REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE YAHIA FARES DE MEDEA



Faculté de Technologie

Département du Génie Electrique

Mémoire de fin d'études de Master

Filière : Télécommunications

Spécialité : Systèmes des Télécommunications

Etude de l'influence des paramètres de la modélisation Markovienne sur les performances d'un système de reconnaissance de la parole

---

Présentées par :

Melle. TEBBAL Meriem

Melle. RABAH SIDHOUM Lamia

Proposé et Dirigé par :

Mr. RAGGAI Djahid

Promotion 2020

Dédicaces

Avant tout, je tiens à remercier le bon dieu, et l'unique qui m'offre le courage et la volonté nécessaire pour affronter les différentes difficultés de la vie.

Je dédie ce modeste travail :

A mes très chers parents, source de vie, d'amour et d'affection Mon père ***Mahfoud****, et ma mère* ***Nadia****, que je n'arriverai jamais à exprimer ma gratitude et ma reconnaissance, pour ses amours ses soutiens tout au long de mes études et leurs prières, leur encouragements et qu'aucune dédicace ne pourra compenser leur sacrifice pour mon bien être et mon bonheur.*

À la famille Tebbal surtout Mes sœurs et Mes frères : ***Amira, Samah, Israa, Mustapha*** *et* ***Yazid*** *sources de joie et de bonheur.*

À mes amis et collègue : ***Samira, Afrah, Dounia, Manel, Noussaiba*** *ET* ***Zineb***

En particulier à ma chère amie avant d'être un binôme ***Rabah Sidhoum Lamia*** *qui n'a jamais manqué de travailler.*

À notre encadreur ***Mr. Raggai Djahid*** *pour son immense patience, sa grande disponibilité et ses conseils qui ont contribué grandement à la réalisation de ce travail. Qu'il trouve ici l'expression de ma profonde gratitude*

À vous cher lecteur

Et à tous ceux qui nous ont soutenus pour l'achèvement de ce projet avec réussite.

Meriem

Dédicaces

Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, c'est tout simplement que : Je dédie cette thèse de magistère à :

A Ma tendre Mère ***Hafida*** *: Tu représentes pour moi la source de tendresse et l'exemple de dévouement qui n'a pas cessé de m'encourager. Tu as fait plus qu'une mère puisse faire pour que ses enfants suivent le bon chemin dans leur vie et leurs études.*

A Mon très cher Père ***Mohammed*** *: Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours pour vous. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être. Ce travail et le fruit de tes sacrifices que tu as consenti pour mon éducation et ma formation le long de ces années.*

A mon très cher Frère ***Ahmed*** *: Tes sacrifices, pour ton soutien moral et matériel. A monsieur* ***RAGGAI Djahid*** *Cette humble dédicace ne saurait exprimer nom grand respect et ma profonde estime, que dieu vous procure bonne santé et long vie.*

A mes amis et collègues : ***Saadi khaoula, Ben rahmoune flora, Merrar feriel, Khamili ilhem*** *et en particulier à ma chère amie et sœur* ***Tebbal Meriem*** *qui n'a jamais manqué de travailler.*

A tous les membres de ma promotion. A tous mes enseignants depuis mes premières années d'études.

Lamia

Remerciements

Tout d'abord je Remercie le bon Dieu le tout puissant pour la bonne santé, la volonté et de la patience qu'il nous a donné pour accomplir ce travail.

Je remercie très sincèrement '' Mr. Raggai Djahid '' L'encadreur de ce travail, pour ses conseils pertinents, Ses orientations judicieuses, et sa patience.

Nous tenons à remercier également les membres de jury

De Bien vouloir accepter d'examiner et de juger ce modeste travail.

Nous remercions aussi tous ceux qui ont contribué de près ou de loin à la réalisation de ce mémoire de fin d'étude, sans oublier d'adresser nos remerciements à tous les enseignants du département Génie électrique notamment la spécialité télécommunication.

Lamia et Meriem

الملخص

يمثل هذا العمل دراسة حول تأثير وسائط النمذجة المركوفية على أداء نظام التعرف الآلي على كلمات عربية معزولة. الوسائط المعنية هي: عدد حالات النماذج، طوبولوجيا النماذج، النقطة الابتدائية للنماذج إضافة الى عدد توزيعات غوس لكل حالة. الهدف المنشود هو تغيير هذه الوسائط مع مراقبة تطور أداء النظام. التجربة اثبتت ان استعمال عدد حالات بين 7 و9 مع ثلاثة توزيعات لكل حالة تعطي نتائج جيدة خاصة بالنسبة للطوبولوجيا يسار يمين رقم 2. أداة العمل المستعملة هي HTK

الكلمات المفتاحية: التعرف الآلي على الكلام، نماذج ماركوف المخفية، أداة HTK.

Résumé

Ce travail est une étude de l'influence des paramètres de la modélisation markovienne sur les performances d'un système de reconnaissance de mots arabes isolés. Les paramètres en question sont : le nombre d'états des modèles, la topologie des modèles, la méthode d'initialisation des modèles ainsi que le nombre de gaussiennes par état. L'objectif est de varier chacun des paramètres précédents tout en observant l'évolution du taux de reconnaissance. L'expérience montre qu'un nombre d'états entre 7 et 9 avec 3 gaussiennes pour chaque état donne de très bons résultats surtout pour la topologie gauche-droite 2. L'outil de manipulation utilisé est le HTK (Hidden Markov Toolkit)

Mots-clés : reconnaissance automatique de la parole, modèles de Markov cachés, l'outil HTK

Abstract

This work is a study of the effect of the Markovian modelization on the performances of recognition of isolated arab spoken words. The studied parameters are the number of states in the models, the topology of the models, initialization method of the models as well as the number of Gaussians per sate of the model. The goal is to variate these parameters and observe the evolution of the performances of the system. The experiment shows that using a number of states between 7 and 9 with three Gaussians lead to good performances especially for topology left-right 2.

Keywords: automatic speech recognition, hidden Markov models, HTK tool

TABLE DES MATIERES

LISTE DES FIGURES I

LISTE DES TABLEAUX II

LISTE DES ACRONYMES III

Introduction générale ..... **Error! Bookmark not defined.**
CHAPITRE I :GENERALITE SUR LE TRAITEMENT DE LA PAROLE 3
1. Introduction ..... **Error! Bookmark not defined.**
2. Historique ..... Error! Bookmark not defined.
3. Le son 4
4. La production de la parole 5
5. La pareil auditif ..... Error! Bookmark not defined.
5.1. La pareil auditif humaine 7
6. Les caractéristiques phonetiques ..... **Error! Bookmark not defined.**
6.1. Phonéme ..... **Error! Bookmark not defined.**
7. Les caracteristiques acoustiques ..... Error! Bookmark not defined.
7.1. Echantillonage 10
7.1.1. Théorème de shannon 10
7.2. La quantification 11
8.La reconnaissance 12
8.1. Les systemes de reconnaissance du locuteur 12
8.2. Les systemes de reconnaissance de la parole 13
9. Les methodes utilisées pour la reconnaissance de la parole 13
9.1. La méthode globale ..... **Error! Bookmark not defined.**
9.1.1. Dynamic Time Warping 14
9.1.2. Hidden Markov Model (les modèles de Markov cachés) 14
9.1.3. Les réseaux neurone 14
9.1.4. Mélange de gaussiennes GMM 14
9.2. La méthode analytique 15
10. analyse et codage de la parole 15
10.1. Analyse temporelle 16
10.2. Analyse fréquentielle 16
10.2.1. Transformée de Fourier Discrète à court terme 16
10.2.2.Le spectrogramme 19
10.3.Bancs de filtre passe bande 20

10.4. Analyse cepstrale 20
11. Conclusion : 24
CHAPITRE II : LES MODELES DE MERKOV CACHES 25
1. Introduction 26
2. Modéles de markov cachés(MMC) 26
2.1. Topologies des HMMs 27
3. Reconnaissance des mots isolés 29
4. Les problémes majeurs en HMMs 31
5.Calcule de la vraissemblance d'une séquence 32
6. L'apprentissage 33
7. La reconnaissance 34
8. Conclusion 34
CHAPITRE III : RESULTATS ET INTERPRETATIONS 35
1. Introduction 36
2. Description de la base de données utilisées 36
2.1. Vocabulaire 36
2.2. système d'étiquetage des fichiers de la base 37
3. Descripition de l'outil de reconnaissance HTK 37
4. Préparation des fichier de la base de données pour la reconnaissance 37
4.1. Le codage 37
4.2. Choix de topologie des modéles (prototypes) 40
4.3. Initialisation des modéles 42
4.4. L'apprentissage 43
4.5. definition de la grammaire et du dictionnaire 44
4.6. La reconnaissance 47
5. Analyse des résultats des tests sur la RAP 48
6. Résultats et interprétations 48
6.1. L'effet de la topologie et du nombre d'états 48
6.2. L'etude de l'effet de la methode d'initialisation sur les performances de la reconnaisance 50
6.3. Etude de l'effet du nombre de gaussiennes par état 51
7. Conclusion 53
Conclusion générale 54
Références bibliographiques IV

LISTE DES FIGURES

Figure I.1. *L'appareil phonatoire* 7
Figure I.2. *L'appareil auditif humaine*..... 8
Figure I.3. *Enregistrement numérique d'un signal acoustique* ....**Error! Bookmark not defined.**
Figure I.4. *Echantillonage d'un signal.* 10
Figure I.5. *Un signal quantifier*..... 11
Figure I.6. *Audiogramme de signal de mot « effacer »*..... 11
Figure I.7. *Exemple de son voisé et non-voisé* 12
Figure I.8. Schéma générique d'un système de reconnaissance de la parole 13
Figure I.9. *Système d'obtention du signal parole*..... 15
Figure I.10. Signal pondéré par une fonction fenêtre symétrique de durée finie N..... 17
Figure I.11. *Analyse par transformée de Fourier* 18
Figure I.12. Les fonctions de fenêtrage……..…….....19
Figure I.13. Spectrogramme à bande étroite (d'un son voisé)……..…….....19
Figure I.14 *Spectrogramme à large bande (d'un son voisé)* 20
Figure I.15 *Analyse de bancs de filtre* 20
Figure I.16. *Etapes de calcule du cepstre* ..... **Error! Bookmark not defined.**
Figure I.17. *Echelle de Mel.* ..... **Error! Bookmark not defined.**
Figure I.18. *Banc de filtre à l'échelle de Mel.* ..... **Error! Bookmark not defined.**
Figure I.19. *Etapes de calcule MFCC* ..... **Error! Bookmark not defined.**
Figure II.1. *Topologie 1(en haut) et topologie 2 (en bas)*..... 28
Figure II.2. *Le modèle de génération de Markov*..... 30
Figure II.3. Utilisation des HMMs dans la reconnaissance des mots isolés 31
Figure III.1. *La procédure du codage* 38
Figure III.2. *Utilisation de HCopy pour un groupe de fichiers* 38
Figure III.3. *Fichier de configuration*..... 39
Figure III.4. Le fichier texte spécifiant le nom et localisation..... 40
Figure III.5. *Topologie choisie pour HMMs*..... 41
Figure III.6. *Fichier de description HMM* 41
Figure III.7. *Fonctionnement de HCompv*. 43
Figure III.8. *Fonctionnement de HRest*..... 44
Figure III.9. *Le fichier de la grammaire* 45
Figure III.10. *Le fichier du dictionnaire* 45
Figure III.11. *Le réseau de connexion entre mot* 46
Figure III.12. *Processus du codage et reconnaissance*..... 47
Figure III.12. Example d'un fichier reco 48
Figure III.14. *Topologie 1* 49
Figure III.15. *Topologie 2* 49

LISTE DES TABLEAUX

Tableau I.1 : *Propriétés de quelques fenêtres connues……..18*

Tableau III.1 : *Le taux de reconnaissance obtenu pour des nombres d'états différents et pour les deux topologies ……49*

Tableau III.2 : *Taux de reconnaissance de l'initialisation de chaque mot avec sa classe d'exemples ……..50*

Tableau III.3 : *Taux de reconnaissance de l'initialisation avec initialisation globale ……..50*

Tableau III.4 : *Utilisation de 2 gaussiennes par état pour les deux topologies..... .51*

Tableau III.5 *: Utilisation de 3 gaussiennes par état pour les deux topologies.....52*

LISTE DES ACRONYMES

ADN: Acide Désoxyribonucléique

ANN: Artificial Neural Network

API: Alphabet Phonétique International

DCT: Discret Cosinus Transform

DFT: Discret Fourier Transform

DTW: Dynamic time Warping

EM: Exaptation Maximization

FFT: Faste Fourier Transform

GMM: Gaussien Mixture Models

HMM: Hidden Markov Models

HTK: Hidden Markov Model Toolkit

LPC: Analyse Prédictives Linéaire

MMC: Modèle De Markov Caché

MFCC: Mel Frequency Cepstral coefficient

RAP: Reconnaissance Automatique De La Parole

TCFT: Transformer De Fourier A Court Terme

TF: Transformer De Fourier

TFD: Transformer De Fourier Discrète

VQ: Quantification Vectorielle

Introduction générale

La parole est l'un des principaux moyens de communication entre êtres humains, sa simplicité en fait d'ailleurs le moyen de communication le plus populaire dans la société humaine (il est plus facile de parler à quelqu'un que de lui écrire ou de lui faire un schéma). Néanmoins, cette simplicité (pour l'être humain) renferme un traitement très complexe fait par notre cerveau, de la production de la parole jusqu'à sa perception et sa compréhension, ce qui rend la parole difficilement automatisable pour une machine. L'avancement technologique et surtout de l'informatique a suscité le besoin de nouveaux moyens de dialogue homme-machine (la parole), des moyens qui libéreraient l'homme d'un contact constant avec la machine limitant ainsi l'utilisation du clavier et autres périphériques qui rendaient la communication avec la machine très difficile et très lente. C'est ce qui a conduit à donner naissance à plusieurs travaux de recherche dans le domaine de la Reconnaissance Automatique de la Parole (RAP).

Un système de RAP est un système qui a la capacité de détecter à partir du signal vocal la parole et de l'analyser dans le but de transcrire ce signal en une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. La RAP Introduite à la fin des années 60, la théorie de la reconnaissance vocale a suscité beaucoup de controverses à l'époque. De nos jours, des taux de reconnaissance avoisinant les 99 % sont observables sur des petits vocabulaires (reconnaissance de mots isolés, reconnaissance de digits) indépendamment du locuteur. Toutefois, à l'heure actuelle, l'impact des systèmes basés sur la reconnaissance de la parole est tout à fait négligeable dans notre vie de tous les jours. Les performances des systèmes de reconnaissances automatiques de la parole (RAP) se sont considérablement améliorées, permettant d'aborder des domaines d'application de complexité croissante. Les travaux actuels les plus avancés concernent des systèmes de dialogue via le téléphone, la reconnaissance de la parole spontanée ou la transcription d'émissions de radio ou télévision. Les performances obtenues dépendent beaucoup du type de tâche considérée (taille et difficulté du vocabulaire, nombre et diversités des locuteurs, conditions d'enregistrement).

Le signal de la parole est l'un des signaux les plus complexes, il n'est pas facile de le caractériser par un modèle simple. L'un des problèmes de la reconnaissance de la parole est la variabilité du signal. Les précurseurs dans le domaine du traitement de la parole, pensaient que la parole était une simple juxtaposition d'éléments appelés "caractéristiques distinctives" ou invariants (phonèmes), et qu'à partir d'échantillons représentant les invariants d'une langue, on pouvait reconstituer ou reconnaître n'importe quelle phrase. Pour surmonter ces difficultés, de

Nombreuses méthodes et modèles mathématiques originaux ou adaptés d'autres domaines ont été développés, parmi lesquelles on peut citer : la comparaison dynamique, les systèmes-experts, les réseaux de neurones, les modèles stochastiques et en particulier les modèles de Markov cachés, etc.

