appliqués pour la reconnaissance automatique de la parole **[1]**, **[2]**. Les MMCs se sont avérés donc comme les modèles les mieux adaptés aux problèmes de reconnaissance et la quasi-totalité des outils de reconnaissance disponibles actuellement sur le marché sont basés sur cette technologie.
L'une des difficultés de la reconnaissance de la parole est sa segmentation en sous séquences stationnaires, pour leur reconnaissance. Les modèles de Markov cachés permettent d'appliquer le paradigme de segmentation-reconnaissance simultanée. Après reconnaissance complète, il est possible de faire une segmentation optimale du signal en zones stationnaires modélisées. C'est l'alignement temporel du signal sur une séquence de labels tels que les phonèmes, les mots, etc. Les MMCs permettent donc de modéliser par des états différents chaque sous-partie statistiquement stable de la séquence d'observations. C'est à dire que toutes les observations modélisées par chaque état sont représentées par des vecteurs de primitives se regroupant dans une certaine région de l'espace des observations. Les transitions du signal entre ces parties stables correspondent aux transitions entre les états du modèle de Markov caché. Les MMCs sont donc des modèles doublement stochastiques dont la première composante est un processus stochastique non observable, donc caché, mais qui peut l'être par l'intermédiaire d'un second processus stochastique. L'évolution dans le temps du premier processus produit une séquence d'états $Q = q_1, q_2, ..., q_T$. Le second processus permet d'observer l'évolution du modèle à travers la séquence d'observations qu'il émet : $X = x_1, x_2, ..., x_T$.

4.2.1 Présentation générale des MMCs

Un modèle de Markov caché est un automate doublement stochastique capable, après entraînement, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle. Idéalement, il faudrait pouvoir associer à chaque phrase un modèle. Il va de soi que ceci est irréalisable en pratique car le nombre de modèles serait beaucoup trop élevé. Des sous-unités lexicales comme le mot, la syllabe, ou le phonème sont utilisés afin de réduire le nombre de paramètres à entraîner. A chacune de ces unités est associé un modèle de Markov caché constitué d'un nombre fini d'état prédéterminé. Formellement, un modèle de Markov cachés peut être défini par les paramètres :

- Un ensemble d'états cachés $S = \{S_1, S_2, ..., S_N\}$.
- La matrice des probabilités de transitions sur l'ensemble des états du modèle $A = \{a_{ij}\}$, où a_{ij} est la probabilité de transiter de l'état S_i vers l'état S_j.

Pour un MMC d'ordre un, cette probabilité ne dépend que de l'état précédent :

$$a_{ij} = P(q_t = S_j \,|\, q_{t-1} = S_i)$$

Elle dépend des deux états précédents dans le cas d'un MMC d'ordre deux :

$$a_{ij} = P(q_t = S_j \,|\, q_{t-1} = S_i \,, q_{t-2} = S_k)$$

En d'autres termes, l'évolution du système entre deux instants *t-1* et *t* ne dépend que de l'état de ce système au temps *t*-1 (ordre 1) ou des deux instants précédents *t*-1 et *t*-2 (ordre deux).

- La matrice des probabilités d'émission $B = \{b_j(x_t)\}$, où $b_j(x_t) = P(x_t \mid q_t = S_j)$ est la probabilité d'observer x_t dans l'état S_j. La forme que prend cette distribution détermine le type du modèle MMC. C'est ainsi qu'on parle de MMC discrets, continus, semi-continus, etc.

- La distribution des probabilités initiales des états.

Les modèles de Markov cachés supposent que la séquence de vecteurs acoustiques représentatifs du signal de parole soit une succession de segments stationnaires. Ainsi, la parole est modélisée par une succession d'états, avec des transitions instantanées entre ces états. Chaque observation est supposée être une fonction probabiliste de l'état. En reconnaissance de la parole, des modèles de Markov gauche–droite d'ordre 1 sont les plus souvent utilisés du fait de l'aspect séquentiel du signal de parole. Un modèle de Markov d'ordre un à N états est donné sur la **Fig. 4.1**.

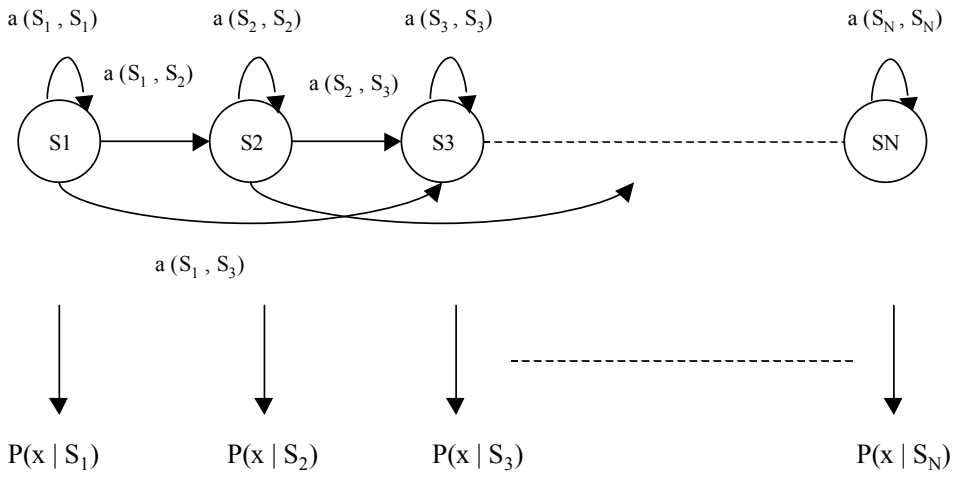


Fig. 4.1 – Modèle de Markov caché gauche-droite à N états

4.2.2 Types de Modèle

On distingue principalement trois types de modèles en fonction de la forme que prend la distribution des probabilités d'émission $P(x_t | q_t = S_i) = b_j(x_t)$: les modèles continus, les modèles discrets et les modèles semi-continus.

- **Les modèles continus :** utilisent des fonctions continues de densité de probabilité pour évaluer les probabilités d'observation directement dans l'espace des primitives. Chaque état modélise ses observations indépendamment des autres états du modèle, par une somme pondérée de fonctions élémentaires. La plus communément utilisée est la somme pondérée de gaussiennes multidimensionnelles (GMM : Gaussian Mixture Model pour l'anglais) :

$$b_j(x_t) = \sum_{m=1}^{M} c_{im} \; N(x_t \; ; \mu_{im} \; , \; \Sigma_{im}) \qquad \textbf{(4.1)}$$

  où μ_{im} et Σ_{im} sont respectivement le vecteur moyenne et la matrice de covariance de la $m^{ième}$ gaussienne de l'état i et c_{im} le coefficient de pondération qui lui est affecté.

- **Les modèles discrets :** les vraisemblances des observations $P(x_t | q_t = S_i) = b_j(x_t)$ sont estimées pour chaque état par des distributions discrètes de probabilité. Ceci nécessite que le nombre de formes soit limité. Les observations sont en général décrites par des vecteurs continus à n dimensions. On fait alors une partition de l'espace de représentation des formes en un nombre fini M de régions par exemple par une quantification vectorielle (QV). Cette étape de QV sera détaillée ultérieurement étant donné son importance dans la performance des systèmes de reconnaissance utilisants les modèles discrets.

- **Les modèles semi-continus :** dans ce cas l'espace des observations est modélisé par un unique jeu de gaussiennes. Ensuite, pour chaque état, les gaussiennes les plus probables sont sélectionnées et pondérées ; c'est cette somme pondérée des sorties des gaussiennes qui permet de calculer les probabilités des observations sachant l'état courant.

4.2.3 Les paramètres des MMCs

Dans le cadre de notre travail, nous avons utilisé des MMCs de premier ordre, discrets et avec émission des observations par les états du modèle. Le formalisme que nous allons présenter maintenant correspond à celui associé à de tels modèles. Pour un formalisme générique, de plus amples détails sur les MMCs sont présentés dans les références **[32]**, **[109]**.

La définition des MMCs fait appel à un certain nombre de variables et paramètres :

T La longueur de la séquence d'observations.

N Nombre d'états du modèle.

S $\{S_1, ..., S_N\}$, l'ensemble des états du modèle.

X $\{x_1, x_2, ..., x_T\}$, une séquence d'observations où x_t est l'observation à l'instant t.

Q $\{q_1, q_2, ..., q_T\}$, une séquence d'état du modèle où q_t est l'état du modèle à l'instant t.

M Le nombre de symboles d'observations possibles.

V $\{V_1, V_2, ..., V_M\}$, l'ensemble des symboles d'observations possibles.

A $\{a_{ij}\}$, la matrice des probabilités de transition sur l'ensemble des états du modèle où $a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i)$ est la probabilité de transiter de l'état S_i vers l'état S_j.

$$\text{avec} \quad a_{ij} \rangle 0 \quad \forall i, j \qquad et \qquad \sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$$

B $\{b_i(k)\}$, la matrice des probabilités d'émission où $b_j(k) = P(V_k \mid q_t = S_i)$ est la probabilité d'émission de l'observation V_k dans l'état S_j.

$$\text{avec} \quad b_j(k) \rangle 0 \quad \forall j, k \qquad et \qquad \sum_{k=1}^{M} bj(k) = 1 \quad \forall j$$

Π $\{\pi_i\}$, la distribution des probabilités initiales des états où $\pi_i = P(q_1 = S_i)$.

Le modèle sera défini par λ = [N, Π, A, B]. Avec des données réelles, trois problèmes fondamentaux doivent être résolus.

Problème 1 : Estimation des probabilités
Ce problème peut être formulé de la manière suivante : étant donné une séquence d'observation *X* et un modèle λ, comment calculer efficacement la vraisemblance de la séquence, c'est à dire *P(X | λ)* la probabilité d'observer la séquence *X* sachant le modèle λ ?

Problème 2 : Décodage ou reconnaissance
Etant donné une séquence d'observations X et un modèle λ, comment trouver la séquence d'états $Q = \{q_1, q_2, ..., q_T\}$ qui a le plus de chance de produire la séquence d'observation X ?

Problème 3 : Apprentissage des modèles
Etant donnés un ensemble de séquences d'observations et un modèle initial λ_0, comment ré-estimer les paramètres du modèle de manière à augmenter sa vraisemblance de génération de l'ensemble des séquences ?

Estimation des probabilités

Soient le modèle λ={N, A, B, Π}, $\{x_1, x_2, ..., x_T\}$ une séquence d'observations et $Q = \{q_1, q_2, ..., q_T\}$ une séquence d'états. La probabilité d'observer la séquence *X* pour une séquence d'états *Q* est :

$$P(X \mid Q, \lambda) = b_{q_1}(x_1).b_{q_2}(x_2).....b_{q_T}(x_T) \qquad \textbf{(4.2)}$$

Or, la probabilité de la séquence d'états Q peut s'écrire sous la forme suivante :

$$P(Q \mid \lambda) = \pi_{q_1}\ a_{q_1 q_2}\ a_{q_2 q_3}\ .....\ a_{q_{T-1} q_T} \qquad \textbf{(4.3)}$$

La probabilité conjointe du chemin Q et des observations X est :

$$P(X, Q \mid \lambda) = P(X \mid Q, \lambda).P(Q \mid \lambda) \qquad \textbf{(4.4)}$$

La probabilité de la séquence d'observations X sachant le modèle λ est obtenue en sommant la probabilité $P(X, Q \mid \lambda)$ sur toutes les séquences d'états possibles. Ainsi la probabilité d'émission des observations est :

$$P(X \mid \lambda) = \sum_{Q} P(X, Q \mid \lambda) = \sum_{Q} P(X \mid Q, \lambda)\ .\ P(Q \mid \lambda) \qquad \textbf{(4.5)}$$

$$P(X \mid \lambda) = \sum_{q_1 q_2 .....q_T} \pi_{q_1} b_{q_1}(x_1)\, a_{q_1 q_2} b_{q_2}(x_2) a_{q_2 q_3}\ .....\ a_{q_{T-1} q_T} b_{q_T}(x_T) \qquad \textbf{(4.6)}$$

Le calcul direct de cette probabilité consiste donc à énumérer toutes les séquences de longueur T possibles. Cependant cette technique nécessite un grand nombre de calculs (de l'ordre de $2TN^T$ multiplications et N^T-1 additions) ce qui la rend trop complexe et impossible à implémenter. Une procédure appelée Forward **[32]** permet, grâce à une factorisation des chemins, de réduire considérablement le nombre d'opérations et de réaliser efficacement le calcul de cette probabilité.

Considérons la variable Forward $\alpha_t(i)$ définie par :

$$\alpha_t(i) = P(x_1, x_2, ...x_t\ , q_t\ = S_i \mid \lambda) \qquad \textbf{(4.7)}$$

C'est-à-dire la probabilité d'observer la séquence partielle $x_1, x_2, \ldots\ldots, x_T$ et d'être à l'état S_i à l'instant t sachant le modèle λ.

Cette probabilité peut être obtenue de manière itérative :

Initialisation : $\alpha_1(i) = \pi_i\, b_i(x_1) \qquad 1 \le i \le N$

Induction : $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(x_{t+1}), \quad t = 1,2,.....,T-1 \quad et \quad 1 \le i \le N$

Terminaison : $P(X \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

De la même façon, on peut définir la fonction de retour (Backward) : $\beta_t(j)$, la probabilité d'observer la séquence partielle x_{t+1}, x_{t+2}, ……, x_T sachant qu'on est à l'état i à l'instant t et qu'on a le modèle λ.

$$\beta_t(j) = P(x_{t+1}, x_{t+2}, ..., x_T \mid q_t = S_j, \lambda) \qquad \textbf{(4.8)}$$

De la même manière cette probabilité peut être calculée d'une manière récursive :

Initialisation :	$\beta_T(i)=1 \qquad 1 \le i \le N$
Induction :	$\beta_t(i)=\sum_{j=1}^{N} a_{ij} b_j(x_{t+1}) \beta_{t+1}(j), \quad t=T-1,T-2,.....,1 \quad et \quad 1 \le i \le N$
Terminaison :	$P(X \mid \lambda)=\sum_{i=1}^{N} \pi_i b_i(x_1) \beta_1(i)=\sum_{i=1}^{N} \alpha_t(i) \beta_t(i)$

Décodage

Le problème de décodage ou de reconnaissance est généralement résolu par l'intermédiaire de l'algorithme de Viterbi **[35], [36]**. C'est une approximation de la fonction Forward, qui calcule la probabilité du meilleur chemin à la place de la somme sur tous les chemins. L'optimisation globale de la recherche du meilleur chemin est basée sur le principe de la décomposition en une succession d'optimisations locales, principe utilisé dans tous les algorithmes de programmation dynamique. Nous cherchons donc le meilleur chemin caractérisé par la séquence d'états $Q=\{q_1, q_2, ..., q_t\}$ qui peut produire la séquence d'observations $X=\{x_1, x_2, ..., x_t\}$. Pour cela définissons la variable $\delta_t(i)$:

$$\delta_t(i) = \max_{q_1\ q_2\ ....q_{t-1}} P(q_1\ q_2\ .....q_t = S_i,\ x_1\ x_2\ .....x_t \mid \lambda) \qquad \textbf{(4.9)}$$

C'est-à-dire la probabilité du meilleur chemin aboutissant à l'état *Si* à l'instant *t*, étant donné le modèle λ.
On peut déterminer les $\delta_t(i)$ de façon itérative. On a en effet :

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i).a_{ij} \right] b_j(x_{t+1}) \qquad \textbf{(4.10)}$$

En conservant ce max. pour chaque t et chaque i on obtient l'algorithme suivant :

Initialisation :

$$\delta_1(i) = \pi_i \, b_i(x_1) \qquad 1 \le i \le N$$
$$\phi_1(i) = 0$$

Récursion :

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) \, a_{ij} \right] b_j(x_t) \qquad 2 \le t \le T \qquad 1 \le j \le N$$

$$\phi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right]$$

Fin :

$$P^* = \max_{1 \le i \le N} \delta_T(i)$$

$$q_T^* = \arg\max_{1 \le i \le N} \delta_T(i)$$

La séquence d'états optimale :

$$q_t^* = \phi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2, \ldots\ldots\ldots, 1$$

Lors du déroulement de l'algorithme, nous devons conserver la trace du meilleur chemin. Pour cela, une fonction particulière est utilisée $\phi_t(i)$ qui conserve l'état correspondant à la meilleure probabilité $\delta_t(i)$ à chaque instant t. Cet alignement temporel est très utile pour initialiser des coefficients ou donner des informations sur le fonctionnement interne du système.
L'algorithme de Viterbi est similaire à la procédure Forward. La seule différence se situe au niveau de la récursion, où cet algorithme effectue une maximisation au lieu d'une somme.
Du point de vue implémentation, il est à noter qu'il est possible de passer en logarithme et ainsi effectuer des additions de probabilités au lieu de multiplications. Le passage au logarithme des probabilités est également réalisé pour une raison de représentation numérique. En effet, la multiplication d'un grand nombre de probabilités conduit à des valeurs numériques faibles, atteignant parfois la précision de certaines machines.

Apprentissage des modèles

Le problème de l'apprentissage des modèles est celui de l'estimation des paramètres à partir d'un ensemble de séquences d'observations. Il peut être réalisé par l'intermédiaire de l'algorithme de Baum-Welch **[78].**

- **Algorithme de Baum-Welch (EM)**

Il s'agit d'une adaptation de l'algorithme EM (Expectation-Maximization) **[108]** qui garantit la convergence vers un maximum local de la probabilité d'observation de l'ensemble des exemples

d'apprentissage, au sens du critère de maximum de vraisemblance (MLE, Maximum Likelihood Estimate)

Pour décrire cet algorithme d'estimation des paramètres du modèle, on définit :

- la variable $\varepsilon_t(i,j)$ qui représente la probabilité que le modèle soit à l'état i à l'instant t et à l'état j à l'instant t+1 étant donnée la séquence d'observation X et le modèle λ

$$\varepsilon_t(i,j) = P(q_t = S_i\ ,\ q_{t+1} = S_j \mid X,\ \lambda) \qquad \textbf{(4.11)}$$

- la variable $\gamma_t(i)$ qui représente la probabilité d'être à l'état i à l'instant t étant donnés la séquence X et le modèle λ.

$$\gamma_t(i) = P(q_t = S_i \mid X,\ \lambda) \qquad \textbf{(4.12)}$$

Après développement, ces quantités peuvent être exprimées en fonction des variables Forward et Backward (voir les équations **4.7** et **4.8**) de la manière suivante :

$$\varepsilon_t(i,j) = \frac{\alpha_t(i)\ a_{ij}\ b_j(X_{t+1})\ \beta_{t+1}(j)}{P(X \mid \lambda)} \qquad \textbf{(4.13)}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \varepsilon_t(i,j) = \frac{\alpha_t(i)\,\beta_t(i)}{P(X \mid \lambda)} \qquad \textbf{(4.14)}$$

Maintenant si nous effectuons une somme de la variable $\gamma_t(i)$ pour tous les instants t, nous obtenons une quantité qui peut être interprétée comme le nombre de fois que l'état S_i a été visité. De même, la somme sur t de la variable $\varepsilon_t(i,j)$ peut être interpréter comme le nombre de fois que le modèle a transité de l'état S_i vers l'état S_j.

L'algorithme de Baum-Welch permet d'estimer les nouveaux paramètres du modèle comme suit :

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad 1 \le i \le N, \quad 1 \le j \le N \qquad \textbf{(4.15)}$$

$$\hat{b}_j(k) = \frac{\sum_{t=1,\ x_t = V_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}, \qquad 1 \le j \le N, \quad 1 \le k \le M \qquad \textbf{(4.16)}$$

$$\hat{\pi}_i = \gamma_1(i), \qquad 1 \le i \le N \qquad \textbf{(4.17)}$$

Afin d'obtenir une estimation fiable des différents paramètres, l'apprentissage des modèles doit se faire à l'aide d'un grand nombre de séquences d'observations. Cette considération conduit à une modification des formules de ré-estimations.

Considérons un ensemble de L séquences d'observations : $X = \{X^1, X^2, ..., X^L\}$, où $X^1 = \{x_1^1, x_2^1, ..., x_{T_1}^1\}$. Nous supposons que chacune des séquences d'observations est indépendante. Le but de l'apprentissage consiste dans ce cas à ajuster les paramètres du modèle λ de manière à maximiser la probabilité :

$$P(X|\lambda) = \prod_{l=1}^{L} P(X^l \,|\, \lambda) \qquad \textbf{(4.18)}$$

Comme les estimations sont basées sur les fréquences d'occurrences des différentes observations, les formules de ré-estimations sont modifiées en prenant en compte les fréquences individuelles de chaque séquence.

$$\bar{a}_{ij} = \frac{\sum_{l=1}^{L} \frac{1}{P(X^l|\lambda)} \sum_{t=1}^{T_l - 1} \alpha_t^l(i)\, a_{ij}\, b_j(X_{t+1}^l)\, \beta_{t+1}^l(j)}{\sum_{l=1}^{L} \frac{1}{P(X^l|\lambda)} \sum_{t=1}^{T_l - 1} \alpha_t^l(i)\, \beta_t^l(j)} \qquad \textbf{(4.19)}$$

$$\bar{b}_j(k) = \frac{\sum_{l=1}^{L} \frac{1}{P(X^l|\lambda)} \sum_{t=1,\ x_t = V_k}^{T} \alpha_t^l(j)\, \beta_t^l(j)}{\sum_{l=1}^{L} \frac{1}{P(X^l|\lambda)} \sum_{t=1}^{T_l - 1} \alpha_t^l(j)\, \beta_t^l(j)} \qquad \textbf{(4.20)}$$

4.3 Structure générale d'un système de reconnaissance par MMC

La structure générale d'un système de reconnaissance par MMC peut être schématisée par la **Fig. 4.2** suivante. Dans cette structure, on distingue deux phases : une phase d'apprentissage dont le but est la construction des modèles acoustiques (les modèles MMC) et une phase de reconnaissance qui, après extraction des paramètres acoustiques, fournit en sortie la phrase la plus probable étant donnés la grammaire imposée et les modèles MMC ou le mot le plus probable étant donné le lexique et les modèles MMC.

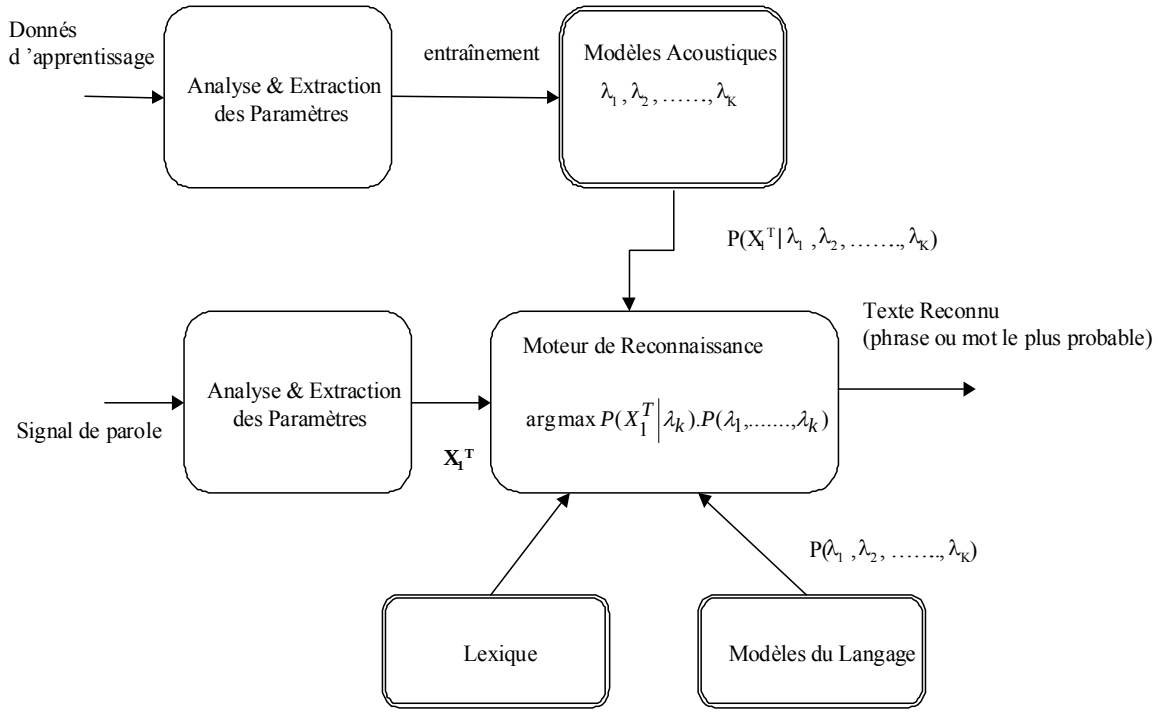


Fig. 4.2 – Structure générale d'un système de reconnaissance par MMC.

Le signal de parole inconnu est d'abord convertit en une séquence de vecteurs acoustiques $X_1^T = x_1, x_2, \ldots\ldots, x_T$ par le module analyse et extraction des paramètres (les différentes analyses et paramètres acoustiques utilisés seront décrits plus loin dans ce chapitre). La reconnaissance de cette séquence s'effectue en recherchant la suite d'unités (l'unité peut être un mot, un phonème, une syllabe, etc.) la plus probable $\lambda_{optimal}$ dans toutes les séquences d'unités possibles $\lambda_1^k = \lambda_1, \lambda_2, \ldots\ldots, \lambda_k$ (k étant inconnu) connaissant la séquence d'observation X.

$$\lambda_{optimal} = \arg\max_{\lambda_1^k} \left[P(\lambda_1^k \middle| X) \right] \quad \textbf{(4.21)}$$

Cette probabilité $P(\lambda_1^k \middle| X)$ appelée probabilité a posteriori n'est malheureusement pas accessible directement par le processus d'entraînement des modèles. En appliquant la règle de Bayes, il est possible de lier cette probabilité à la vraisemblance $P(X \middle| \lambda_1^k)$ par la relation donnée par l'équation **4.22** suivante :

$$\lambda_{optimal} = \arg\max_{\lambda_1^k} \left[\frac{P(X \middle| \lambda_1^k).P(\lambda_1^k)}{P(X)} \right] \quad \textbf{(4.22)}$$

Comme *P(X),* probabilité a priori de la séquence d'observations X, est constante sur la séquence λ_1^k alors :

$$\lambda_{optimal} = \arg\max_{\lambda_1^k} \left[P(X \middle| \lambda_1^k)\ P(\lambda_1^k) \right] \quad \textbf{(4.23)}$$

Ainsi la reconnaissance revient à trouver la séquence $\lambda_{optimal}$ qui maximise le produit $P(X \middle| \lambda_1^k).P(\lambda_1^k)$.

- ✓ **La probabilité $P(\lambda_1^k)$** : la probabilité a priori du modèle est déterminée par le module «modèles du Langage ». Dans ce modèle sont spécifiées les contraintes imposées sur les suites d'unités prononcées. Cette probabilité est généralement prise égale au produit des probabilités conditionnelles à savoir :

$$P(\lambda_1^k) = \prod_{n=1}^{k} P(\lambda_k \middle| \lambda_{k-1}, ..., \lambda_1) \quad \textbf{(4.24)}$$

Pour diminuer la complexité du modèle l'historique des probabilités conditionnelles est limité à un nombre N d'unités, $P(\lambda_1^k)$ s'écrit alors $P(\lambda_1^k) = P(\lambda_k \middle| \lambda_{k-1}, ..., \lambda_{k-N})$. Le modèle de langage est dans ce cas appelé modèle N-gramme. Pratiquement N est pris égale à 2 ou à 3, ainsi un modèle de langage bigramme (respectivement trigramme) permet d'obtenir la probabilité d'une unité étant donné l'unité (respectivement, les deux unités) précédente(s). Un bon modèle du langage permet d'augmenter le taux de reconnaissance et une réduction de la complexité de la procédure de recherche.

- ✓ **La probabilité P(X | λ_1^k)** : la vraisemblance de la séquence étant donnée la suite λ_1^k est déterminée par le module "Modèles Acoustiques". Comparativement à la procédure de calcul de la probabilité $P(\lambda_1^k)$, le calcul de la vraisemblance $P(X \mid \lambda_1^k)$ est plus compliqué, des méthodes de calcul efficaces sont impératives. Ces méthodes sont l'algorithme Forward (Backward) ou l'algorithme de Viterbi décrits précédemment.

4.4 Système MMC à QV conventionnelle

Après avoir introduit les différents outils de la modélisation Markovienne, nous présentons le système de reconnaissance utilisant les Modèles de Markov Cachés combinés à une Quantification Vectorielle Conventionnelle (MMC/QVC) mis en œuvre. La **Fig. 4.3** montre le synoptique de ce système. On distingue deux phases dans le processus de fonctionnement de ce système : une phase d'apprentissage permettant la construction du dictionnaire de référence (Dictionnaire) et l'estimation des paramètres des modèles (Modèles acoustiques) et une phase de reconnaissance dans laquelle les unités à reconnaître (dans notre cas les phonèmes) sont identifiées.

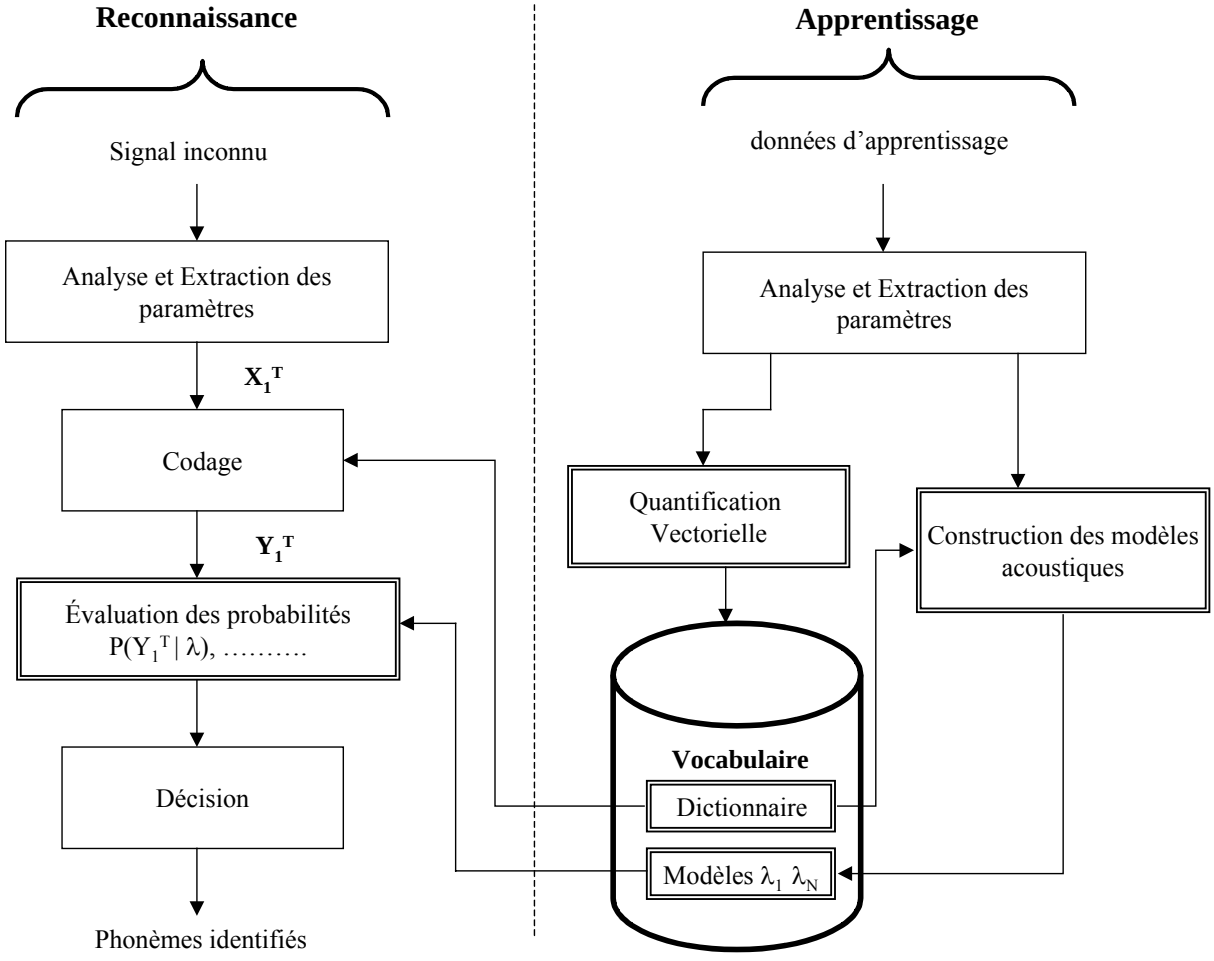


Fig. 4.3 - Schéma du système de reconnaissance MMC/QVC.