Actuellement, les modèles de référence sont les réseaux à apprentissage profond ainsi que les modèles de Markov cachés (HMM pour Hidden Markov Models) qui représentent l'état de l'art en RAP. La construction de ces modèles dont dépendent les performances du système de RAP nécessite l'existence de grandes bases d'apprentissage où les segments des différents phonèmes sont délimités et étiquetés, ceci est d'autant plus vrai pour les réseaux à apprentissage profond. Cette étude est basée sur une approche probabiliste, où on utilise les modèles de Markov cachés (HMM) pour modéliser les vecteurs des paramètres acoustique (MFCC), et donner la décision après avoir calculé le maximum de vraisemblance qui va introduire une erreur de reconnaissance. L'outil HTK va nous servir à manipuler les modèles de Markov avec une grande souplesse.

Ce mémoire a pour objectif de mettre la lumière sur l'effet des paramètres de la modélisation markovienne sur les performances d'un système de reconnaissance de la parole.

· Le premier chapitre représente une vue générale sur la parole. Nous décrivons l'ensemble du système vocal chez l'être humain, ainsi que les différentes étapes de prétraitement du signal parole. Les méthodes et les techniques de la reconnaissance de la parole et sa modélisation ainsi que les outils d'analyse de la parole seront exposés.

· Le deuxième chapitre, est totalement dédié aux Modèles de Markov Cachés. Le formalisme mathématique y est déployé ainsi que l'application à la reconnaissance vocale. Le chapitre sera terminé par une conclusion.

· Le chapitre 3 expose les résultats et les interprétations d'une série d'expérimentations à travers laquelle il sera question de faire varier un certain nombre de paramètres concernant la modélisation markovienne et d'observer leur influence sur les performances de reconnaissance. Ce chapitre sera terminé par une conclusion.

· Ce mémoire sera naturellement clos par une conclusion générale.

CHAPITRE I :

Généralité sur le traitement de la parole

1. Introduction :

La parole est le principal support de la langue, comme pour l'écriture, le message doit donc être bien structuré selon des règles connues par tous ceux qui parlent la même langue (grammaire, syntaxe, vocabulaire…).

Bien que l'évolution de la langue mue par les tendances actuelles rend de plus en plus difficile le fait de fixer des règles statiques une bonne fois pour toute, l'être humain grâce à l'ordinateur le plus sophistiqué au monde (son cerveau), pressé de faire passer son message, transgresse le plus souvent ces règles et arrive sans difficultés à le comprendre, ce qui montre le caractère dynamique du cerveau humain qui arrive très bien à s'adapter à de nouvelles situations, caractère qui devrait caractériser tout système de traitement de la parole.

Ce chapitre expose les notions de base en relation directe avec les besoins de notre étude.

2. Historique :

La reconnaissance de la parole est une discipline quasi contemporaine de l'informatique. Vers 1950 apparut le premier système de reconnaissance de chiffres, appareil entièrement câblé et très imparfait. Vers 1960, l'introduction des méthodes numériques et l'utilisation des ordinateurs changent la dimension des recherches. Néanmoins, les résultats demeurent modestes car la difficulté du problème avait été largement sous-estimée, en particulier en ce qui concerne la parole continue. Vers 1970, la nécessité de faire appel à des contraintes linguistiques dans le décodage automatique de phrases apparaît clairement, alors que la reconnaissance de la parole avait été jusque-là considérée comme un problème d'ingénierie. La fin de la décennie 1970 voit se terminer la première génération des systèmes commercialisés de reconnaissance de mots. Les générations suivantes, mettant à profit les possibilités sans cesse croissantes de la micro-informatique, posséderont des performances de plus en plus grandes (systèmes multi locuteurs, parole continue). Actuellement, les HMMs et les réseaux à apprentissage profond (deep learning) profitent de la disponibilité de bases de données volumineuses et de machines performantes pour atteindre des performances prometeuses.

3. Le son :

Le son est ce que l'oreille perçoit de la vibration d'un corps. Il est donc une vibration, une

Sorte d'onde (produite par un objet) qui se propage par et à travers des corps physiques (air, Eau, métal, bois...), il est caractérisé par 4 paramètres :

- La hauteur (un son grave / aigu).
- La durée (un son court / long).
- Le volume, son intensité (un son doux / fort).
- Le timbre (la carte d'identité du son).

4. La production de la parole :

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique et pragmatique. [1]

- **Le niveau phonétique :** en phonétique, on essaie d'explorer la façon dont le signal est produit par le système articulatoire, l'analyse est effectuée sur trois plans complémentaires, perceptifs, articulatoires, et acoustiques.
- **Le niveau phonologique :** la phonologie a comme objectif d'étudier les variantes phonétiques contextuelles, elle introduit la notion d'unité abstraite du discours le « phonème ». En reconnaissance de la parole, la phonologie regroupe l'ensemble des modules de traitement des altérations possibles d'un phonème (allophones ou variantes) ou d'un mot dans un contexte donné.
- **Le niveau acoustique :** les acousticiens s'intéressent aux traits acoustiques de la parole : sa fréquence fondamentale, son énergie, et son spectre. L'analyse est effectuée sur un signal électrique, un transducteur (le microphone) est utilisé pour réaliser le passage de l'acoustique à l'électrique.
- **Le niveau morphologique :** la morphologie est la branche de la linguistique qui étudie comment les formes lexicales sont obtenues à partir d'un ensemble réduit d'unités porteuses de sens, appelées morphèmes.
- **Le niveau syntaxique :** la syntaxe est l'ensemble des règles qui définissent l'exactitude des phrases, en effet une suite de mots du lexique ne forme pas forcément une phrase correcte.
- **Le niveau sémantique :** la sémantique est l'étude des significations des mots, et la façon dont ils sont liés les uns aux autres. Une phrase peut être correcte du point de vue syntaxique, sans l'être du point de vue sémantique.

- **Le niveau pragmatique :** la pragmatique est l'étude des aspects du langage qui font référence aux relations entre locuteur et interlocuteur, d'une part et entre interlocuteurs et situation concrète, d'autre part, le sens pragmatique est défini comme dépendant du contexte. [2]

Le processus de production de parole est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique. Il y a une grande quantité d'organes et de muscles qui entrent dans la production de sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre les poumons, le larynx, et les cavités supra-glottiques.

Les poumons, le larynx fournissent ce qui est essentiel pour la production de n'importe quel son, qu'il soit musical ou langagier : une source d'air et une source de bruit.

Les cavités Supra-glottiques renferment les organes qui permettent de modifier le son qui est émis par le travail conjoint des poumons, larynx.

Lorsque l'air est expulsé des poumons, il passe à travers un tube formé de plusieurs cartilages appelé le larynx. Le larynx contient des muscles et des cartilages. Les cartilages les plus importants et les plus connus sont les cordes vocales qui peuvent s'ouvrir et se refermer très rapidement (jusqu'à 400 fois par seconde chez les enfants, par exemple), produisant ainsi des variations de pressions dans l'air. Ces variations de pression sont perçues comme du son par l'oreille humaine.

Les cordes vocales sont gardées ouvertes ou fermées par les aryténoïdes (cartilages en forme de pyramide situés à l'arrière des cordes vocales). Une voix typique d'un homme résulte de mouvements d'ouverture de 100 à 120 fois par seconde (un cycle d'ouverture est appelé un Hertz : Hz) alors que celle d'une femme est produite par entre 175 et 250 vibrations des cordes vocales par seconde. Ce bruit, sera modifié par les divers organes de la parole qui font partie des cavités supra-glottiques. Ces cavités servent à faire résonner le son et à lui donner une « couleur » particulière qui permettra de différencier les voyelles entre elles par exemple, ou les consonnes.

Donc la majorité des sons du langage sont le fait du passage d'une colonne d'air venant des poumons, qui traverse un ou plusieurs résonateurs de l'appareil phonatoire. Les résonateurs principaux sont : le pharynx, la cavité buccale, la cavité labiale, les fosses nasales comme illustré dans la figure I.1 :

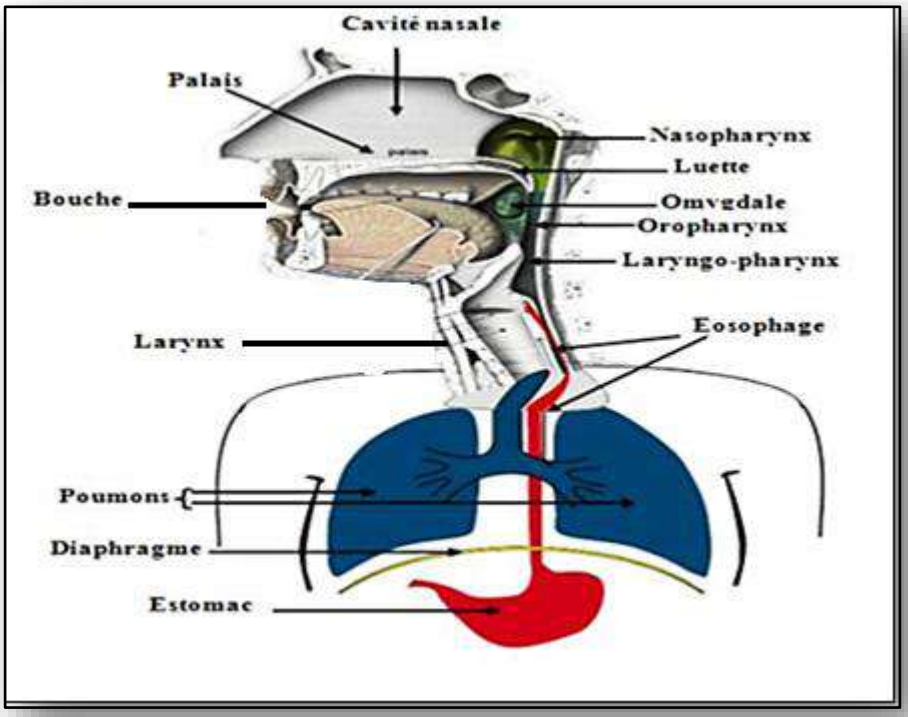


Figure I.1 *: L'appareil phonatoire. [3]*

Un signal parole est représenté phonétiquement comme un ensemble fini de symboles appelés « phonème ».

5. L'appareil auditif :

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur.

5.1 L'appareil auditif humain :

L'oreille est divisée en trois parties distinctes (Figure I.2), cette division se faisant en Fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne.

Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement.

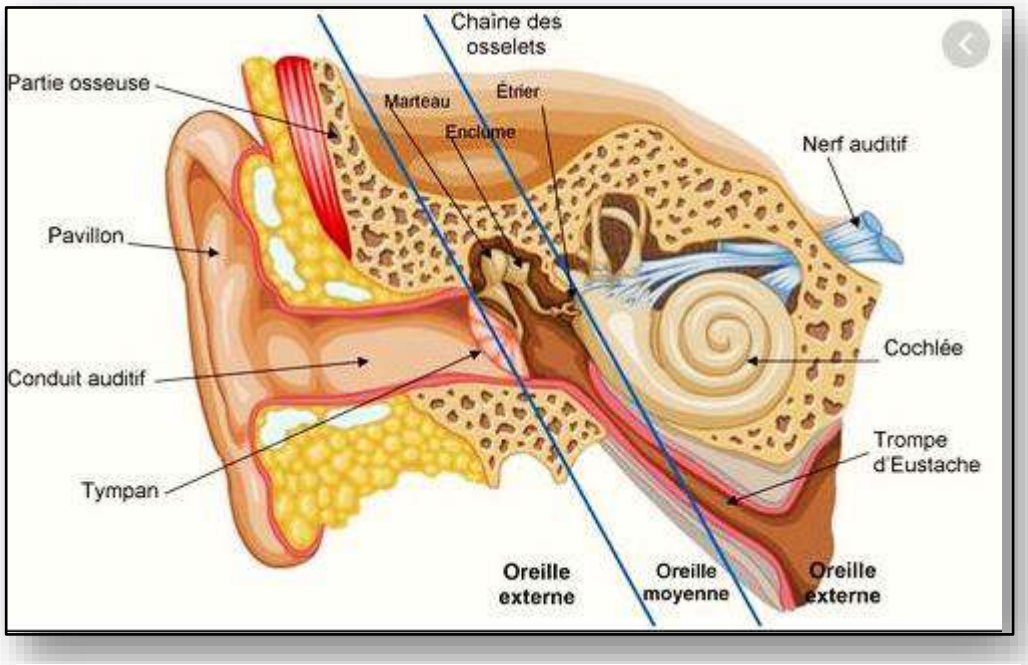


Figure I.2 : *L'appareil auditif humain.* [4]

6. Caractéristiques phonétiques :

6.1 Phonème :

Le phonème est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot. Le nombre de phonèmes est toujours très limité, en générale il est inférieur à 50. [2]

Il est en effet susceptible d'être prononcé de façon différente selon les locuteurs ou selon sa position et son environnement au sein du mot. Les phones sont d'ailleurs les différentes réalisations d'un phonème. L'arabe classique standard a 34 phonèmes parmi lesquels 6 sont des

voyelles et 28 sont des consonnes. Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode articulatoire.

On distingue : [1]

- **Les voyelles** : [i.o.e.u.a…] différent de tous les autres sons par le degré d'ouverture du conduit vocale (et non, comme on l'entend souvent dire, par le degré d'activité des cordes vocales, déjà mentionné sous le terme de **voisement**). Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y'a la production d'une voyelle, [1] les voyelles peuvent être rangées selon : La nasalité, l'ouverture du conduit vocal, la position de la constriction du conduit vocal et l'arrondissement des lèvres.
- **Les consonnes :** Les consonnes se prononcent avec un rétrécissement du passage de 1' air et sont classées selon : le voisement (selon que les cordes vocales vibrent ou non à leur passage), le mode d'articulation (occlusif, nasal, fricatif) et le lieu d'articulation (labiale, dentale, palatale).

L'alphabet phonétique international(API) associe des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations. [1]

7. Les caractéristiques acoustiques :

La parole est produite par l'articulation des membres phonatoires de l'homme et prend une forme analogique apériodique ; ce qui est impossible pour que la machine puisse l'interpréter ou le prédire car elle ne comprend que du numérique. Pour cela on doit faire un traitement de numérisation sur ce signal. L'une des méthodes les plus utilisés dans la numérisation est la méthode Delta ou MIC (Modulation par impulsion et codage) qui consiste en trois étapes : filtre de garde, l'échantillonnage, la quantification, et le codage cette opération est schématisée à la (Figure I.3).

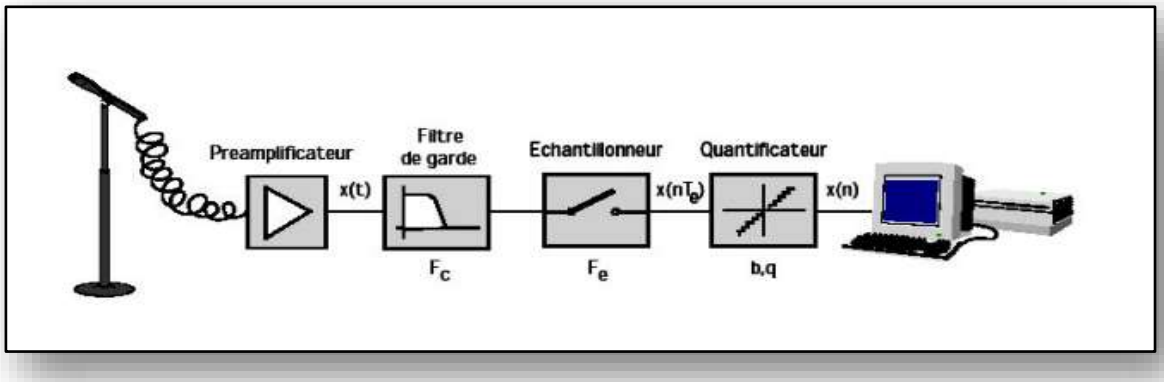


Figure I.3 : *Enregistrement numérique d'un signal acoustique.* [1]

7.1 Echantillonnage :

Transforme le signal à temps continu x(t) en signal à temps discret x(nTe) défini aux instants d'échantillonnage, multiple entiers de la période d'échantillonnage (Te) : pour ce qui concerne le signal vocal, le choix de Fe résulte d'un compromis. Son spectre peut s'étendre jusqu'à 12 KHz. Il faut donc en principe choisir une fréquence Fe égale à 24KHz au moins pour satisfaite raisonnablement au théorème de Shannon.

7.1.1 Théorème de Shannon :

L'information véhiculée par un signal dont le spectre est à support borné, n'est pas modifiée par l'opération d'échantillonnage, à condition que la fréquence d'échantillonnage soit au moins deux fois plus grande que la plus grande fréquence contenue dans le signal.

Les impulsions représentent les amplitudes instantanées du signal à chaque instant. Plus le nombre d'échantillons est grand et plus le signal sera représenté finement, comme illustré dans la figure ci-dessus :

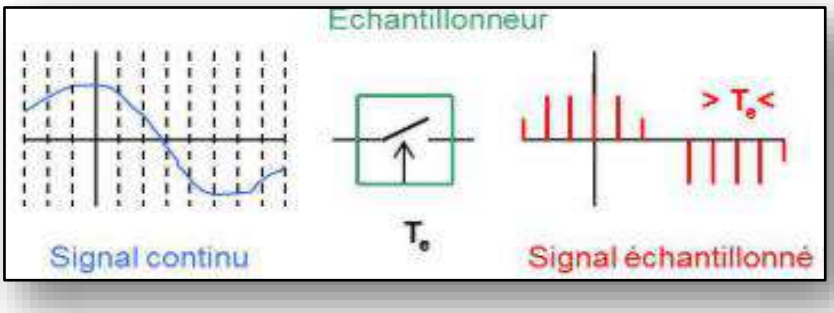


Figure I.4 : *Echantillonnage d'un signal.* [5]

7.2 La quantification :

Quantifier un signal consiste à placer les amplitudes des échantillons sur une échelle de valeurs à intervalles fixes. Chaque impulsion correspond donc à un nombre binaire unique. La quantification ne retient qu'un nombre fini 2^b de valeurs (b étant le nombre de bits de quantification), espacées du pas de quantification q. La transformation d'une valeur physique (en volts) en une valeur binaire introduit donc une distorsion. De même lorsque l'impulsion dépasse la valeur maximale prévue. Le signal numérique résultant est noté x(n).

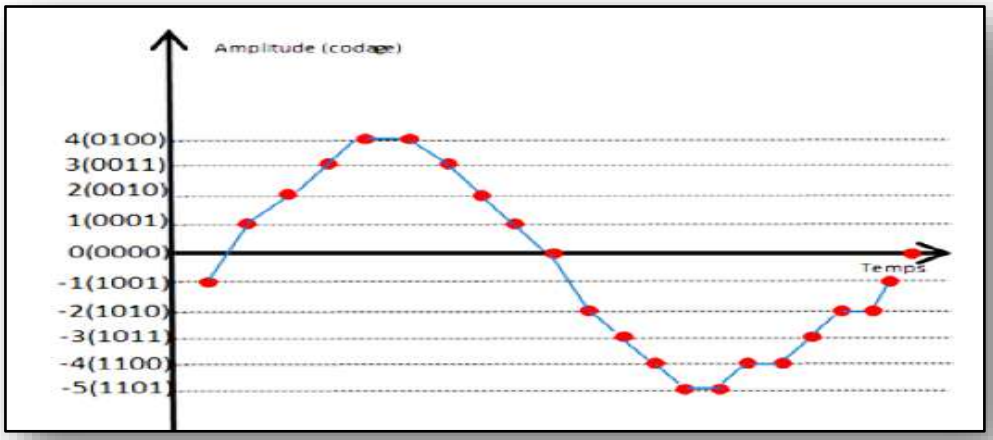


Figure I.5 : *Un signal quantifier.* [6]

La figure I.6 donne la forme temporelle du signal vocal issu d'une prononciation du mot « effacer ». On y constate une alternance de zones assez périodiques et de zones bruitées appelées zones voisées et non-voisées

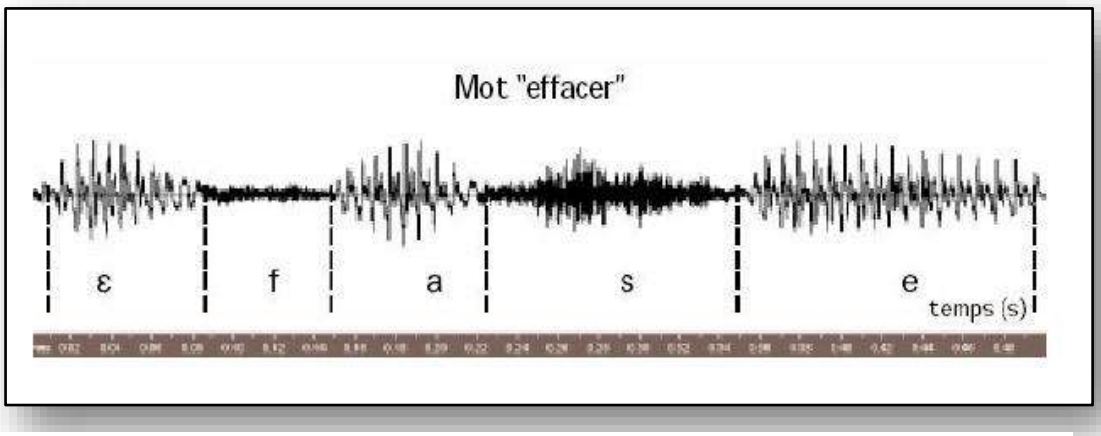


Figure I.6 : *Signal temporel pour le mot « effacer ».* [1]

La figure (I.7) donne une représentation plus fine de tranche de signaux voisés et non-voisés. **Les sons voisés**, tels que des voyelles, sont produits par le passage de l'air de poumons à travers la trachée qui met en vibration les cordes vocales. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité, une énergie élevée et une fréquence fondamentale (pitch). Typiquement, la période fondamentale des différents sons voisés varie entre 2ms et 20ms.

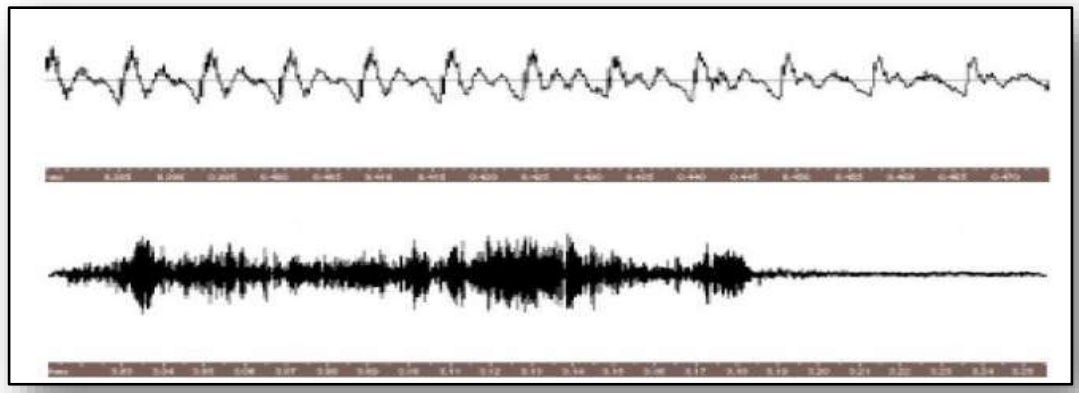


Figure I.7 : *Exemple de son voisé (haut) et non-voisées (bas).* [1]

Pour **les sons non voisés**, comme certaines consonnes, dans ce cas les cordes vocales ne vibrent pas, l'air passe à haute vitesse entre les cordes vocales. Le signal produit est équivalent à un bruit blanc. L'énergie des zones voisées est cependant le plus souvent supérieure à celle des zones non voisées. [1]

8. La reconnaissance :

Consiste à extraire à l'aide d'un ordinateur, l'information lexicale contenue dans un signal parole, et éventuellement de l'interpréter. D'autre manière, c'est le décodage de l'information portée par le signal vocal en séquence de mots par des algorithmes. Parmi les systèmes de reconnaissance, nous pouvons citer :

8.1 Les systèmes de reconnaissance du locuteur :

Les systèmes de reconnaissance du locuteur ont pour but de reconnaître le locuteur, c'est-à-dire la personne qui parle, et trouve un grand champ d'application dans le domaine des serveurs Vocaux mais également de contrôle d'accès.

8.2 Les systèmes de reconnaissance de la parole :

Les systèmes de reconnaissance de la parole ont pour but de reconnaître ce qui est dit (ordres à exécuter par exemple). Ces systèmes sont mis en œuvre dans les applications de dictée vocale, de commande vocale de machines. Ce domaine est appelé reconnaissance de la parole. L'appellation se réfère à plusieurs types de systèmes, c'est une technique qui permet d'analyser et comparer une onde sonore captée au moyen d'un microphone à un ensemble des échantillons, composé généralement de phonème, et la transcrire sous la forme d'un texte exploitable par une machine.

Parmi les techniques utilisées en reconnaissance on peut citer : les modèles de Markov cachés, les réseaux de neurones, la programmation dynamique (DTW).

La figure (I.8) représente le schéma bloc d'un système typique de reconnaissance de la parole.

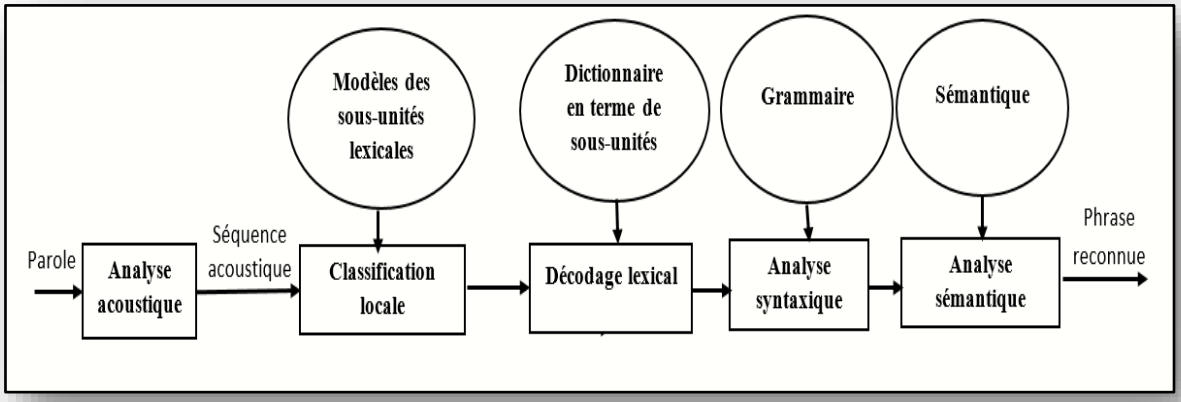


***Figure I.8** : Schéma générale d'un système de reconnaissance de la parole. [1]*

9. Les méthodes utilisées pour la reconnaissance de la parole :

On peut classer les méthodes de reconnaissance en deux :

9.1 La méthode globale :

Dans cette approche l'unité de base, est le mot considéré Comme entité globale non décomposable. Elle a pour avantage d'éviter les effets de coarticulation, dans ce type de méthode on compare globalement le mot ou la phrase, aux différentes références stockées dans un dictionnaire. Dans cette catégorie on trouve principalement la méthode DTW (Dynamic Time Warping), la méthode HMM (Hidden Markov Models), la méthode des réseaux neurones (ANN), la méthode des mélanges de gaussiennes GMM (Gaussian Mixture Models).

9.1.1 Dynamic Time Warping :

Appelée aussi Alignement de Viterbi, introduite par H.Sakoe et Chiba [7]. La reconnaissance basée sur la DTW est plus fiable dans la reconnaissance de parole continue car elle tient compte des compressions et extensions temporelles. Qui sont observées lors de la prononciation plus ou moins rapide d'un mot. Le principe de base est d'essayer de trouver le chemin optimal à parcourir parmi l'ensemble des distances entre les vecteurs (chaque mot du lexique est représenté par une réalisation de référence.

9.1.2 Hidden Markov Model (les modèles de Markov cachés) :

C'est une méthode statistique puissante pour caractériser les échantillons de données observés d'un processus à temps discret. Elle apporte non seulement un moyen efficace de construction de modèles paramétriques, mais elle incorpore aussi le principe de programmation dynamique pour unifier la segmentation et la classification de séquence de données variant dans le temps [8]. Notons que c'est cette approche qui sera adopté de notre travail.

9.1.3 Les réseaux neurones :

Les réseaux de neurones artificiels (ANN), sont un système dont la conception est à l'origine schématiquement inspirée du fonctionnement des neurones biologiques, et qui par la suite s'est rapproché des méthodes statistiques. Ils ont pour principe de s'inspirer (très grossièrement) du fonctionnement des neurones et des synapses du cerveau dans lequel le codage et le traitement de l'information se font au travers du potentiel de sortie (on/off) d'un grand ensemble de neurones. [1]

Une particularité des ANN est qu'ils sont dotés d'algorithmes d'apprentissage qui leur permettent d'apprendre les formes, les classes à reconnaitre et à classer ou bien les problèmes à résoudre. [9]

9.1.4 Mélange de gaussiennes GMM :

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour Gaussian Mixture Models*)* consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes. Ainsi une large gamme de distribution peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble être bien prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, et il est relativement simple à mettre en œuvre. On peut assimiler un modèle GMM à un HMM à un seul état, on ne modélise donc pas

les aspects temporels du signal. Cette méthode est la plus utilisée en reconnaissance du locuteur en mode indépendant du texte [10] [11].

9.2 La méthode analytique :

Dans cette approche est exploitée la structure linguistique du mot. En effet dans ce cas les unités de base sont les phonèmes, les syllabes, et les demi-syllabes, la reconnaissance dans cette méthode passe par la segmentation du signal de la parole en unités de décision, puis par 1'identification de ces unités en utilisant des méthodes de reconnaissance des formes. La première méthode est utilisée pour la reconnaissance des mots isolés, des mots enchaînés, alors que la deuxième méthode est beaucoup mieux adaptée pour les systèmes à grands vocabulaires et pour la parole continue.

10. Analyse et codage de la parole :

L'analyse des signaux de parole tient en considération les façons par lesquelles les sons de parole sont produits. On admet que les signaux de parole résultent de l'interaction de deux facteurs : la source glottique et conduit vocale. Par conséquent, le signal de parole est obtenu par une source (représente le flux d'air au niveau des cordes vocales) passant dans un filtre linéaire à temps variable. Le filtre peut dériver des modèles de production de la parole basés sur la théorie acoustique, ce dernier représente les résonnances du conduit vocal qui change dans le temps. La figure (I.9) représente le processus source-filtre de parole. [12]

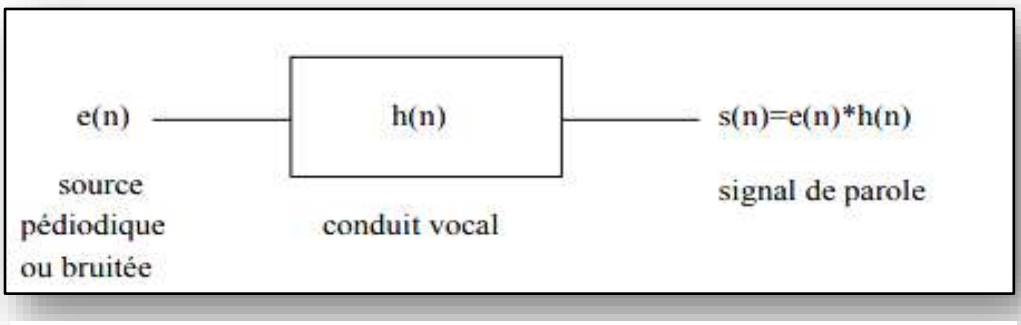


Figure I.9 : *Système d'obtention du signal parole.*

L'analyse des signaux de parole est le processus d'estimation des paramètres d'un modèle, variant dans le temps, dans le but d'extraire des informations ce signal. L'analyse peut être menée dans le domaine temporel ou bien dans le domaine fréquentiel.