4.4.1 Analyse du signal et extraction des paramètres

Le module d'analyse et d'extraction des paramètres acoustiques permet une représentation moins redondante de la parole tout en gardant l'information utile contenue dans le signal de parole. L'analyse commence par une mise en forme du signal de parole, le signal est tout d'abord filtré puis échantillonné à une fréquence donnée Fe (dans notre cas Fe est égale à 10KHz), une pré-accentuation est ensuite effectuée afin de relever les hautes fréquences, le signal est enfin fragmenté en trames par l'utilisation d'une fenêtre de Hamming de 25.6ms glissante avec un déplacement de 10ms. Pour chaque trame (chaque fenêtre d'analyse) sont alors calculés : un nombre P de coefficients LPC (linear Predictif Coding) par l'algorithme de Levinson-Durbin **[79]** et à partir de ces coefficients LPC, un nombre Q de coefficients cesptraux (Q > P) par la procédure récursive suivante :

$$\ln\left[\frac{1}{A_p(z)}\right]=\sum_{n=1}^{\infty}C_q(n)\,z^{-n} \qquad \textbf{(4. 25)}$$

$C_0=\ln\sigma^2$, σ^2 est le gain du modèle LPC **(4.26)**

$$C_m=a_m+\sum_{k=1}^{m-1}\left(\frac{k}{m}\right)C_k\,a_{m-k} \quad \text{, pour } 1 \le m \le P \qquad \textbf{(4.27)}$$

$$C_m=\sum_{k=1}^{m-k}\left(\frac{k}{m}\right)C_k\,a_{m-k} \quad , \qquad \text{pour } m > P \qquad \textbf{(4.28)}$$

Les coefficients cepstraux de bas ordre sont sensibles à la pente spectrale globale et les coefficients cepstraux d'ordre supérieur sont sensibles au bruit. Ainsi, il est devenu une technique standard de pondérer les coefficients cepstraux par une fenêtre conique (équation **4.29**) afin de réduire au minimum ces sensibilités.

Pour $1 \le m \le Q$

$$W(m)=1+\frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right) \qquad \textbf{(4.29)}$$

$$Cep_m=C_m\,W(m) \qquad \textbf{(4.30)}$$

Les dérivées cepstrales sont ensuite calculées suivant la relation :

$$\Delta Cep_{pl}(m)=G\,x\sum_{k=-K}^{K}k\,Cep_{p\,(l-k)}(m) \qquad \textbf{(4.31)}$$

$$E\ =20Log_{10}\left(\sum_{i=1}^{N}x_i^2\right) \qquad \textbf{(4.32)}$$

Dans notre travail, K (nombre de trames de part et d'autre de la trame courante) est pris égal à 2 et le gain G est égal à 0.38. Pour chaque trame, le vecteur acoustique x_t est donc constitué des coefficients cepstraux, ses dérivées et de l'énergie (équation **4.32**).

$$x_t = \left\{ Cep_p(m), \Delta Cep_p(m), E \right\} \qquad \textbf{(4.33)}$$

4.4.2 Quantification vectorielle conventionnelle

Les MMC utilisés pour la modélisation des phonèmes sont de nature discrète, i.e., leurs densités de probabilités d'observations sont discrètes, ce qui nécessite l'utilisation d'un quantificateur vectoriel pour faire correspondre chaque vecteur acoustique continu (représentant une trame) à un indice discret d'un dictionnaire de référence (CodeBook). Une fois le dictionnaire de référence obtenu, cette correspondance entre les vecteurs caractéristiques des trames et les indices du dictionnaire devient un simple calcul de type plus proche voisin. Ainsi, le point essentiel dans la procédure de quantification vectorielle est la conception d'un dictionnaire de référence approprié pour la quantification. Divers travaux ont été effectués dans ce sens pour développer des procédures itératives pour la conception des dictionnaires. Dans notre travail nous avons utilisé pour la construction du dictionnaire de référence l'algorithme des K-moyennes dans sa version LBG (Linde-Buzo-Gray). Dans ce qui suit, une description de la procédure de QV ainsi que l'algorithme utilisé sont présentés.
Considérons $x = (x_1, x_2, \ldots\ldots, x_N)^t$ un vecteur de dimension N, dont toutes les composantes (x_i , $1 \le i \le N$) sont des variables aléatoires à valeurs réelles et amplitude continue. La QV consiste à faire correspondre au vecteur x un nouveau vecteur y, de même dimension N, à valeurs réelles et d'amplitude discrète (ou continue). En notant *quant(.)* l'opérateur de quantification, nous pouvons écrire :

$$y = quant(x) \qquad \textbf{(4.34)}$$

Le vecteur résultant y prend sa valeur parmi un ensemble fini de vecteurs prédéterminés $Y = (y_i , 1 \le i \le L)$ où $y_i = (y_{i1}, y_{i2}, \ldots\ldots, y_{iN})^t$. L'ensemble Y est appelé dictionnaire (CodeBook) et L est la taille de ce dictionnaire. La construction du dictionnaire est obtenue en divisant l'espace à N dimension du vecteur aléatoire x en L régions et en associant à chacune des régions R_i, un vecteur y_i. La quantification consiste alors à attribuer au vecteur x la valeur y_i suivant sa position dans l'espace :

$$quant(x) = y_i \qquad Si \quad x \in R_i \qquad \textbf{(4.35)}$$

Une telle opération entraîne naturellement une perte d'information qui est appelée généralement erreur de quantification. Une mesure de distorsion d(x , y) est alors utilisée afin de quantifier cette perte. Du fait que cette mesure peut être assimilée à une distance, il existe un grand nombre de possibilités lors de sa mise en œuvre **[111].** La mesure de distorsion la plus utilisée est celle des moindres carrés :

$$d(x, y) = \frac{1}{N} \sum_{j=1}^{N} (x_j - y_j)^2 \qquad \textbf{(4.36)}$$

L'utilisation de cette mesure de distorsion suppose que chaque composante des vecteurs à la même importance. Il est possible pour une application donnée que cette contrainte ne soit pas satisfaisante, en particulier si les différentes composantes du vecteur ont des moyennes et/ou des variances hétérogènes.

Il est alors conseillé de centrer et réduire les données et d'utiliser l'inverse de la matrice de covariance des données Σ pour pondérer les différentes composantes :

$$d(x,y) = (x-y)^t \sum^{-1} (x-y) \qquad \textbf{(4.37)}$$

Cette mesure de distorsion est connue sous le nom de distance de Mahalanobis.
Le processus de quantification est optimal lorsque la distorsion globale *D*, c'est-à-dire pour les L classes, est minimale. Cette dernière peut être définie par :

$$D = E\left[d(x,y)\right] = \sum_{i=1}^{N_s} D_i \qquad \textbf{(4.38)}$$

où E[.] désigne l'espérance mathématique et D_i la distorsion moyenne pour la région R_i. Si la distorsion globale *D* est utilisée comme critère lors de la définition de l'alphabet *Y*, deux conditions sont nécessaires à l'optimalité de la solution. La première est que le processus de quantification doit être réalisé en utilisant la loi de sélection du plus proche voisin. Cette dernière s'écrit :

$$q(x) = y_i \ , \quad si\ et\ seulemnt\ \ si \quad d(x,y_i) \le d(x,z_j) \qquad \textbf{(4.39)}$$

$$\text{pour } j \neq i \quad \text{et} \quad 1 \le j \le N_s$$

Cela signifie que le processus de quantification choisit le vecteur y_j qui conduit à la plus petite distorsion étant donné le vecteur x. La deuxième condition nécessaire concerne l'obtention des vecteurs y_i de l'alphabet Y. Ces derniers doivent être choisis de manière à minimiser la distorsion moyenne de chaque région R_i.

Considérons un ensemble de N_v vecteurs d'apprentissage (x_k , $1 \le k \le N_v$) et N_{vi} le nombre de ces vecteurs contenu dans la région R_i. La distorsion moyenne associée à chaque région est définie de la manière suivante :

$$D_i = \frac{1}{N_{vi}} \sum_{x \in R_i} d(x,y_i) \qquad \textbf{(4.40)}$$

Le vecteur qui minimise la distorsion au sein de la région R_i est appelée centre de gravité (Cdg) de la région R_i et est noté :

$$y_i = Cdg(R_i) \qquad \textbf{(4.41)}$$

La technique la plus utilisée pour minimiser itérativement la mesure de la distorsion moyenne est l'algorithme des K-moyennes (K-means, pour l'anglais) **[80]**, **[81].** L'idée de base est de diviser

l'ensemble des vecteurs d'apprentissage en *L* régions R_i de telle manière que les deux conditions nécessaires décrites précédemment soient respectées.

Algorithme des K-moyennes

étape 1 : Choisir un ensemble de vecteurs initiaux y_i, $1 \le i \le N_s$.

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

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

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

Cette procédure divise la tâche de minimisation de la distorsion globale en deux étapes. La première suppose que les centres de gravité soient connus, elle réparti les échantillons d'apprentissage parmi les L régions possibles en fonction de la mesure de distorsion. La seconde suppose que les régions où classes sont connues et calcule la nouvelle position du centre de gravité qui minimise la distorsion intra-classe. Il a été démontré que cet algorithme ne conduit qu'à un minimum local de la fonction de distorsion **[82]**. Une solution optimale peut être approximée en appliquant l'algorithme plusieurs fois avec des valeurs initiales différentes et en choisissant l'alphabet qui conduit à la plus faible distorsion.

Algorithme LBG

Un autre algorithme utilisé pour la QV est celui de Linde, Buzo et Gray appelé LBG **[82].** C'est une version étendue de l'algorithme K-moyennes. Il permet de résoudre l'un des problèmes associés à l'algorithme des K-moyennes, celui de l'initialisation des centres de gravité. En fait, l'algorithme LBG divise itérativement l'ensemble des échantillons d'apprentissage en 2, 4, …, 2^P régions, et calcule le centre de gravité de chacune des régions.

étape 1 : On attribue la valeur 1 à N_s (nombre de classes ou régions) et on Calcule le centre de gravité y_1 associé à l'ensemble des échantillons d'apprentissage.

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

étape 3 : On perturbe à nouveau les deux Cdg (CodeWords) obtenus pour obtenir un dictionnaire initial de 4 éléments et on relance l'algorithme des K-moyennes …….

étape 4 : On arrête l'algorithme lorsque l'on a un dictionnaire de taille M souhaitée.

Il faut signaler que la perturbation des CodeWords peut conduire parfois à des classes vides qui peuvent se multiplier au cours des itérations. Pour éviter cela, après chaque perturbation, il faut vérifier la présence éventuelle d'une classe vide et refaire la perturbation.

4.4.3 Estimation des modèles

Les unités à reconnaître sont modélisées par des sources de Bakis, modèle gauche-droite stationnaire d'ordre 1 à trois états émetteurs (**Fig. 4.4**). L'apprentissage des modèles permet d'associer à chaque unité un modèle optimisé. L'estimation des paramètres des modèles optimisés est faite en utilisant l'algorithme de Baum-Welch.

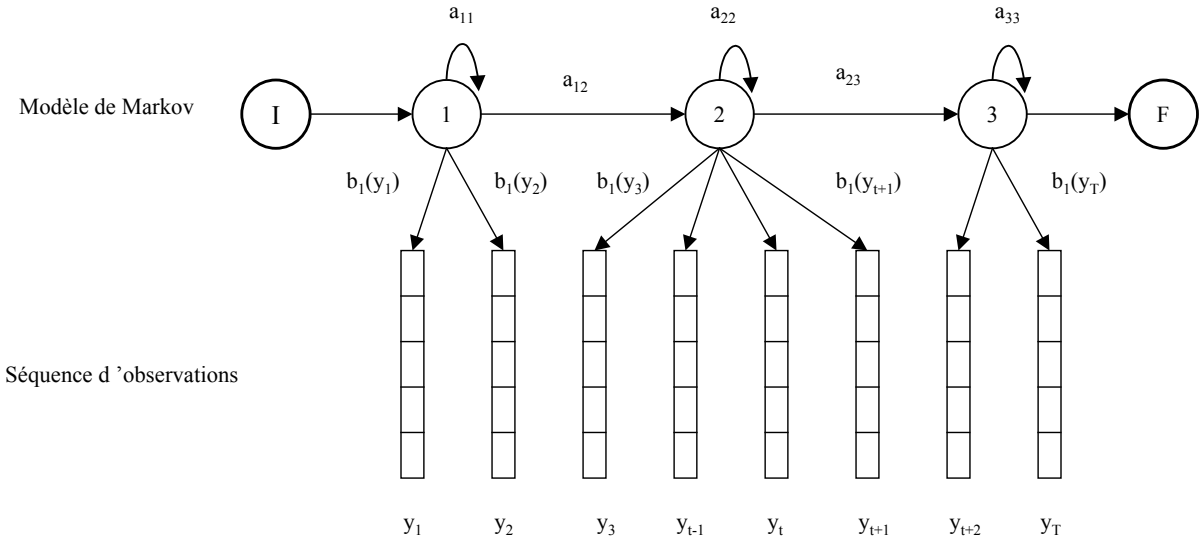


Fig. 4.4 – Modèle gauche-droite d'ordre un à trois états émetteurs.

4.4.4 Reconnaissance

Pour chaque unité à reconnaître le système calcule la vraisemblance pour tous les modèles et sélectionne le modèle possédant la plus grande vraisemblance. Le calcul de la vraisemblance plus précisément le logarithme de la vraisemblance $\alpha\ (t,i)$ est réalisé en utilisant l'algorithme de Viterbi **[35]**. Cette vraisemblance est dans ce cas donnée par l'équation suivante :

$$\alpha\ (t,i)\ =\ \max_{i}\left[\alpha(t-1,j)\ +\ \log a_{ij}\ \right]\ +\ \log b_i\ (y_t\ ) \qquad \textbf{(4. 42)}$$

avec $\quad 1 \le t \le T \quad$ et $\quad 1 \le i\ ,\ j \le N$

4.4.5 Expériences et résultats comparatifs

Des expériences de reconnaissance des phonèmes arabes, les consonnes arrières et les consonnes emphatiques, ont été réalisées. Le but est d'évaluer le système MMC à QV conventionnelle et de comparer les résultats obtenus avec ceux de l'approche analytique système expert mise en œuvre dans le chapitre 3. A cet effet, le corpus de mots (**annexe C**) répétés cinq fois par 20 locuteurs dans des

conditions semblables a été utilisé pour l'apprentissage des modèles et le corpus de phrases (**annexe D**) pour la reconnaissance. Le vecteur acoustique utilisé est de la forme $x_t = \{Cep_p(12), \Delta Cep_p(12), E\}$ et le dictionnaire de référence généré est de taille égale à 128. Le **Tableau 4.1** donne les résultats obtenus avec le système MMC à QV conventionnelle comparés avec ceux de l'approche analytique.

Phonèmes	**Système Expert T.R en (%)**	**MMC à QV conventionnelle T.R en (%)**
/d./	79	86
/t./	85	90
/s./	87	87
/D./	78.1	79
/A/	70	81
/H/	83	86
/E/	68	80
/h/	69	84
global	**77.4**	**84.1**

Tableau 4.1: Comparaison des résultats de reconnaissance.

Commentaires

L'analyse des résultats obtenus montre que nous avons une amélioration des taux de reconnaissance avec l'approche statistique basée sur les modèles de Markov cachés à quantification vectorielle conventionnelle.

Conclusion

Nous avons présenté les résultats de reconnaissance des consonnes emphatiques et arrières de l'arabe par deux systèmes utilisant des approches très différentes : le premier basé sur des règles acoustico-phonétiques gérées par un système expert et le second basé sur une approche statistique utilisant les modèles phonémiques de Markov discrets. Le choix de ces consonnes est motivé par le fait qu'en plus qu'elles soient spécifiques à la langue arabe, ces consonnes se caractérisent par un aspect spectral très variable en fonction du locuteur, du mode d'élocution et de l'environnement phonétique d'où la difficulté de traduire aisément l'ensemble de cette variabilité par des règles acoustico-phonétiques. Dans le cas du système expert, un ensemble de 76 règles a été déduit après un long travail d'observation et d'analyse spectrographique. Mais, au regard des résultats obtenus, nous pouvons déduire que l'approche stochastique MMC en plus qu'elle soit moins coûteuse s'affranchit bien de ces contraintes et elle est mieux adaptée pour l'identification des phénomènes de forte distorsion temporelle et spectrale telle que la parole.

4.5 Performance et paramètres des modèles

Notre objectif à travers cette étude est de discuter l'influence que peuvent avoir les paramètres des modèles MMC sur la performance globale du système de reconnaissance. En effet, nous pensons que cette discussion est privilégiée dans la conception des systèmes de reconnaissance au lieu d'avoir recours à un enrichissement du corpus d'apprentissage pour compenser toute insuffisance au niveau de l'architecture du modèle. Cette assertion est encore plus vraie dans le cas de la langue arabe. Nous traiterons dans cette étude : la résolution acoustique (nature du vecteur acoustique), l'approche multi-variables du vecteur acoustique, l'approche multi-dictionnaires qui assigne un dictionnaire spécifique à chaque composante du vecteur acoustique, le type de modèles utilisés à savoir les modèles indépendants et dépendants du contexte et enfin les modèles à dictionnaires multiples.

4.5.1 Résolution acoustique

Dans les expériences précédemment réalisées, nous avons utilisé un vecteur acoustique x_t de la forme $x_t = \{Cep_p(12), \Delta Cep_p(12), E\}$ avec un dictionnaire de taille égale à 128. Afin d'étudier l'influence de la résolution acoustique sur la performance du système nous calculons à la place des coefficients cepstraux pondérés dérivés des coefficients LPC, les coefficients cepstraux dans l'échelle Mel à savoir les coefficients MFCC (*MFCC : Mel Frequency Cepstral Coefficient*) et nous élargissons le vecteur acoustique par l'introduction de l'information dynamique portée par la vitesse *(ΔMFCC* : dérivée d'ordre 1) et par l'accélération *(ΔΔMFCC* : dérivée d'ordre 2).

Les coefficients MFCC

Les coefficients MFCC **[83]** sont basés sur l'estimation de l'enveloppe spectrale dans une échelle perceptive. Les échelles perceptives les plus utilisées sont l'échelle Mel ou l'échelle Bark. Dans notre cas nous avons fait le choix d'utiliser l'échelle Mel. Celle-ci peut être définie par la relation suivante entre la fréquence en Hertz et sa correspondance en Mel :

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \qquad \textbf{(4.43)}$$

Le processus de calcul des coefficients MFCC peut être décrit par les étapes du schéma de la **Fig. 4.5**. On applique d'abord une transformée de Fourier discrète (DFT : Discret Fourier Transform), en particulier FFT (Fast Fourier Transform) pour passer dans le domaine fréquentiel et extraire le spectre du signal. Ensuite, un filtrage est effectué en multipliant le spectre obtenu par les gabarits des filtres répartis linéairement sur l'échelle Mel. Ces filtres sont en général de forme triangulaire ou sinusoïdale. Dan notre travail nous avons choisi d'utiliser des filtres triangulaires. Les sorties du banc de filtres (énergies) subissent alors une analyse homomorphique par l'application de la transformée en cosinus discrète (DCT : Discret Cosinus Transform) aux valeurs logarithmiques des énergies. Cette analyse homomorphique a pour effet de rendre les coefficients obtenus plus discriminants, plus robustes au bruit ambiant et moins corrélés entre eux. La formule de la transformée en cosinus discrète est la suivante :

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi_i}{N}(j-0.5)\right) \quad , \quad i = 1 , \ldots\ldots\ldots, P \qquad \textbf{(4.44)}$$

où :
m_j : valeurs logarithmiques des énergies à la sortie des filtres, N : le nombre de filtres et P : le nombre de coefficient MFCC.

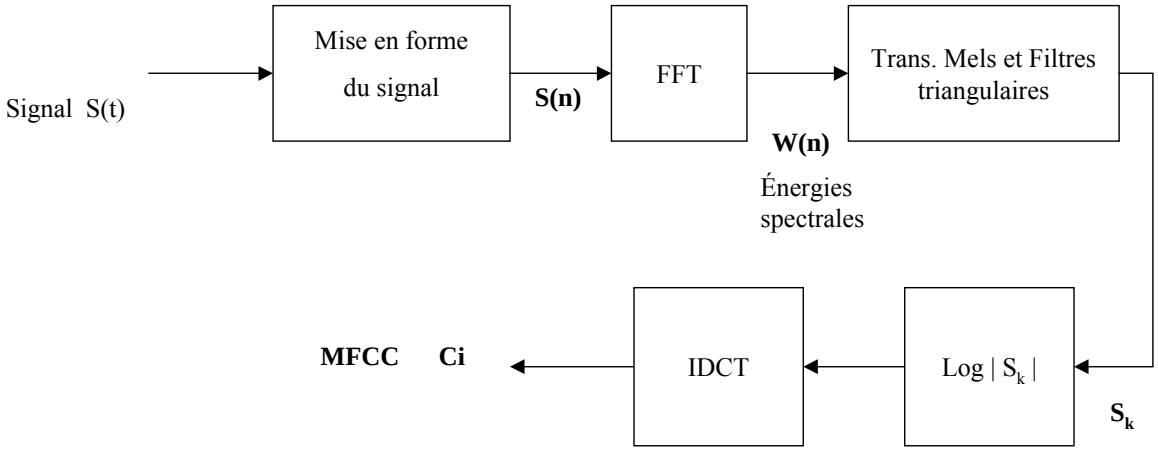


Fig. 4.5 – Processus de calcul des coefficients MFCC

Les formules de calcul des coefficients ΔMFCC et ΔΔMFCC sont données respectivement par les équations :

$$\Delta MFCC_l(m) = \left[\sum_{k=-K}^{K} k(MFCC_{l-k}(m))\right] \qquad \textbf{(4.45)}$$

$$\Delta\Delta MFCC_l(m) = \left[\Delta MFCC_{l+1}(m) - \Delta MFCC_{l-1}(m)\right] \qquad \textbf{(4.46)}$$

où l est l'index de la trame courante et k le nombre de trames de part et d'autre de la trame courante (dans notre cas k est pris égal à 2).

Type du vecteur d'entrée	Unités d'entrées	Taux global en %
$Cep_p(12) + \Delta Cep_p(12) + E$	25	84.1
$MFCC(12) + \Delta MFCC(12) + E$	25	86.1
$MFCC(12) + \Delta MFCC(12) + E + \Delta E$	26	87
$MFCC(12) + \Delta MFCC(12) + \Delta\Delta MFCC(12)$	36	87.2
$MFCC(12) + \Delta MFCC(12) + \Delta\Delta MFCC(12) + E$	37	88.1
$MFCC(12) + \Delta MFCC(12) + \Delta\Delta MFCC(12) + E + \Delta E$	38	89.7

Tableau 4.2: Résultats de reconnaissance pour différents types de vecteurs acoustiques

Commentaires

Les résultats obtenus montrent que les coefficients MFCC donnent de meilleurs résultats de reconnaissance par rapport aux coefficients cepstraux linéaires. De plus, ces coefficients combinés avec leurs dérivées et l'énergie donnent les meilleurs taux de reconnaissance.

Conclusion

Ces résultats nous permettent de dire que les coefficients MFCC représentent mieux l'enveloppe spectrale du signal. L'introduction de l'information dynamique portée par la vitesse et l'accélération *(ΔMFCC, ΔΔMFCC, ΔE*) augmente la performance du système par conséquent, l'utilisation d'un vecteur d'entrée multi-variables est un facteur d'amélioration de la performance globale du système. Seulement, dans ce choix il faut aussi tenir compte de la capacité du vecteur à réaliser un compromis entre la rapidité d'apprentissage et la capacité de généralisation

4.5.2 Modèle à vecteurs multi-variables et dictionnaires spécifiques

Dans le modèle MMC à dictionnaires spécifiques (**Fig. 4.6**), le vecteur acoustique d'entrée x utilise différents types d'indices statiques et dynamiques. Chaque type d'indice véhicule un type d'information. Un moyen d'introduire ces différents types d'informations dans le modèle de reconnaissance est de modéliser chaque type d'information par une composante du vecteur acoustique. Pour cela, en posant $x = \{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(k)}\}$, la composante (sous-vecteur) $x^{(1)}$ représente par exemple l'énergie E, la composante $x^{(2)}$ les coefficients MFCC, la composante $x^{(3)}$ les ΔMFCC, etc. Chaque composante est alors codée par son propre dictionnaire $m^{(k)}$ telle que $QV(x) = \{ m^{(1)}(x), m^{(2)}(x), m^{(3)}(x), \ldots, m^{(k)}(x)\}$, chaque dictionnaire possède une taille $M^{(k)}$.

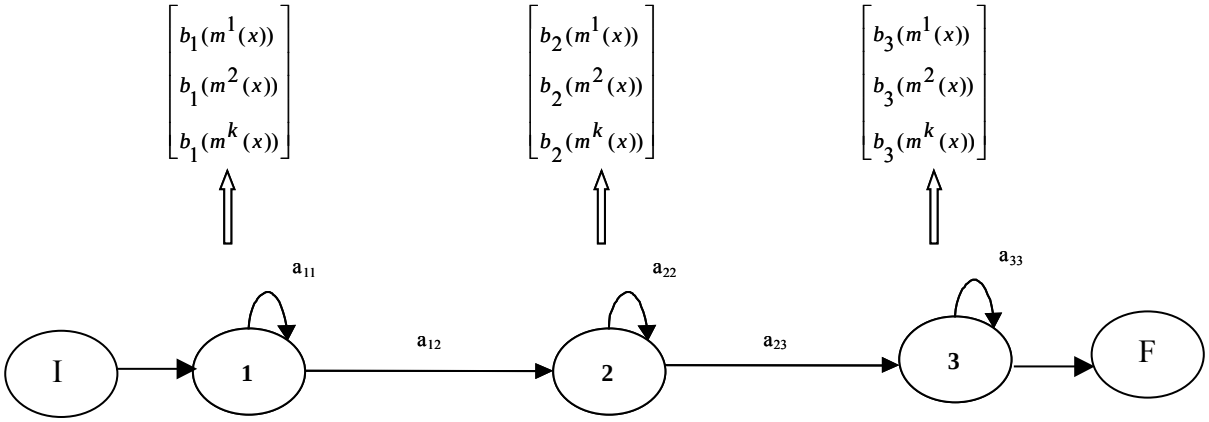


Fig. 4.6 – Modèle gauche-droite à vecteurs multi-variables et dictionnaires spécifiques.

Afin de réduire le nombre de paramètres des modèles et de simplifier les calculs, les probabilités de sortie par état $b_j(m^{(1)}(x))$, $b_j(m^{(2)}(x))$,……, $b_j(m^{(k)}(x))$ sont supposées statistiquement indépendantes. Dans ce cas, la probabilité de sortie $b_j(x)$ peut être considérée comme étant égale au produit des probabilités. Cette probabilité est dans ce cas donnée par la relation (**4.47**) suivante :

$$b_j(x) = \prod_K b_j(m^{(k)}(x)) \qquad \textbf{(4.47)}$$

Pour mesurer l'apport de l'utilisation de dictionnaires spécifiques à chaque composante du vecteur acoustique sur la performance du système, nous avons réalisé les expériences précédentes mais avec un dictionnaire propre à chaque type de composante. La taille de chaque dictionnaire est spécifique à chaque composante (128 pour les MFCC, 128 pour les ΔMFCC, 128 pour les ΔΔMFCC, 32 pour E et 32 pour ΔE) . Les résultats de reconnaissance comparatifs sont résumés dans le **Tableau 4.3**.

Type du vecteur d'entrée	**Dictionnaire Unique** T.R (%)	**Dictionnaires Spécifiques** T.R (%)
Cep(12) + ΔCep(12) + E	84.1	86.9
MFCC(12) + ΔMFCC(12) + E	86.1	87.7
MFCC(12) + ΔMFCC(12) + E + ΔE	87	88
MFCC(12) + ΔMFCC(12) + ΔΔMFCC(12)	87.2	88.3
MFCC(12) + ΔMFCC(12) + ΔΔMFCC(12) + E	88.1	90.5
MFCC(12) + ΔMFCC(12) + ΔΔMFCC(12) + E + ΔE	89.7	91.2

Tableau 4.3: Résultats comparatifs du système MMC à Dictionnaire unique vs. système MMC à dictionnaires spécifiques.

Conclusion

L'analyse des résultats obtenus montre que l'utilisation de dictionnaires spécifiques à chaque type de composante du vecteur acoustique permet d'améliorer la performance du système de reconnaissance basé sur les MMC discrets. Ceci peut s'expliquer par le fait que le vecteur acoustique multi-variables véhiculant des informations de nature différente. Celles-ci sont mieux introduites dans le système lorsque chaque type d'information est quantifiée par son propre dictionnaire.

4.5.3 Modèles dépendants du contexte

Il apparaît clairement de l'étude spectrographique réalisée dans le cadre de l'élaboration de la base de connaissances pour la mise en œuvre de l'approche de reconnaissance fondée sur un système expert que les caractéristiques acoustiques des phonèmes sont affectées par le contexte. Afin de modéliser au mieux cette variabilité du signal de parole de nombreux travaux **[84], [85], [86],** ont montré l'intérêt d'utiliser des modèles dépendants du contexte. Ces modèles sont normalement plus adaptés parce que les effets de coarticulation sont explicitement pris en compte. Deux types de modèles dépendants du contexte sont généralement utilisés : les modèles de diphones et les modèles de triphones. Un diphone est un modèle de phonème prenant en considération les contextes gauches ou droits immédiats alors qu'un triphone est un modèle de phonème prenant en compte les contextes gauches et droits

immédiats. Dans cette étude expérimentale nous nous sommes intéressés à l'utilisation des diphones dans le contexte précédent immédiat comme modèles dépendants du contexte. La **Fig. 4.7** illustre un exemple de modèles de diphones dans le contexte précédent immédiat.
Dans le cas d'une modélisation indépendante du contexte (IC) le mot "Eilm" (savoir) est construit par une concaténation des phonèmes /ε/, /i/, /l/, /m/. Dans le cas d'une modélisation dépendante du contexte (DC) par diphones, le mot "Eilm" est construit par concaténation des différents modèles "# – ε", "ε – i " , "i – l " et "l – m " (rappelons que le phonème " # " correspond au silence).

mot	علم (Savoir)			
monophones	/m/	/l/	/i/	/ε/
diphones	« m – l »	« l – i »	« i - ε »	«ε - # »

Fig. 4.7 – Exemples de diphones

L'entraînement de ces modèles a été effectué de la manière suivante :

- Initialisation de la segmentation en diphones à partir de celles des phonèmes indépendants du contexte. Il suffit d'assigner au label correspondant au phonème Y celui correspondant au diphone X-Y de la segmentation présentée sur la **Fig. 4.8.** Durant l'entraînement, les paramètres du modèle de diphone X-Y sont estimés directement à partir des données disponibles pour le phonème Y dans ce contexte particulier.

- Entraînement des modèles par l'algorithme de Baum-Welch et itération du processus jusqu'à convergence.

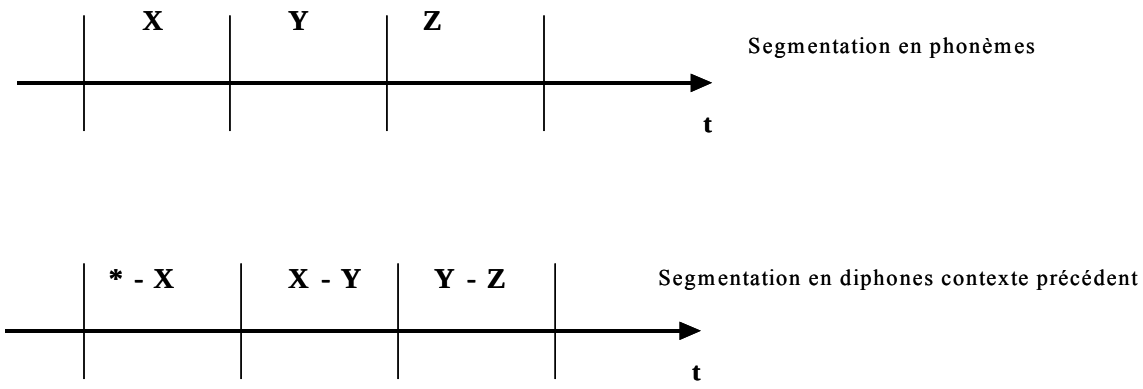


Fig. 4.8 – Passage d'une segmentation phonèmes IC à phonèmes DC

Notre intérêt s'est focalisé sur le contexte vocalique étant reconnu que ce contexte est le plus prépondérant dans la langue arabe. Chaque consonne d'étude est dans ce cas représentée par six modèles sachant que la langue arabe possède six voyelles, trois voyelles brèves /a/, /u/, /i/ et trois voyelles longues /aa/, /uu/, /ii/. Les résultats de reconnaissance obtenus pour un vecteur d'entrée de type x_t *={ MFCC(12), ΔMFCC(12), ΔΔMFCC(12)}* sont résumés par le **Tableau 4.4** suivant :

	MMC indépendant du contexte (IC)	**MMC dépendant du contexte (DC)**
Taux de reconnaissance en (%)	87.2	89.2

Tableau 4.4: Taux de reconnaissance globale obtenu en utilisant des modèles (IC) et des modèles (DC).