10.1 Analyse temporelle:

Les fonctions constituant le modèle source- filtre de la parole sont toutes variables dans le temps. L'excitation est souvent considérée comme un bruit blanc, un signal périodique ou un mélange des deux. Un signal périodique résulte d'une excitation voisée qui se produit lorsque les cordes vocales vibrent. Le bruit aléatoire se manifeste dans le cas d'une excitation non voisée au niveau d'une construction dans le conduit vocal créant un flux d'air turbulent.
On appelle voisement le mélange de ces deux excitations. L'excitation périodique est caractérisée par une fréquence fondamentale appelée « pitch ». [12]
L'analyse temporelle est une analyse qui permet d'obtenir la réponse de la structure à une excitation imposée dans un intervalle de temps déterminé.

10.2 Analyse fréquentielle (spectrale) :

10.2.1 Transformée de Fourier Discrète à court terme :

Les propriétés spectrales du signal vocal présentent un intérêt majeur pour la perception auditive, l'oreille procède en effet à une analyse spectrale de l'onde acoustique reçue. Toutefois, le concept de densité spectrale n'est valable en toute rigueur que pour un signal stationnaire, par contre le signal vocal normal est non stationnaire et il faut introduire la notion de propriétés statistique court-terme [1]. Le signal parole est plus facile à analyser et plus systématique en fréquence qu'en temps. Le spectre de sortie du modèle de production de la parole est le produit de la réponse fréquentielle du conduit vocale et le spectre de l'excitation. Donc il est évident que cette sortie spectrale reflète les propriétés des réponses fréquentielles de l'excitation et du conduit vocal. Un des processus utilisée pour obtenir le spectre d'un signal temporel est la transformée de Fourier (TF). Mathématiquement, on définit la TF comme suit :

$$\mathbf{S}\left(\mathbf{e}^{\mathbf{j2\pi f}}\right) = \sum_{\mathbf{n}=-\infty}^{\infty} \mathbf{S(n)e}^{-\mathbf{j2\pi fn}} \qquad \textbf{(I.1)}$$

Où **S(n)** représente l'échantillon de la séquence.
On va tenir compte du caractère non stationnaire d'un signal x en le pondérant par une fonction fenêtre w repéré par rapport à un instant particulier n comme présenter dans la figure (I.10)
Cette fenêtre peut être de durée infinie à décroissance exponentielle (figure I.10 (b) ou de durée finie, et en général symétrique (fig. I.10 c)

Vu la nature non stationnaire du signal parole, il est impératif de considérer des séquences courtes en temps telle que les caractéristiques du conduit vocale restent inchangées. L'analyse porte sur des tranches de 30ms prélevées dans des zones stables, on dira que le signal est quasi stationnaire et on définit alors une nouvelle forme de la transformée de Fourier nommée la transformer de fourrier court-terme.

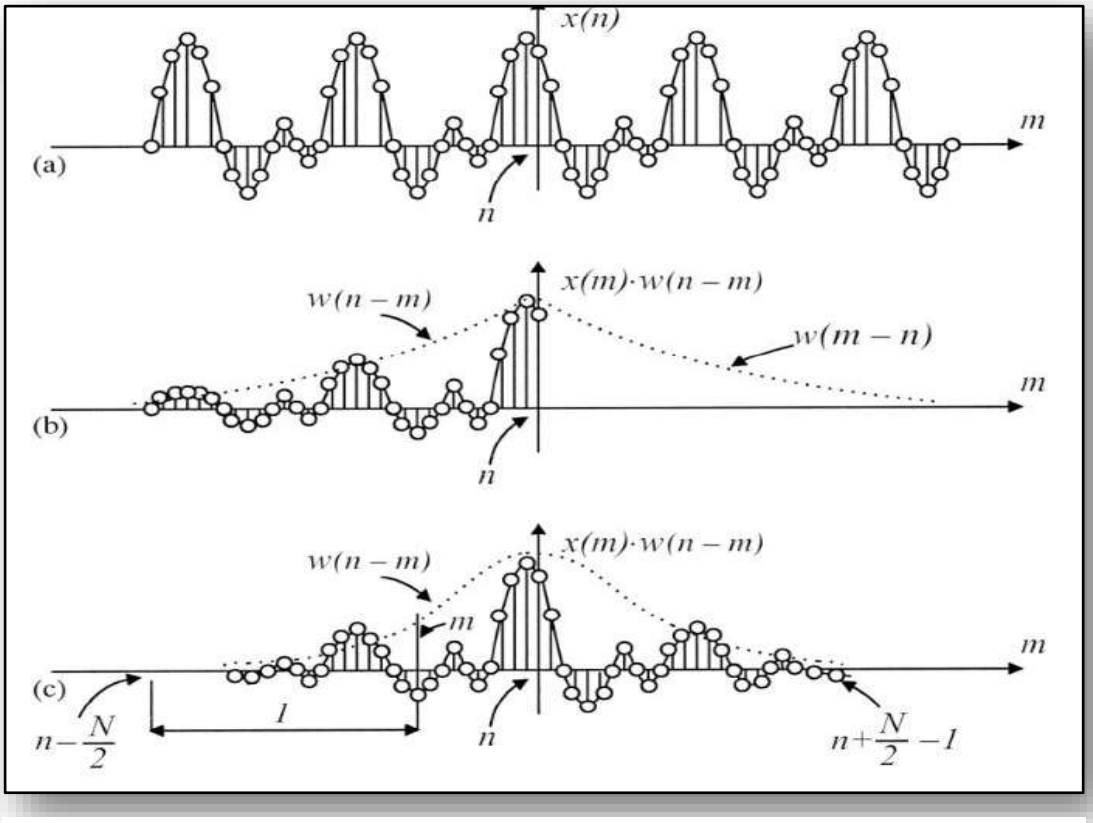


Figure I.10 : *Signal pondéré par une fonction fenêtre :(a) signal, (b) fenêtre de duré infinie, (c) fenêtre symétrique de durée finie N.* [1]

La transformer de fourrier court-terme d'un signal échantillonné est par définition la transformée du signal pondéré : [1]

$$\mathbf{S_m\,(e^{j2\pi f}) = \sum_{n=-\infty}^{\infty} w\,(m-n).\,s(n)e^{-j2\pi fn}} \qquad \mathbf{(I.2)}$$

Où W(n) est la fonction de fenêtrage qui est nulle sauf pour une durée finie. La variable m est utilisé pour sélectionner la partie de séquence à analyser. La fonction S_m est appelé Transformée de Fourier à court terme(TFCT). Pour des raisons de calcul, on échantillonne la fréquence de l'équation (1.2), on obtient par conséquent, la transformé de Fourier discrète comme suit :

$$\mathbf{Sm\,(k) = \sum_{n=0}^{N-1} w\,(m-n)\,.S[n]e^{\frac{-j2\pi nk}{N}},\ pour\ k=0,\ldots,N-1} \qquad \mathbf{(I.3)}$$

L'amplitude du spectre du signal parole définie par la TFD, permet d'extraire plusieurs caractéristiques utiles comme la fréquence fondamentale d'un son voisé à travers la séparation des pics. Les pics les plus larges sont appelés « formant », ils résultent des résonnances du conduit vocal. On désigne par formant (acoustique) d'un son de parole l'un des maxima d'énergie du spectre sonore de ce son de parole. Il y a plusieurs définitions du mot « formant » (résonances du conduit vocal, pôles, etc.). Ils constituent le facteur fondamental de la caractérisation du timbre. La figure I.11 illustre le principe de l'analyse fréquentielle.

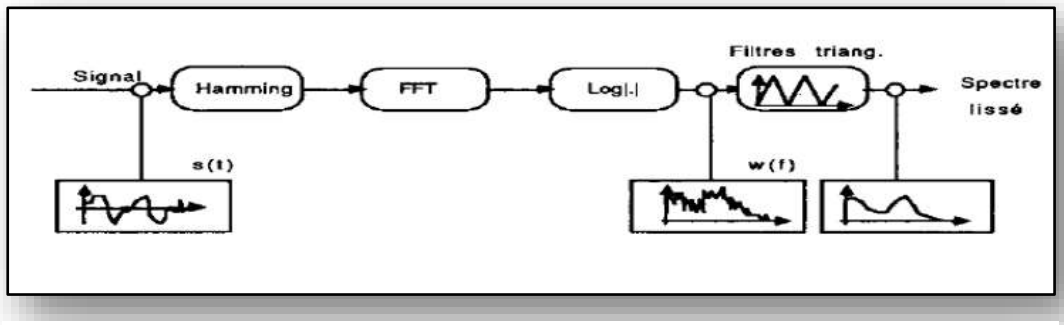


Figure I.11 : *Analyse par transformée de Fourier.* [13]

Pour réduire la distorsion spectrale due au fenêtrage, la fonction W(n) doit satisfaire deux conditions : un lobe principal étroit et fort et, une large atténuation des lobes secondaire. Comme c'est deux conditions sont contradictoires, un compromis doit être fait, ce qui a donné lieu à plusieurs fenêtres. Le tableau I.1 résume les propriétés de fenêtres utilises dans la pratique. Et la figure I.12 donne l'apparence temporelle de chaque fenêtre.

Nom de fenêtre	W[n] dans $\frac{-N-1}{2} \leq n \leq \frac{N-1}{2}$	Largeur du lobe principale (x (π/N)	Niveau des lobes secondaire [dB]
Rectangle	1	4	-13.3
Hanning	$\frac{1}{2}(1+\cos\frac{2\pi n}{N}))$	8	-31.5
Hamming	$0.54+0.46\cos\frac{2\pi n}{N}$	8	-42.7
Blackman	$0.42+0.5\cos\frac{2\pi n}{N}+0.08\cos\frac{4n\pi}{N}$	12	-58.1

Tableau I.1 : *Propriétés de quelques fenêtres connues.*

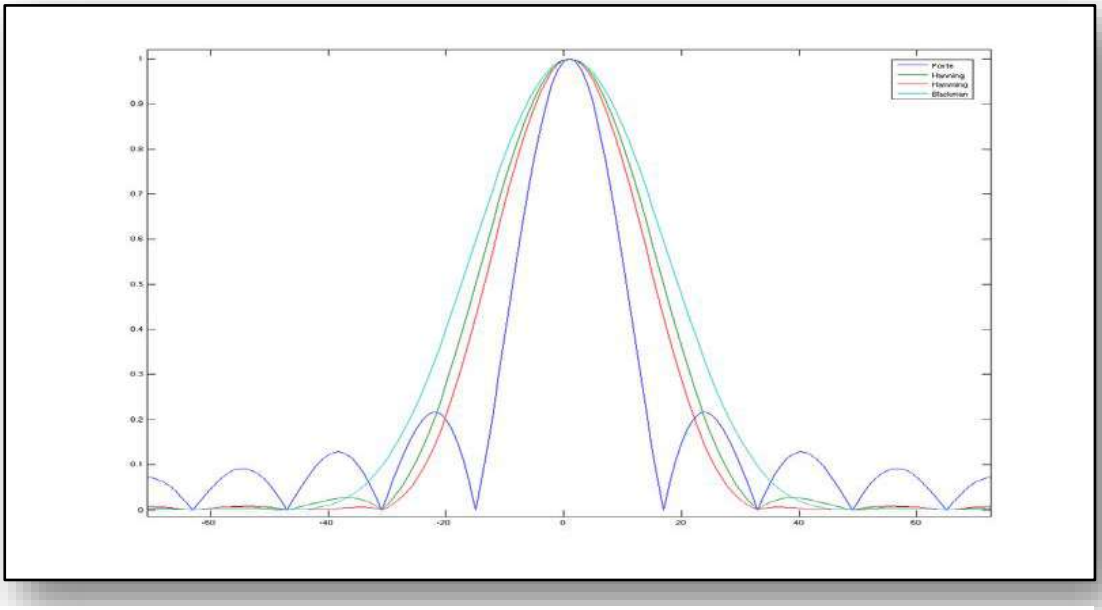


Figure I.12 *: Les fonctions de fenêtrage. [13]*

Dans le domaine de parole La fenêtre la plus utilise est celle de Hamming elle permet une grande atténuation en dehors de la bande passante comparativement à la fenêtre rectangulaire d'où son avantage.

10.2.2 Spectrogramme :

L'intérêt du spectrogramme est de pouvoir représenter le spectre en évoluant dans le temps. Le principe du spectrogramme est de « découper » le son en trames. Pour chacune de ces trames on calcule une transformée de Fourier. L'amplitude du spectre y apparait sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme à bande étroite (figure I.13) ou à large bonde (figure I.14) selon la dure de la fenêtre de pondération.

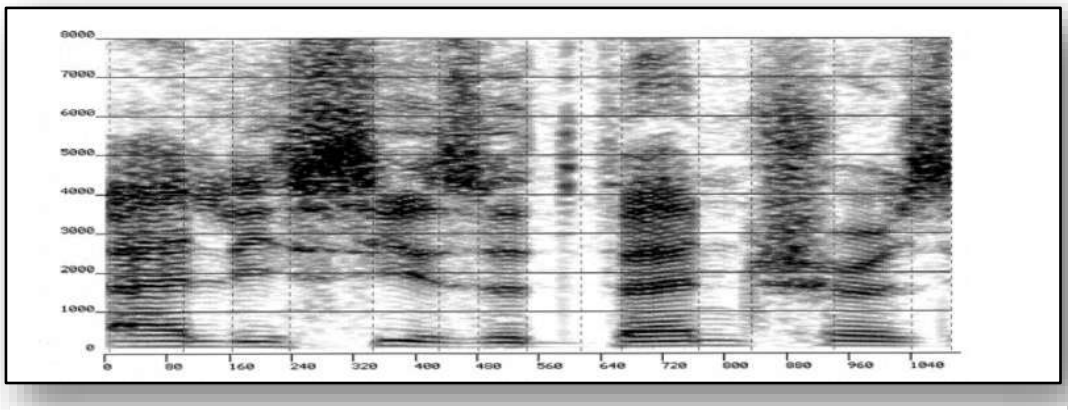


Figure I.13 : *Spectrogramme à bande étroite (d'un son voisé).* [1]

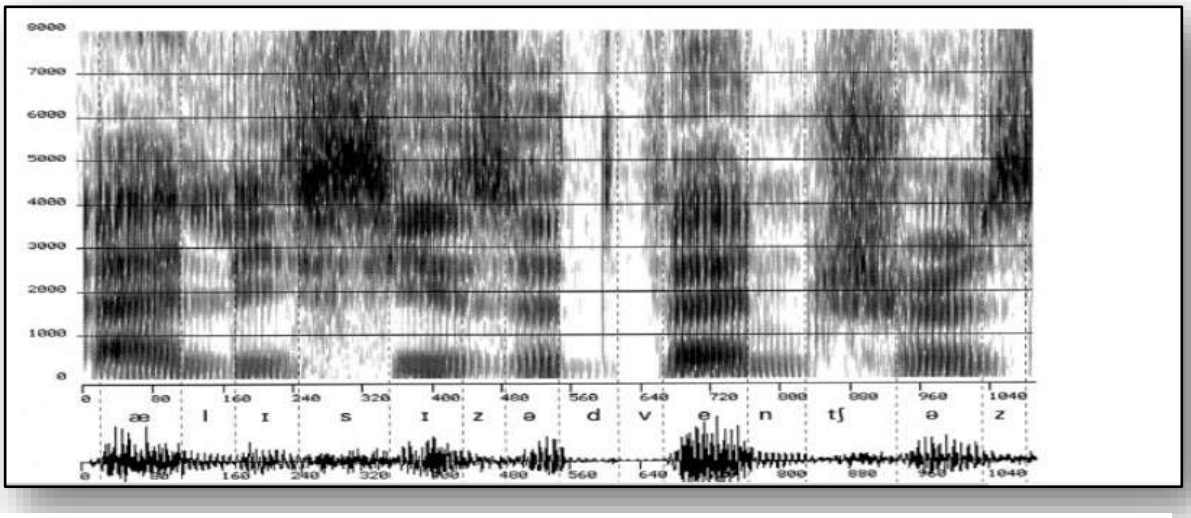


Figure I.14 : *Spectrogramme à large bande (d'un son voisé).* [1]

Les spectrogrammes à large bande sont obtenus avec des fenêtres de faible durée (typiquement 10ms) ; ils mettent en évidence l'enveloppe spectrale du signal, et permettent de visualiser l'évolution temporelle des formants. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine su spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales. [1]

10.3 Bancs de filtre passe bandes :

Le signal électrique obtenu à la sortie du microphone (après avoir été amplifié et pré accentué) passe à travers un banc de filtres passe-bande (figure I.15). Le signal à la sortie de ces filtres passe ensuite à travers un circuit détecteur d'enveloppe, on obtient, ainsi, une approximation du module de la TF à court terme dans différentes bandes de filtres.

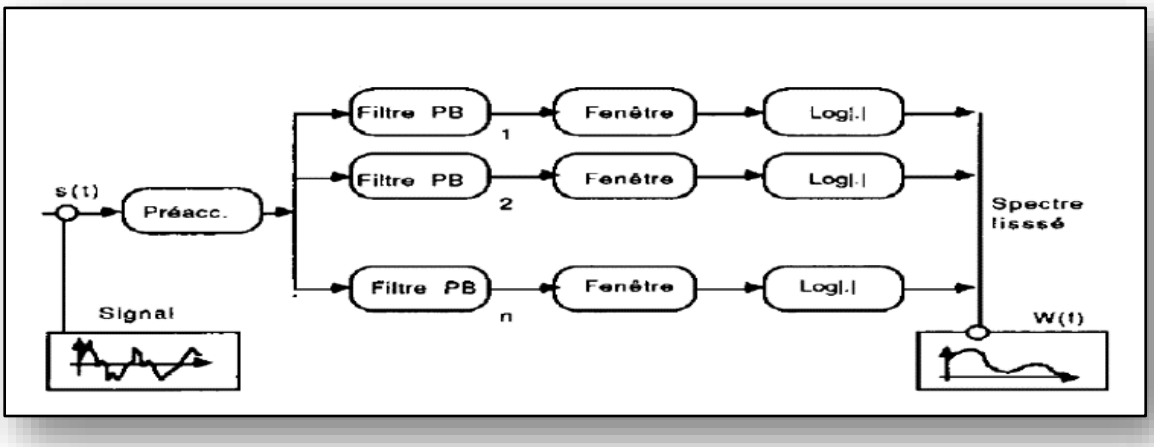


Figure I.15 : *Analyse par bancs de filtre.* [13]

10.4 Analyse cepstrale :

Parmi les paramètres les plus couramment adopté pour l'analyse acoustique, nous citons ici le cepstre. Il s'agit d'une méthode permettant d'étudier séparément la source et le conduit vocal. Le cepstre est basé sur une connaissance du mécanisme de la production de la parole. On part de l'hypothèse que la suite d'échantillons constituant le signal vocal est les résultats de la convolution du signal de la source par le filtre correspondant au conduit :

$$s = u * h \quad \textbf{(I.4)}$$

Où **s** est le signale de la parole, u le signal d'excitation (source) et h la fonction de transfert du conduit vocal (filtre).

Le but est de séparer ces deux contributions (source et filtre)[3] par application de la dé-convolution à travers une Transformée en Cosinus Discret :

$$S = TFD(s) = U.H \quad \textbf{(I.5)}$$

Le logarithme de l'amplitude transforme le produit de la TFD en somme. On obtient alors :

$$\log |S(v)| = \log |U| + \log |H| \quad \textbf{(I.6)}$$

Par transformation en cosinus discret (DCT), on obtient le cepstre, l'expression du cepstre réel est :

$$C = DCT(\log(TFD(s))) \quad \textbf{(I.7)}$$

Les résultats du calcul cepstral est une séquence temporelle comme le signal d'entrée lui-même. Si le signal d'entrée possède une période de hauteur fondamental forte, elle Apparaît dans le cepstre sous forme de pic, en mesurant la distance entre le temps zéro et le temps pic on trouve la période fondamentale de cette hauteur [14]. Les coefficients cepstraux permettent la séparation des effets du conduit vocal et de la source périodique. En effet, la contribution du conduit vocal est représentée par l'enveloppe du spectre de fréquence et affecte en conséquence les premiers coefficients cepstraux. Les caractères plus fins du spectre de fréquence, qui représentent la source périodique, sont décrits par les coefficients cepstraux d'ordre supérieur.

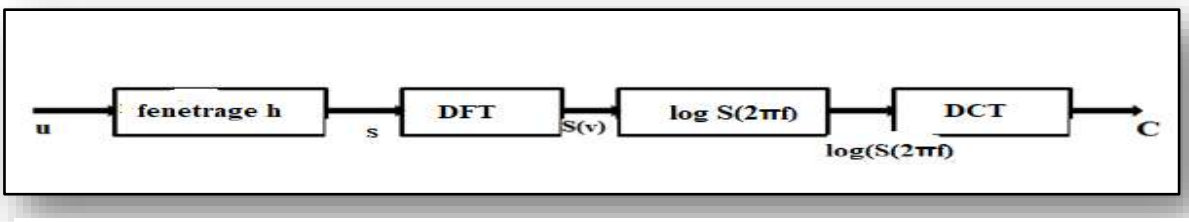


Figure I.16 : *Etapes de calcul du cepstre.*

Le type de coefficients cepstraux le plus utilisé est appelé MFCC (Mel Frequency Cepstral Coefficients). Leur calcul utilise une echelle fréquentielle non- linéaire ou échelle Mel. La fréquence Mel est définie à partir de la fréqence abituelle par :

$$B(f) = 2595 \log_{10}(1+\frac{f}{700}) \quad \textbf{(I.8)}$$

Où *f* est la fréquence en Hz et B*(f)* est la fréquence correspondante sur l'echelle de Mel. L'interet de l'echelle Mel est d'étre assez proche de l'échelle issue des études sur la perception sonore et sur les bandes passantes critiques de l'oreille. L'idée est de moyenner le spectre dans des bandes de fréquences correspondant grossiérement au filtrage effectué par le membrane basiliaire. L'echelle Mel est en général approchée par un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1 KHz, puis espacés logarithmiquement jusqu'aux fréquences maximales. Elle a été concue de telle facon que 1000 Hz corespondonde à 1000 Mels. La figure I.17 représente l'echelle de Mel et la figure I.18 le banc de filtres correspondant.

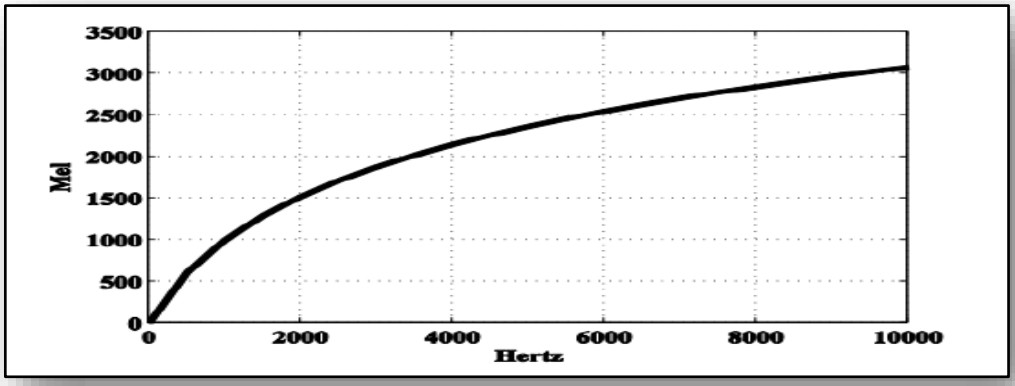


Figure I.17 : *Echelle de Mel.* [15]

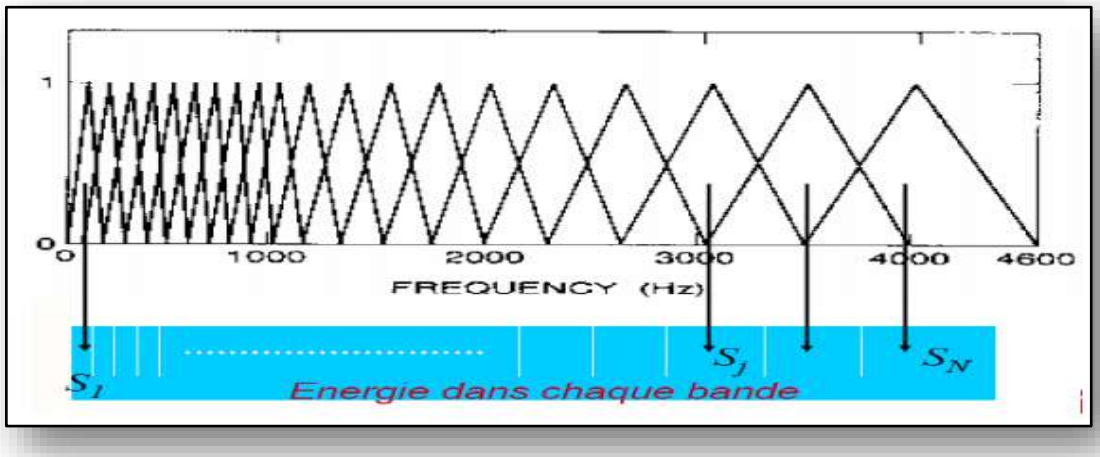


Figure I.18 : *banc de filtre à l'échelle de Mel.* [15]

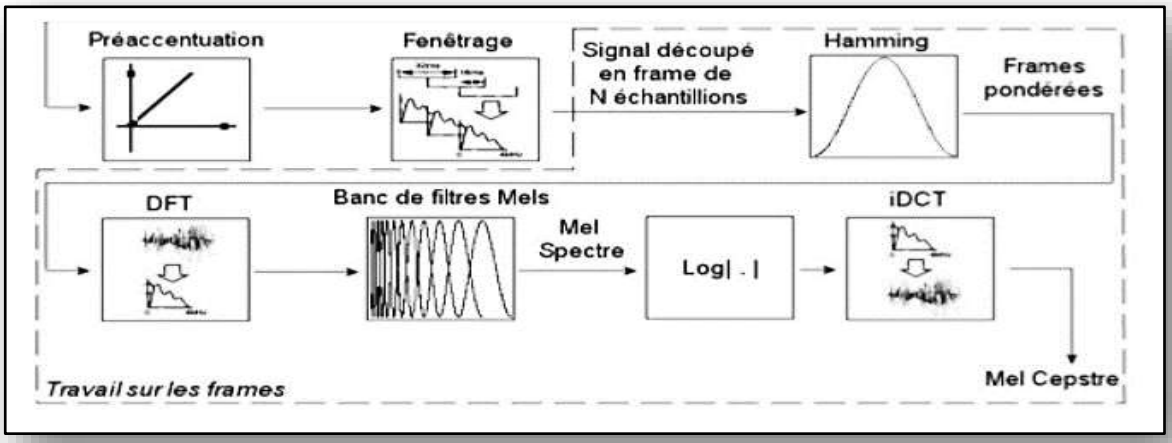


Figure I.19 : *Etapes de calcule MFCC.* [16]

La figure I.19 illustre les étapes détaillées de la façon de calcule des MFCCs :

- **Préaccentuation** : Le rôle de la préaccentuation est d'augmenter les hautes fréquences. Et ce pour se caler sur notre perception des aiguës. Il serait dommage de négliger ces zones qui contiennent beaucoup d'information (c'est à cet endroit que l'on retrouve l'énergie des fricatives).
- **Fenêtrage :** découpage du signal en trames. Il se fera en petits morceaux de 20 à 30 ms toutes les 10-15 ms à partir desquels nous obtiendrons nos coefficients pour chacun d'entre eux. Pourquoi découper le signal ? Par ce que le signal de parole est considéré comme stationnaire sur des intervalles de temps de 20 à 30 ms.
- **Pondération de notre fenêtre carrée par une fenêtre de Hamming** : ceci a été expliqué dans la section 10.2.
- **Estimation du spectre** : Passage du domaine temporel vers le domaine spectral ! Ici on calcule une estimation du spectre de chaque trame à l'aide de la transformée de Fourier discrète (DFT – Discrete Fourier transform).
- **Bancs de filtres Mels :** nous allons appliquer un banc de filtres dont les largeurs de bandes et l'espacement correspondent à l'échelle Mel.
- **Calcul du logarithme de chaque énergie :** Ceci est également motivé par l'ouïe humaine : nous ne percevons pas l'intensité sonore au travers d'une échelle linéaire. En

gros, pour percevoir un son deux fois plus fort, il faut multiplier par huit sa puissance d'émission.

- **Calcul de la transformé en cosinus discrète inverse (IDCT) :** Cette transformée nous fait revenir dans un pseudo-domaine temporel (d'où le nom de cepstre, quéfrence, …). En général on ne garde que les 13 premiers coefficients. Le premier étant proportionnel au log de l'énergie moyenne il est tout simplement remplacé par l'énergie de la trame. A ces 13 coefficients s'ajoutent leur première et deuxième dérivée pour avoir enfin 39 coefficients.

11.Conclusion :

Le présent chapitre a été exposé dans l'esprit de représenter une brève introduction au traitement de la parole en se limitant aux notions essentielles en relation avec notre travail. Ainsi ont été exposées les propriétés du signal de parole, les techniques de reconnaissance les plus connues et la méthodologie d'analyse du signal vocal en vue de sa préparation pour l'étape de reconnaissance. Le prochain chapitre sera entièrement consacré aux modèles de Markov Cachés (MMC) et à leur application dans la reconnaissance automatique de la parole.

CHAPITRE II :

Les modèles de Markov cachés

1. Introduction :

Les modèles de Markov cachés sont des outils statistiques permettant de modéliser des phénomènes stochastiques. Ces modèles sont utilisés dans de nombreux domaines tels que la reconnaissance et la synthèse de la parole, la biologie, l'ordonnancement, l'indexation de document et la prédiction de séries temporelles. Ce chapitre a pour objectif d'établir les principes théoriques, les notations utiles ainsi que les principaux algorithmes qui constituent la théorie des modèles de Markov cachés (MMC). [17]

2. Modèles de Markov cachés MMC (ou HMM) :

Les HMMs (Hidden Markov Models) ont été introduits par Baun Welch à la fin des années 60 et utilisés à partir des années 80 en reconnaissance de la parole. Un HMM peut-être décrit comme un automate probabiliste à états finis comportant deux processus : un processus caché et un processus d'émission. Le premier processus est dit caché car il est non observable, il concerne la suite (chaine) d'états visités alors que par le second est observable car il concerne la suite d'observations émises par chaque état de la chaine cachée [18]. A chaque transition est associé un symbole (observation) d'un alphabet fini ou continu. Un HMM se définit donc par une structure composée d'états et de transitions et par un ensemble de distributions de probabilité sur les transitions.