Le problème principal inhérent à ce type de modélisation est le nombre de paramètres à estimer. Dans notre cas pour huit consonnes étudiées, la modélisation indépendante du contexte nécessite l'utilisation de 8 modèles donc 3072 (3072 = 8*3*128) paramètres à estimer, les vecteurs acoustiques ont été quantifiés avec un dictionnaire de taille 128 et des modèles à trois états sont considérés. Dans le cas d'une modélisation dépendante du contexte par diphones il faut utiliser 48 (8*6) modèles donc 18432 paramètres à estimer, ce qui est relativement important. Afin de réduire le nombre de paramètres à estimer dans le cas des modèles dépendants du contexte plusieurs méthodes de quantification sont utilisées **[87]**, **[88].** La quantification consiste à regrouper les unités possédant des caractéristiques similaires. Cette quantification peut se faire aussi bien au niveau acoustique qu'au niveau phonétique. Il est donc possible par quantification de réduire le nombre de paramètres à estimer en partageant les

distributions des différents modèles, c'est-à-dire en regroupant, sous un même groupe, les distributions ayant des propriétés similaires. La méthode utilisée dans cette étude, pour effectuer ce regroupement, est basée sur l'utilisation des connaissances sur la langue afin de regrouper les contextes phonétiques dont les effets sur le phonème examiné sont similaires donc c'est une quantification au niveau phonétique. Dans la grammaire arabe traditionnelle une voyelle longue est perçue comme la juxtaposition de deux voyelles brèves. Dans notre expertise en lecture de spectrogrammes nous avons noté que l'opposition voyelle brève/voyelle longue est sémantiquement pertinente mais en termes d'influence contextuelle la voyelle brève et sa correspondante longue sont globalement identiques. Ceci nous a amener à regrouper ces contextes et donc de passer d'une modélisation à 48 modèles à une modélisation à 24 modèles donc 9216 (24*3*128) paramètres à estimer. Malgré que le nombre de paramètres ait été divisé par deux, des taux de reconnaissance globalement similaires ont été obtenus.

4.5.4 Modèles à dictionnaires multiples

L'approche Modèles de Markov Cachés à Dictionnaires Multiples (MMC/DM) utilise non pas un dictionnaire unique commun à l'ensemble des modèles MMC mais plutôt un dictionnaire pour chaque modèle représentant une unité du vocabulaire de reconnaissance. L'utilisation de dictionnaire propre à chaque modèle permet de mieux caractériser les différentes productions acoustiques de l'unité à reconnaître et de produire une information supplémentaire qui est la distorsion globale engendrée lors de la quantification de la séquence d'observations inconnue par les différents dictionnaires. Cette distorsion traduite par une probabilité comme montré dans **[89]** peut être utilisée dans la procédure de reconnaissance de la séquence inconnue.

4.5.4.1 Description du système

La **Fig. 4.9** représente le synoptique du système de reconnaissance basé sur les modèles de Markov cachés à dictionnaires multiples mis en œuvre. Les particularités de ce système par rapport au système MMC à QV conventionnelle sont d'une part, la génération de dictionnaire propre à chaque unité du vocabulaire de reconnaissance donc à chaque modèle et d'autre part, la procédure d'évaluation combine la distorsion de quantification de la séquence inconnue traduite par la probabilité de quantification et la probabilité de génération de cette séquence. Dans ce qui suit, nous allons développer les différentes phases de ce système qui sont la phase d'apprentissage et la phase de reconnaissance. L'intérêt de cette approche réside dans la procédure d'évaluation ainsi donc nous le formalisme mathématique relatif au calcul des probabilités utilisées par cette procédure d'évaluation sera développé séparément.

Phase d'apprentissage : cette phase consiste en la génération des dictionnaires de références et en l'estimation des paramètres des modèles. Pour chaque unité du vocabulaire, un dictionnaire propre est généré par l'utilisation de l'algorithme des K-moyennes dans sa version LBG et un modèle est estimé en utilisant l'algorithme de Baum-Welch.

Phase de reconnaissance : dans cette phase la séquence de vecteurs inconnue est codée par les différents dictionnaires. L'évaluation utilise la combinaison de deux probabilités, la probabilité

$P(X_1^T \mid Y_1^T, \lambda)$ **,** qui représente la probabilité de quantification et la probabilité $P(Y_1^T \mid \lambda)$ qui représente la probabilité de génération de la séquence codée Y_1^T par le modèle λ.

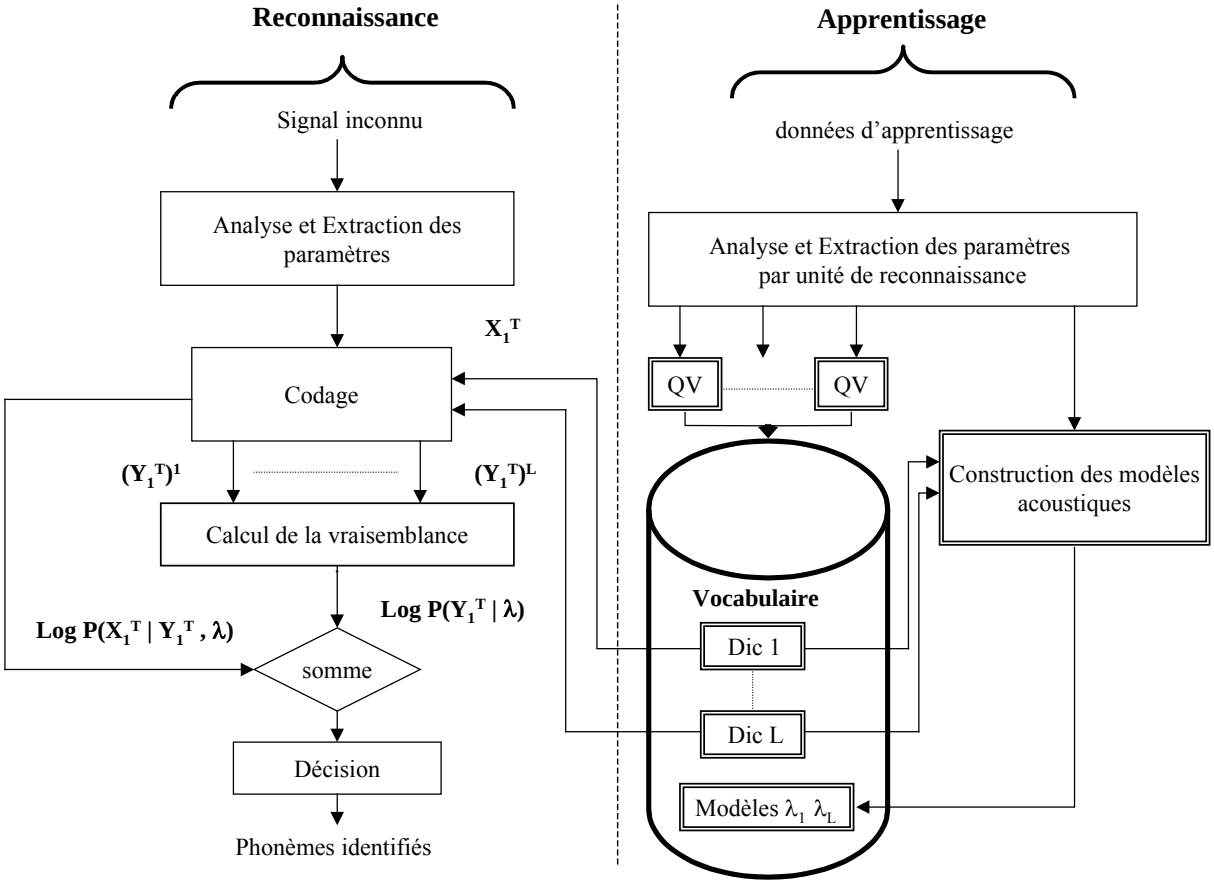


Fig. 4.9 – Schéma du système de reconnaissance MMC à dictionnaires multiples**.**

4.5.4.2 Formalisme mathématique

Nous allons décrire ici le formalisme mathématique relatif à la procédure d'évaluation. Etant donné une séquence d'observation $X_1^T = x_1, x_2, ..., x_T$ et $Q = q_1, q_2, ..., q_T$ une séquence d'états, la probabilité de génération de la séquence X par le modèle λ est :

$$P(X_1^T \mid \lambda) = \sum_{Q = q_1, ..., q_T} P(X_1^T \mid Q, \lambda)\, P(Q \mid \lambda) \qquad \textbf{(4.48)}$$

$$P(Q \mid \lambda) = \prod_{t=1}^{T} P(q_t \mid q_{t-1}, \lambda) \quad \textbf{(4.49)}$$

$$P(X_1^T \mid Q, \lambda) = \prod_{t=1}^{T} P(x_t \mid q_t, \lambda) \quad \textbf{(4.50)}$$

$$P(x_t \mid q_t, \lambda) = \sum_{y_t \in M(q_t, \lambda)} P(x_t \mid y_t, q_t, \lambda) P(y_t \mid q_t, \lambda) \quad \textbf{(4.51)}$$

Dans le cas ou l'ensemble des prototypes du dictionnaire sont indépendants aussi bien de l'état que du modèle, nous avons :

$$P(x_t \mid q_t, \lambda) = \sum_{y_t \in M} P(x_t \mid y_t, q_t, \lambda) P(y_t \mid q_t, \lambda) \quad \textbf{(4.52)}$$

En considérant les prototypes disjoints et dépendants uniquement du modèle λ considéré, nous avons alors :

$$P(x_t \mid q_t, \lambda) = P(x_t \mid y_t^*, q_t, \lambda) P(y_t^* \mid q_t, \lambda) \quad \textbf{(4.53)}$$

$$y_t^* = \arg\max_{y_t \in M(\lambda)} \left[P(x_t \mid y_t, \lambda) \right] \quad \textbf{(4.54)}$$

A partir des équations **4.48**, **4.50** et **4.53** on peut déduire l'équation suivante :

$$P(X_1^T \mid \lambda) = P(X_1^T \mid Y_1^{T*}, \lambda) P(Y_1^{T*} \mid \lambda) \quad \textbf{(4.55)}$$

Dans cette équation, la probabilité $P(X_1^T \mid Y_1^{T*}, \lambda)$ représente la probabilité de quantification et $P(Y_1^{T*} \mid \lambda)$ la probabilité de génération de la séquence d'observations.

- **Calcul de la probabilité $P(X_1^T \mid Y_1^{T*}, \lambda)$:** pour évaluer cette probabilité il faut choisir un modèle paramétrique pour les codewords. Dans notre cas les codewords sont modélisés par une gaussienne à matrice de covariance identité $\Sigma_\lambda = \sigma_\lambda^2 I$

$$P(X_1^{\,T} \mid Y_1^{\,T^*}, \lambda) = \prod_{t=1}^{T} P(x_t \mid y_t^*, \lambda) \qquad \textbf{(4.56)}$$

$$P(x_t \mid y_t, \lambda) = (2\pi)^{\frac{P}{2}} (\sigma_\lambda)^{\frac{P}{2}} \exp\left\{ -\frac{\left\| x_t - \mu_{y_t,\lambda} \right\|^2}{2\sigma_\lambda^2} \right\} \qquad \textbf{(4.57)}$$

$$y_t^{\,*} = \underset{o_t \in V(\lambda)}{\arg\max} \left\{ P(x_t \mid y_t) \right\} = \underset{y_t \in M(\lambda)}{\arg\min} \left\{ \left\| x_t - \mu_{y_t,\lambda} \right\|^2 \right\} \qquad \textbf{(4.58)}$$

$$Log\, P(X_1^{\,T} \mid Y_1^{\,T^*}, \lambda) = T\left[-\frac{P}{2} Log 2\pi - \frac{P}{2} Log \sigma_\lambda^{\,2} - \frac{D_\lambda(X_1^{\,T})}{2\sigma_\lambda^{\,2}} \right] \qquad \textbf{(4.59)}$$

avec

$$D_\lambda(X_1^{\,T}) = \frac{1}{T} \sum_{t=1}^{T} \left\| x_t - \mu_{y_t^*,\lambda} \right\|^2 \qquad \textbf{(4.60)}$$

où

$D_\lambda(X_1^{\,T})$: représente la distorsion moyenne de quantification de la séquence de vecteur $X_1^{\,T}$ par le dictionnaire du modèle λ.

$\mu_{y_t^*,\lambda}$ **:** le vecteur moyen correspondant au prototype présentant la plus faible distance Euclidienne avec le vecteur y_t.

$\sigma_\lambda^2 = \frac{\bar{D}_\lambda}{P}$**,** P étant la dimension du vecteur acoustique et $\bar{D}_\lambda$ la distorsion moyenne produite lors de l'entraînement du modèle λ.

- **Calcul de la probabilité $P(Y_1^{\,T^*} \mid \lambda)$:** le calcul de cette probabilité se fait par l'algorithme conventionnel de Viterbi.

4.5.4.3 Expériences et Résultats

Les expériences précédentes concernant la résolution acoustique, le vecteur multi-variables et les modèles dépendants du contexte ont été réalisées uniquement pour le mode multi-locuteurs. Le modèle à dictionnaires multiples a été implémenté et testé pour le mode multi-locuteurs mais

également pour le mode indépendant du locuteur. C'est ainsi que pour le mode multi-locuteurs, le corpus de mots a servi à l'apprentissage des modèles et le corpus de phrases aux expériences de reconnaissance alors que pour le mode indépendant du locuteur, les prononciations de dix locuteurs ont servi à l'apprentissage et les prononciations des dix autres locuteurs ont servi aux tests. Ces expériences ont été réalisées pour différentes tailles de dictionnaires. Le vecteur d'entrée utilisé est un vecteur de la forme x_t *={ MFCC(12), ΔMFCC(12), ΔΔMFCC(12)}*. Les résultats de reconnaissance obtenus pour le mode multi-locuteurs sont résumés par les **Fig. 4.10, 4.11** et **4.12.** Les figures **4.13**, **4.14** et **4.15** représentent les taux d'identification pour le mode indépendant du locuteur. Ces figures représentent le taux de reconnaissance en fonction de la taille du dictionnaire.

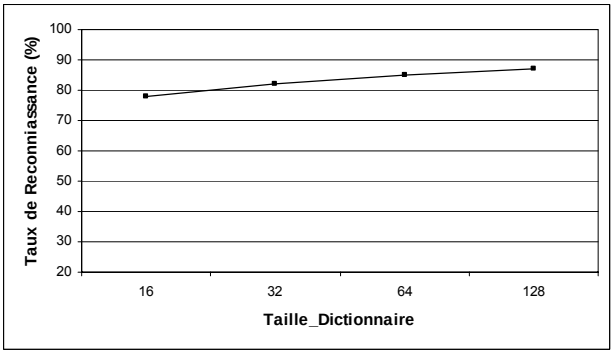


Fig. 4.10 – Taux de reconnaissance du système MMC/QVC.

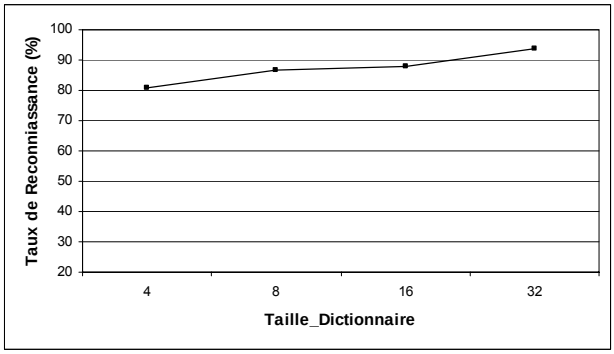


Fig. 4.11 – Taux de reconnaissance du système MMC à dictionnaire multiples

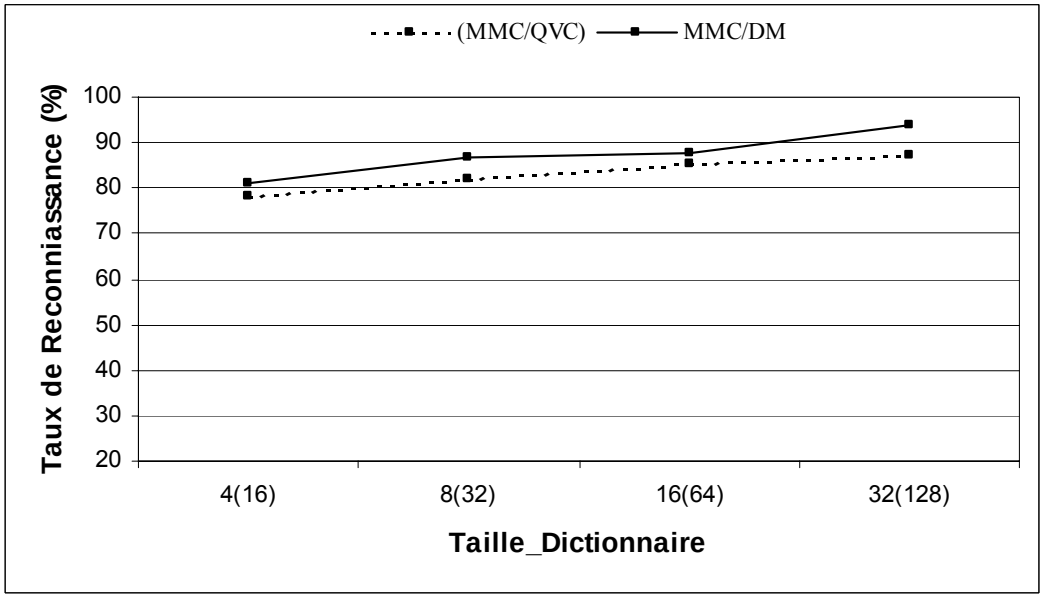


Fig. 4.12 – Comparaison des taux de reconnaissance en mode multi-locuteurs.

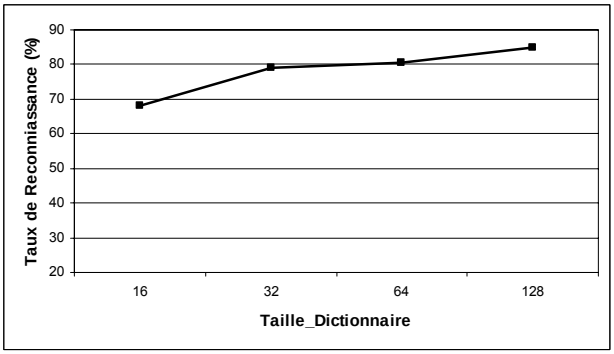


Fig. 4.13 – Taux de reconnaissance du système MMC/QVC.

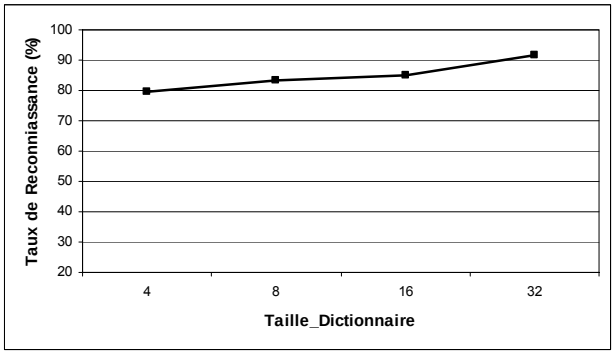


Fig. 4.14 – Taux de reconnaissance du système MMC à dictionnaire multiples.

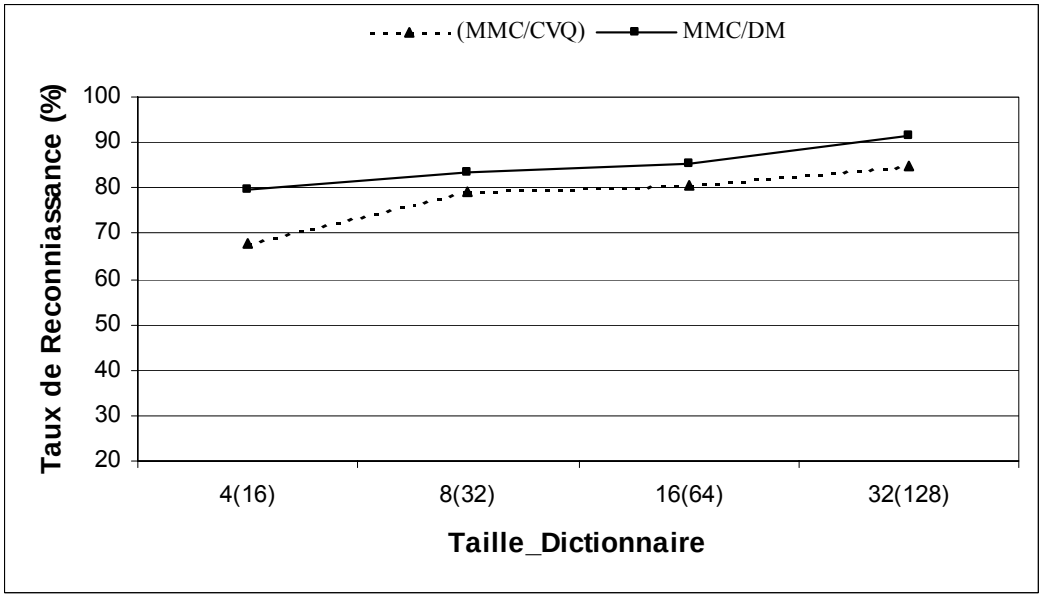


Fig. 4.15 – Comparaison des taux de reconnaissance en mode indépendant du locuteur.

Commentaires

Les résultats de reconnaissance obtenus montrent en premier lieu que le taux de reconnaissance augmente avec la taille du dictionnaire pour les deux approches. L'approche MMC à dictionnaires multiples est plus performante en terme de taux d'identification que l'approche MMC à QV conventionnelle dans les deux modes multi-locuteurs et indépendant du locuteur. Ceci peut s'expliquer

d'une part, par le fait que l'utilisation de dictionnaire spécifique à chaque unité de reconnaissance permet une meilleure prise en compte de la variabilité interlocuteurs et que d'autre part, l'introduction de l'erreur de quantification conduit à un critère de classification optimal.

4.6 Conclusion

Un système de reconnaissance basé sur les MMC combinés à la quantification vectorielle conventionnelle a été mis en œuvre et comparé au système analytique fondé sur les connaissances. Au regard des résultats obtenus, nous avons conclu à la suprématie de l'approche MMC. Elle est plus performante, moins coûteuse et mieux adaptée pour l'identification des phénomènes de forte distorsion temporelle et spectrale telle que la parole.

Avec ce système utilisant les MMC discrets, nous avons introduit plusieurs modifications afin d'augmenter la discrimination des modèles. C'est ainsi que des études sur la nature du vecteur acoustique d'entrée, la topologie des modèles et l'architecture des modèles ont été menées aboutissant à certaines conclusions. La paramètrisation par MFCC représente mieux l'enveloppe spectrale du signal. En incluant la vitesse (paramètre différentiel d'ordre 1) et l'accélération (paramètre différentiel d'ordre 2), les taux de reconnaissance augmentent. L'utilisation de modèles dépendants du contexte (prise en compte explicite du phénomène de co-articulation) et l'amélioration de la représentation de l'information au sein du système permet aussi d'augmenter la discrimination des modèles. Les expériences de reconnaissance réalisées en tenant compte de tous ces aspects ont montré qu'un gain appréciable (de l'ordre de 5%) peut être obtenu pour le taux de reconnaissance global.

Dans le cas du système MMC à dictionnaires multiples où chaque modèle est doté de son propre dictionnaire, les expériences réalisées ont montré que l'utilisation de dictionnaires spécifiques à chaque unité améliore la performance globale. Un gain de l'ordre de 3% en terme de taux de reconnaissance est obtenu. Cet état de fait peut s'expliquer, d'une part, par une meilleure prise en compte de la variabilité interlocuteurs et, d'autre part, l'introduction de l'erreur de quantification optimise le critère de classification.

Modèles de Markov Cachés Multibandes

5.1 Introduction

Dans ce chapitre, nous présentons les modèles de Markov cachés multibandes. Les modèles multibandes sont un modèle parallèle de reconnaissance automatique de la parole. Les fondements à l'origine de ces modèles sont principalement de deux types : des fondements conceptuels et des fondements psycho-acoustiques. Les fondements d'ordre conceptuels sont liés au paradigme multibandes qui permet de traiter l'information par des sous-reconnaisseurs partiels travaillant indépendamment les uns des autres sur chaque partie de l'information. Les fondements psycho-acoustqiues sont en relation avec les travaux de H. Fletcher **[90]** et de J.B. Allen **[91]** qui, suggèrent que la perception auditive humaine soit basée sur des sous-bandes de fréquence traitées indépendamment les unes des autres.

5.2 Fondements des multibandes

5.2.1 Fondements conceptuels

Dans le modèle MMC classique (modèle monobande), le signal de parole est converti en une séquence de vecteurs de paramètres où chaque élément de ce vecteur décrit une fraction de l'information véhiculée par le signal. Pour réaliser le décodage, le modèle monobande considère chaque vecteur de paramètres comme une seule entité. Par conséquent, même si un élément de ce vecteur est bruité (voir **Fig.5.1)**, tout le vecteur est contaminé et les performances du décodeur en sont ainsi fortement affectées. Afin de palier à cette dégradation de la performance pour cause de bruitage d'un des éléments du vecteur (ou de quelques éléments de ce vecteur), le modèle multibandes est développé dans un paradigme lui donnant la capacité de traiter indépendamment les unes des autres les informations partielles (éléments du vecteur) du signal de parole.

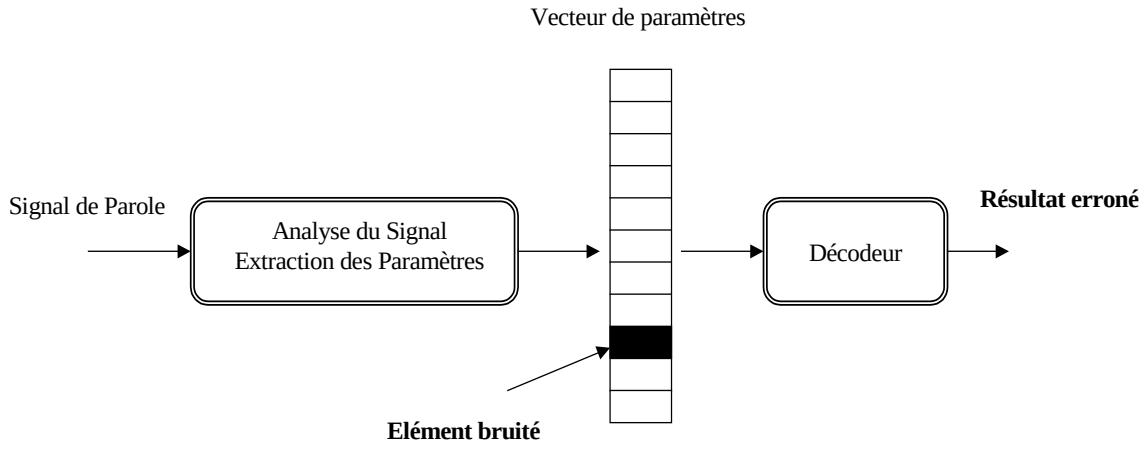


Fig. 5.1 – Modèle MMC monobande.

5.2.2 Fondements psycho-acoustiques

Afin de quantifier la qualité auditive des sons de la parole dans le but d'augmenter l'intelligibilité de la parole téléphonique H. Fletcher a entrepris des expériences très intéressantes sur la perception

humaine de la parole. Dans ces expériences, Fletcher a défini un terme appelé "articulation" qui mesure la probabilité de reconnaissance correcte des unités de parole dénuées de sens. Il a déterminé de manière empirique la relation qui lie l'articulation d'une syllabe (AS) de type CVC dénuée de sens et les articulations (C) et (V) respectivement de la consonne et de la voyelle. C'est ainsi que pour toutes les valeurs du gain α qui est le rapport signal sur bruit (SNR, Signal-to-Noise Ratio pour l'anglais), cette relation est donnée par :

$$AS(\alpha) = C^2(\alpha).V(\alpha) \approx s^3(\alpha) \qquad \textbf{(5.1)}$$

avec $s(\alpha)$, l'articulation moyenne des sons (des phones).

Cette relation, équation **5.1,** traduit le fait que les phonèmes de la syllabe soient décodés de manière indépendante par le système auditif humain.

Afin de mieux comprendre le mécanisme humain de la perception, Fletcher entreprend alors l'étude de l'influence de la réponse en fréquence du canal sur la perception des phonèmes en filtrant le signal étudié par un filtre passe-bas et par un filtre passe-haut. Il déduit que les articulations partielles S_L et S_H respectivement pour le filtre passe-bas et pour le filtre passe-haut ne correspondent pas à l'articulation dans la bande totale, ainsi :

$$s(\alpha) \neq s_L(f_c, \alpha) + s_H(f_c, \alpha) \qquad \textbf{(5.2)}$$

où f_c représente la fréquence de coupure des filtres.

Suite à cette étude, Fletcher fait l'hypothèse de l'existence d'une transformation non linéaire sur S, appelée A(s), qui satisfait la relation d'additivité suivante pour toutes les valeurs de α et de f_c.

$$A(s(\alpha)) = A(s_L(f_c, \alpha)) + A(s_H(f_c, \alpha)) \qquad \textbf{(5.3)}$$

Cette transformation non linéaire A(s) est définie par le terme "indice d'articulation", elle est déterminée de manière empirique en trouvant, pour toutes les valeurs du gain α, la fréquence de coupure $f_c = f_c^*$ qui réalise l'égalité entre $s_L(f_c^*, \alpha)$ *et* $s_H(f_c^*, \alpha)$.

Ainsi donc, pour cette fréquence f_c^* les articulations passe-bas et passe-haut sont donc égales et reliées à l'articulation de la bande totale par la relation :

$$A(s_L(f_c^*, \alpha)) = A(s_H(f_c^*, \alpha)) = 0.5\, A(s(\alpha)) \qquad \textbf{(5.4)}$$

Aussi, en déterminant empiriquement A(s) pour toutes les valeurs de α, Fletcher trouve que pour une syllabe dénuée de sens de type CVC, A(s) est donnée par une fonction de la forme :

$$A(s) = \frac{\log_{10}(1-s)}{\log_{10}(1-0.985)} \qquad \textbf{(5.5)}$$

Où la valeur 0.985 correspond à l'articulation maximale observable pour des conditions idéales, c'est à dire α=1. En remplaçant cette relation de A(s) dans l'équation **5.4,** on aboutit à la relation :

$$(1-s) = (1-s_L).(1-s_H) \qquad \textbf{(5.6)}$$

et $$(e) = (e_L).(e_H) \qquad \textbf{(5.7)}$$

Où le terme *(e)* représente l'erreur d'articulation donnée par *(1 – s)*. En se basant sur la nature additive de l'indice d'articulation donné par l'équation **5.3**, Fletcher généralise l'équation **5.7** au modèle multibandes de perception où le signal de parole est filtré par un nombre K de filtres passe-bande. Il déduit alors que l'erreur d'articulation *(e)* dans le cas de ce modèle est donnée par l'équation **5.8** suivante :

$$(e) = (e_1).(e_2).....(e_K) \qquad \textbf{(5.8)}$$

Cette relation entre l'articulation de la bande totale et les articulations partielles suggère que l'erreur dans une bande de fréquence donnée soit indépendante des erreurs dans les autres bandes de fréquences. B.J Allen **[92]** interprète aussi, dans ces travaux, cette équation **5.8** par le fait que les phones sont traités dans des bandes de fréquences indépendantes et les résultats des estimateurs indépendants correspondants sont par la suite combinés afin de produire la reconnaissance finale.

Un autre résultat concernant le traitement de l'information de la bande bruitée peut être déduit de cette équation. Considérons par exemple le cas d'un modèle à deux bandes de fréquences avec le signal filtré passe-haut contaminé par du bruit et possédant une erreur de reconnaissance égale à 90% et le signal filtré passe-bas propre avec une erreur de reconnaissance égale à 10%. Au regard de l'équation **5.7**, l'erreur de reconnaissance de la bande totale est égale à 0.09 (0.1 x 0.9 = 0.09), c'est une erreur qui est largement dictée par l'erreur du filtrage passe-bas malgré que cette bande ne soit pas affectée par le bruit. Ainsi, l'information provenant de la sous bande contaminée ne doit pas être utiliser lors de la reconnaissance.