Les modèles de Markov cachés sont utilisés pour modéliser (générer) des séquences d'observations. Ces observations peuvent être de nature discrète (par exemples les caractères d'un alphabet fini) ou continue (fréquence d'un signal, température). Dans notre cas de figure les symboles émis seront des vecteurs spectraux (MFCC), chaque vecteur étant composé de 39 valeurs réelles. Pour le signal de parole un HMM servira à modéliser une succession d'unités acoustiques. Un HMM est caractérisé par les paramètres suivants:

- **$S= \{S_1, S_2, \ldots, S_N\}$** : L'ensemble des états du modèle avec **N** le nombre d'états, contenant un état initial et état final et des états émettant des symboles. On note que **q_t** L'état courant à l'instant **t**.
- **$V= \{V_1, V_2, \ldots, V_M\}$** : un alphabet des observations avec **M** nombre fini ou non de symboles d'observation par état. On note **Q_t** l'observation à l'instant t.

- **A= {a_{ij}} :** une matrice des probabilités de transition entre états, dont a_{ij} est la probabilité de transition de l'état i à l'état j. sachant que n'existe pas de transition pour aller vers l'état initial et qu'il n'existe pas de transition partant de l'état final. on a :

$$a_{ij} = P((q_{t+1} = S_j) / (q_t = S_i)),\ 1 \leq i, j \leq N \quad \textbf{(II.1)}$$

- **B = {$b_j(k)$} :** une matrice des probabilités d'émission des observations dans chaque état, dont **$b_j(k)$** est la probabilité d'émission de l'observation **V_k** dans l'état **S_j**. On a :

$$b_j(k) = P((v_k = O_t) / (q_t = s_j)),\quad 1 \leq j \leq N,\ 1 \leq k \leq M \quad \textbf{(II.2)}$$

- **π= {π_i} :** une matrice de distribution de l'état initial, ou **π_i** est la probabilité d'etre dans l'état S_i à l'instant initial. On a :

$$\pi_i = P(q_1 = s_i),\quad 1 \leq i \leq N \quad \textbf{(II.3)}$$

La définition complète d'un HMM nécessite donc la spécification des paramètres cités plus haut [18] et qui seront désignés par le triplet ***λ= {A, B, π}***

Une chaine de Markov est un type particulier de processus stochastique qui vérifie deux conditions :

- Hypothèse de Markov : l'état au temps t dépend uniquement de son état au temps t−1 :

$$P(q_t = S_j / q_1 = S_k, \ldots, q_{t-1} = S_i) = P(q_t = S_j / q_{t-1} = S_i) \quad \textbf{(II.4)}$$

- Propriété stationnaire : la probabilité de passage d'un état **i** à un état **j** est constante, ne varie pas avec le temps :

$$P((q_t = S_j) / (q_{t-1} = S_i)) = a_{ij},\ \forall t\ \ 0 < t \leq T \quad \textbf{(II.5)}$$

2.1 Topologies des HMM :

La structure du graphe reliant les états est appelée topologie du HMM. Dans le choix d'une topologie, il y a un compromis entre simplicité et complexité. Un modèle trop simple ne sera probablement pas en mesure de générer l'ensemble de données avec une forte probabilité, tandis qu'un modèle trop complexe apprendra facilement l'ensemble de données, mais il est peu probable qu'il se généralise bien. La complexité dépend du nombre d'états et de la structure. Les structures les plus complexes sont des graphes entièrement connectés, tandis que les plus simples sont des chaînes linéaires [21].

La topologie du modèle de Markov est fixée à priori et n'est pas modifiée lors de l'apprentissage. En effet, seules les probabilités de transition entre les différents états et les probabilités d'observation à l'intérieur de ces derniers sont ajustées à l'aide des algorithmes classiques d'apprentissage dans les modèles de Markov. Le choix de la topologie initiale est donc un problème délicat.

Il existe de nombreux types des topologies dans le modèle de Markov caché. Les deux qui nous intéressent dans ce travail sont illustrées dans la figure ci-dessous :

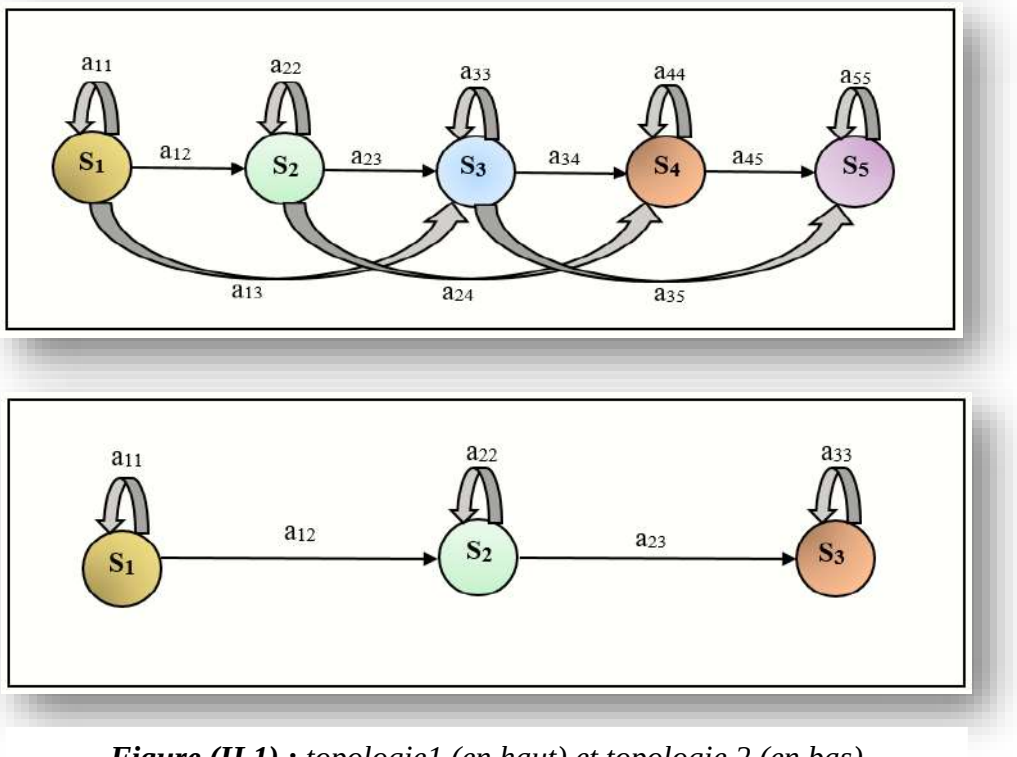


Figure (II.1) : *topologie1 (en haut) et topologie 2 (en bas)*

Ce sont des modèles dits Gauche-Droite. Chaque état se connecte à lui-même (auto-transitions) et à l'état suivant et peut se connecter à l'état plus lointain pour la topologie 1. Les auto-transitions expliquent les variations de la durée des motifs dans un état [20], dans notre cas il exprime la durée d'une unité phonétique. Le Mouvement entre les états est de gauche à droite uniquement, on ne peut pas retourner à l'état précédent. La topologie 1 est communément appelée modèle de BAKIS.

3. Reconnaissance des mots isolés :

Les HMMs peuvent être utilisés de plusieurs façons en RAP, selon l'importance de l'application (taille du vocabulaire et type de parole : mots isolés ou parole continue). Pour la reconnaissance de mots isolés, il est possible de modéliser chaque mot par un HMM, La reconnaissance revient alors à calculer la vraisemblance de la suite d'observations acoustiques constituant le mot à reconnaître par rapport à chacun des modèles. Le modèle présentant la plus grande vraisemblance d'avoir émis cette suite d'observations fournit le mot reconnu. L'algorithme permettant d'optimiser ce calcul est fondé sur la programmation dynamique, mais dans un cadre stochastique, l'algorithme de Viterbi. [19]

Chaque mot prononcé doit être représenté par une séquence de vecteurs de parole ou d'observations O, définie comme

$$\mathbf{O = O_1, O_2, ..., O_T} \qquad \textbf{(II.6)}$$

Où $\mathbf{O_t}$ est le vecteur de parole observé au temps t. Le problème de reconnaissance des mots isolés peut alors être considéré comme celui du calcul de

$$\mathbf{\underset{i}{argmax}\{P\ (w_i\ /\ O)\}} \qquad \textbf{(II.7)}$$

Où $\mathbf{W_i}$ est le $i^{ème}$ mot de vocabulaire. Cette probabilité n'est pas directement calculable mais l'utilisation de la règle de Bayes donne :

$$\mathbf{P\ (w_i\ /\ O) = P\ (O\ /\ w_i)\ P\ (w_i)\ /\ P\ (O)} \qquad \textbf{(II.8)}$$

Ainsi, pour un ensemble donné de probabilités antérieures $\mathbf{P\ (W_i)}$, le mot prononcé le plus probable ne dépend que de la vraisemblance $\mathbf{P\ (W_i\ /O)}$.

Dans la reconnaissance vocale basée sur HMM, il est supposé que la séquence de vecteurs spectraux observés correspondant à chaque mot est générée par un modèle de Markov comme indiqué sur la figure II.2. Dans notre contexte, un modèle de Markov sera donc une machine à états finis qui change d'état une fois à chaque unité de temps et à chaque fois qu'un état j est entré, un vecteur de parole O_t est généré à partir de la densité de probabilité bj (O_t). De plus, la transition de l'état i à l'état j est également probabiliste et est régie par la probabilité discrète a_{ij}. La figure II.2 montre un exemple de ce processus dans lequel la séquence d'états ayant généré notre mot est X = 1, 2, 2, 3, 4, 4, 5, 6. Le mot en question est $O=O_1O_2O_3O_4O_5O_6$.

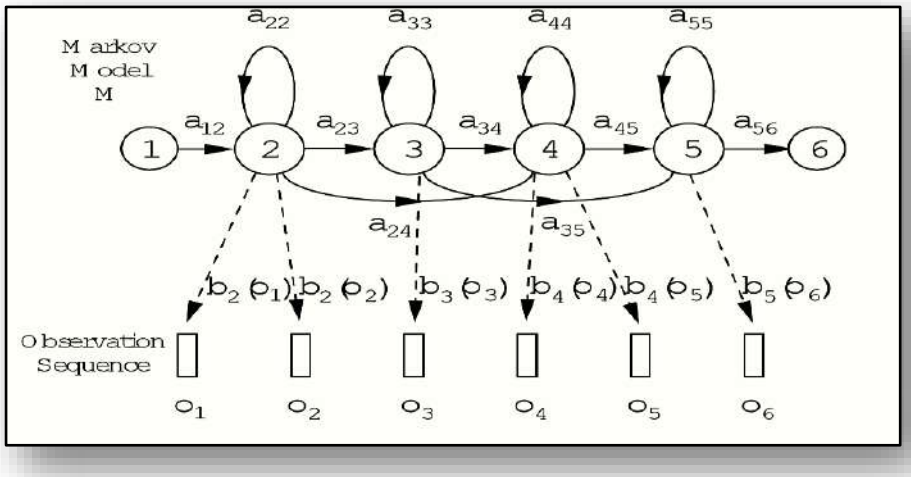


Figure II.2 *: Le modèle de génération de Markov* [17]

La probabilité conjointe que O soit généré par le modèle M se déplaçant dans la séquence d'états X est calculée simplement comme le produit des probabilités de transition et des probabilités de sortie. Donc, pour la séquence d'états X de la figure II.2.

$$\boldsymbol{P(O,X/M)} = \boldsymbol{a_{12}b_2(O_1)a_{22}b_2(O_2)a_{23}b_3(O_3)a_{34}b_4(O_4)a_{44}b_4(O_5)a_{45}b_5(O_6)a_{56}} \ldots. \quad \textbf{(II.9)}$$

Cependant, en pratique, seule la séquence d'observation O est connue et la séquence d'états sous-jacente X est masquée, c'est pourquoi on l'appelle un modèle de Markov caché.
Étant donné que X est inconnu, la vraisemblance requise est calculée en sommant toutes les séquences d'états possibles X = x (1), x (2), x (3), ..., x (T), c'est-à-dire

$$\boldsymbol{P(O/M)} = \sum_X \boldsymbol{a_{X(0)X(1)}} \prod_{t=1}^{T} \boldsymbol{b_{X(t)}(O_t)a_{X(t)X(t+1)}} \quad \textbf{(II.10)}$$

Où x(0) est contraint à être l'état d'entrée du modèle et x (T +1) est contraint à être l'état de sortie du modèle.

Comme alternative à l'équation II.10, la probabilité peut être approximée en considérant uniquement la séquence d'états la plus probable, c'est-à-dire

$$\boldsymbol{P(O/M)} = \max_x \{\boldsymbol{a_{X(0)X(1)}} \prod_{t=1}^{T} \boldsymbol{b_{X(t)}(O_t)a_{X(t)X(t+1)}}\} \quad \textbf{(II.11)}$$

Bien que le calcul direct des équations II.10 et II.11 ne soit pas traitable, il existe des procédures récursives simples qui permettent de calculer les deux quantités très efficacement. Avant d'aller plus loin, cependant, notez que si l'équation II.7 est calculable, le problème de reconnaissance

est résolu. Étant donné un ensemble de modèles M_i correspondant aux mots W_i, l'équation II.8 est résolue en utilisant II.12 et en supposant que

$$\boldsymbol{P(W_i/O) = P(O/M_i)} \quad \textbf{(II.12)}$$

La figure II.3 résume l'utilisation des HMM pour la reconnaissance de mots isolés. Tout d'abord, un HMM est formé pour chaque mot de vocabulaire en utilisant un certain nombre d'exemples de ce mot. Dans cet exemple, le vocabulaire est celui utilisé dans notre travail. Deuxièmement, pour reconnaître un mot inconnu, la probabilité que chaque modèle génère ce mot est calculée et le modèle le plus probable identifie le mot. [17]

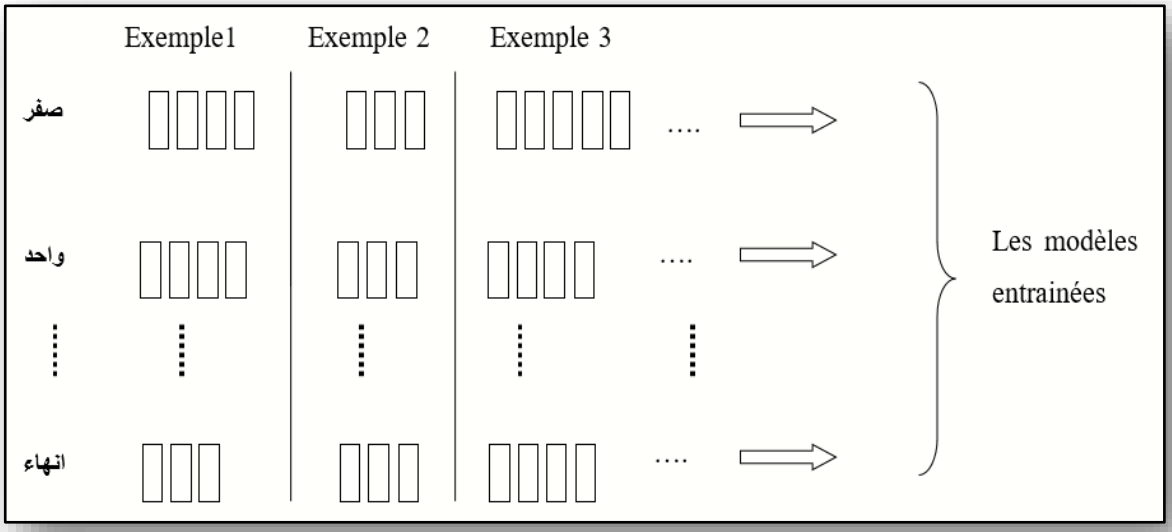


Figure II.3 : *Utilisation des HMMs dans la reconnaissance des mots isolés.*

4. Les problèmes majeurs en HMMs :

La manipulation des HMMs soulève trois problèmes principaux :

- **Calcul de la vraisemblance** : C'est l'évaluation de la probabilité de l'observation d'une séquence. Etant donnée une suite d'observations $O = \{O_1, O_2, O_3,..., O_t\}$ et un modèle $\lambda=(A,B, \pi)$, on cherche à évaluer la probabilité d'apparition de cette séquence connaissant le modèle. Pour résoudre ce problème on utilise l'algorithme Forward-Backward.
- **Le problème d'apprentissage :** qui cherche à ajuster les paramètres du modèle

$\lambda = (A,B,\pi)$, pour maximiser $P(O/\lambda)$, à partir d'un ensemble de séquences d'apprentissages $O = \{O_1, O_2, O_3,..., O_t\}$ qui ont été émises par ce modèle. L'algorithme de Baum Welch permet de résoudre ce problème.

- **Le problème de la reconnaissance :** C'est la recherche du chemin le plus probable, ou l'estimation de la partie cachée. Etant donnée une suite d'observation $O=\{O_1,O_2,O_3,...,O_t\}$ et un modèle $\lambda = (A, B, \pi)$, on cherche à trouver la suite d'états, appartenant à $X = \{x_1, x_2, x_3,..., x_T\}$, qui explique le mieux l'observation. Pour ce faire, on fait appel à l'algorithme de Viterbi qui permet de montrer la séquence d'états la plus probable produisant cette séquence d'évènements.

5. Calcul de la vraisemblance d'une séquence:

Une séquence donnée peut en général être engendrée de plusieurs façons distinctes par un HMM. La vraisemblance d'une séquence O peut s'exprimer sous la forme suivante :

$$P(O(t)/\lambda) = \sum P(O(t), q(t)/\lambda) \qquad \textbf{(II.13)}$$

La somme est faite pour toutes les séquences possibles d'états q(t). Il s'agit là d'un calcul direct de la probabilité $P(O(t)/\lambda)$. Mais cette façon de procéder nous mène à des calculs d'ordre exponentiel ($T \times N^T$), qui sont très coûteux en temps de calcul.

Par la suite, nous décrivons un procédé récursif plus astucieux et plus rapide qui permet réduire le temps de calcul de la probabilité II.13. Il s'agit de l'algorithme Forward-Backward ou progressif-rétrogressif [21]**.** La vraisemblance de observation est évaluée par :

$$\mathbf{P(O/\lambda) = \sum_{i=1}^{n} \alpha_t(i)\beta_t(i)} \qquad \textbf{(II.14)}$$

Où $\alpha_t(i)$ est la probabilité d'émettre le début $O_{(1:t)}$ et d'être à q_i à l'instant t, et $\beta_t(i)$ est la probabilité d'émettre la fin $O_{(t+1:T)}$ sachant que l'on part de q_i à l'instant t, le calcul de α se fait avec t croissant tandis que le calcule de β se fais avec t décroissant, d'où l'appellation Forward-Backward. La probabilité d'observation d'une séquence est obtenue en prenant les valeurs de α et β à un instant t quelconque par l'équation II.18. Cependant, on utilise le plus souvent les valeurs obtenues pour deux cas particulier (t=0) ou (t=T).

6. L'apprentissage :

Il s'agit de modifier les paramètres d'un HMM à partir d'un certain nombre d'exemples pour que le modèle représente de manière fiable cette classe d'exemples. Les formules de ré-estimation sont :

$$\widehat{a}_{ij} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^r(t)a_{ij}b_j(O_{t+1}^r)\beta_j^r(t+1)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^r(t)\beta_j^r(t)} \quad \textbf{(II.15)}$$

$$\widehat{a}_{1j} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{P}\alpha_j^r(1)\beta_j^r(1) \quad \textbf{(II.16)}$$

$$\widehat{a}_{iN} = \frac{\sum_{r=1}^{R}\frac{1}{P_r}\alpha_i^r(T)\beta_i^r(T)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T}\alpha_i^r(t)\beta_i^r(t)} \quad \textbf{(II.17)}$$

Pour les gaussiennes simples, il est plus efficace dans ce cas d'utiliser **:**

$$l_j^r(t) = \frac{1}{p_r}\alpha_j(t)\beta_j(t) \quad \textbf{(II.18)}$$

Où P_r**= P (O/M).**

Les formules de ré-estimation peuvent maintenant être exprimées en terme de $l_j^r(t)$ **:**

$$\widehat{\mu}_J = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} l_j^r(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} l_j^r(t)} \quad \textbf{(II.19)}$$

$$\widehat{\sum}_j = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} l_j^r(o_t^r-\widehat{\mu}_j)(o_s^t-\widehat{\mu}_j)'}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} l_j^r(t)} \quad \textbf{(II.20)}$$

Avec :

R : nombre d'exemples utilisés pour chaque mot.

Tr : longueur du mot en termes de nombre de vecteurs spectraux.

$\hat{\mu}_J$ **:** Moyenne estimée de la gaussienne utilisée pour modéliser la distribution d'émission de l'état j.

$\widehat{\sum}_j$ **:** Matrice de covariance de la gaussienne utilisée pour modéliser la distribution d'émission de l'état j.

7. La Reconnaissance :

A partir d'une séquence d'observation représentant un mot à reconnaitre, on cherche le modèle qui donne le maximum de probabilité de génération de la séquence d'observation. L'algorithme utilisé est l'algorithme de Viterbi. Sa particularité réside dans le fait que, dans le calcul de la vraisemblance, au lieu de considérer toutes les séquences d'états possibles, on garde seulement la séquence d'états donnant la plus grande probabilité. Tous les algorithmes précédemment abordés sont implémentés dans la plateforme HTK que nous utiliserons pour manipuler les modèles de Markov cachés.

8. Conclusion :

Dans ce chapitre nous avons présenté le concept des modèles de Markov cachés (HMMs), leur fonctionnement ainsi que leur emploi dans le domaine de la reconnaissance de la parole. D'après ce qu'on vient de voir, la modélisation Markovienne repose sur un formalisme à la fois simple et rigoureux. A ces deux points forts viendra s'ajouter un troisième qui est la robustesse comme on le verra à travers les résultats de reconnaissance exposés dans le dernier chapitre. Le chapitre suivant, sera consacré à décrire notre mise en pratique des modèles de Markov cachés pour de la RAP en mode des mots isolés, sous la plate-forme HTK, l'une des plus utilisée actuellement dans le domaine de la RAP.

CHAPITRE III :

Résultats et Interprétations

1. Introduction :

Ce chapitre est consacré aux résultats et à leurs interprétations. Une description brève de la base de données utilisée sera suivie par la description des principaux outils HTK utilisés pour manipuler les modèles de Markov cachés. Seront ensuite exposées en détail toutes les étapes de réalisation du système de reconnaissance par HTK à savoir : le codage, le choix de la topologie du HMM, l'apprentissage et la reconnaissance. La stratégie de l'étude sera ensuite abordée et les résultats seront exposés avec les interprétations. Le chapitre sera achevé par une conclusion.

2. Description de la base de données utilisée :

La base comporte 20 mots arabes isolés, chaque mot est prononcé 10 fois par un ensemble de 50 locuteurs arabes natifs. Théoriquement, le nombre total de prononciations devrait être 10 000, mais il existe 9 fichiers corrompus ce qui rend la taille de la base de données égale à 9 991 fichiers. Le corpus a été enregistré avec une fréquence d'échantillonnage de 44100 Hz et une résolution de 16 bits. Notons que notre base de données a été divisée en deux ensembles : corpus d'apprentissage (20%) de la base et corpus de test (80%) de la base. [12]

2.1 Vocabulaire

La liste des mots est la suivante :

1. صفر (zero)
2. واحد (one)
3. اثنان (two)
4. ثلاثة (three)
5. أربعة (four)
6. خمسة (five)
7. ستة (six)
8. سبعة (seven)
9. ثمانية (eight)
10. تسعة (nine)
11. التنشيط (activation)
12. التحويل (transfert)
13. الرصيد (balance)
14. التسديد (payement)
15. نعم (yes)
16. لا (no)
17. التمويل (funding)
18. البيانات (data)
19. الحساب (account)
20. انهاء (end)

2.2 Système d'étiquetage des fichiers de la base :

Les fichiers ont été étiquetés en utilisant le système de codage suivant :

S (numéro de locuteur). (Nombre de répétition). (Numéro du mot)

Exemple :

S01.01.01.wav représente le premier locuteur (parmi 50) prononce pour la première foi (parmi 10) le premier mot (parmi 20).

3. Description de l'outil de reconnaissance HTK :

Le Hidden Markov Model Toolkit (HTK) est une boite à outils portable permettant la création et la manipulation des modèles de Markov cachés. HTK est principalement utilisé pour la recherche sur la reconnaissance vocale bien qu'il soit tout à fait utilisable dans nombreuses autres applications telles que la synthèse vocale, la reconnaissance de l'écriture ou la reconnaissance de séquences d'ADN (acide désoxyribonucléique). [19] La plate-forme logicielle HTK était originalement mise au point à l'université de Cambridge, dédiée au développement de systèmes à base de HMMs.

HTK consiste en un ensemble de modules de bibliothèque et d'outils disponibles sous langage source C. Ces différents outils facilitent l'analyse vocale, l'apprentissage des HMMs, la réalisation des tests et l'analyse des résultats. Il est à noter, que ce qui a contribué au succès de HTK, est qu'il est accompagné d'une bonne documentation. HTK prend en charge les HMMs à observation continues à densité de distribution avec mélange de gaussiennes ainsi que les HMMs discrets et peut être utilisé pour créer des systèmes HMM complexes. [22] Le logiciel contient une documentation complète et des exemples.

4. Réalisation du système de reconnaissance par HTK :

4.1 Le codage :

Cette étape appelé extraction de paramètres, consiste à extraire les paramètres MFCC à partir du signal vocal, les détails théoriques sont exposés dans le chapitre(I). La figure(III.1) illustre la procédure de codage d'un seul fichier (une seule prononciation), il s'agit de multiplier chaque trame du signal parole par une fenêtre de Hamming puis de calculer un vecteur de paramètres (vecteur de coefficients MFCC dans notre cas), par la suite on obtient une suite de vecteurs de

paramètres pour chaque prononciation. Le codage peut être réalisé sur un groupe de fichier à la fois au lieu d'un seul fichier, ceci est illustré sur la figure (III.2) où on a utilisé un fichier script contenant la liste de tous les fichiers à coder (fichiers sources) et la liste des noms des fichiers après le codage (fichiers cibles).

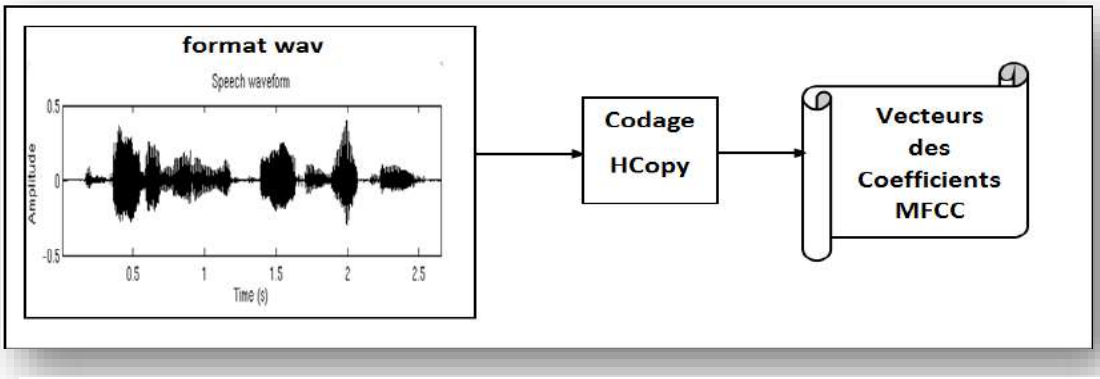


Figure III.1 : *La procédure du codage*

La fonction **HCopy** est un outil à usage général pour copier, coder et manipuler des fichiers de parole, cette fonction est invoquée avec la syntaxe suivante :

Syntax: HCopy –A –D –C config.txt –S script.txt

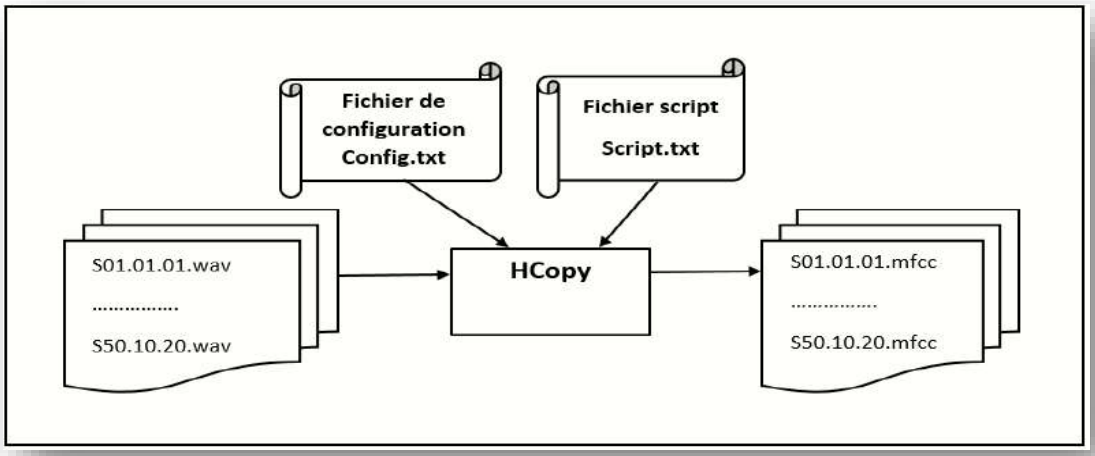


Figure III.2 : *Utilisation de HCopy pour un groupe de fichiers.*

Le fichier qui définit les paramètres de l'extraction des coefficients acoustiques et spécifie tous les paramètres de conversion, nommé config.txt est illustré dans la figure suivante :

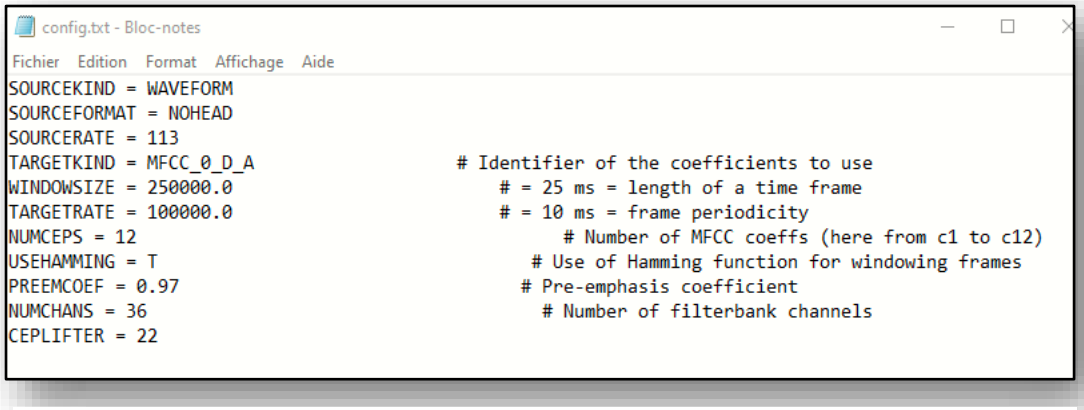


Figure III.3 : *Fichier de configuration*

Avec un tel fichier de configuration, l'analyse MFCC est effectuée (préfixe « MFCC » dans l'identifiant TARGETKIND). Pour chaque période de signal, les coefficients suivants sont extraits :

- Les 12 premiers coefficients MFCC [c1,….c12] (depuis NUMCEPS=12).
- Les coefficients MFCC « nul » c0, indique que le spectre 0 est ajouté (suffixe«-0»dansTARGETKIND).
- 13 « coefficients Delta », estimant la dérivée du premier ordre de [c0, c1 ...c12] (suffixe « -D » dans TARGETKIND).
- 13 coefficients d'accélération, estimant le second ordre dérivatif de [c0, c1,…c12] (suffixe « -A »dans TARGETKIND.
- Le symbole # est utilisé pour les commentaires.