Ainsi donc, les résultats de cette étude suggèrent que la perception humaine de la parole obéisse au processus suivant :

- Traitement de l'information dans des bandes de fréquence indépendantes.
- L'information provenant de la bande bruitée est exclue de la reconnaissance.
- Recombinaison des estimateurs indépendants pour produire la reconnaissance finale.

Ces résultats sont à la base de l'approche multibandes. Ils sont aussi motivés par d'autres travaux dans le domaine de la perception humaine que nous présenterons dans la section suivante.

5.2.3 Autres motivations

D'autres travaux ont également contribués à mettre en évidence l'intérêt de l'approche multibandes. Nous pouvons citer les travaux de Arai et *al* **[93]** qui ont montré la robustesse du système auditif humain lorsque différentes bandes de fréquences sont désynchronisées. Ceci corrobore l'hypothèse selon laquelle chacune de ces bandes est traitée indépendamment dans le cerveau.

Miller et Nicelly **[94],** dans leurs travaux sur la perception humaine par bandes de fréquences ont montré que l'audition humaine est robuste au filtrage fréquentiel. Ils ont montré que pour des syllabes de type CVC dénuées de sens passées à travers un filtre passe-bas de fréquence de coupure égale à 800Hz, l'intelligibilité des consonnes est autour de 40%. Lorsque le signal est largement filtré dans les moyennes fréquences (800Hz – 4000Hz), ne laissant subsister pratiquement que les basses et hautes fréquences, l'intelligibilité passe à 90%. Ceci montre que les humains ont la capacité d'extraire toute l'information contenue dans une bande de fréquence, même étroite.

Les travaux de Kryter **[95]** qui a étudié l'effet du filtrage sur l'intelligibilité de la parole par l'utilisation de plusieurs filtres de type passe-bande placés successivement à différents points de l'échelle des fréquences. Kryter est arrivé au résultat que l'intelligibilité d'une seule bande passante est autour de 30%. L'utilisation de deux bandes fait passer l'intelligibilité de 30% à une valeur comprise entre 50% et 75% en fonction de la manière dont les fréquences centrales sont combinées. L'intelligibilité passe à 85% dans le cas où trois filtres passe-bande seraient utilisés successivement.

Les résultats de ces quelques travaux cités montrent que les humains perçoivent la parole avec une relative grande précision même avec un nombre limité de paramètres spectraux (pour cause de filtrage). Ils suggèrent donc que le spectre de parole contienne une quantité appréciable d'information redondante et que les humains sont dotés de capacité d'intégration facile de paramètres acoustiques provenant de différentes régions fréquentielles. Il faut noter que d'autres travaux que nous n'avons pas citer ici se sont intéressés à l'étude de la perception humaine. Ces travaux ont abouti globalement à des conclusions presque similaires.

5.3 Principe de l'approche multibandes

Dans l'approche MMC multibandes (**Fig. 5.2),** la bande de fréquence totale du signal de parole est divisée en un nombre 'N' de sous-bandes sur les quelles des analyses sont faites afin d'extraire les paramètres acoustiques. Pour chaque sous-bande un sous-reconnaisseur (estimateur de probabilités) est développé, chaque sous-reconnaisseur possède ses propres modèles acoustiques. Les sorties de tous les sous-reconniasseurs sont par la suite fusionnées afin de produire la décision globale de reconnaissance.

Sachant que 'N' représente le nombre de sous-bandes, considérons :

- Y_b, le vecteur de paramètres d'une trame de la sous-bande b ou la séquence de vecteurs de paramètres de la même sous-bande b caractérisant une unité de reconnaissance. Cette unité peut être soit un phonème, une syllabe, un mot, etc.
- M, le nombre de classes (phonème, syllabes, mots, ..) à reconnaître.
- λ_1 , λ_2 , ……., λ_M , les symboles ou modèles associés respectivement aux M classes.

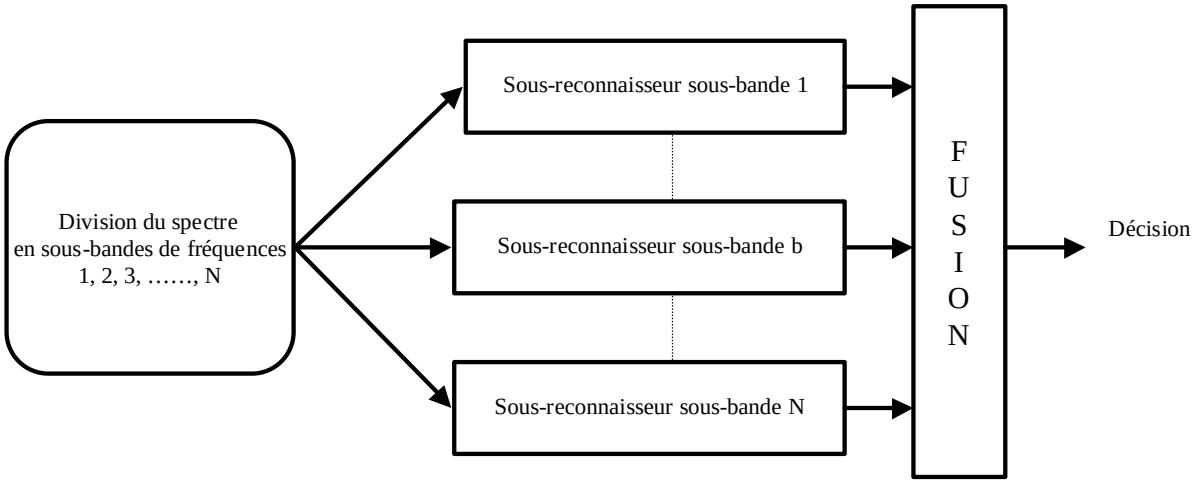


Fig. 5.2 – Schéma général d'une modélisation multibandes pour la R.A.P.

La sortie du sous-reconnaisseur de la sous-bande b est un vecteur de probabilités dont les éléments sont les vraisemblances $P(Y_b | \lambda_m)$.

La sortie désirée du module de recombinaison est alors un vecteur constitué des probabilités $P(Y_1, ..., Y_b, ..., Y_N | \lambda_m)$.

Pour une recombinaison au niveau de l'étage de décodage lors d'une tâche de reconnaissance de mots isolés par exemple, après décodage au niveau des sous-bandes, la décision est basée sur le calcul du modèle qui maximise la probabilité $P(Y_1, ..., Y_b, ..., Y_N | \lambda_m)$ soit :

$$\lambda^* = \max_j P(Y_1, ..., Y_b, ..., Y_N \mid \lambda_j), \quad \text{pour } j = 1, \ldots\ldots, M \qquad \textbf{(5.9)}$$

5.4 Problèmes inhérents à l'approche multibandes

Les principaux problèmes inhérents à cette approche et qui doivent donc être résolus afin de concevoir des modèles multibandes sont principalement :

- La définition des sous-bandes de fréquences.
- Le choix des paramètres acoustiques à utiliser dans chaque sous-bande de fréquence.
- La stratégie de recombinaison ou de fusion.
- Détermination du niveau temporel de fusion.

Après cette énumération des problèmes principaux inhérents à la structure multibandes, nous donnons dans la section suivante une sorte de l'état de l'art relatif à chaque problème.

5.4.1 Définition des sous-bandes de fréquences

Le premier problème lié à l'approche multibandes est la définition des sous-bandes à utiliser et les limites fréquentielles de chacune de ces sous-bandes. Plusieurs travaux se sont intéresses à ce problème. Nous pouvons citer en particulier les travaux de Duchnowsky **[96]** et les travaux de H. Bourlard **[97]**. Duchnowsky a utilisé dans ses travaux quatre bandes de fréquences [100-700], [700-1500], [1500-3000] et [3000-4500Hz]. Le choix de ces bandes a été motivé par le fait que chacune englobe un formant des voyelles de l'anglais. H. Bourlard *et al* ont utilisé dans leur système 3, 4 et 6 bandes de fréquences. Les meilleurs résultats qu'ils ont obtenus correspondent à l'utilisation de quatre bandes de fréquences dont les limites sont [0-901], [797-1661], [1493-2547] et [2298-4000Hz]. Ces limites sont calculées à partir des limites des bandes critiques proposées par B.J Allen **[92].**

5.4.2 Le choix des paramètres acoustiques

Un autre problème dans la conception des modèles multibandes est le choix des paramètres à utiliser dans chaque sous-bande de fréquences. Les meilleurs résultats de reconnaissance obtenus par les modèles MMC classiques sont le fait de l'utilisation des coefficients de bancs de filtres dont les fréquences centrales sont réparties dans une échelle Mel. Ces coefficients sont désignés sous le terme MFCC (Mel Frequency Cepstral coefficients). Principalement, tous les travaux qui se sont intéressés au problème du choix des paramètres acoustiques dans l'architecture multibandes **[96], [97], [98], [99]** ont également conclu en faveur de l'utilisation des coefficients cepstraux comme paramètres acoustiques.

5.4.3 La méthode de fusion

Le problème le plus crucial dans la mise en œuvre du modèle multibandes est la méthode de fusion à utiliser pour combiner les sorties des reconnaisseurs partiels. La recombinaison de ces sorties multiples est basée sur le critère suivant : les reconnaisseurs doivent produire une information complémentaire afin d'espérer atteindre une meilleure performance que le reconnaisseur unique (modèle classique). Plusieurs techniques de fusion relatives à la reconnaissance des formes sont proposées dans la littérature. Les techniques les plus utilisées dans le cas qui nous intéresse ici, le modèle MMC multibandes, peuvent être classées en deux principales méthodes : la méthode linéaire et la méthode dite non linéaire. Pour mieux cerner ce problème de fusion dans le modèle multibandes et avant de développer chacune de ces deux méthodes, nous donnons à titre d'exemple dans la **Fig. 5.3** suivante le synoptique général schématisant une fusion dans un modèle à trois sous-reconnaisseurs permettant l'identification de quatre classes (quatre unités à reconnaître).

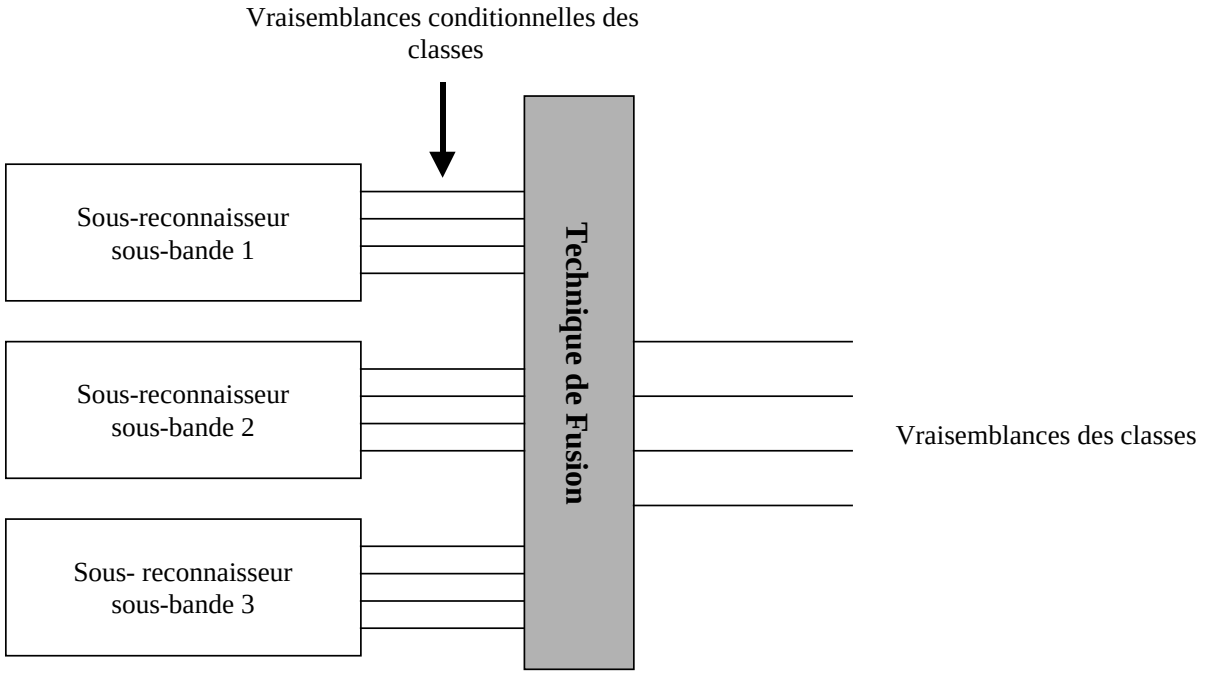


Fig. 5.3 – Schéma synoptique de la fusion dans un modèle à trois sous-bandes et quatre classes.

Méthode de fusion linéaire

La méthode de fusion linéaire peut être considérée comme une simple moyenne des sorties des différents reconnaisseurs ou comme une combinaison linéaire de ces sorties pondérées **[100]**, **[101]**, **[102]**. Dans l'étude **5.3** par exemple la sortie du module de fusion est une estimation des probabilités conditionnelles conjointes des classes, dans le cas d'une fusion linéaire et avec des sous-bandes considérées indépendantes les unes des autres, cette sortie peut être exprimée par :

$$P(Y_1, \ldots\ldots, Y_b, \ldots\ldots\ldots, Y_N \mid \lambda_m) = \prod_{b=1}^{N} P(Y_b \mid \lambda_m) \qquad \textbf{(5.10)}$$

Dans le cas d'une combinaison linéaire pondérée, la sortie du module de fusion est alors :

$$P(Y_1, \ldots\ldots, Y_b, \ldots\ldots\ldots, Y_N \mid \lambda_m) = \prod_{b=1}^{N} w_b \, . P(Y_b \mid \lambda_m) \qquad \textbf{(5.11)}$$

où w_b représente le facteur de confiance accordé à la sous-bande b.

Méthode de fusion non linéaire

La méthode de fusion linéaire est basée sur l'hypothèse de l'indépendance des sous-bandes. Néanmoins, cette hypothèse peut ne pas être vraie, une des raisons à cela : les sous-bandes bien qu'elles soient définies pour des régions de fréquences différentes il existe toujours une étendue de recouvrement entre les bandes due au chevauchement naturel des bandes critiques. Il a été montré que dans le cas ou les reconnaisseurs d'un système sont dépendants les uns des autres, la règle de fusion optimale est de nature non linéaire **[103].** Les techniques de fusion non linéaires utilisées dans les modèles MMC multibandes sont principalement la recombinaison de type "vote majoritaire" qui n'utilise que l'étiquette de la classe gagnante, la recombinaison utilisant une information basée sur le rang des classes et la recombinaison de type neuronale qui utilise les scores retournés par tous les reconnaisseurs pour chaque classe.

- **Technique des votes majoritaires :** cette technique est très souvent utilisée du fait de la simplicité de sa mise en œuvre. La classe (unité à reconnaître) gagnante correspond dans ce cas à la classe qui est le plus souvent proposée (i.e. ayant le plus nombre de votes) par les différents reconnaisseurs. En cas d'égalité, des solutions sont proposées comme de considérer le test d'identification comme un échec, de choisir la classe la plus probable, ou encore de pondérer les votes par des fonctions de probabilités des classes. Cette technique s'avère généralement bien adaptée au cas où le nombre de sous-bandes utilisées serait grand.

- **Technique de recombinaison sur les rangs :** plusieurs variantes de cette technique existent, nous citons dans ce qui suit les deux variantes principales. La première variante qui permet de recombiner les rangs des classes (unités à reconnaître) retournés par les différents reconnaisseurs consiste à calculer une intersection d'ensemble des classes. Une fois que les rang des classes ont été obtenus pour tous les exemples du corpus d'apprentissage. Le rang maximum (i.e. le moins bon) retourné par chaque reconnaisseur pour chaque classe effectivement prononcée est conservé. Ces rangs maximums constituent alors un seuil au-dessus du quel, en phase de test, la classe est éliminée des réponses possibles. Un autre processus de décision permet alors de sélectionner une classe unique parmi les classes restantes. La deuxième variante consiste à calculer une union d'ensembles de classes. Le meilleur rang obtenu pour chaque exemple du corpus d'apprentissage par la classe à laquelle appartient cet exemple, est calculé pour chaque reconnaisseur. Le maximum de ces rangs (i.e. le moins bon) est alors utilisé comme un seuil pour chaque reconnaisseur et chaque classe. Pendant le test, seules les classes ayant un rang inférieur (i.e. meilleur) à ces seuils sont conservées. Pour plus de détails sur cette technique le lecteur pourra consulter les travaux de T.K. Ho **[104]**.

- **Technique de recombinaison neuronale sur les scores :** cette technique de recombinaison utilise toute l'information produite par les différents reconnaisseurs en sous-bandes. Tous les scores issus des reconnaisseurs en sous-bandes sont recombinés par un réseau de neurones (système non linéaire). Les avantages impliqués par l'utilisation d'un réseau de neurones sont l'affranchissement de l'hypothèse d'indépendance entre les bandes de fréquences et la capacité du réseau à apprendre à peu près n'importe quel type de fonction de recombinaison.

5.4.4 Le niveau temporel de fusion

Le modèle multibandes réalise une reconnaissance indépendante sur chaque sous-bande de fréquences, puis fusionne les résultats des reconnaisseurs. Le problème est donc essentiellement de savoir à quel niveau réaliser la fusion. Nous allons ici distinguer les deux niveaux principaux de fusion :

- **La fusion après des segments de plusieurs trames :** Ces segments peuvent aussi bien être une succession de trames de longueur T donnée, des phonèmes, des syllabes ou des mots. Dans la section **5.4.3** (méthode de fusion linéaire), nous avions traité un exemple de fusion au niveau du mot étant donné que celle-ci est réalisée après estimation des vraisemblances des classes dans chaque reconnaisseur en sous-bandes. La **Fig. 5.4** illustre le synoptique de la fusion au niveau du mot. Le problème inhérent à ce type de fusion est l'asynchronisme entre les bandes. Parmi les modèles multibandes développés avec ce type de fusion, nous pouvons citer le modèle multibandes développé par H. Bourlard **[98]** qui a utilisé une fusion au niveau de la syllabe.

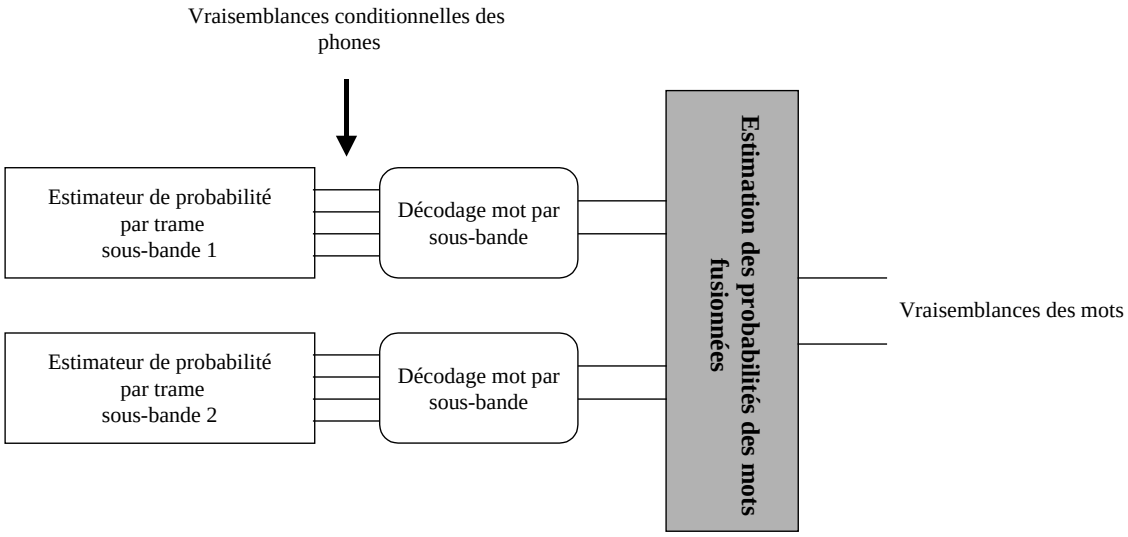


Fig. 5.4 – Schéma synoptique d'une fusion au niveau du mot, un modèle avec deux sous-bandes, deux mots à reconnaître et quatre classes phonétiques.

- **La fusion au niveau de la trame :** Cette fusion est équivalente à une recombinaison au niveau des états des MMC. Les sorties des reconnaisseurs en sous-bandes sont ainsi recombinées après chaque trame du signal. L'avantage de cette méthode est la résolution du problème de l'asynchronisme entre les bandes. La **Fig. 5.5** illustre le synoptique de la fusion au niveau de la trame. Plusieurs systèmes développés utilisent cette fusion au niveau des trames, nous pouvons citer le système développé par Duchnowsky **[96]** et le système développé par Tibrewala **[99]**.

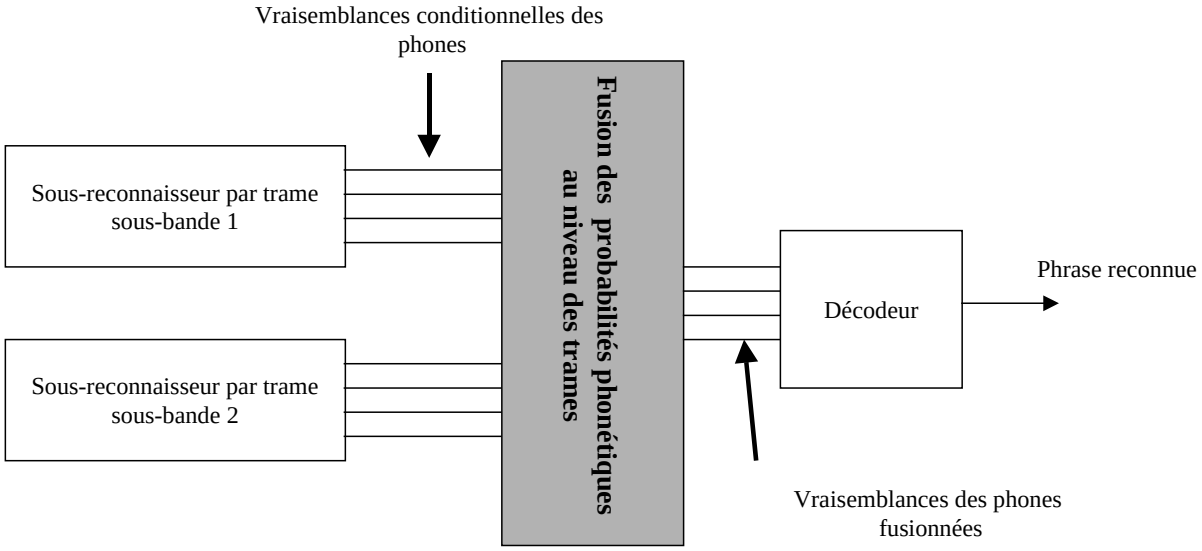


Fig. 5.5 – Schéma synoptique d'une fusion par trame pour un modèle à deux sous-bandes et de quatre classes.

5. 5 Système MMC multibandes

Le système MMC multibandes mis en œuvre est donné par la **Fig. 5.6** suivante. Après une description des différentes étapes de ce système, nous présenterons les différentes études réalisées avec ce modèle. Ces études ont concerné le choix des bandes de fréquences et leurs limites, la recombinaison et le comportement du système en milieu bruité. Il faut préciser ici que dans toutes ces études le niveau de recombinaison adopté est situé après chaque phonème, en mode isolé. Donc c'est une étude du système multibandes en l'absence de synchronisme entre les bandes.

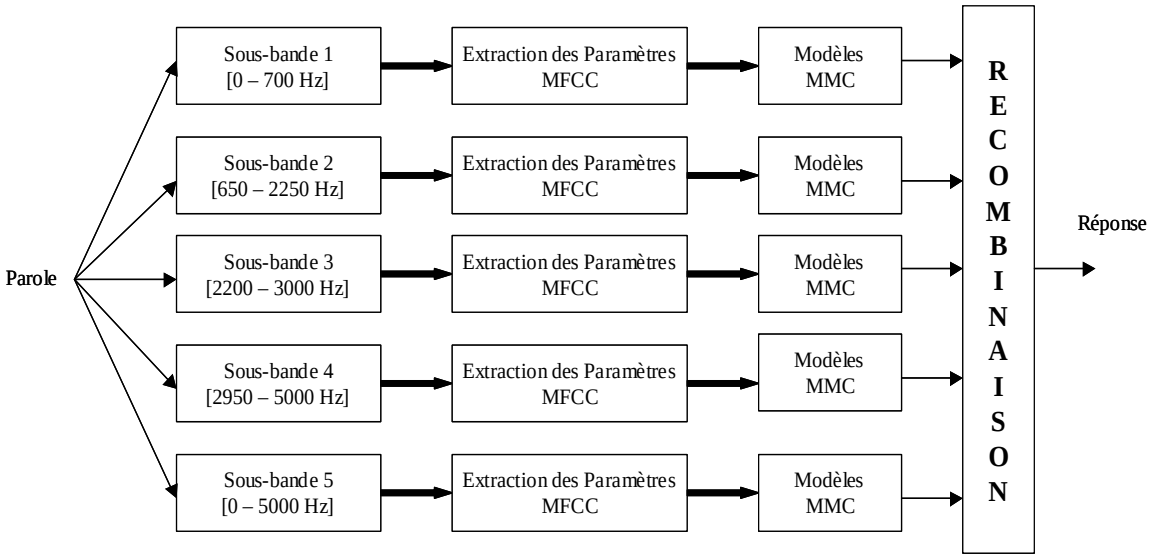


Fig. 5.6 – Schéma du système multibandes.

5.5.1 Description du système

Le spectre du signal de parole est divisé en un nombre 'L' de sous-bandes de fréquences sur les quelles des analyses sont effectuées afin d'extraire les paramètres acoustiques. Pour chaque sous-bande, un sous-reconnaisseur est développé. Chaque sous-reconnaisseur possède ses propres modèles acoustiques qui sont des modèles MMC. Les sorties de tous les sous-reconnisseurs sont par la suite fusionnées au niveau du phonème en mode isolé afin de produire la décision finale de reconnaissance.

- **Paramètres acoustiques :** Les paramètres acoustiques utilisés sont composés de 4 coefficients MFCC ((*Mel Frequency Cepstral Coefficients*) auxquels sont ajoutées les dérivées d'ordre 1 et les dérivées d'ordre 2. Le nombre total de coefficients utilisés est donc de 12.
- **Nombre et limites des bandes :** cette partie a fait l'objet d'expériences avec différentes bandes de fréquences. Elle sera développée plus loin dans la partie étude sur les bandes de fréquence.
- **Reconnaisseurs dans chaque bande :** les reconaisseurs utilisés dans chaque bande de fréquence sont des modèles de Markov cachés discrets d'ordre 1 à trois états émetteurs de même topologie que ceux utilisés dans le modèle classique (chapitre 4). La taille du dictionnaire de référence est égale à 128.
- **Module de recombinaison :** Le module de recombinaison utilise les scores retournés pour chaque phonème par tous les reconnaisseurs c'est donc une recombinaison sur les scores issus de tous les reconnaisseurs en sous-bandes. Deux types de recombinaison ont été testés, la recombinaison linéaire pondérée et la recombinaison non linéaire à base de réseau de neurones. Ce module sera détaillé plus loin dans la partie étude sur la recombinaison.
- **Apprentissage des modèles :** nous avons réalisé l'ajustement des paramètres des modèles (MMC) dans chaque bande de fréquence, en isolant ceux-ci du reste du système et en utilisant l'algorithme d'apprentissage de Baum-Welch. Cela revient à entraîner les modèles séparément sur chaque bande de fréquence.
- **Corpus :** dans cette étude nous avons utilisé le corpus de mots donné en annexe B prononcés par 20 locuteurs pour l'apprentissage des modèles et le corpus de phrases pour la reconnaissance.

5.5.2 Etude sur les bandes de fréquence

Les différents travaux ayant traités le problème du choix du nombre de bandes et de leurs limites dans le modèle multibandes se sont principalement appuyés sur les travaux de H. Fletcher et de J.B Allen. Les meilleurs résultats rapportés ont été obtenus avec l'utilisation de quatre bandes de fréquences englobant grossièrement les formants des voyelles de l'anglais **[98]**, **[99]**. Pour notre part nous avons réalisé dans un premier temps des expériences avec 2, 4 et 7 bandes de fréquences dont les limites sont résumées dans le **Tableau 5.1**. Le modèle à deux bandes correspond à une division du spectre vers la fréquence de 1000Hz, fréquence à partir de laquelle les bandes critiques sont dans une échelle logarithmique. Le modèle à quatre bandes couvrant les formants des voyelles arabes et le modèle à sept bandes qui est en relation avec les travaux de J.B Allen qui suggèrent qu'une bande d'articulation représente en moyenne deux bandes critiques. Etant donné qu'il y a 15 bandes critiques donc sept bandes d'articulation ce qui correspond à 7 bandes de fréquence.

Bandes de fréquences	Limites des bandes en Hertz
B.T	0 - 5000
B1	0 - 1050
B2	1000 - 5000
B1	0 – 700
B2	650 – 2250
B3	2200 – 3000
B4	2950 - 5000
B1	0 – 375
B2	325 – 630
B3	565 – 950
B4	870 – 1370
B5	1250 – 1950
B6	1800 – 2700
B7	2500 - 5000

Tableau 5.1: Limites des bandes de fréquences.

Bandes de fréquences	Taux de reconnaissance par bande en (%)	Taux de reconnaissance global en %
B.T	85.2	87.2
B1	60.2	66.2
B2	72.2	
B1	59.8	60.4
B2	62.5	
B3	57.4	
B4	61.8	
B1	48.2	52.8
B2	55.6	
B3	56.0	
B4	50.8	
B5	48.0	
B6	51.2	
B7	60.1	

Tableau 5.2: Comparaison des taux de reconnaissance pour les différentes bandes de fréquences en mode multi-locuteurs.

Commentaires

Les résultats montrent que le taux de reconnaissance diminue avec le nombre de bandes du modèle. Ceci peut s'expliquer par le fait que plus le nombre de bandes augmente plus la largeur de ces bandes devient étroite par conséquent une faible plage fréquentielle est couverte par la bande donc, une mauvaise description de l'information. De plus, comparativement au taux de reconnaissance du modèle de référence (B.T, Bande Totale), les taux multibandes sont plus faibles. Ces taux globaux sont obtenus sans recombinaison, se sont juste les moyennes des taux obtenus par bande.

5.5.3 Etude sur la recombinaison

La méthode de recombinaison qui a donné les meilleurs résultats est la recombinaison par réseaux de neurones qui s'est avéré être une règle de fusion optimale dans le cas de sous-reconnaisseurs indépendants. Nous avons utilisé pour la recombinaison des réseaux de neurones de type PMC (Perceptron Multi- Couches) à une couche cachée. L'apprentissage de ces réseaux a été réalisé grâce à l'algorithme de la rétro-propagation du gradient **[38]**, **[105]**.

Modèle	**Taux de reconnaissance en (%)**
B1	59.8
B2	62.5
B3	57.4
B4	61.8
B1+B2+B3+B4 (après recombinaison de type PMC)	78.6
B.T	87.2

Tableau 5.3: Taux de reconnaissance pour un modèle à quatre bandes avec recombinaison.

Commentaires

Les résultats de reconnaissance résumés dans le Tableau 5.3 sont ceux du modèle à quatre bandes comparés au modèle de référence (B.T), étant donné que c'est le modèle à 4 bandes qui a donné les meilleurs résultats comparativement aux modèles 2 et 7 bandes. Nous remarquons que même avec une recombinaison, le modèle à une seule bande (B.T) donne de meilleurs résultats de reconnaissance par rapport au modèle multibandes.

Les résultats de reconnaissance du modèle multibandes obtenus après recombinaison sont inférieurs à ceux du modèle de référence (B.T). Nous avons pensé, comme certains travaux **[109], [110]** l'ont suggéré, à ajouter au modèle à quatre bandes la bande totale. Ceci nous amène donc à un modèle à cinq bandes. Avec ce modèle, nous avons réalisé des expériences de reconnaissance avec recombinaison par réseaux de neurones de types PMC possédant 40 (8x5) entrées, 8 sorties et 16 cellules pour la couche cachée. Les résultats obtenus sont résumés dans le **Tableau 5.4** suivant.

Modèle	**Taux de Reconnaissance en (%)**
B1+B2+B3+B4+B.T (après recombinaison de type PMC)	89.1
B.T	87.2

Tableau 5.4: Taux de reconnaissance pour un modèle à cinq bandes avec recombinaison.

Commentaires

Les résultats de reconnaissance obtenus montrent que l'ajout de la bande totale améliore la performance du modèle multibandes. Ceci peut s'expliquer par le fait que la bande totale est porteuse d'information supplémentaire qu'on ne retrouve pas dans les sous-bandes de fréquences.

5.5.4 Etude en milieu bruité

Pour évaluer la robustesse de l'approche multibandes aux conditions bruitées, nous avons dégradé artificiellement les signaux de parole suivant deux protocoles différents, bruit bande étroite (protocole 1) et bruit réparti sur la bande totale (protocole 2). Des expériences de reconnaissance avec le modèle à cinq bandes de fréquence (B1+B2+B3+B4+B.T) avec une recombinaison neuronale sur les scores ont été réalisées. Les résultats obtenus sont reportés sur les **Fig. 5.7** et **5.8** respectivement pour le bruit bande étroite et pour le bruit réparti sur la bande totale.

- **Protocole 1 :** bruit bande étroite

Nous avons ajouté aux signaux un bruit blanc que nous filtrons pour qu'il affecte uniquement une bande bien précise du spectre (nous avons choisi arbitrairement la bande quatre). Nous multiplions ensuite l'amplitude de ce bruit par un coefficient variable qui nous permet de tester le système pour différents niveaux de bruit. Celui-ci est ensuite ajouté au signal original.

le rapport signal/bruit du signal obtenu est calculé par la relation suivante :

$$RSB_{dB} = 10 \log_{10}\left(\frac{E_S}{E_B}\right) \qquad \textbf{(5.12)}$$

E_S représente l'énergie moyenne du signal dans une bande *b* que l'on désire bruiter et E_B représente l'énergie moyenne du bruit souhaité sur la bande.

- **Protocole 2 :** bruit réparti sur toute la bande de fréquence

Il est bien évident qu'un bruit réel n'est pas toujours à bande étroite ; pour aller au-delà du cas particulier précédent et mieux approcher l'environnement réel nous dégradons les signaux de parole par un bruit réparti en moyenne sur la bande de fréquence totale pour différents RSB.

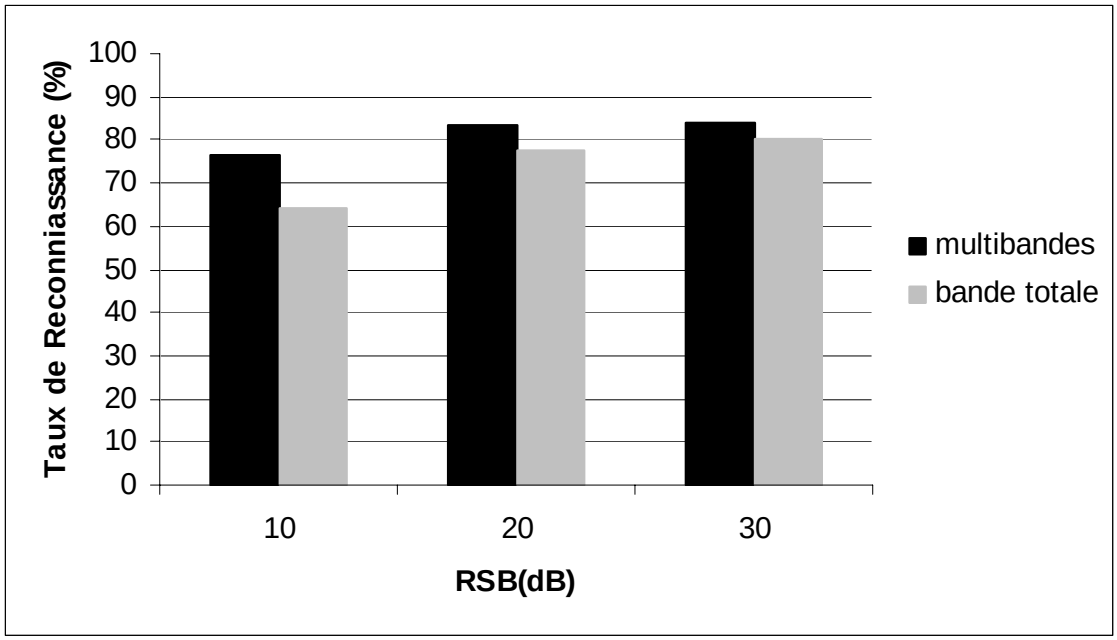


Fig. 5.7 – Comparaison des taux de reconnaissance (bruit limité) pour différents RSB.

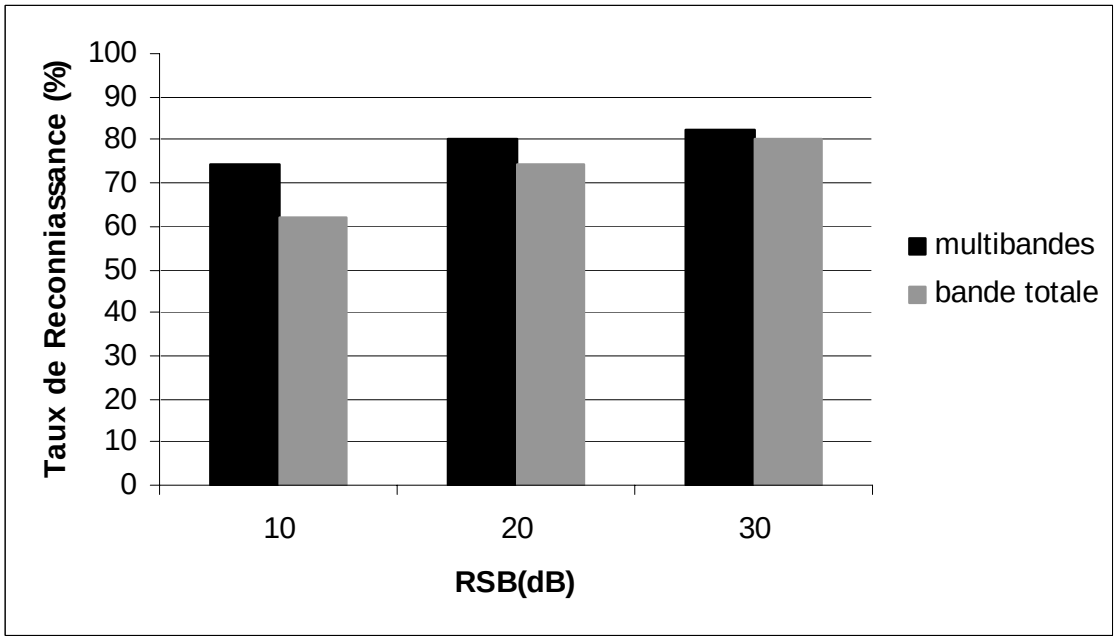


Fig. 5.8 – Comparaison des taux de reconnaissance (bruit réparti) pour différents RSB.

5.6 Conclusion

Dans ce chapitre, nous avons étudié un modèle parallèle de reconnaissance de la parole, le modèle multibandes. Des expériences en milieu calme et bruité ont été réalisées. Dans le cas de données non bruitées (clean data), le taux de reconnaissance du système monobande utilisant tout le spectre (MMC à QV conventionnelle) est nettement supérieur aux taux de reconnaissance des systèmes sous-bandes qui n'utilisent qu'une partie du spectre. La recombinaison améliore les taux du système multibandes. Celle-ci utilisant les réseaux de neurones s'est avérée être une règle de fusion optimale. L'introduction de la bande totale dans le système multibandes améliore le taux de reconnaissance de ce dernier qui devient supérieur au taux de reconnaissance du système de référence (B.T). Ceci montre l'importance de l'utilisation du spectre complet (bande totale) dans le système multibandes. Dans le cas de données bruitées artificiellement, deux catégories de bruits ont été utilisées, un bruit limité fréquentiellement et un bruit réparti sur toute la bande de fréquences. Les résultats des expériences en environnement bruité montrent que le système multibandes est plus performant que le système à bande totale, il offre donc une meilleure robustesse au bruit.

MMC à Quantification Vectorielle Distribuée

6.1 Introduction

Les modèles de Markov cachés discrets sont attractifs de par leur relative complexité algorithmique et leur vitesse de décodage c'est pourquoi plusieurs travaux leurs sont consacrés **[3]**, **[4]**, **[5]**. Plus récemment, dans le contexte de la croissance prodigieuse des applications réseaux, les systèmes de reconnaissance basés sur les modèles discrets qui utilisent la quantification vectorielle (QV) constituent une solution utile et peu coûteuse **[6]**, **[7]**. Cependant, ces modèles souffrent de problèmes inhérents à leur structure qui limitent ainsi leur performance dans le domaine de la reconnaissance automatique de la parole. Ces problèmes sont principalement liés à la perte d'information sur le signal de parole original engendrée par la procédure de quantification vectorielle et au manque de données d'apprentissage qui ne permet pas une bonne estimation des paramètres de ces modèles. Dans le but de réduire ces insuffisances inhérentes à la structure des modèles discrets, nous proposons l'utilisation d'une nouvelle approche de QV que nous avons appelé Quantification Vectorielle Distribuée (QVD) **[106]**, **[107]**. Cette nouvelle approche de QV, basée sur le principe d'une distribution optimale des composantes du dictionnaire sur les états du modèle, permet l'estimation des paramètres des modèles par unification entre les sources acoustique et phonétique. Dans les paragraphes suivants, nous présentons le système QVD proposé et les deux variantes de cette approche à savoir l'approche hybride K-moyennes-QVD et la variante hybride réseaux de neurones artificiels à quantification vectorielle distribuée (RNA-QVD).

6.2 Description du système proposé

Le synoptique du système de reconnaissance MMC à quantification vectorielle distribuée (MMC/QVD) mis en œuvre est donné par la **Fig. 6.1** suivante :

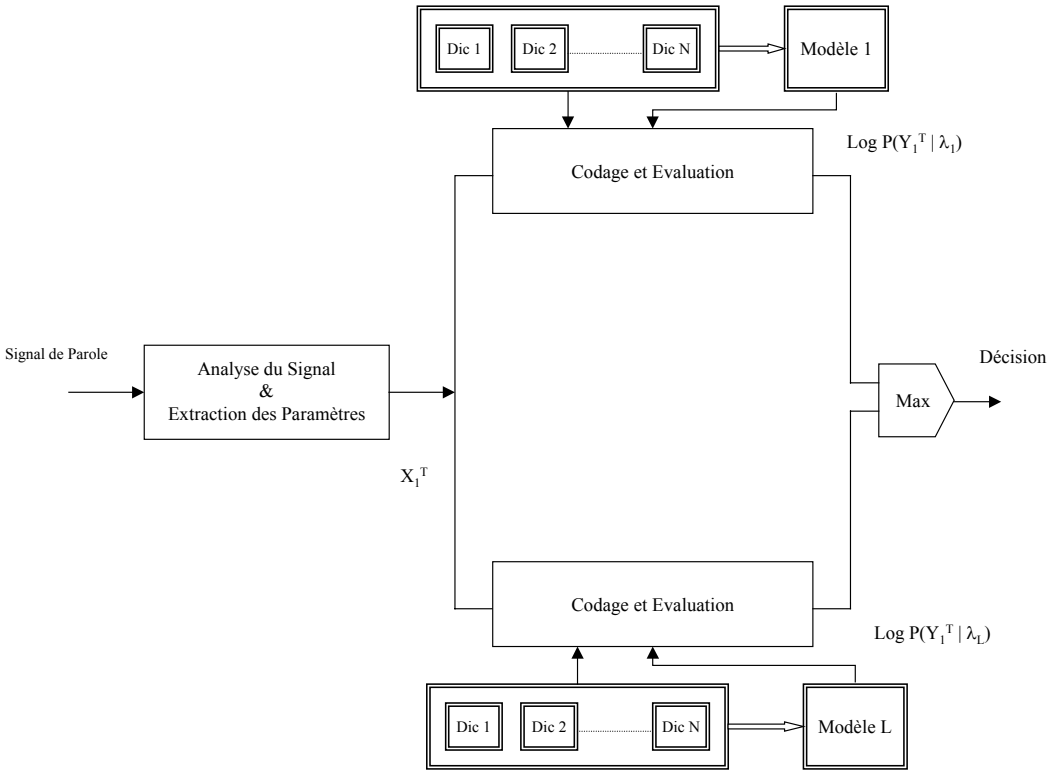


Fig. 6.1 - Système de Reconnaissance par MMC/QVD.

Les différentes étapes de ce système sont : l'analyse acoustique, l'apprentissage distribuée (génération des dictionnaires distribués sur les états des modèles et estimation des paramètres) et l'identification des unités à reconnaître (codage et évaluation).

L'étape analyse du signal et l'extraction des vecteurs acoustiques, est similaire à celle du système MMC à QV conventionnelle. En revanche, la phase de reconnaissance (identification phonémique) consiste à calculer la probabilité d'émission de la séquence d'observation (l'unité à reconnaître codée) en utilisant l'algorithme de Viterbi dans sa version logarithmique modifiée. La vraisemblance de la séquence est dans ce cas donnée par l'équation **6.1** suivante :

$$P(Y_1^T \mid \lambda) = \alpha(t, j) = \max_j \left[\alpha(t-1, j) + \log a_{ji}\right] + \log b_i(Y_t^{(i)}) \qquad \textbf{(6.1)}$$

avec $1 \le t \le T$ et $1 \le i, j \le N$

La **Fig. 6.2** illustre le synoptique de construction des dictionnaires distribués sur les états des modèles et l'estimation de leurs paramètres. Deux variantes ont été implémentées: l'approche K-moyennes-QVD et l'approche RNA-QVD décrites dans les sections suivantes.

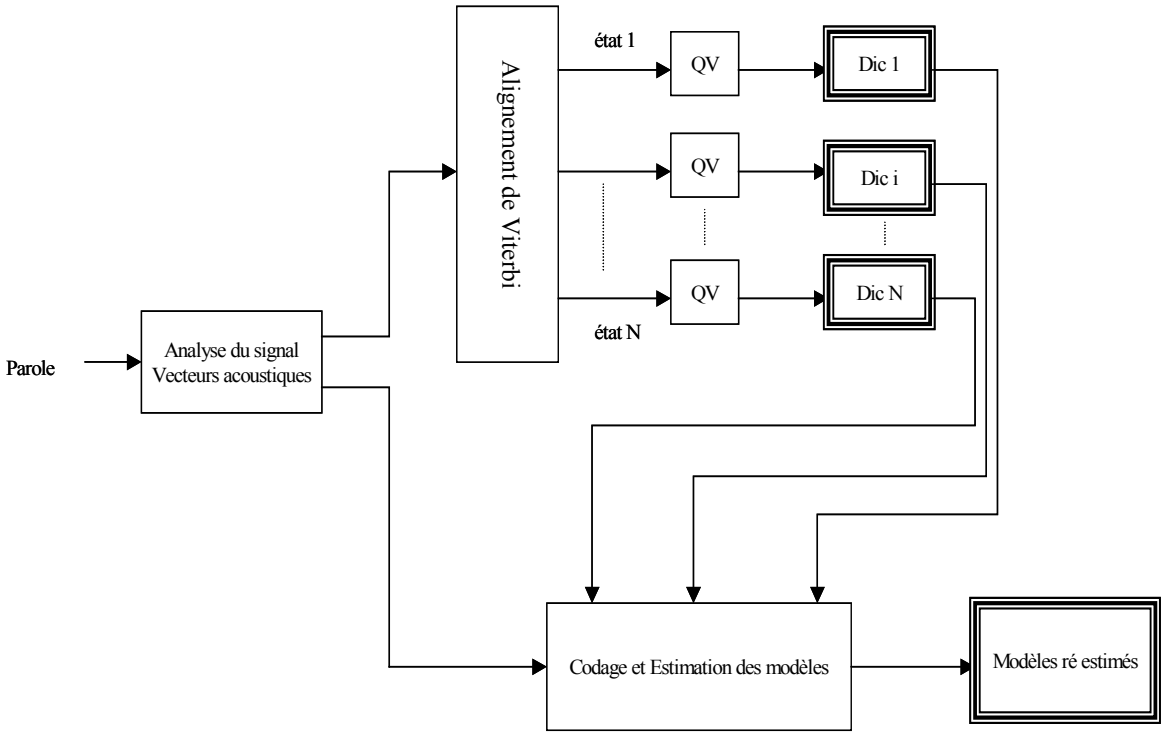


Fig. 6.2 – Synoptique de la phase d'apprentissage dans le système MMC/QVD.

6.3 L'approche hybride K-moyennes-QVD

Dans cette approche, nous utilisons l'algorithme des K-moyennes dans sa version LBG pour générer les dictionnaires distribués sur les états. Et, de cette distribution des dictionnaires nous estimons les paramètres des modèles.

- **Génération des dictionnaires :**

L'algorithme de génération des dictionnaires se fait à travers les étapes de suivantes :

étape 1 : Prendre les unités à reconnaître prononcées plusieurs fois par un certain nombre de locuteurs (base d'apprentissage) sous forme de séquences d'observations (séquences de vecteurs acoustiques).

étape 2 : Déterminer la suite d'états optimale q* en appliquant l'algorithme de Viterbi **[35]** à l'ensemble des séquences d'observations et par un " backtracking " récupérer les observations (vecteurs acoustiques) par état du modèle, soit N_j le nombre d'observations de l'état j.

étape 3 : Grouper les vecteurs acoustiques en régions, chaque région contient les vecteurs acoustiques d'un état donné.

étape 4 : Appliquer la quantification vectorielle (algorithme LBG) pour regrouper la population de chaque région en M classes, soit N_k le nombre d'observations de la classe k.

- **Estimation des modèles :**

Les modèles utilisés dans cette étude sont de topologie similaire à celle des modèles utilisés tout au long de ce travail à savoir des modèles gauche-droite à trois états émetteurs. L'estimation de leurs paramètres en particulier les densités de probabilités discrètes de sortie est faite en utilisant l'algorithme décrit par les étapes suivantes faisant suite aux étapes de génération des dictionnaires.

N, le nombre d'états du modèle
N_j, le nombre d'observations de l'état j
N_k, le nombre d'observations de la classe k

étape 5 : Ré estimer les éléments de la matrice densités de probabilités de sortie B = { b_{jk} } en utilisant la formule suivante :

$$b_{jk} = \frac{N_k}{N_j} \qquad \textbf{(6.2)}$$

avec $1 \le j \le N$ et $1 \le k \le Lj$

étape 6 : Lisser les probabilités.

étape 7 : Affiner les paramètres des modèles en utilisant les formules de ré estimation standard.

6.4 L'approche hybride RNA-QVD

L'algorithme des K-moyennes est basé uniquement sur un critère de minimisation de la distorsion globale. L'information sur la classe phonétique du vecteur acoustique n'est pas prise en compte lors de la génération des dictionnaires, elle est donc perdue dans le processus acoustique des modèles discrets. Afin de garder cette information phonétique sur le vecteur acoustique, nous utilisons dans cette approche une configuration basée sur les réseaux de neurones artificiels (RNA) pour réaliser la quantification vectorielle distribuée. L'algorithme des K-moyennes est donc remplacé par un algorithme de quantification utilisant les réseaux de neurones artificiels. Ces réseaux sont entraînés en mode non-supervisé basé sur un critère de maximisation de l'information mutuelle (MIM) **[58]**. Ce mode d'apprentissage permet d'intégrer l'information phonétique dans le processus de QV et permettre ainsi une meilleure estimation des paramètres liés aux modèles en particulier, les densités de probabilités discrètes. Avant de décrire l'algorithme relatif à cette approche de QV, les définitions en relation avec la théorie de l'Information Mutuelle (I.M) sont d'abord présentées.

6.4.1 Théorie de l'information mutuelle

Nous présentons dans cette étude les notions essentielles nécessaires à la définition de certaines probabilités utilisées dans les formulations relatives à la théorie de l'information mutuelle. Pour plus de détails concernant la théorie de l'information le lecteur peut consulter entre autres les références **[127]**, **[128]**.
Le premier concept de la théorie de l'information est celui de mesure quantitative de l'information, c'est ainsi que la quantité d'information h(x) associée à la réalisation d'un événement x de probabilité P(x) est définie par la fonction h(x) suivante :

$$h(x) = \log\left[\frac{1}{P(x)}\right] = -\log P(x) \qquad \textbf{(6.3)}$$

Le choix de la base du logarithme définit l'unité d'information. Shannon **[105]** a proposé de prendre la base égale à 2 et de nommer l'unité le « bit » pour *binary unit*.

Si l'on considère deux événements x et y. On peut associé au couple (x , y) la quantité d'information :

$$h(x,y) = -\log P(x,y) \qquad \textbf{(6.4)}$$

où P(x , y) désigne la probabilité conjointe des deux événements. On peut également mesurer la quantité d'information associée à l'événement x conditionné par la réalisation de l'événement y par :

$$h(x\,|\,y) = -\log P(x\,|\,y) \qquad \textbf{(6.5)}$$

où $P(x\,|\,y)$ est la probabilité conditionnelle de l'événement x étant donné l'événement y.

En utilisant la règle de Bayes nous pouvons écrire :

$$P(x,y) = P(x\,|\,y).P(y) = P(y\,|\,x).P(x) \qquad \textbf{(6.6)}$$

On en déduit :

$$h(x,y) = h(x|y) + h(y) = h(y|x) + h(x) \qquad \textbf{(6.7)}$$

Dans le cas où les deux événements x et y sont indépendants on a : $h(x|y) = h(x)$, on retrouve ainsi la condition d'additivité de la quantité d'information à savoir $h(x,y) = h(x) + h(y)$.

Une autre mesure intéressante est la quantité d'information qu'un événement, y par exemple, apporte sur un autre événement, x en l'occurrence. Cette mesure est donnée par :

$$i(x,y) = \log\left[\frac{P(x|y)}{P(x)}\right] \qquad \textbf{(6.8)}$$

où P(x) correspond à la probabilité a priori que l'événement x se réalise et $P(x|y)$ à la probabilité a posteriori que x ait été réalisé, sachant que l'événement y a été produit. La quantité $i(x,y)$ mesure l'accroissement de la probabilité de x que l'observation de y apporte. A partir de la règle de Bayes, cette quantité peut s'écrire :

$$i(x,y) = \log\left[\frac{P(x,y)}{P(x).P(y)}\right] \qquad \textbf{(6.9)}$$

Cette quantité est appelée information mutuelle par opposition à h(x) qui est appelée information propre de x.

Nous avons vu jusqu'à présent des mesures d'informations d'événements individuels, intéressons nous maintenant aux estimations moyennes de tels événements. Considérons une source discrète, finie et stationnaire, les événements x_i peuvent alors être interprétés comme le choix d'un symbole parmi un alphabet $\Delta = \{x_1, x_2, ..., x_N\}$. Supposons de plus que les événements successifs sont mutuellement indépendants. Chaque émission de la source est alors décrite par une variable aléatoire *X* prenant ses valeurs dans l'alphabet considéré. A chacun de ces symboles est associé une probabilité :

$$P_i = P(X = x_i), \qquad i = 1, 2, ..., N \qquad avec \sum_{i=1}^{N} P_i = 1$$

La quantité d'information moyenne associée à chaque symbole est la moyenne de l'information propre de chacun des événements $X = x_i$, c'est à dire :

$$H(X) = E[h(X)] = \sum_{i=1}^{N} p_i \log\left[\frac{1}{P_i}\right] = -\sum_{i=1}^{N} P_i \log P_i \qquad \textbf{(6.10)}$$

où $E[.]$ désigne l'espérance mathématique. Cette quantité *H(X)* est également appelée entropie.

Considérons maintenant deux variables aléatoires *X* et *Y* prenant leurs valeurs dans deux alphabets distincts : $\mathrm{X} = \{x_1, x_2, ..., x_N\}$ et $\mathrm{Y} = \{y_1, y_2, ..., y_M\}$, on obtient alors l'entropie conjointe moyenne à partir de la définition précédente.

$$H(\mathrm{X},\mathrm{Y}) = E[h(\mathrm{X},\mathrm{Y})] = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log P(x_i, y_j) \qquad \textbf{(6.11)}$$

De la même façon nous obtenons l'entropie conditionnelle d'une variable aléatoire *X* étant donné la variable aléatoire *Y* :

$$H(\mathrm{X}\,|\,\mathrm{Y}) = E[h(\mathrm{X}\,|\,\mathrm{Y})] = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log P(x_i \mid y_j) \qquad \textbf{(6.12)}$$

Il est à noter que dans cette dernière équation, la probabilité en argument du logarithme est une probabilité conditionnelle. Elle est obtenue à partir des probabilités conjointes de x_i et y_j et de la probabilité marginale de y_j .

$$P(x_i \,|\, y_j) = \frac{P(x_i, y_j)}{P(y_j)} \qquad avec \quad P(y_j) = \sum_{i=1}^{N} P(x_i, y_j) \qquad \textbf{(6.13)}$$

Pour ce couple de variables aléatoires, une autre mesure peut être définie, il s'agit de l'information mutuelle moyenne définie de la manière suivante :

$$I.M(X\,;\,\mathrm{Y}) = E[i(X\,;\,\mathrm{Y})] = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log\left[\frac{P(x_i, y_j)}{P(x_i).P(y_j)}\right] \qquad \textbf{(6.14)}$$

où $\qquad P(x_i) = \sum_{j=1}^{M} P(x_i, y_j) \quad et \quad P(y_j) = \sum_{i=1}^{N} P(x_i, y_j)$

6.4.2 Génération de dictionnaires et estimation des modèles

Après avoir donné les principales définitions relatives à la théorie de l'information mutuelle, nous allons décrire maintenant l'algorithme d'apprentissage basé sur la maximisation de l'information

mutuelle. Cette description concerne essentiellement les étapes 4 et 5 de l'algorithme, les autres étapes sont identiques à celles de l'algorithme décrit précédemment relatif à l'approche hybride K-moyennes-DVQ. Soient les variables *Y* et *W*, l'information mutuelle $I.M(Y;W)$ peut être considérée comme la diminution de l'incertitude sur la variable *Y* provoquée par la connaissance de l'autre variable *W*, elle est ainsi donnée par :

$$I.M(Y;W) = H(Y) - H(Y|W) \qquad \textbf{(6.15)}$$

La variable *Y* représente dans notre cas la séquence de prototypes produite par le quantificateur vectoriel (le réseau de neurones) et *W* l'unité correspondante (dans notre cas le phonème).

H(Y) est l'entropie de la sortie du réseau désignée par la variable $Y=(y_1, y_2, ..., y_M)$, elle est donnée par :

$$H(Y) = -\sum_{m=1}^{M} P(y_m)\log P(y_m) \qquad \textbf{(6.16)}$$

$H(Y|W)$ est l'entropie conditionnelle, elle est donnée par :

$$H(Y|W) = -\sum_{n=1}^{N}\sum_{m=1}^{M} P(y_m, W_n)\log P(y_m|W_n) \qquad \textbf{(6.17)}$$

L'information mutuelle peut alors s'écrire :

$$F = I.M(Y;W) = \sum_{n=1}^{N}\sum_{m=1}^{M} P(y_m|W_n)\,P(W_n)\log\frac{P(y_m|W_n)}{P(y_m)} \qquad \textbf{(6.18)}$$

Différentes topologies de réseaux de types PMC à trois couches (une couche d'entrée, une couche cachée et une couche de sortie) ou des réseaux à deux couches (une couche d'entrée et une couche de sortie) peuvent être utilisées. Nous allons décrire l'algorithme pour le cas d'un réseau à deux couches dont la topologie est donnée par la **Fig. 6.3**.
La couche d'entrée du réseau utilisé possède un nombre de cellules égale à la taille D du vecteur acoustique $X=(x_1, x_2, ..., x_D)$ et la couche de sortie possède un nombre de cellules M égale à la taille désirée du dictionnaire.

La fonction d'activation est exprimée par la distance Euclidienne entre le vecteur acoustique et le prototype d'ordre m représenté par les poids de connexion w_{dm} de la cellule m de la couche de sortie, soit :

$$Z_m(k) = \|w_m - X\| = \sum_{d=1}^{D} (w_{dm} - x_d)^2 \qquad \textbf{(6.19)}$$

Pour chaque présentation d'un vecteur acoustique *k*, la cellule de sortie gagnante (celle qui présente la distance minimale) est activée à 1.0, toutes les autres cellules sont mises à 0.0.
La probabilité conditionnelle $P(y_m|W_n)$ du label m dans le flot Y résultant de la présentation des vecteurs acoustiques (L_n) du phonème W_n est calculée par la relation suivante :

$$P(y_m|W_n) = \frac{1}{L_w} \sum_{l=1}^{L_n} Z_m(l) \qquad \textbf{(6.20)}$$

la probabilité $P(y_m)$ du label m dans le flot y résultant de la présentation de tous les vecteurs acoustiques est donnée par l'équation suivante :

$$P(y_m) = \frac{1}{K} \sum_{k=1}^{K} Z_m(k) \qquad \textbf{(6.21)}$$

Ces probabilités et la probabilité $P(W_n)$ (probabilité a priori du phonème W_n) sont utilisées pour calculer l'information mutuelle donnée par l'équation **6.18**. La procédure d'entraînement modifie de manière itérative les poids w_{dm*} du label m*, label avec la plus grande fréquence d'occurrence, en utilisant l'équation suivante :

$$w_{dm*} = w_{dm*}(j-1) + \Delta w \qquad \textbf{(6.22)}$$

La variation de la fonction d'activation est alors calculée par l'équation **6.23** suivante :

$$\Delta Z_{m*}(k) = \Delta w(\Delta w + 2(w_{dm*} - x_d(k))) \qquad \textbf{(6.23)}$$

Le calcul des variations $\Delta Z_{m*}(k)$ de la fonction d'activation et des nouvelles probabilités $P(y_m)$ et $P(y_m|W_n)$ donne la variation de l'information mutuelle (ΔF). Si cette variation est positive, la modification des poids (équation **6.22**) est acceptée, si non la procédure est répétée avec la valeur négative (-*Δw*). L'entraînement du réseau est arrêté une fois tous les poids ont été visités pour modification. La procédure d'entraînement terminée, les poids de connexions des cellules représentent les prototypes du dictionnaire et les probabilités $P(y_m)$ représentent les densités de probabilité discrètes b_{jk}.

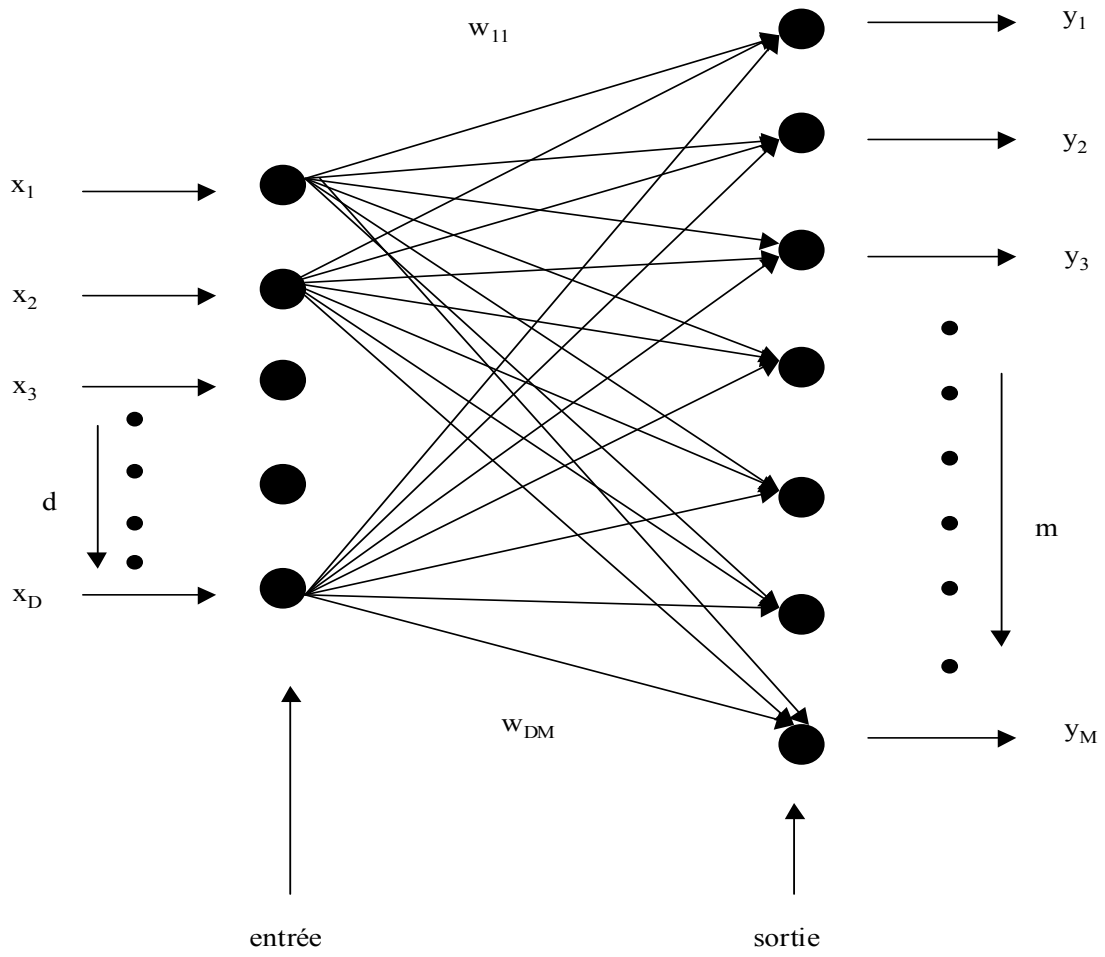


Fig. 6.3 - Topologie du réseau quantificateur à deux couches.

6.5 Expériences et résultats

Des expériences de reconnaissance des phonèmes arrières et emphatiques dans les deux modes, multi-locuteurs et indépendant du locuteur, ont été réalisées en utilisant ces deux approches de QVD. Pour le mode multi-locuteurs, le corpus de mots a servi à l'apprentissage des modèles et le corpus de phrases aux expériences de reconnaissance alors que pour le mode indépendant du locuteur, les prononciations de dix locuteurs ont servi à l'apprentissage et celles des dix autres aux tests. Le vecteur acoustique utilisé dans les deux approches est un vecteur à 36 composantes $\{MFCC(12), \Delta MFCC(12), \Delta\Delta MFCC(12)\}$. Le pas d'apprentissage qui a conduit aux meilleurs résultats et donc retenu pour la modification des poids du réseau dans l'approche RNA-QVD est $\Delta g = 0.05$. Différentes tailles de dictionnaires ont été utilisées pour une comparaison avec l'approche MMC à QV conventionnelle (MMCQVC). Les résultats comparatifs obtenus en mode indépendant du locuteur et en mode multi-locuteurs sont résumés respectivement par la **Fig. 6.4** et la **Fig. 6.5**. Ces figures représentent les taux de reconnaissance en fonction de la taille du dictionnaire.

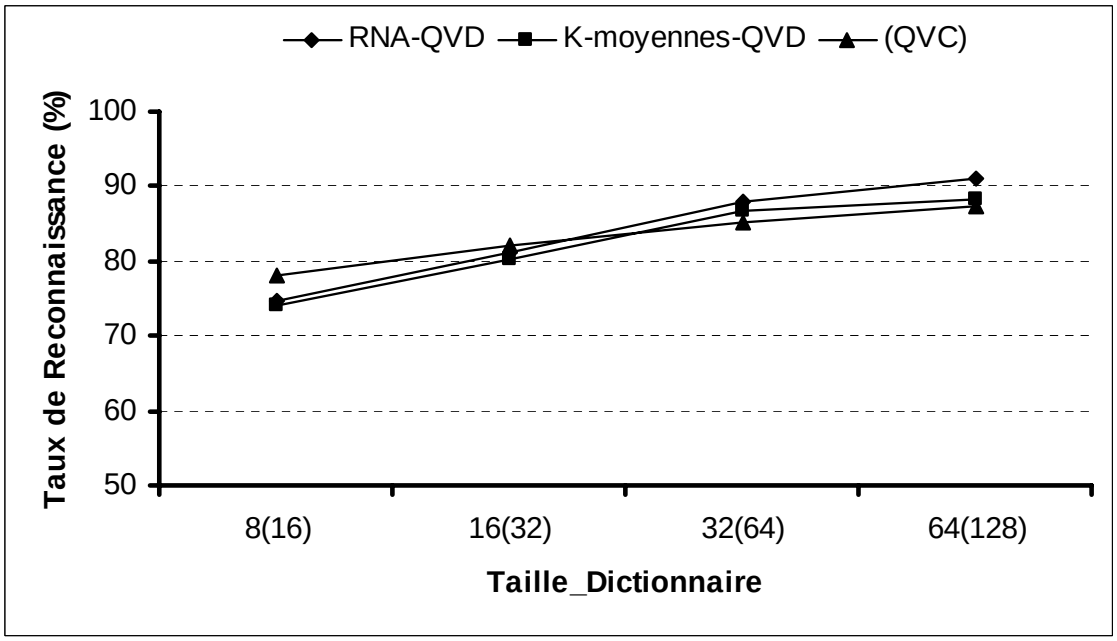


Fig. 6.4 – Comparaison des taux de reconnaissance en mode multilocuteurs : approches QVC, K-moyennes QVD et RNA-QVD.

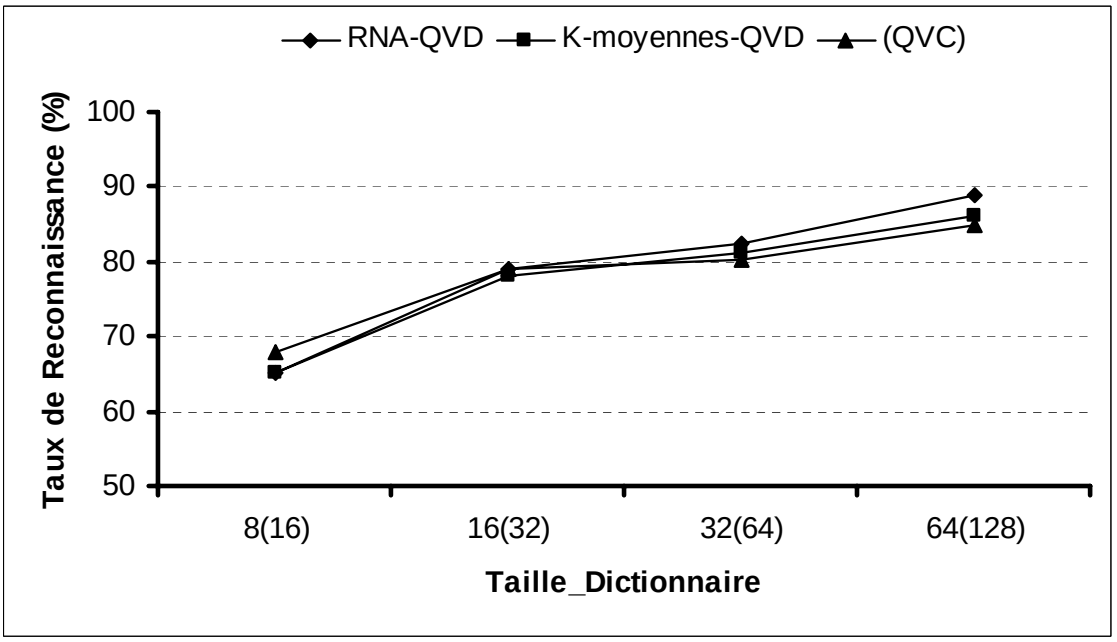


Fig. 6.5 – Comparaison des taux de reconnaissance en mode indépendant du locuteur : approches QVC, K-moyennes-QVD et RNA-QVD.

Nous avons également étudier le comportement de l'approche proposée avec le nombre de données d'apprentissage utilisées. C'est ainsi que des expériences de reconnaissance mode multi-locuteurs utilisant des bases de données d'apprentissage de plus en plus grandes ont été réalisées, le but étant de suivre l'évolution du taux de reconnaissance en fonction du nombre de données d'apprentissage. Quatre bases d'apprentissage (B.A) ont été utilisées B.A 1 (5 locuteurs x 5 répétitions), B.A 2 (10 locuteurs x 5 répétitions), B.A 3 (15 locuteurs x 5 répétitions) et B.A 4 (20 locuteurs x 5 répétitions). La taille du dictionnaire utilisé pour l'approche QVD est de 64 alors qu'elle est de 128 pour l'approche QVC. Les résultats de ces expériences sont résumés par les **Fig. 6.6**, **Fig. 6.7** et **Fig. 6.8** respectivement pour l'approche conventionnelle QVC, l'approche K-moyennes-QVD et l'approche RNA-QVD.

Commentaires

Une première constation est que l'augmentation de la taille du dictionnaire conduit souvent à une amélioration du taux de reconnaissance quelle que soit l'approche de QV utilisée.

Si nous analysons les résultats de ces expériences, nous pouvons aussi remarquer que pour des dictionnaires de taille inférieure à 32 (taille 64 pour l'approche QVC), les deux approches classique et distribuée donnent des taux de reconnaissance approximativement identiques. Au-delà de 32, un gain en terme de taux de reconnaissance est obtenu, de l'ordre de 2% pour le cas de la variante K-moyennes et de l'ordre de 4% pour le cas de la variante RNA.

Une autre constation est que l'augmentation du nombre de données d'apprentissage conduit bien souvent à une amélioration du taux de reconnaissance quelle que soit l'approche de QV utilisée. Cependant, pour l'approche QVC, à partir d'un certain nombre de données d'apprentissage (B.A 3, dans notre cas) le taux de reconnaissance n'évolue plus et reste pratiquement constant alors que pour l'approche QVD proposée, dans ces deux variantes, le taux de reconnaissance continue à augmenter. Ceci nous conduit à déduire que les résultats de cette approche restent encore perfectibles.

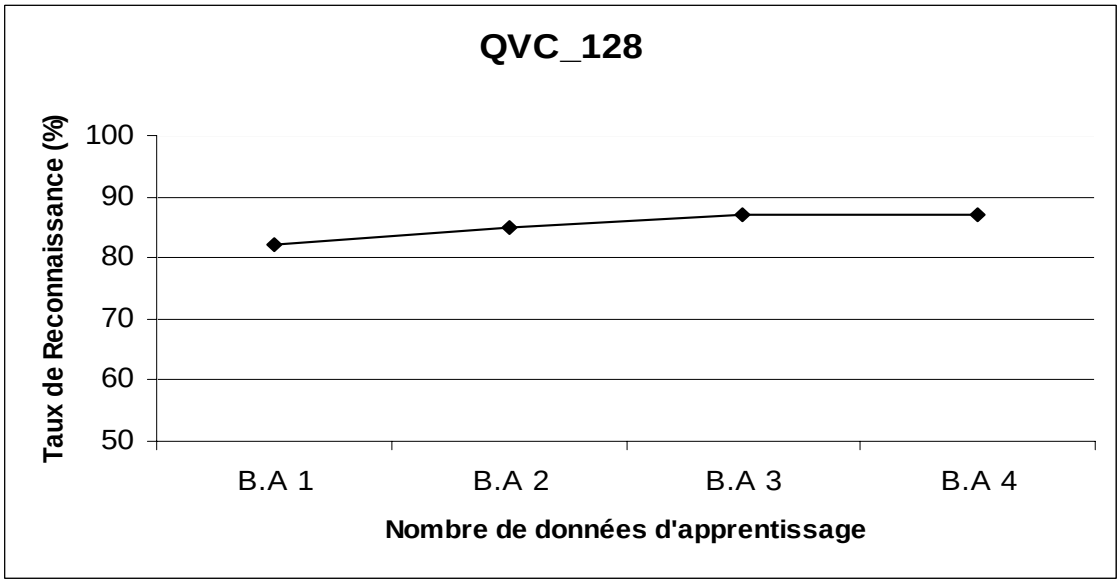


Fig. 6.6 - Taux de reconnaissance en fonction des données d'apprentissage: approche QVC.

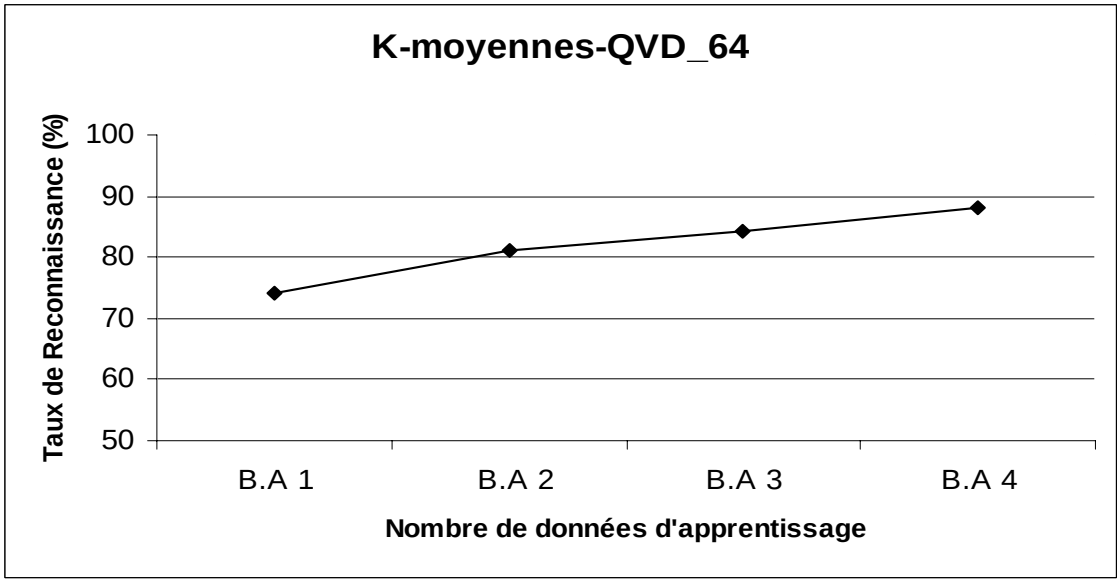


Fig. 6.7 - Taux de reconnaissance en fonction des données d'apprentissage: approche K-moyennes-QVD

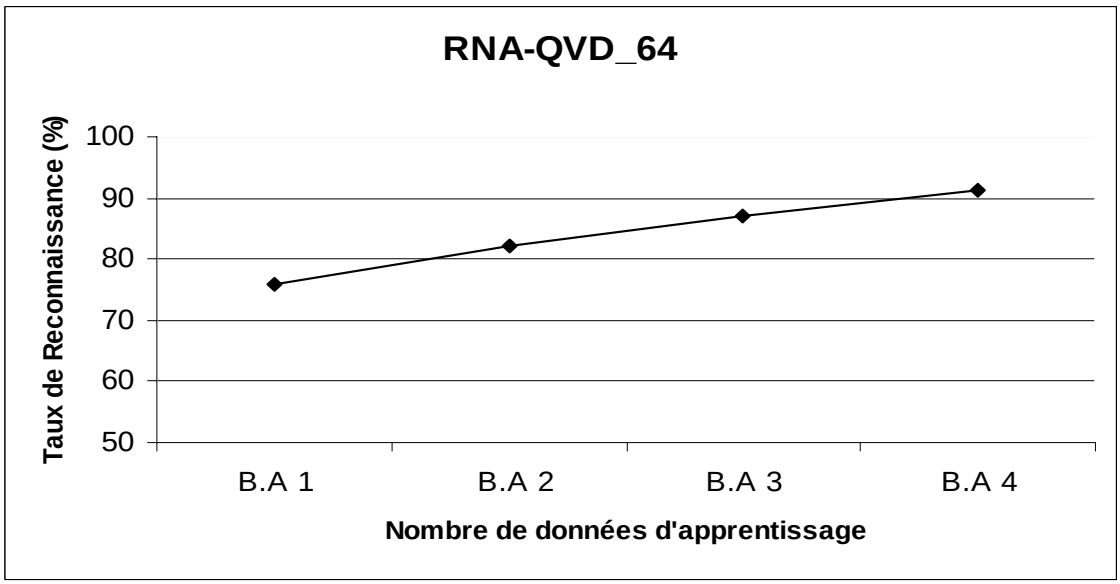


Fig. 6.8 - Taux de reconnaissance en fonction des données d'apprentissage: approche RNA-QVD

6. 6 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche de quantification vectorielle pour les modèles de Markov cachés discrets. Cette approche tente d'apporter des solutions aux problèmes des modèles discrets en particulier les problèmes inhérents à leur structure. Nous avons proposé deux variantes de quantification vectorielle distribuée, l'une utilisant l'algorithme des K-moyennes et l'autre des réseaux de neurones artificiels entraînés sur le principe de maximisation de l'information mutuelle. Des expériences de reconnaissance dans les deux modes multi-locuteurs et indépendant de locuteur ont été réalisées. L'analyse des résultats des expériences réalisées a montré que l'approche proposée augmente la performance des modèles discrets en terme de taux de reconnaissance et de vitesse de décodage étant donné le nombre de distances à calculer lors du processus de reconnaissance. De plus, de meilleurs taux de reconnaissance sont attendus si le nombre de données d'apprentissage augmente.

Conclusions et Perspectives

7.1 Conclusions

Dans le cadre de cette thèse, nous avons abordé le problème de la reconnaissance automatique de la parole en langue arabe standard. Nous nous sommes intéressés à la reconnaissance des phonèmes spécifiques à la langue arabe, des phonèmes reconnus unanimement comme responsables des limites des systèmes de reconnaissance dédiés à la langue arabe. Deux approches de reconnaissance ayant des stratégies complètement différentes ont été mises en œuvre.

La première approche est fondée sur des connaissances traduites par des règles de production gérées par un système expert. Le choix de cette approche était dicté par l'intérêt scientifique de disposer d'un système analytique de reconnaissance automatique de la parole. En effet, les systèmes à base de règles permettent, tout en validant des hypothèses théoriques issues de la linguistique et de la phonétique, d'établir une correspondance directe entre le signal acoustique et l'information linguistique véhiculée. Cette approche conditionnée par une caractérisation précise et complète de la langue d'étude possède néanmoins des retombées indéniables sur le plan académique. Cependant, à travers cette étude, nous avons constaté la difficulté à mettre en œuvre les systèmes à base de règles par le fait que l'expertise est rarement explicite, souvent loin d'être exhaustive et que l'exploitation des connaissances phonétiques nécessite la gestion d'un grand nombre de seuils empiriques, ce qui compromet l'utilisabilité du système.

Afin de nous affranchir de la gestion complexe du nombre élevé de paramètres et de règles, nous nous sommes tout naturellement orientés vers la mise en œuvre de l'approche globale basée sur la modélisation statistique des formes notamment les modèles de Markov cachés (MMC). C'est ainsi qu'un système de reconnaissance basé sur les MMCs combinés à la quantification vectorielle a été mis en œuvre. Au regard des résultats comparatifs obtenus, nous avons conclu à la suprématie de l'approche MMC. Elle est plus performante, moins coûteuse et mieux adaptée pour l'identification des phénomènes de forte distorsion temporelle et spectrale.

En utilisant les MMCs, nous avons également tenté d'en améliorer la performance par une étude complète portant sur le type de modèle (modèles dépendants et indépendants du contexte), la nature du vecteur acoustique d'entrée (différents types de paramétrisation et introduction de l'information dynamique) et l'utilisation de dictionnaire spécifique à chaque type de composante du vecteur d'entrée. Les conclusions aux quelles nous avons aboutit sont :

- La paramètrisation MFCC représente mieux l'enveloppe spectrale et l'introduction de la vitesse (paramètre différentiel d'ordre 1) et de l'accélération (paramètre différentiel d'ordre 2) augmente la discrimination des modèles.
- Les modèles dépendants du contexte où les effets de la coarticulation sont explicitement pris en compte confèrent également plus de discrimination au modèle.
- L'utilisation de dictionnaires spécifiques à chaque composante du vecteur d'entrée introduit mieux l'information au sein du système permettant ainsi d'augmenter sa performance.
- Les expériences de reconnaissance réalisées en tenant compte de tous ces aspects ont montré qu'un gain appréciable (de l'ordre de 5%) peut être obtenu pour le taux de reconnaissance global.

Nous avons aussi étendu notre étude au cas du système MMC à dictionnaires multiples où chaque modèle est doté de son propre dictionnaire. Les expériences réalisées avec ce type de système ont montré que l'utilisation de dictionnaire spécifique à chaque modèle donc à chaque unité à reconnaître améliore la performance du système, un gain de l'ordre de 3% en terme de taux de reconnaissance est

obtenu. Ceci peut s'expliquer d'une part, par une meilleure prise en compte de la variabilité interlocuteurs et d'autre part, l'introduction de l'erreur de quantification conduit à un critère de classification optimal. Mais nous devons reconnaître que le gain en terme de taux de reconnaissance est en fait obtenu au détriment de la vitesse de décodage des modèles.

Nous avons également étudié le paradigme multibandes des modèles de Markov cachés. Des expériences sur le choix des bandes de fréquence, leur nombre et leurs limites ont été réalisées. Le choix d'une stratégie de recombinaison des scores issus de chaque sous-bande a été aussi étudié étant donné qu'architecture et fusion sont intimement liées. Enfin, le comportement du modèle multibandes a été abordé dans un environnement bruité artificiellement. Les expériences réalisées ont conduit à des résultats qui confirment la robustesse des modèles multibandes par rapport aux modèles MMCs classiques pour une reconnaissance dans un tel environnement. Ces résultats rejoignent globalement ceux des travaux ayant traité ce paradigme pour des d'autres langues.

Nos recherches ont été, tout au long de ce travail, guidées par le souci permanent d'amélioration de la capacité de discrimination des modèles de Markov cachés. C'est ainsi que l'approche MMC à quantification vectorielle distribuée a été proposée. Elle tente de remédier aux insuffisances inhérentes à la structure des modèles discrets. Elle constitue à notre sens une tentative originale qui peut être généralisée pour une reconnaissance de parole continue grand vocabulaire.

7.2 Perspectives

Ce travail peut se prolonger sur de nombreuses directions qui recoupent les préoccupations actuelles de la communauté de l'ingénierie des langues.
D'abord, nous sommes conscients que ce travail présente des limites. La principale limite concerne la base de données utilisée. C'est une base de données locale, orientée en fonction des objectifs assignés à notre travail. L'utilisation d'une base de données normalisée de la langue arabe est donc nécessaire de notre point de vue afin d'asseoir véritablement les conclusions auxquelles nous avons aboutit. C'est un de nos objectifs à court-terme.

Dans ce travail nous avons testé dans le cas du système MMC à QV conventionnelle plusieurs types de vecteurs acoustiques d'entrées en particulier le vecteur multivariables. Ces techniques peuvent être généralisées et appliquées au paradigme multibandes et au modèle MMC à QV distribuée.

Un autre point qui peut être utile au modèle multibandes est l'établissement d'un lien avec l'approche analytique dans les stratégies de recombinaison et les pondérations associées à chaque sous-reconnaisseur en sous-bande. Les stratégies de recombinaison et les pondérations seraient alors dépendantes du phonème considéré : par exemple, poids de recombinaison plus élevé dans les hautes fréquences pour les fricatives.

Nous pensons également qu'il serait intéressant de faire coopérer dans le même système plusieurs sources de connaissances par exemple l'intégration dans le système MMC des connaissances prosodiques.

Notre intérêt pour les modèles de Markov cachés discrets trouve sa motivation dans la capacité de ces modèles à pouvoir être appliquer dans les nouvelles technologies de la communication (IP, GSM,..) étant donné le nombre de paramètres nécessaire à leur mise en œuvre. Dans ce sens, des travaux sur la reconnaissance des signaux transcodés GSM sont entamés.

Bibliographie

[1] Huang X.D., Hon H.W., Hwang M.Y. & Lee K.F., "A comparative study of discrete, semi continuous, and continuous hidden Markov models" *Computer Speech and Language*, Vol. 7, pp.359-368, 1993.

[2] Morgan N. & Bourlard H., "Continuous speech recognition" *IEEE Signal Processing Magazine*, Vol.12, N° 3, 1995.

[3] Q. Huo & C. Chan, "Contextual vector quantization for speech recognition with discrete hidden Markov model" *Pattern Recognition*, Vol. 28, N° 4, pp. 513-517, 1995.

[4] V. Digalakis, S. Tsakalidis, C. Harizakis & L. Neumeyer, "Efficient speech recognition using sub vector quantization and discrete-mixture HMMs" *Computer Speech and Language*, Vol. 14, pp. 33-46, 2000.

[5] F. Lefevre, "Non parametric probability estimation for HMM-based automatic speech recognition" *Computer Speech and Language*, Vol. 17, pp113-136, 2003.

[6] A. Bernard & A. Alwan, "Low-Bit-rate Distributed Speech Recognition for Packet-Based and Wireless Communication" *IEEE Transactions on Speech and Audio Processing*, Vol. 10, N° 8, pp. 570-580, 2002.

[7] R. Ethan, D. A. Subramaniam & B. D. Rao, "Improved quantization structures using generalized HMM modeling with application to wideband speech coding" *IEEE International Conference on Audio Speech and Signal Processing*, Vol. 1, Montreal, pp. 161-164, 2004.

[8] S. Dupont, H. Bourlard, and C. Ris, "Multi-stream speech recognition" *Tech. Rep IDIAP-PR 96-07*, IDIAP, Martigny, 1996.

[9] Glottin H., Tessier E., Bourlard H. et Berthommier F., "Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes" *IX ème Journées Neurosciences et Sciences de l'Ingenieur*, Munster, France, Mai 1998.

[10] Cerisara C., Haton J.P. et Fohr D., "Robust Behavior of Multi-band Paradigm" *Workshop on Robust Methods for Speech Recognition in Adverse Condictions*, Tampere, Finlande, Mai 1999.

[11] P. Flandrin, "Temps-fréquence, traité des nouvelles technologies" *Série Traitement du Signal*, Hermès, 1993.

[12] J.D Markel, A.H Gray Jr, "Linear prediction of speech" *Springer Verlag*, New York, 1976.

[13] H. Wakita, "Linear prediction of speech and its application to speech processing" *Speech Technology Laboratory, Phonetica*, Vol. 37, N° 1-2, 1980.

[14] R. Boite et M. Kunt, "Traitement de la parole" *Presses Polytechniques ROMANDES*, 1997.

[15] Calliope, "La parole et son traitement automatique" *Masson*, 1999.

[16] H. Hermansky, B.A Hauson et H. Wakita, "Low-dimentional representation of vowels based on all-pole modeling in the psychophysical domain" *Speech Communication*, Vol. 4, pp. 181-187, 1985.

[17] H. Hermansky, "Perceptual linear predictive analysis of speech" *Journal of the Acoustical Society of America*, Vol. 87, N° 4, pp. 1738-1752, 1990.

[18] M.R Schroeder, "Recognition of complex acoustic signal" *Life Science Research Report edited by T.H. Bullock*, 1977.

[19] Hermansky H. et Morgan N., "RASTA processing of speech" *IEEE Trans. on Speech and Audio Processing*, Vol. 2, N° 4, pp. 578-589, 1994.

[20] V.W. Zue, "The use of phonetic rules in automatic speech recognition" *Speech Communication*, Vol. 2, pp. 181-186, 1983

[21] D. Fohr, "APHODEX un système expert en décodage acoustico-phonétique de la parole continue" *Thèse de Docteur d'université Nancy I*, 1986.

[22] S.M O'Brien, "Knowledge-based system in speech recognition : a survey" *Int. J. Man-Machine Studies*, Vol. 38, pp. 71-95, 1993.

[23] F. Itakura, "Minimum production residual principle applied to speech recognition" *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 23, pp. 67-72, 1975.

[24] H. Sakoe and S. Chiba, "Dynamic programming algorithms optimization for spoken word recognition" *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 26, N° 1, pp. 43-49, 1978.

[25] J.S. Bridle, M.D. Brown, and R.M. Chamberlain, "An algorithm for connected word recognition" *In Proc. Int. Conf. Acoust.. Speech and Signal Processing*, pp. 899-902, 1982.

[26] H. Sakoe, "Two level DP-matching- A dynamic programming based pattern matching algorithm for connected word recognition" *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 27, pp. 588-595, 1979.

[27] C.S. Myers et L.R Rabiner, "Connected digit recognition using a level building DTW algorithm" *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 29, pp. 351-363, 1981.

[28] C.S. Myers et L.R Rabiner, "A comparative study of several dynamic time warping algorithms for connected-word recognition" *The Bell System Technical Journal*, Vol. 60, N° 7, pp. 1389-1409, 1981.

[29] L.R Rabiner et al, "Speaker independent recognition for isolated word using clustring techniques" *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 77, N° 4, pp. 336-349, 1977.

[30] K.L. Lee et H.W. Hon, "speaker-independent phoneme recognition using hidden markov models" *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-37, pp. 1641-1648, 1989.

[31] Euler S., "Clustring of gausian densities in hidden markov models" *In Laface, P. & De Mori, R., editors, Speech Recognition and Understanding*, Vol. F35 of NATO ASI Series, pp. 83-88, Springer Verlag, 1992.

[32] Rabiner L.R., Juang B., "A tutorial on hidden Markov Models and Selected applications in speech recognition" *Proceedings of the IEEE Trans. Speech Proces* , Vol. 77, N° 2 , pp. 257-285, 1989.

[33] Lee, C.H., Rabiner, L.R., & Pieracini, R., "Speaker-independent continuous speech recognition using continuous density hidden markov models" *In laface, P. & De Mori, R., editors, Speech Recognition and Understanding*, Vol. F35 of NATO ASI Series, pp. 135-8163, Springer Verlag, 1992.

[34] Juang, B.H., et Rabiner, L.R., "Issues in using hidden Markov models for speech recognition" *Technical report, Speech Research Department, AT&T Bell laboratories*, 1990.

[35] A.J Viterbi, "Error bounds for convolution codes and an asymptotically optimum decoding algorithm" *Proceedings of IEEE Transaction on Information Theory*, Vol. 13, N° 2, pp. 260-269, 1967.

[36] G.D Forney, "The Viterbi algorithm" *Proceedings of IEEE*, Vol. 61, pp. 268-278, March 1973.

[37] R.P Lippmann, "Review of neural networks for speech recognition" *Neural Computation*, Vol. 1, pp. 1-38, 1989.

[38] Widrow, B., Lehr, M.A., "30 Years of Adaptaive Neural Network : Perceptron, Madaline, and Backpropagation" *Proc. of IEEE*, Vol. 78, N° 9, pp. 1445-1442, 1990.

[39] D. Rumelhart, G. Hinton, and R. Williams, "Leraning internal representation by error propagation ," *In Rumelhart, D. and McClelland, J., editors, Parallel Distributed Processing*, Vol. 1, Chapter 8, pp. 318-362. MIT Press, Cambridge.

[40] A. Waibel, et al, "Phoneme recognition using time-delay neural networks" *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 328-339, March 1989.

[41] A. Waibel, H. Sawai, and K. Shikano, "Modularity and scaling in large phonemic neural networks" *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 37, pp. 1888-1898, 1989.

[42] H. Bourlard and C.J. Wellekens, "Speech dynamics and recurrent neural network" *Proceedings of Int. Conf. On Acoustics, Speech and Signal Processing*, Vol.1, pp. 33-36, Glasgow U.K, 1989.

[43] A.J. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system" *Computer Speech and Language*, Vol. 5, N° 3, pp. 259-274, July 1991.

[44] Y.D. Cho, et al "Extende Elman's recurrent neural networks for syllabe recognition" *Proceedings of The int. Conf. On Spoken Language Processing*, pp. 1057-1060, 1990.

[45] M. Gori, Y. Bengio, and R. de Mori, "BPS : A learning algorithm for capturing the dynamic nature of speech" *Proceedings of the Int. Joint Conference on Neural Networks*, Vol. 2, pp. 417-423, 1989.

[46] H. Bourlard and C. Welleekens, "Links between hidden Markov models and multilayer perceptron" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, pp. 1167-1178, 1990.

[47] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods" *IEEE Trans. on Neural Networks*, Vol. 4, N° 6, pp. 893-909, 1993.

[48] H. Bourlard and N. Morgan, "Connectionit Speech Recognition – A Hybrid approach" *Kluver Academic Publisher*, 1994.

[49] Y. Yan, M. Fanty, and R. Cole, "Speech recognition using neural networks with forward-backward probability generated targets" *In Int. Conf. On Acoustics, Speech and Signal Processing*, pp. 3241-3244, Munich, 1997.

[50] S. Renalds et al, "Connectionist probability estimators in HMM speech recognition" *IEEE Trans. on Speech and Audio processing*, Vol. 2 N° 1, pp. 1161-1174, 1994.

[51] T. Robinson, "An application of recurrent nets to phone probability estimation" *IEEE Trans. on Neural Networks*, Vol. 4, N° 2, pp. 298-305, 1994.

[52] M. Hochberg et al, "Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system" *In Proc. Of ICSLP*, pp. 1499-1502, 1994.

[53] H. Franco *et al*, "Context-dependent connectionist probability estimation in a hybrid hidden Markov model-neural net speech recognition system" *Computer Speech and Language*, Vol. 8, pp. 211-222, 1994.

[54] D. Kimber et al, "Speaker-independent vowel classification using hidden Markov models and LVQ2" *In Int. Conf. On Acoustics, Speech and Signal Processing*, pp. 497-500, 1990.

[55] H. Iwamada, S. Katagiri, and E. McDermott, "Speaker independent large vocabulary word recognition using an LVQ/HMM hybrid algorithm" *In Int. Conf. On Acoustics, Speech and Signal Processing*, pp. 553-556, Toronto 1991.

[56] T. Kohonen, "Self-Organisation and Associative Memory" *Springer-Verlag*, Berlin, 1989.

[57] H.P Hutter, "Comparison of a new hybrid connectionist-SCHMMM approach with other hybrid aproaches for speech recognition" *In Int. Conf. On Acoustics, Speech and Signal Processing*, pp. 3311-3314, Detroit 1995.

[58] C. Neukirchen and G. Rigoll, "Advanced training methods and new network topologies for hybrid MMI-connectionist/HMM speech recognition systems" *In Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 3257-3260, Munich, 1997.

[59] C.S. Jang and C.K. Un, "A new parameter smoothing method in the hybrid TDNN/HMM architecture for speech recognition" *Speech Communication*, Vol. 19, N° 4, pp. 317-324, 1996.

[60] P.L Cerf, W. Ma, and D.V. Compernolle, "Multilayer perceptrons as labelers for hidden Markov models" *IEEE Trans. on Speech and Audio Processing*, Vol. 2, N° 1, pp. 185-193, 1994.

[61] Y. Bengio et al, "Global optimization of a neural network-hidden Markov model hybrid" *IEEE Trans. on Neural Networks*, Vol. 3, N°2, pp. 252-259, 1992.

[62] F.T. Johansen, "Global optimization of HMM input transformations" *In Proc. Of ICSLP*, Vol. 1, pp. 239-242, Yokohama, 1994.

[63] S.K. Riis and A. Krogh, "Hidden Neural networks : a framework for HMM/NN hybrids" *In Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 3233-3236, Munich, 1997.

[64] Bonnot J.F, "Etude expérimentale de certains aspects de la gémination et de l'emphase en arabe" *Travaux de l'institut phonétique de Strasbourg*, Vol. 11, pp. 109-118, 1979.

[65] Debyeche. M & al, "Knowledge Based Approach for Arabic back consonant recognition in continuous speech" *Proceeding of the First KFUPM Workshop on Information & Computer Science*, Dhahran, Saudi Arabia, pp. 137-141, June 9 1996.

[66] Debyeche. M *& al*, "A Knowledge-Based Approach for Arabic Emphatic Consonant Identification Based on Speech Spectrogram reading" *30th IEEE Southeastern Symposium on System Theory (SSST)*, March 8-10, West Virginia University, Morgantown West Virginia, 1998, ISBN: 0-7803-4547-9

[67] El-ani S.H, "Arabic phonology : an acoustical and physiological investigation" *Eds. Mouton, The Hague*, 1970.

[68] Dellatre P., "Pharyngeal feature in consonants of Arabic, German, Spanish, French an American English" *Phonetica*, vol.23, pp. 129-155, 1971.

[69] Ghazali S., "Elements of Arabic phonetics" *Applied Arabic linguistics and Signal Information Processing*, pp. 51-58, 1987.

[70] Ghazali S. "Etude emg préliminaire sur les consonnes arrières de l'arabe" *16ème Journées d'Etude sur la Parole J.E.P SFA* Hammamet Tunisie, pp. 286-289, 1987.

[71] Betari A. "Caractérisation des phonèmes de l'arabe en vue d'une reconnaissance automatique de la parole" *Thèse de Doctorat*, Aix-en-Provence, 1993.

[72] Emam O. "Speech Recognition of Arabic" *Technical report of IBM* Cairo Scientific Center, 1997.

[73] Selouani S.A et Caelen J. "Recognition of phonetic features using neural networks and knowledge-based system" *International Journal on Artificial Intelligence Tools, World Scientific Publishing ed*. ISSN. 0218-2130. Vol.8, N°.1, pp. 73-103, 1999.

[74] K. Kirchhoff & al, "Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Workshop" *Proceedings of the Int. Conf. on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003.

[75] Debyeche. M *& al*, "APHAK: Un logiciel Interactif d'Analyse du Signal de Parole" *17ème journées d'électrotechnique et d'automatique de Tunis*, 5-6 Novembre 1997.

[76] Léon L., "Traitement d'algorithmes par ordinateurs" *Cepadues Edition*, Tome 2, pp. 321-333, 1983.

[77] Dubnowsky J.J, Shaffer R.W and Rabiner L.R., "Real-time digital harqware pitch detector" *IEEE Trans. Acoust. and Signal Processing*, Vol. ASSP 24, pp. 2-6, 1976.

[78] L.E Baum, "An unequally and associated maximization technique in statistical estimation of probabilistic functions of Markov processes*"* *Inequalities*, pp. 1-8, 1972.

[79] J.D. Markel et A.H. Gray, "Linear prediction of speech" *Springer Verlag*, New York, 1976.

[80] Makhoul J., Roucos I. and Gish H., "Vector quantization in speech coding" *Proceedings IEEE*, Vol. 73, pp. 1551-1588, 1985.

[81] A. Likas, N. Vlassis & J. J. Verbeek, "The global k-means clustering algorithm" *Pattern Recognition, 36(2)*, pp451-461, 2003.

[82] Y. Linde, A. Buzo & R.M. Gray, "An algorithm for Vector Quantizer" *IEEE Trans. on Communication , Vol.* 28, N° 1 , 1980.

[83] S. Davis & P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*"* *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 28, N° 4, pp. 357-366, 1980.

[84] Darouault A.M., "Context-dependent phonetic Markov models for large vocabulary speech recognition" *In Proceedings of the Int. Conf. on Acoustics, Speech and Signal Processing*, Dallas, pp. 360-363, 1987.

[85] K.F. Lee, "Automatic speech recognition. The development of the SPHINX system*"* *Kluver Publishers*, Boston, USA, 1989.

[86] J.J Odell, "The use of context in large vocabulary speech recognition" *PhD Thesis*, University of Cambridge, U.K, March 1995.

[87] R. Schwartz, Y.L. Chow, O.A. Kimball, S. Rouscos and, L. Makhoul, "Context dependent modeling for acoustic-phonetic recognition of continuous speech" *Proceedings of the Int. Conf. On Acoustics, Speech and Signal Processing*, Vol. 1, pp. 1203-1208, April 1985.

[88] K.F. Lee, "Context dependent phonetic hidden Markov models for speaker independent continuous speech recognition" *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 38, pp. 347-365, 1990.

[89] A. M Peinado, & al, "Discriminative Codebook Design Using Multiple Vector Quantization in HMM-Based Speech Recognition" *IEEE Trans. on Speech and Audio Processing*, Vol. 4, N° 2, 1996.

[90] H. Fletcher, "Speech and Hearing in communication" *Krieger* , 1953, New York.

[91] J.B Allen, "How do human process and recognize speech ?" *IEEE Transaction ASSP*, Vol. 2, N°. 4, pp. 567-577, 1994.

[92] J .B. Allen, "Harvey Fletcher ‘s role in the creation of communication acoustics" *Journal of the Acoustical Society of America*, Vol. 99, N° 4, pp. 1825-1839, 1996.

[93] T. Arai et al, "Intelligibility of speech with filtered time trajectories of spectral envelopes" *In Proceedings of the Int. Conf. on Spoken Language Processing*, Vol. 4, pp. 2490-2493, Philadelphia 1996.

[94] G. Miller et P. Nicely, "An analysis of perceptual confusion among some English consonants" *Journal of the Acoustical Society of America*, Vol. 27, N° 2, pp. 338-352, 1995.

[95] K. Kryter, "Speech bandwidh compression through spectrum selection" *Journal of the Acoustical Society of America*, Vol. 32, N° 5, pp. 547-556, 1960.

[96] P. Duchnowsky, "A new structure for automatic speech recognition" *PhD. Thesis, Massachusetts Institute of Technology*, Cambridge, USA, 1993.

[97] Bourlard, H., Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands" *In Proceedings ICSLP*, Vol. 1, pp. 426-429, Philadelphia, USA, 1996.

[98] Bourlard H., Dupont S., "Subband-based speech recognition" *In Proc. IEEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1251-1254, Munich, Germany, April 1997.

[99] S. Tibrewala et H. Hermansky, "Sub-band based recognition of noisy speech" *In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1255-1258, Munich, Germany, 1997.

[100] L. Xu, A. Kryzak et C.Y. Suen, "Methods of combining multiple classifiers and their applications" *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 22, N° 3, pp. 418-435, 1992.

[101] S. Hashem, "Optimal linear combination of neural networks" *Neural Networks*, Vol. 10, N° 4, pp. 599-614, 1994.

[102] A. Adjouani et C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR" *In Speech Reading by Humans and Machines*, Vol.150 of NATO ASI Series, Series F : Computer and Systems Sciences, pp. 461-471, Springer Verlag Berlin, 1996.

[103] M. Pavel et H. Hermansky, "Information fusion by human and machine" *Proceedings of the First European Conference on Signal Analysis and Prediction*, pp. 350-353, 1997.

[104] T.K. Ho, J.J. Hull et S.N. Srihari, "Decision combination in multiple classifier system*" IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 16, N° 1, 1994.

[105] Amrouche A., Debyeche M, Adoul. K, Amrouch. K, Rouvaen J.M, "Reconnaissance des phonèmes par réseaux de neurones et normalisation temporelle: Application aux consonnes pharyngales et glottales Arabes" *XXIIèmes Journées d'étude sur la parole (JEP'98)*, 15-19 Juin, Matigny-Valais-Suisse, 1998.

[106] Debyeche. M *& al*, "Phoneme Recognition System based on HMM with Distributed VQ Codebook" *6th European Conference on Speech Communication and Technology*, EUROSPEECH, 6-9 September 1999, Budapest.

[107] Debyeche M., Haton J.P, and A. Houacine "A New Vector Quantization front-end Process for Discrete HMM Speech Recognition System" *IJSP: International Journal of Signal Processing,* Vol. 3, N°. 1, pp. 46-51, 2006, ISSN 1304-4478.

[108] Dempster A.P., Laird N.M., and Rubin D.B., "Maximum likelihood from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society*, Series B, Vol. 39, N° 1, pp. 1-39, 1977.

[109] C. Cerisara, J.P. Haton, J.F. Mari, et D. Fohr, "A recombinaison model for multi-band speech recognition" *ICASSP'98,* Seattle, USA, 1998.

[110] H. Hermansky et S. Sharma, "TRAPS- Classifiers of Temporal Patterns" *ICSLP'98*, Sidney, 1998.

[111] Jelinek F., "Statistical Methods for Speech Recognition" *MIT Press, Cambridge Massachusetts*, 1998.

[112] L.R. Bahl, F. Jelinek, and P.L. Mercer, "Amaximum likelihood approach to continuous speech recognition" *IEEE Trans. on Pattern Analysis and Machine Recognition*, Vol. 5, N° 2, pp. 179-190, 1983.

[113] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification" *John Wiley & Sons*, 2nd edition, New York, USA, 2001.

[114] Rabiner L.R, Huang B.H., "An introduction to hidden Markov models" *ASSP Magazine*, pp. 4-16, 1986.

[115] Rabiner L.R., Juang B.H., "Fundamentals of speech recognition" *PTR Prentice Hall*, 1993.

[116] Bahl et al, "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy" *Trans. IEEE Speech and Audio Processing*, Vol.1 N° 1, pp. 77-83, 1993.

[117] S.D.Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum" *IEEE Trans on Acoustics, Speech, and Signal Processing*, ASSP Vol. 34, N° 1, February 1986.

[118] J.W. Picone, "Signal Modeling Techniques in Speech Recognition" *Proceedings of the IEEE*, Vol. 81, N° 9, September 1993.

[119] A. V. Oppenheim and R. W. Schafer, "Discrete Time Signal Processing" *Prentice-Hall, Inc. Englewoods cliffs*, New Jersey, 1988.

[120] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals" . *Prentice-Hall, Inc., Englewood Cliffs*, New Jersey 07632, 1978.

[121] A. Gersho and R. Gray, "Vector quantization and signal compression" *Kluwer Academic Publishers,* Boston, 1997.

[122] OROS "Manuel et guide d'utilisation" 1990.

[123] N. Gilbert "Statistiques" *Livre , Editions* HRW, 1978.

[124] Stéphanie S. Mc candless, "An algorithm for automatic formant extraction using linear prediction spectral" *IEEE Trans.* Vol. 22 N° 2, pp. 135-141, 1974.

[126] Y. Laprie, "Notice d'utilisation de SNORRI" *Technical Report* CRIN, 1988.

[127] T. Cover et J. Thomas, "Elements of information theory" *John Wiley & Sons,* Inc, New York, 1991.

[128] G. Battail, "Théorie de l'information : application aux techniques de communication" *Ed. Masson,* Paris, France, 1997.

APHAK : Un Logiciel Interactif d'Analyse de la Parole

Résumé

Le présent travail **APHAK** (**A**cquisition **PH**onetic **A**coustic **K**nowledge pour l'anglais) est un logiciel d'analyse du signal de parole. Il s'adresse aux chercheurs qui souhaitent acquérir des connaissances acoustico-phonétiques. Il a été développé en C++ sous Windows. APHAK offre au chercheur un ensemble de fonctions telles que la manipulation du signal, l'analyse spectrale, l'extraction des paramètres acoustiques, la segmentation & l'étiquetage et l'étude statistique. Par sa modularité, APHAK est un produit évolutif, d'utilisation facile**.**

1 Introduction

Les études sur la parole telles que la reconnaissance, la synthèse, l'identification de locuteurs, … nécessitent des outils informatiques permettant d'analyser finement le signal vocal. Le présent travail se veut comme une réponse à cette nécessité. APHAK est un ensemble de programmes C++ sous l'environnement Windows permettant de réaliser des acquisitions/restitutions de signaux de parole, la visualisation temporelle de ces signaux, des traitements spécifiques énergie, contour d'énergie, taux de passage par zéro, formants, pitch, coefficients MFCC, etc. Il est possible aussi de réaliser le suivi temporel des résultats de ces différents traitements. Ce logiciel permet également l'observation et la manipulation de spectrogrammes numériques calculés par FFT, par LPC et par Cepstre. La segmentation et l'étiquetage de phrases ainsi que l'étude statistique sur les résultats des différents traitements disponibles sont des fonctions qui peuvent égalent être réalisées avec ce logiciel.

2 Configuration matérielle

La réalisation d'un logiciel d'analyse de la parole nécessite une puissance de calcul et d'espace mémoire considérable que se soit pour le développement ou l'utilisation. La configuration matérielle minimale pour l'exploitation du logiciel APHAK serait un micro-ordinateur (compatible IBM) i80486, 4Mb de mémoire vive et éventuellement une carte de traitement du signal de type OROS AU21 **[122]** pour l'acquisition et la restitution des signaux de parole.
APHAK est un logiciel développé en C++, l'approche orientée objet est un ensemble de concepts de programmation visant un objectif commun, le développement de logiciel ou plus précisément de modules réutilisables.
Les fonctions du logiciel APHAK sont accessibles avec la souris ou avec le clavier, APHAK exploite les ressources du multi-fenêtrage en associant une fenêtre au signal temporel, au spectrogramme, au zoom, etc. L'utilisateur peut réduire, agrandir, sauvegarder et restaurer en cas de besoin une image donnée.

3 Schéma Fonctionnel

La réalisation d'un logiciel d'analyse des signaux de parole est complexe, sa structuration en plusieurs modules est donc nécessaire. Cela donne une meilleure vision des différentes parties qui le constituent. La conception modulaire accroît donc les performances du logiciel tout en réduisant le temps nécessaire à son développement. D'où le schéma fonctionnel suivant :

4 Présentation du logiciel

Il s'agit dans cette partie d'une présentation des différents menus du logiciel.

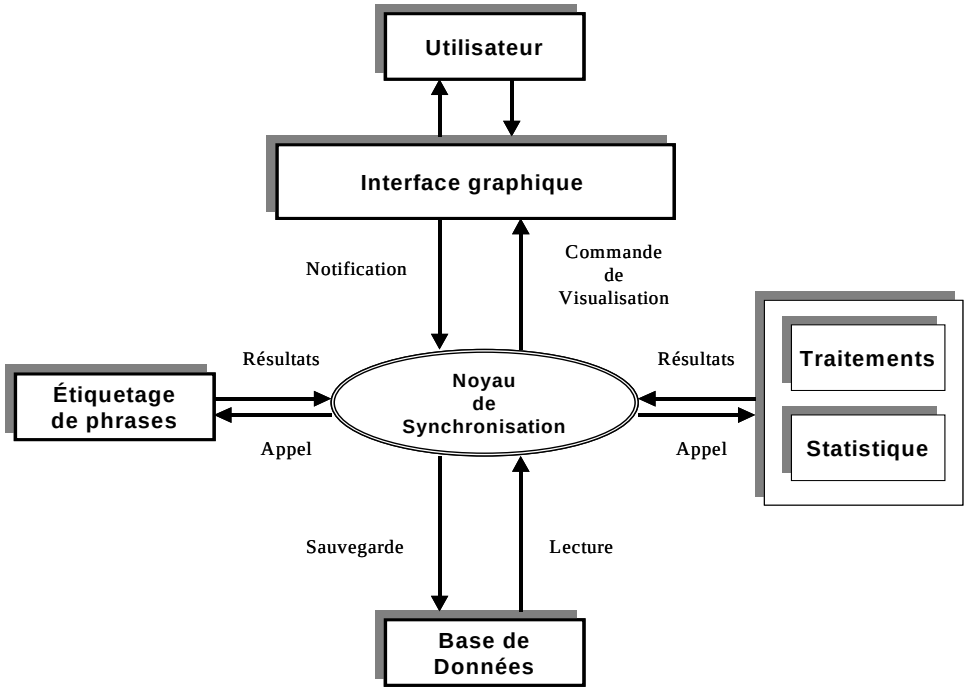
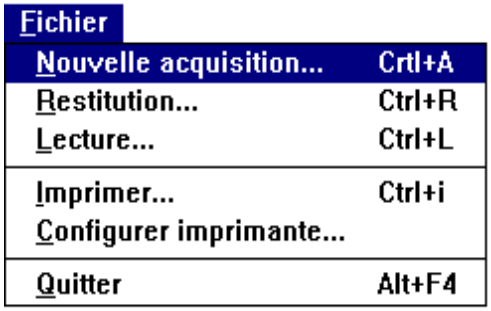


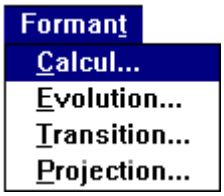
Fig. A.1- Schéma fonctionnel du logiciel

4.1 Les menus du logiciel

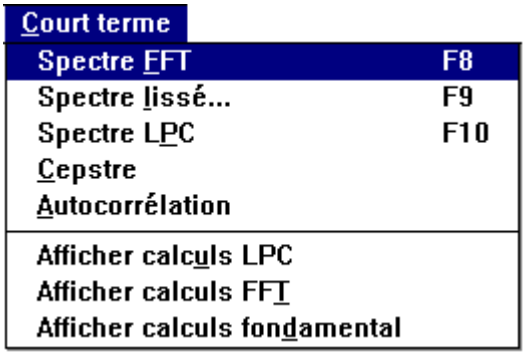
Le menu Fichier



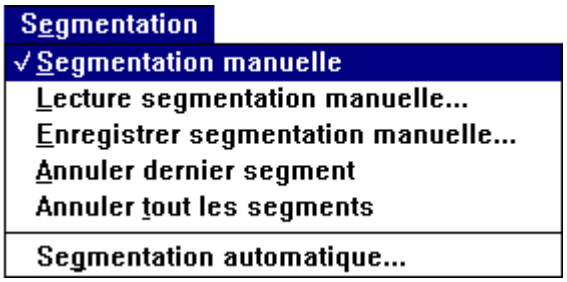
Le menu Formant



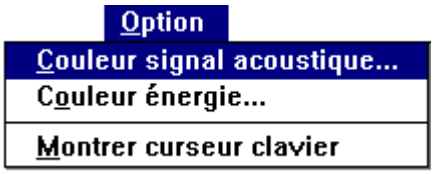
Le menu Court terme



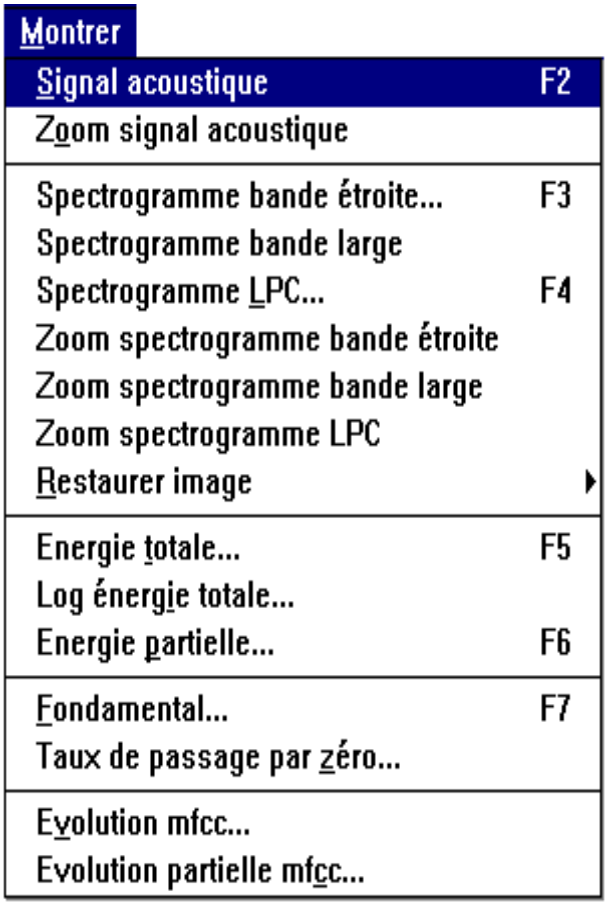
Le menu Segmentation



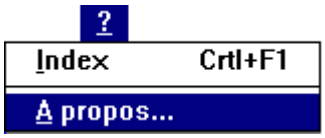
Le menu Option



Le menu Montrer



Le menu Help (?)



La ligne d'états

C'est une bande statique (l'utilisateur ne peut intervenir sur cette barre) qui visualise

- La description des différentes commandes des menus ainsi que les messages du logiciel destinés à l'utilisateur.
- La fréquence d'échantillonnage du fichier de parole en cours d'analyse.
- La durée du fichier de parole (millisecondes) en cours d'analyse.
- L'heure courante qui permet d'estimer la durée des calculs.

4.2 Convention du menu

- Le nom de commande est grisé : Vous ne pouvez pas exécuter cette commande. Vous devez activer une autre commande avant de l'utiliser.

- Une coche (√) apparaît à gauche du nom de commande : La commande ou la fenêtre est active.

- Des points de suspensions apparaissent à droite du nom de commande : Cette commande fait apparaître une boite de dialogue, demandant l'utilisateur la saisie des paramètres nécessaires à l'exécution de la commande.

- Un triangle (▷) apparaît à droite du nom : La rubrique mène à un sous menu en cascade, qui propose à son tour une autre liste de commandes.

- Le texte se trouvant à droite de chaque commande indique que celle-ci est aussi actionnable à partir du clavier en combinant les touches indiquées. (**Ex :** Ctrl+A pour une acquisition).

4.3 Le menu système

Le menu système apparaît dans toutes les fenêtres ouvertes, les icônes sélectionnées, les boites de dialogue ainsi que dans la fenêtre principale. Il est activé par la combinaison des touches **"Alt-Espace"**. Ce menu permet :

- De redonner à une fenêtre sa taille initiale.
- De déplacer la fenêtre ou de la fermer.
- De changer la taille de la fenêtre.
- De réduire la fenêtre en icône ou de l'agrandir.
- De basculer vers une autre application.

4.4 Boite de dialogue

Une boite de dialogue est une fenêtre contenant des champs de saisies, des options au choix multiples, des boutons de validation et d'annulation. Pour se déplacer dans une boite de dialogue, il suffit :

- D'appuyer sur la touche Tabulation à chaque déplacement.
- De cliquer sur le contrôle de la boite de dialogue.
- D'appuyer sur la touche "Alt" suivi en même temps de la lettre soulignée.
- d'appuyer sur les touches de direction pour se déplacer dans une zone.

5 Fonctions principales du logiciel

5.1 Signal temporel

La **Fig. 2** montre le signal temporel tel qu'il est acquis par APHAK à travers la carte de traitement du signal OROS AU21. La fréquence d'échantillonnage est de 10 KHz. Mais il est possible de l'augmenter jusqu'à 40 KHz. Chaque acquisition est automatiquement stockée sur disque sous un format propre. Le fichier de sauvegarde sera composé de la fréquence d'échantillonnage suivie des données de parole. L'utilisateur peut restituer le signal stocké. Il peut aussi examiner avec précision une partie du signal grâce à un zoom dont il précise l'étendue avec la souris ou avec le clavier.

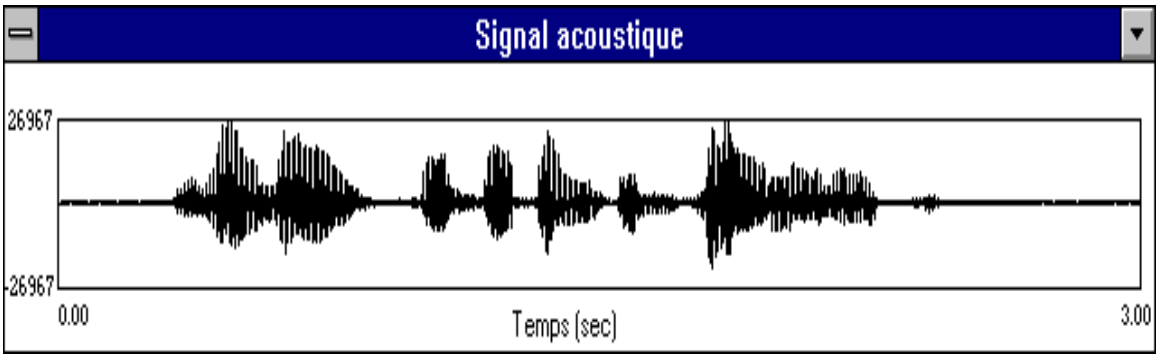


Fig. A.2- Signal temporel.

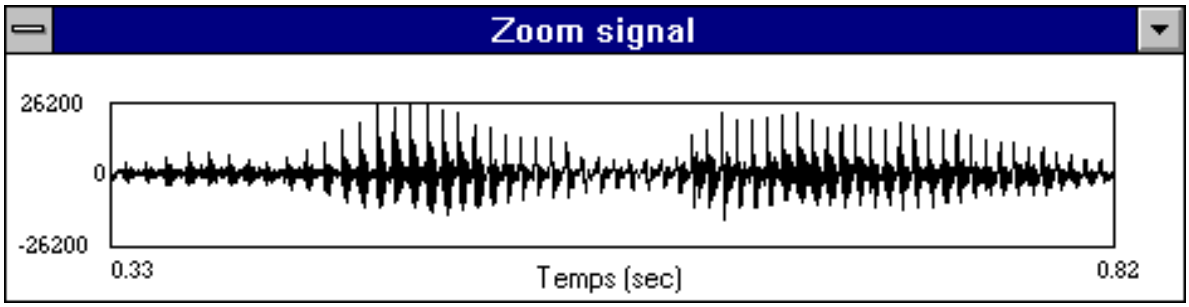


Fig. A.3- Zoom d'un signal temporel.

5.2 Analyse Spectrale

La **Fig. 4** montre des spectres calculés par FFT et par LPC. Ces spectres peuvent être obtenus sur tout le signal temporel, il suffit à l'utilisateur de choisir le début de la partie à analyser, une fois ce choix effectué une fenêtre de 256 (ou 128) échantillons est prélevée et le calcul est activé.

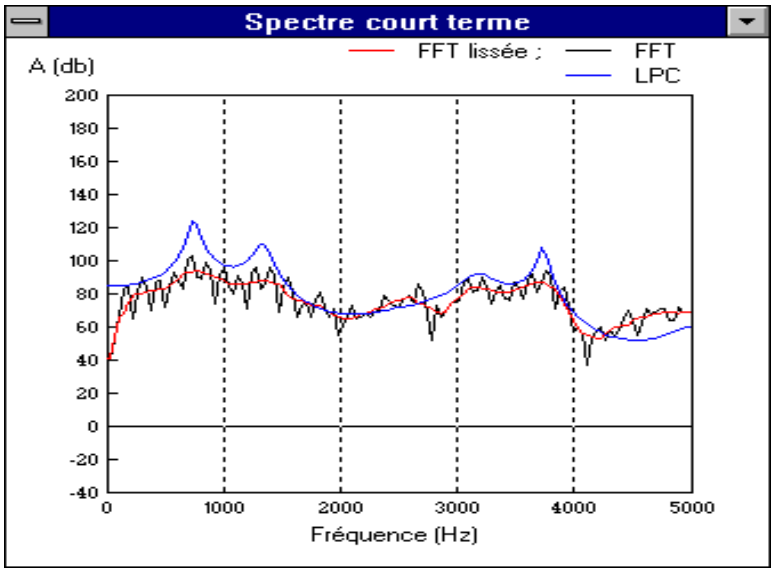


Fig. A.4 - Spectres à court-terme par FFT, LPC et FFT lissée.

La représentation par spectrogramme (évolution temporelle du spectre à court-terme) est possible dans les deux modes (bande étroite et bande large). Les **Fig. 5** et **6** en donnent des exemples. L'utilisateur peut aussi observer un spectrogramme local pour une meilleure définition grâce à un zoom dont il précise l'étendu.

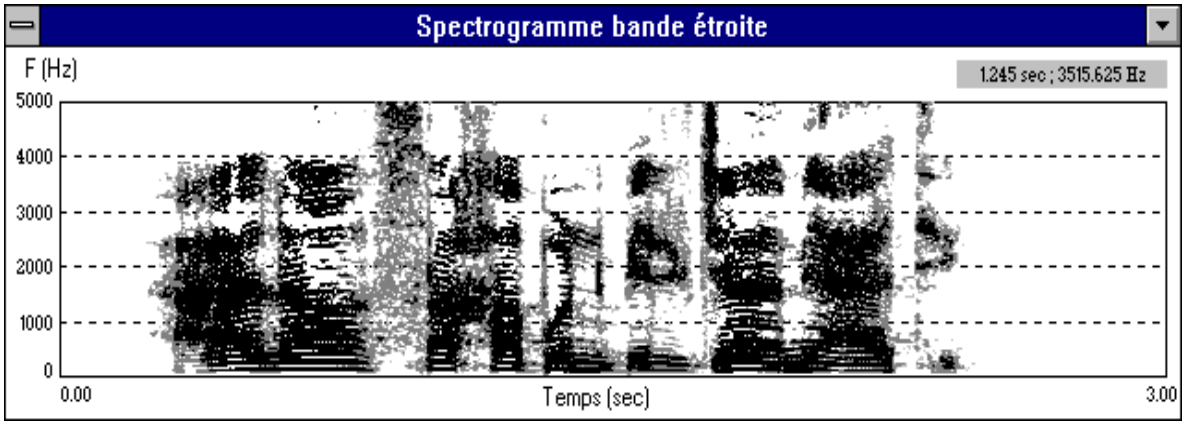


Fig. A.5 – Spectrogramme obtenu par FFT à bande étroite.

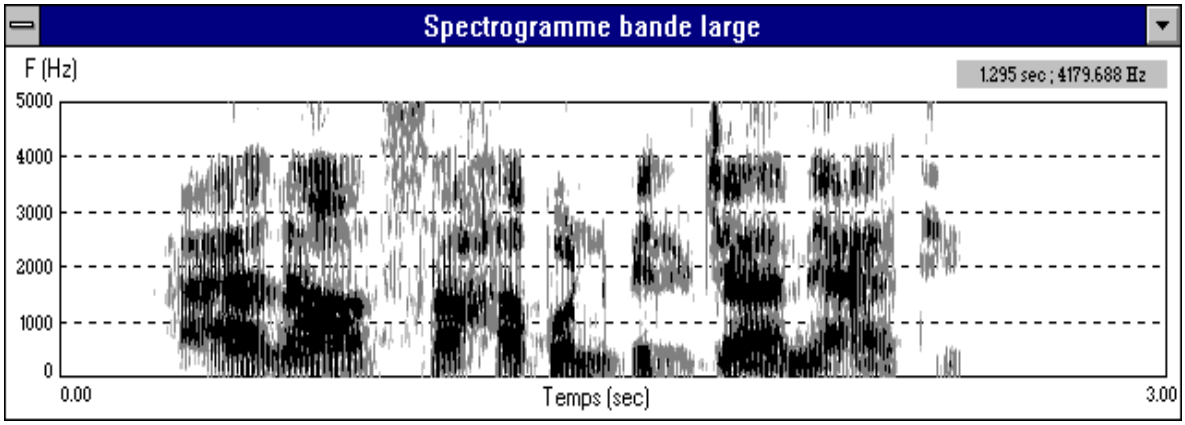


Fig. A.6 – Spectrogramme obtenu par FFT à bande large.

5.3 Paramètres Extraits

APHAK permet l'extraction d'un certain nombre de paramètres acoustiques pertinents pour la recherche dans le domaine du traitement automatique de la parole. Ces paramètres sont principalement les formants, le pitch, les coefficients MFCC (Mel Frequency Cepstral Coefficient). Pour les formants, l'utilisateur peut effectuer leur calcul, leurs évolutions temporelles sur le spectrogramme, faire des projections et estimer par une méthode de régression linéaire **[125]** leurs transitions.

De même que l'utilisateur peut effectuer un calcul du pitch et des coefficients MFCC sur l'ensemble du signal ou uniquement une partie. Il peut observer leurs évolutions temporelles.

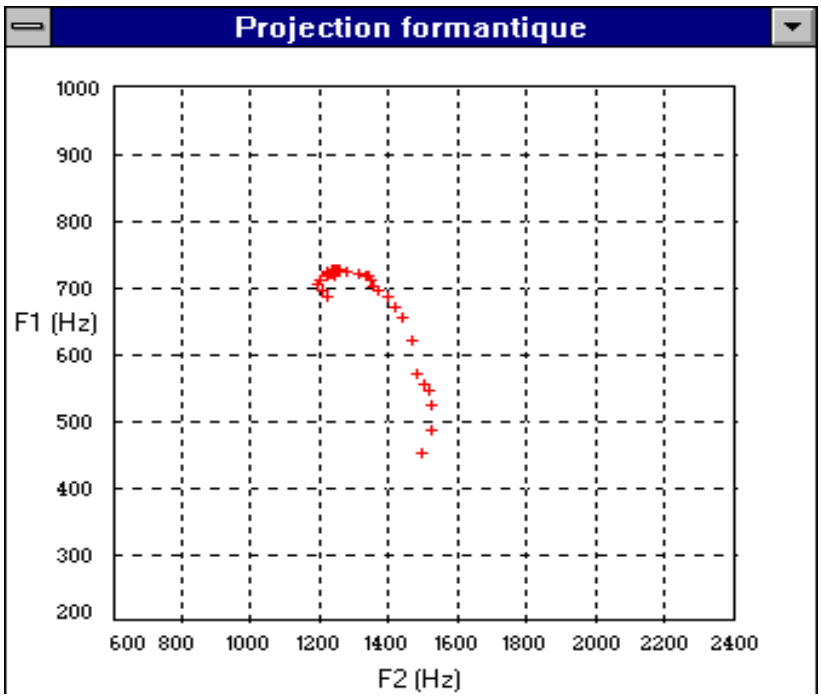


Fig. A.7 – Projection formantique dans le plan F1/F2.

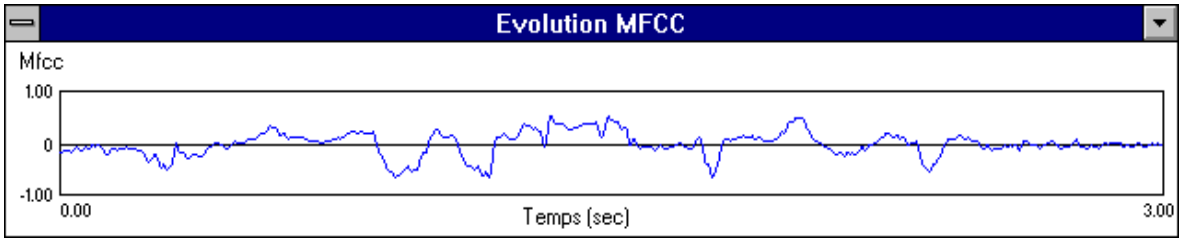


Fig. A.8 – Evolution temporelle de coefficients MFCC(1).

5.4 Segmentation et étiquetage

APHAK est aussi un outil de segmentation et d'étiquetage. L'utilisateur pose les marques temporelles de segmentation soit sur le spectrogramme soit sur le zoom du signal temporel. Pour affiner la segmentation il peut écouter la partie délimitée autant de fois qu'il le désire. Une batterie de labels par classe phonétique est à sa disposition pour étiquetage. La **Fig. 9** montre un exemple de segmentation-étiquetage d'une phrase.

Une fois la segmentation et l'étiquetage terminée. Il est possible à l'utilisateur de sauvegarder cette segmentation dans un fichier d'extension " .Seg ". Les fichiers ".Seg" sont constitués d'une entête et d'un ensemble de structures (segments). L'entête contient le nom du fichier signal ainsi que le nombre total de chaque classe de phonème. Chaque structure ou segment du fichier est composée de champs suivants :

- La classe phonétique à laquelle appartient le phonème.
- Le code ou label du phonème.
- Le début du segment dans le fichier signal correspondant.
- La fin du segment.
- La durée du segment en milliseconde (ms).

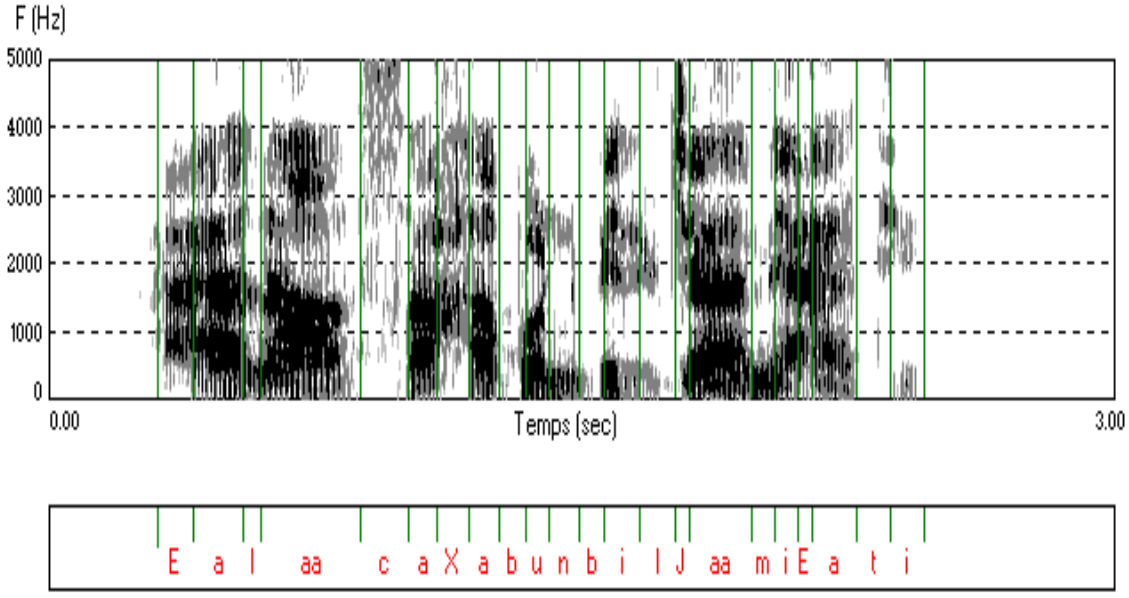


Fig. A.9 – Segmentation et étiquetage d'une phrase.

5.5 Etude Statistique

L'étude statistique peut être effectuée sur les résultats de tous les traitements disponibles. Elle consiste principalement en le tracé de l'histogramme et l'estimation de la loi de probabilité à laquelle obéit le résultat d'un traitement donné.

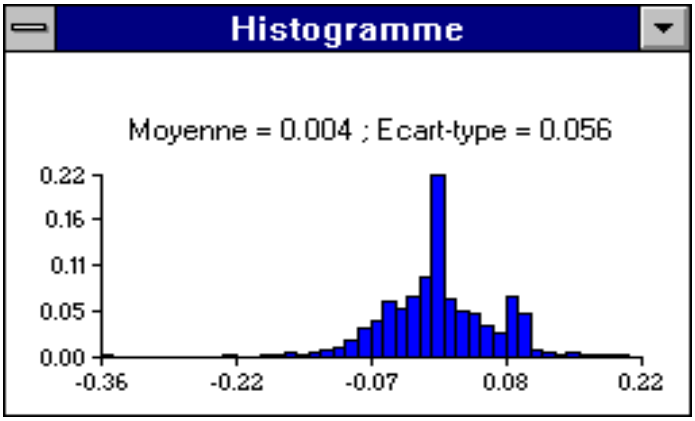


Fig. A.10 – Histogramme de MFCC(1).

6 Conclusion

APHAK a été développé pour tourner sur micro- ordinateur (compatible IBM). Il est doté d'un ensemble de fonctions de traitement constituant une aide performante pour les investigations dans le domaine de la parole en particulier le décodage acoustico-phonétique. La segmentation telle que développée est orientée de manière à permettre des études statistiques sur la durée des phonèmes ainsi que sur tous les paramètres acoustiques calculés. Elle permet aussi un apprentissage dans le cadre d'une reconnaissance phonémique globale.
APHAK constitue également un outil d'analyse des paramètres prosodiques (évolution du pitch, durée des phonèmes et énergie) éléments essentiels dans le développement des systèmes de synthèse de la parole.
APHAK possède aussi un caractère pédagogique, il peut être utilisé pour des travaux pratiques sur le traitement du signal en général.

Corpus de Mots

سألَ	وؤد	سِئم	مأكِلْ	تؤود	وئيد	مأخوذْ
مؤدب	ـأمرؤُ	سِئلَ	سؤالْ	رؤوس	يئِين	مؤمِن
ـفِئتُ	ـفِؤلٌ	ـبإدِنِه	ـونَّامُ	ـفِؤولٌ	ـزِئْبق	
ساءَلَ	تفاؤلْ	خائِنُ	تراءى	داؤوبُ	مرائيلْ	
مملُوءَةٌ	مملوءٌ	دوئب	مملُوأت	مملوؤونْ	مملوئين	رديئاً
رديئي	مليءٌ	مسألةٌ	مسؤولةٌ	مرئِس	مِراب	مرؤوس

بحثَ	فَحلَ	فَحِلَ	محاسِن	فَحولْ	وحيد
محلف	فُحلْ	جحِد	محامِ	وحوش	محيلْ
مِحك	فَصِحُ	بحمايةِ	زِحامُ	ريحُ	مكافِحين
ناحب	تباحثَ	ناحلَ	لماحات	مناحوك	سماحيق
ممنوحتُ	ممنوحُ	جوحِد	ممنوحات	ممنوحونْ	
كديحين	متحف	ممحُو	متحِفُ	مِمحات	مسحور

وعد	وعد	وعِد	معامِلْ	وعود	وعيد	معملْ
معلمُ	وعد	وعِد	معادِ	وعود	معين	معتمِر
سِعلٌ	فِعلٌ	فِعلْمِه	مِعادُ	فِعولٌ	متعين	بِعثَ
صاعق	تفاعلْ	مساعِف	جماعات	جماعوهم	مفَاعِيلْ	ممنوعةٌ
ممنوعُ	ووعِد	ممنوعات	ممنوعونْ	ممنوعين	ممنوعُ	نبعمُ
ربيعُ	ربيعِيةٌ	يستطْلِعانْ	منيعُ	مسعوف	مسعِفُ	مِفْعالٌ

فَصلَ	فَصر	نصِر	فَصالٌ	فَصولِيةٌ	فَصِيحُ	مصلٌ
حصلٌ	فُصرُ	فُصِلَ	مصابُ	فُصولٌ	مصِير	مصحف
فِصلٌ	فِصلٌ	ـبصِفَةِ	عِصامُ	مِصورُ	ـنصِير	مِصدق
فاصلَ	فاصلْ	فاصِل	فاصولْ	فَاصِلٌ		
فُوصلَ	فُوصلٌ	فُوصِلَ	فوصالْ	فوصولْ	فوصيلَ	
فيصلَ	فيصلُ	فيصِلْ	فيصولْ	فيصالْ	فيصيلْ	

مطْرُ	بطُلَ	قَطِر	مطارُ	سطُورُ	قَطيرُ	مطْوي
حطَمُ	سطْرٌ	لُطِم	حطامُ	لُطُوم	مطِيع	بطْؤنْ
حِطَمُ	فِطُمُ	فِطِيم	شِطانٌ	فِطونْ	قِطين	قِطْنُ
لاطَمَ	لاطُم	لاطِمُ	لاطوم	لاطيم	فِيطِين	
لوطَمُ	لوطُمُ	لوطِم	لوطُوم	لوطام	لوطيم	
فيطَمُ	فيطُمُ	فِيطِم	حِيطام	فيطوم	فِيطيم	

نظَر	نظُر	كَظِم	نظارُ	نظُورُ	نظير	نظْرتُ
يقُظَ	نظُمُ	نظِم	مظالِمُ	كُظُوم	مظِيف	مظْلِم
مِظَل	نِظُم	فِظِلْ	نِظامُ	نِظُوم	فِظِيعُ	مظْلُم
كاظَم	كاظُم	كاظِمُ	كاظوم	كاظيم	يظْلُم	
عوظَمُ	عوظُمُ	عوظِمُ	عوظام	كُوظُوم	عوظيم	
خيظَم	عيظُمُ	عيظِم	خيظام	خيظوم	خِيظِيم	

رضخ	وضر	خضِر	خضارُ	حضور	رضيعُ	رضعتُ
خضر	فُضلٌ	نضِر	خضارُ	حضور	مضير	مضفِر
فِضب	قِضم	نِضم	خِضار	فِضولٌ	قِضِيب	مِضمار
فاضل	فاضل	فاضِل	قاضا	فاضو	فاضِيق	
فوضلَ	فوضل	فوضِلَ	فوضالْ	فوضولْ	فوضيلْ	
فيضلَ	فيضلْ	فيضِلْ	فِيضالْ	فِيضولْ	فِيضِيلْ	

ذَهب	فَهل	فَهم	نهاك	فَهولْ	فَهيمُ	ترهبن
مهرب	فُهلَ	نهب	تهامِ	جهود	مهين	رهبانْ
مِهنُ	قَهقَهُ	بِهوايةِ	رِهانْ	نِهين	مِهور	مِهنةٌ
جاهد	مِياهُ	مجاهِد	تاها	الناهور	تلاهِين	أنهاكُم
فُوهةٌ	موهب	جوهِد	موهوب	بوهيب	تنهيكُم	تنهونَ
مجهر	مِنهم	مرهِق	أنهاك	مرهوك	تمهيلْ	

Corpus de Phrases

01 لَقَدْ احْتَفَظوا بِعَادَاتِ إفْطَارِهِمْ
02 تَهُبُ الرِّيَاحُ مِنَ الشَّمَال
03 حَظَرَ الأبُ بَعْدَ انْتِهَاءِ الْحَقْل
04 هُنَاكَ حَقْل بَهِيجْ هَذَا المَسَاء
05 فَصلَ القاضي الأمر
06 فِصالُ عَظِيمُ
07 الأبُ مطاعُ قَظولُ
08 الرّضيعُ مُصَابْ مَريضْ
09 عِصَامْ فُضُولِي
10 خُضَرُ طازَجَهُ
11 كاظَمَ غَيْضَهُ
12 حُضورُ مَطْلوبْ
13 الشّاعِرُ بَطَلُ فَصيحُ
14 فوصِلَ في أمْرِهِ
15 الإنْسَانُ الفاضِلُ
16 خابَ من طَغَى
17 صاحَتْ العَصافيرْ البطيّهْ
18 ضُرِبَ ضَرْباً
19 أصْبَحَ يَقِظاً
20 تَصَرَّفَ الطالِبُ
21 الوَطَنُ المَظْلومْ
22 فاضولْ كاظِمُ الغَيْض
23 امسَكَ الوَضيعُ مَطْوى
24 أصْبَحَت حُطَمُ و حُطامُ
25 فاصَلَ القَاضِي
26 عُوِّضَ تَعْويضاً
27 العُصْفورْ مُطارَدُ
28 أجْرَى المُعَلِّمُ امْتِحاناً
29 عُلِمَ حُظُورْ الشَّاعِرْ هذا المَساء
30 حاوَلَ دائِماً تَجَنُّبَ الأخْطاءِ
31 المَشْرُوع الوَحِيدْ في هَذِهِ السَّاعَةِ
32 يَنْسَحِبُ القائِدْ مِنَ المَعْرَكَةِ مُنْهَزِمُ
33 رَئِيسْ الحكومَةِ يَعودُ مِنَ الهِنْدِ
34 أهْتَمَ المُؤْمِنُ بِتَحليلْ القُرْانْ الكَرِيمْ
35 تَمْرَحُ الأطْفال في العيدِ
36 كَتَبَ حَرْفا واحِدَ على الوَرَقَةِ
37 تَعاهَدُ بِتَحْقِيقْ الوَحْدَةِ
38 عُقِدَة جَمْعِيةُ عامةُ
39 وُعِدَ الشّعْبُ بِحياةِ أفْضل
40 قَدْ يَمْنَعونَهُمْ عَنِ الطّعامْ
41 أحْتاجَ الأستاذُ إلى مُعينْ لتَصْحيحْ
42 نَوّهَ الأستاذُ بِجُهودْ أحْمَدْ
43 تُعْتَقَدُ جَمْعِيّةُ عامّةُ في المَصْنَع
44 هذا البُسْتانْ مَمْنوعْ لِلْوحوشْ
45 كانَ وَلَدُهُ مَشْؤومْ

Règles de Production

1 REGLES GÉNÉRALES

1.1 Règles sur le voisement

R (1)
Si
(Pitch_Actuel)
PHONEMES [A -10 H-10 h -10 E -10 s. -10 D. -10 t. –10 a 10 u 10 i 10 aa 10 uu 10 ii 10]
FIN.

R (2)
Si
NON (Pitch_Actuel)
PHONEMES [A 10 H 10 d. 10 h -5 E -10 D. -5]
FIN.

2 REGLES SUR LES CONSONNE /A/, /t./ et /d./

2.1 Règles sur le Burst

R (3)
CONTEXT-DROIT [a aa]
Si
(Burst_présent &
Burst-fréq-Act 3800 5500)
PHONEMES [A 10]
FIN.

R (4)
CONTEXT-DROIT [a aa]
Si
(Burst_présent &
Burst-fréq-Act 1500 2400)
PHONEMES [t. 10]
FIN.

R (5)
CONTEXT-DROIT [a aa]
Si
(Burst_présent &
Burst-fréq-Act 3500 4000)
PHONEMES [d. 10]
FIN.

R (6)
CONTEXT-DROIT [a aa]
Si
(Burst_présent &
centre-gravité 3500 5000)
PHONEMES [A 10]
FIN.

R (7)
CONTEXT-DROIT [i ii]
Si
(Burst_présent &
Burst-fréq-Act 4500 5500)
PHONEMES [A 10]
FIN.

R (8)
CONTEXT-DROIT [i ii]
Si
(Burst_présent &
Burst-fréq-Act 1700 2400)
PHONEMES [t. 10]
FIN.

R (9)
CONTEXT-DROIT [i ii]
Si
(Burst_présent &
burst-fréq-Act 4000 5000)
PHONEMES [d. 10]
FIN.

R (10)
CONTEXT-DROIT [u uu]
Si
(Burst_présent &
burst-fréq-Act 1500 2200)
PHONEMES [t. 10]
FIN.

R (11)
CONTEXT-DROIT [u uu]
Si
NON (Burst_présent)
PHONEMES [A 5]
FIN.

R (12)
CONTEXT-DROIT [u uu]
Si
NON (Burst_présent)
PHONEMES [d. 5]
FIN

R (13)
CONTEXT-DROIT [a aa i ii]
Si
NON (Burst_présent)
PHONEMES [A 5]
FIN.

R (14)
CONTEXT-DROIT [a aa i ii]
Si
NON (Burst_présent)
PHONEMES [t. 5]
FIN.

R (15)
CONTEXT-DROIT [a aa i ii]
Si
NON (Burst_présent)
PHONEMES [d. 5]
FIN.

R (16)
CONTEXT-DROIT [autre que voyelle]
Si
(Burst_présent &
burst-fréq-Act 3500 6000)
PHONEMES [A 10]
FIN.

2.2 Règles sur le suivi de formants

R (17)
CONTEXT-DROIT [a aa i ii u uu]
Si
(F1-droit-plat &
F2-droit-plat
PHONEMES [A 10]
FIN.

R (18)
CONTEXT-DROIT [a aa i ii u uu]
Si
F1-droit-plat
PHONEMES [A 10]
FIN.

R (19)
CONTEXT-DROIT [a aa i ii u uu]
Si
F1-droit-montant
PHONEMES [A -10]
FIN.

R (20)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F2-droit-plat)

PHONEMES [d. 10]
FIN.

R (21)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F2-droit-plat)
PHONEMES [t. 10]
FIN.

R (22)
CONTEXT-DROIT [i ii]
Si
(F1-droit-descendant &
F2-droit-montant)
PHONEMES [t. 10]
FIN.

R (23)
CONTEXT-DROIT [i ii]
Si
(F1-droit-montant &
F2-droit-montant)
PHONEMES [d. 10]
FIN.

R (24)
CONTEXT-DROIT [u uu]
Si
(F1-droit-plat &
F2-droit-montant)
PHONEMES [t. 10]
FIN.

R (25)
CONTEXT-DROIT [u uu]
Si
(F1-droit-descendant &
F2-droit-montant)
PHONEMES [d. 10]
FIN.

3 REGLES SUR LES CONSONNES /H/, /h/, /s./ et /D./

3.1 Règles sur le seuil de friction

R (26)
CONTEXT-DROIT [a aa]
Si
(Durée-act > 40 &
seuil-fric-inf 600 1000)

PHONEMES [H 10]
FIN.

R (27)
CONTEXT-DROIT [a aa]
Si
(Durée-act > 60 &
seuil-fric-inf 2800 3000)
PHONEMES [s. 10]
FIN.

R (28)
CONTEXT-DROIT [a aa]
Si
(Durée-act > 50 &
seuil-fric-inf 3000 3300)
PHONEMES [D. 10]
FIN.

R (29)
CONTEXT-DROIT [u uu]
Si
(Durée-act > 40 &
seuil-fric-inf 500 1300)
PHONEMES [H 10]
FIN.

R (30)
CONTEXT-DROIT [u uu]
Si
(Durée-act > 60 &
seuil-fric-inf 3000 3200)
PHONEMES [s. 10]
FIN.

R (31)
CONTEXT-DROIT [u uu]
Si
(Durée-act > 50 &
seuil-fric-inf 2800 3000)
PHONEMES [D. 10]
FIN.

R (32)
CONTEXT-DROIT [i ii]
Si
(Durée-act > 40 &
seuil-fric-inf 1800 2200)
PHONEMES [H 10]
FIN.

R (33)
CONTEXT-DROIT [i ii]

Si
(Durée-act > 60 &
seuil-fric-inf 3000 3300)
PHONEMES [s. 10]
FIN.

R (34)
CONTEXT-DROIT [i ii]
Si
(Durée-act > 50 &
seuil-fric-inf 2200 2400)
PHONEMES [D. 10]
FIN.

R (35)
CONTEXT-DROIT [autre que voyelle]
Si
(Durée-act > 40 &
seuil-fric-inf 1800 2000)
PHONEMES [H 10]
FIN.

R (36)
CONTEXT-DROIT [autre que voyelle]
Si
(Durée-act > 60 &
seuil-fric-inf 2800 3000)
PHONEMES [s. 10]
FIN.

3.2 Règles sur le centre de gravité

R (37)
CONTEXT-DROIT [a aa]
Si
(Durée-act > 40 &
centre-gravité-norm 4000 4300)
PHONEMES [s. 10]
FIN.

R (38)
CONTEXT-DROIT [i ii]
Si
(Durée-act > 40 &
centre-gravité-norm 2600 3800)
PHONEMES [h 10]
FIN.

R (39)
CONTEXT-DROIT [i ii]
Si
(Durée-act > 60 &

centre-gravité-norm 4000 4500)
PHONEMES [s. 10]
FIN.

R (40)
CONTEXT-DROIT [u uu]
Si
(Durée-act > 40 &
centre-gravité-norm 500 2200)
PHONEMES [h 10]
FIN.

R (41)
CONTEXT-DROIT [u uu]
Si
(Durée-act > 60 &
centre-gravité-norm 4200 4600)
PHONEMES [t. 10]
FIN.

3.3 Règles sur l'évolution des formants

R (42)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F2-droit-plat)
PHONEMES [H 10]
FIN.

R (43)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F2-droit-montant)
PHONEMES [s. 10]
FIN.

R (44)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F2-droit-plat)
PHONEMES [D. 10]
FIN.

R (45)
CONTEXT-DROIT [u uu]
Si
(F1-droit-plat &
F2-droit-descendant)
PHONEMES [H 10]

FIN.

R (46)
CONTEXT-DROIT [u uu]
Si
(F1-droit-plat &
F2-droit-descendant)
PHONEMES [s. 10]
FIN.

R (47)
CONTEXT-DROIT [u uu]
Si
(F1-droit-descendant &
F2-droit-montant)
PHONEMES [D. 10]
FIN.

R (48)
CONTEXT-DROIT [i ii]
Si
(F1-droit-plat &
F2-droit-montant)
PHONEMES [H 10]
FIN.

R (49)
CONTEXT-DROIT [i ii]
Si
(F1-droit-montant &
F2-droit-montant)
PHONEMES [D. 10]
FIN.

R (50)
CONTEXT-DROIT [i ii]
Si
(F1-droit-montant &
F2-droit-montant)
PHONEMES [s. 10]
FIN.

R (51)
CONTEXT-DROIT [a aa u uu]
Si
(F3-droit-montant)
PHONEMES [H 10]
FIN.

R (52)
CONTEXT-DROIT [a aa u uu i ii]
Si
(F1-droit-plat &

F3-droit-plat)
PHONEMES [h 10]
FIN.

3.4 Règles sur les formants

R (53)
CONTEXT-DROIT [a aa]
Si
(formant 1-Act 500 650 &
formant 2-Act 1450 1550)
PHONEMES [h 10]
FIN.

R (54)
CONTEXT-DROIT [a aa]
Si
(formant 1-Act 250 275 &
formant 2-Act 900 1000 &
formant3-Act 2300-2350)
PHONEMES [D. 10]
FIN.

R (55)
CONTEXT-DROIT [u uu]
Si
(formant 1-Act 280 400)
PHONEMES [h 10]
FIN.

R (56)
CONTEXT-DROIT [u uu]
Si
(formant 1-Act 250 275 &
formant 2-Act 900 1000 &
formant3-Act 2300-2350)
PHONEMES [D. 10]
FIN.

R (57)
CONTEXT-DROIT [i ii]
Si
(formant 1-Act 280 350 &
formant 2-Act 2000 2200)
PHONEMES [h 10]
FIN.

R (58)
CONTEXT-DROIT [i ii]
Si
(formant 1-Act 250 275 &
formant 2-Act 900 1000 &

formant3-Act 2300-2350)
PHONEMES [D. 10]
FIN.

4 REGLES DUR LA CONSONNE /E/

4.1 Règles sur les formants

R (59)
CONTEXT-DROIT [a aa]
Si
(formant 1-Act 650 790)
PHONEMES [E 10]
FIN.

R (60)
CONTEXT-DROIT [u uu]
Si
(formant 1-Act 380 500)
PHONEMES [E 10]
FIN.

R (61)
CONTEXT-DROIT [i ii]
Si
(formant 1-Act 400 490)
PHONEMES [E 10]
FIN.

R (62)
CONTEXT-DROIT [autre que voyelle]
Si
(formant 1-Act 600 690)
PHONEMES [E 10]
FIN.

R (63)
CONTEXT-DROIT [a aa]
Si
(formant 1-Act 740 790 &
formant 2-Act 1440 1520 &
formant 3-Act 2400 2480)
PHONEMES [E 10]
FIN.

R (64)
CONTEXT-DROIT [i ii]
Si
(formant 1-Act 445 490 &
formant 2-Act 1740 1850 &
formant 3-Act 2600 2700)
PHONEMES [E 10]
FIN.

R (65)
CONTEXT-DROIT [u uu]
Si
(formant 1-Act 440 480 &
formant 2-Act 1200 1850 &
formant 3-Act 2100 2220)
PHONEMES [E 10]
FIN.

R (66)
CONTEXT-DROIT [autre que voyelle]
Si
(formant 1-Act 600 696 &
formant 2-Act 1550 1720 &
formant 3-Act 2400 2496)
PHONEMES [E 10]
FIN.

R (67)
CONTEXT-DROIT [a aa]
Si
(formant 1-Act 200 450)
PHONEMES [E -10]
FIN.

R (68)
CONTEXT-DROIT [u uu]
Si
(formant 1-Act 150 320)
PHONEMES [E -10]
FIN.

R (69)
CONTEXT-DROIT [i ii]
Si
(formant 1-Act 100 300)
PHONEMES [E -10]
FIN.

R (70)
CONTEXT-DROIT [autre que voyelle]
Si
(formant 1-Act 100 550)
PHONEMES [E -10]
FIN.

4.2 Règles sur l'évolution des formants

R (71)
CONTEXT-DROIT [a aa]
Si

(F1-gauche-plat)
PHONEMES [E 10]
FIN.

R (72)
CONTEXT-DROIT [a aa]
Si
(F1-droit-plat &
F3-droit-montant)
PHONEMES [E 10]
FIN.

R (73)
CONTEXT-DROIT [u uu]
Si
(F2-droit-descendant)
PHONEMES [E 10]
FIN.

R (74)
CONTEXT-DROIT [i ii]
Si
(F1-droit-descendant &
F2-droit-montant)
PHONEMES [E 10]
FIN.

R (75)
CONTEXT-GAUCHE [a aa u uu i ii]
Si
(F1-gauche-montant)
PHONEMES [E 10]
FIN.

R (76)
CONTEXT-GAUCHE [a aa u uu i ii]
Si
(F1-gauche-descendant)
PHONEMES [E -10]
FIN.

5 REGLES SUR LES VOYELLES

R (77)
Si
(formant 1-Act 580 620 &
formant 2-Act 1500 1600 &
formant 3-Act 2350 2455)
PHONEMES [a 10]
FIN.

R (78)
Si
(formant 1-Act 280 310 &
formant 2-Act 835 900 &
formant 3-Act 2100 2275)
PHONEMES [u 10]
FIN

R (79)
Si
(formant 1-Act 285 320 &
formant 2-Act 1900 2200 &
formant 3-Act 2550 2650)
PHONEMES [i 10]
FIN.

R (80)
Si
(formant 1-Act 580 620 &
formant 2-Act 1900 2200 &
formant 3-Act 2550 2600)
PHONEMES [aa 10]
FIN.

R (81)
Si
(formant 1-Act 280 310 &
formant 2-Act 750 850 &
formant 3-Act 2250 2400)
PHONEMES [uu 10]
FIN.

R (82)
Si
(formant 1-Act 285 320 &
formant 2-Act 2150 2250 &
formant 3-Act 2600 2750)
PHONEMES [ii 10]
FIN.

R (83)
Si
(formant 1-Act 285 320 &
formant 2-Act 1900 2200 &
formant 3-Act 2550 2650)
PHONEMES [a 10]
FIN.

R (84)
Si
(formant 1-Act 600 680 &
formant 2-Act 1150 1250 &

formant 3-Act 2550 2655)
PHONEMES [a 10]
FIN.

R (85)
Si
(formant 1-Act 300 340 &
formant 2-Act 835 900 &
formant 3-Act 2200 2375)
PHONEMES [u 10]
FIN

R (86)
Si
(formant 1-Act 310 370 &
formant 2-Act 1550 1680 &
formant 3-Act 2650 2750)
PHONEMES [i 10]
FIN.

R (87)
Si
(formant 1-Act 600 650 &
formant 2-Act 1150 1250 &
formant 3-Act 2600 2700)
PHONEMES [aa 10]
FIN.

R (88)
Si
(formant 1-Act 320 370 &
formant 2-Act 750 850 &
formant 3-Act 2250 2400)
PHONEMES [uu 10]
FIN.

R (89)
Si
(formant 1-Act 300 350 &
formant 2-Act 2000 2150 &
formant 3-Act 2700 2800)
PHONEMES [ii 10]
FIN.

R (90)
Si
(Durée-act > 120 &
formant 2-Act 750 850)
PHONEMES [uu 10]
FIN.

R (91)
Si
(Durée-act > 120 &
formant 2-Act 2000 2150)
PHONEMES [ii 10]
FIN.

R (92)
Si
(Durée-act > 120 &
formant 2-Act 1150 1250)
PHONEMES [aa 10]
FIN.