Chaque vecteur contient 39 coefficients extraits de chaque période de signal

- **Script.txt** est le fichier texte spécifiant le nom et la localisation de chaque onde acoustique à transformer (fichier signal à traiter), avec le nom et la localisation des fichiers de coefficients des cibles.

La figure ci-dessous représente un exemple d'un **script** :

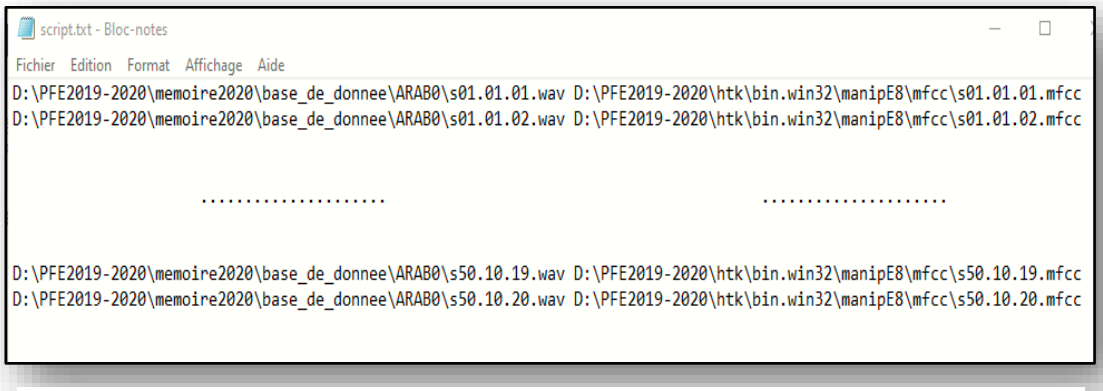


Figure III.4 : le fichier texte spécifiant le nom et localisation

Remarque : si le chemin des fichiers n'est pas spécifier dans le Script, tous les fichiers (WAV et MFCC) seront dans le répertoire en cours.

Les fichiers audio sont maintenant sous forme d'une suite de vecteurs de paramètres (fichier à une extension MFCC).

4.2 Choix de la topologie des modèles (prototypes) :

La topologie des modèles HMMs a une incidence importante sur les performances des systèmes de reconnaissance automatique de la parole à base de HMM. Nous adopterons dans notre expérimentation deux topologies différentes. Nous les appellerons : topologie1 et topologie 2. Les deux topologies sont représentées par un modèle gauche-droite (figure (III.5)) mais diffèrent dans les transitions entre états. Dans l'exemple de la figure (III.5) on a un HMM avec :

- Un nombre d'états égale à 8 dont 6 états émetteurs. Les probabilités de transition sont données par la matrice de probabilité de transition.
- Chaque état émetteur est caractérisé par une loi d'émission continue à une ou plusieurs gaussiennes caractérisées par un vecteur moyen μ (39composantes) et une matrice de covariance ∑diagonale.

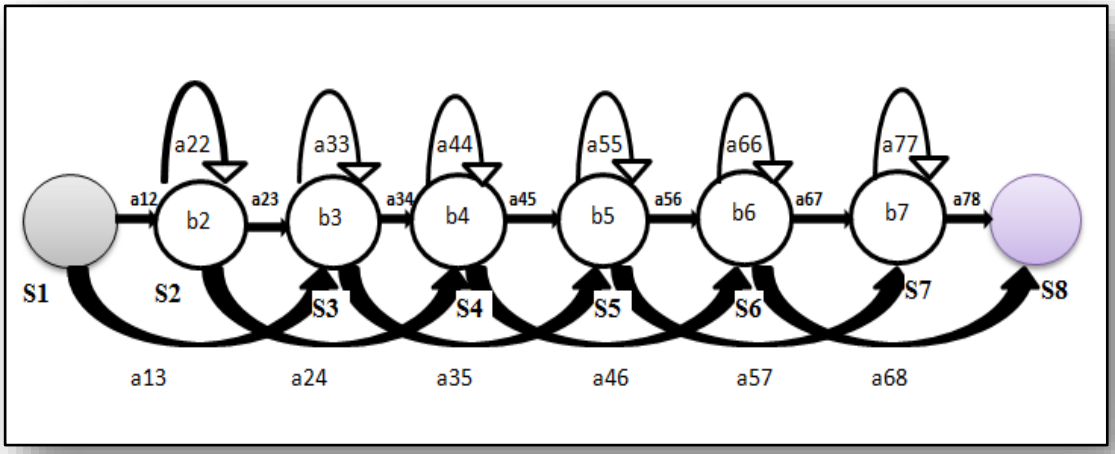


Figure III.5 : *Topologie choisie pour HMMs*

Remarque : généralement la topologie la plus communément adoptée est celle de **BAKIS**, il n'y a pas une loi fixe pour le choix de la topologie.

Chaque mot est représenté par un HMM. Dans notre cas on doit écrire 20 HMM (20mots).

La figure suivante représente un exemple du fichier de description du HMM, ce dernier est un fichier texte qui comporte tous les éléments du HMM.

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "zero"
<BeginHMM>
        <NumStates> 8
        <State> 2
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <State> 3
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <State> 4
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <State> 5
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <State> 6
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <State> 7
                <Mean> 39
                        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                <Variance> 39
                        1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        <TransP> 8
                0 0.5 0.5 0 0 0 0 0
                0 0.3 0.4 0.3 0 0 0 0
                0 0 0.3 0.4 0.3 0 0 0
                0 0 0 0.3 0.4 0.3 0 0
                0 0 0 0 0.3 0.4 0.3 0
                0 0 0 0 0 0.3 0.4 0.3
                0 0 0 0 0 0 0.5 0.5
                0 0 0 0 0 0 0 0
<EndHMM>
```

FigureIII.6 : *Fichier de description HMM*

Les éléments du fichier de description HMM sont les suivants :

- **~o <VecSize> 39 <MFCC_0_D_A> :** est l'entête du fichier, donnant la taille du vecteur coefficient (ici 39 coefficients), et le types de coefficient ici (MFCC-0-D-A).
- **~h "zero" <BeginHMM>…<EndHMM> :** nom du HMM (ici zéro), et commencer et finis la description d'un HMM.
- **<NumStates> 8** donne le nombre total d'états dans le HMM incluant les deux non-émettant (1et 8).
- **<State> 2** introduit la description de la fonction d'observation de l'état 2, ici nous avons choisi des fonctions d'observation gaussienne unique, avec des diagonale matrice diagonale, une telle fonction est entièrement décrite par un vecteur de moyenne et une variance (les éléments diagonaux de la matrice d'autocorrélation). Les états 1 et 6 ne sont pas décrits dans la mesure où ils n'ont pas de fonction d'observation.
- **<Mean> 39**

  0.0(…) 0.0 (×39) donne le vecteur de moyenne (dans un espace d'observation de dimension 39) de la fonction d'observation courante, chaque élément est arbitrairement initialisé à 0 : le fichier ne donne que les prototypes du HMM (sa topologie globale), les valeurs de ces coefficients seront obtenus plus tard après l'apprentissage.
- **<Variance> 39**

  1.0(…) (×39) donne le vecteur de variance de la fonction d'observation courante. Chaque élément est arbitrairement initialise à 1.
- **<TransP> 8** donne la transition 8×8 du HMM, ou a_{ij} est la probabilité de transition de l'état i a l'état j. les valeurs nuls indiquent que les transitions correspondante ne sont pas autorisées. Les autres valeurs sont arbitrairement initialise (la somme sur chaque ligne de la matrice doit être égale à 1), les valeurs seront modifier plus tard au cours du processus d'apprentissage.

4.3 Initialisation des modèles :

Avant de commencer le processus d'apprentissage les paramètres HMM doivent être initialisés correctement avec les données d'apprentissage. Afin d'obtenir une convergence rapide de l'algorithme d'apprentissage. HTK propose l'outil d'initialisation : **HCompv.** [5] Ce dernier donne une initialisation plate du modèle. Chaque état du HMM reçoit le même vecteur

de variance et de moyenne : ceux-ci sont calculés globalement sur le corpus entier d'entrainement. L'exemple d'initialisation des modèles est illustré dans la figure (III.7).

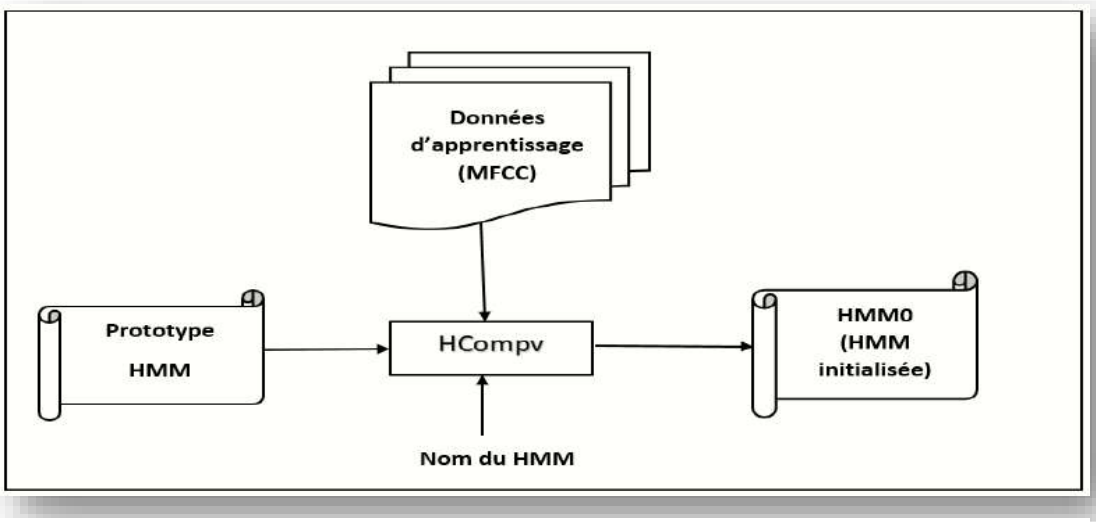


Figure III.7 : *Fonctionnement de HCompv*

HCompv doit être invoque comme suit :

Syntaxe : HCompv -S manip/aprent.txt -M manip/5hmm0 -H manip/prototypes5/zero.txt zero

Avec:

- **manip/aprent.txt :** représente la liste complète des fichiers MFCC formant le corpus d'apprentissage pour chacun des 20 mots de la base de données.
- **HMM0 :** représente le nom du dossier qui contient les descriptions des HMMs (résultats) après l'initialisation seront placées.
- **manip/prototypes5/zero.txt :** représente le chemin (manip) et le fichier de description contenant les prototypes HMMs (des 20mots).
- **Zero :** est le nom du HMM à initialise (zero,one…end).

Ce fichier de sortie HMM par **HCompv** à le même nom que le prototype d'entre.

Cette procédure doit être répète pour chacune des 20modele.

4.4 Apprentissage :

La ré-estimation des modèles est faite avec l'outil **HTK HRest** comme illustre dans la figure (III.8) utilisant l'algorithme de **Baum-welch**, qui calcule les valeurs optimales pour les

paramètres HMM (probabilités de transition, plus les vecteurs de moyenne et de variance pour chacune des fonctions d'observation)

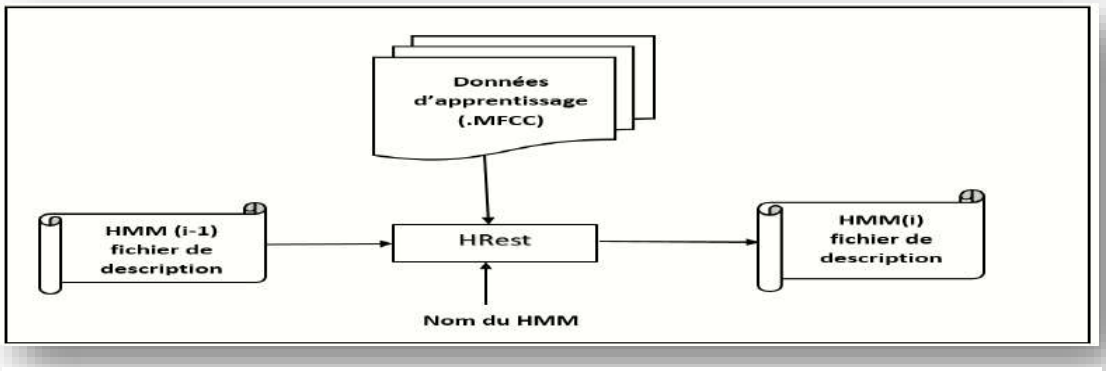


Figure III.8 : *Fonctionnement de HRest*

La ligne de commande suivante permet une itération de ré-estimation des modèles :

Syntaxe : HRest -H manip/script/aprent01.txt -M manip/HMM0 -H manip/prototypes8/zero.txt zero

Avec:

- **HMM0 :** représente le répertoire de sortie, indique l'indice de l'itération courante i.
- **manip/prototypes8/zero.txt :** est le fichier de description de HMM appelé names of HMMs il est stocker dans un répertoire dont le nom indique l'indice de la dernière itération (ici c'est manip/HMM0)
- **zero :** représente le nom du HMM à apprendre (entraîner).

Finalement on dispose maintenant les 20modeles entrainés qui sont prêtes à être utilisés pour la reconnaissance.

4.5 Définition de la grammaire et du dictionnaire :

Avant d'utiliser nos modèles de mots entrainés, on doit définir la grammaire du système. En HTK, la grammaire est décrite dans un fichier texte selon certaines règles syntaxiques. [12] Dans notre cas la grammaire est assez simple, elle décrit un ensemble de 20 mots isolés. Par isolés on entend que le système s'attend à chaque fois à ce qu'un seul mot soit prononcé par le locuteur. Le résultat de reconnaissance sera l'un des 20 mots possibles appartenant au vocabulaire.

Dans notre cas la grammaire est assez simple :

```
$word=zero|one|two|three|four|five|six|seven|eight|nine|activatio
n|transfert|balance|payement|yes|no|funding|data|account|end;
($word)
```

Figure III.9 : *Le fichier de la grammaire*

La variable **word** peut être remplace par l'un des 20 mots du corpus (zero ou one….. end)

Le système doit savoir bien sûr à quel HMM correspond chacune des variables de grammaire zero, one, … end. Cette information est stockée dans un fichier texte appelé dictionnaire. Dans une telle tâche simple, la correspondance est simple, et la tache dictionnaire renferme tout simplement les 20 entrées. Comme illustre dans la figure(III.10) :

```
zero            zero
one             one
two             two
three           three
four            four
five            five
six             six
seven           seven
eight           eight
nine            nine
activation      activation
transfert       transfert
balance         balance
payement        payement
yes             yes
no              no
funding         funding
data            data
account         account
end             end
```

Figure III.10 : *Le fichier du dictionnaire*

Les éléments gauches se référant aux noms des variables de grammaire. Les éléments droits se référant aux noms des HMMs (introduit par ~h dans les fichiers de description HMM).

Pour compléter la définition des taches on a besoin de créer un réseau de connexion reliant les 20 mots (comme illustré dans la figure suivante) dans lequel chaque mot, et chaque transition mot-à-mot est explicitement mentionné. Ce réseau de mots peut être créé automatiquement par la compilation du fichier grammaire avec l'outil HParse :

Syntaxe : HParse gram.txt reseau.slf

Tel que

- **gram.txt :** est le fichier grammaire.
- **Reseau.slf :** est le fichier obtenu qui représente le réseau.

Un fichier SLF contient une liste de nœuds représentant des mots et une liste d'arcs représentant les transitions entre les mots. Chaque définition de nœud et d'arc qui est écrit sur une seule ligne, se compose d'un certain nombre de domaines.

Maintenant, notre tâche de reconnaissance est complétement définie par son réseau, son dictionnaire est son ensemble de HMM (les 20 HMMs de la dernière itération), est prête à l'emploi).La figure ci-dessous représente le réseau de connexion entre les mots :

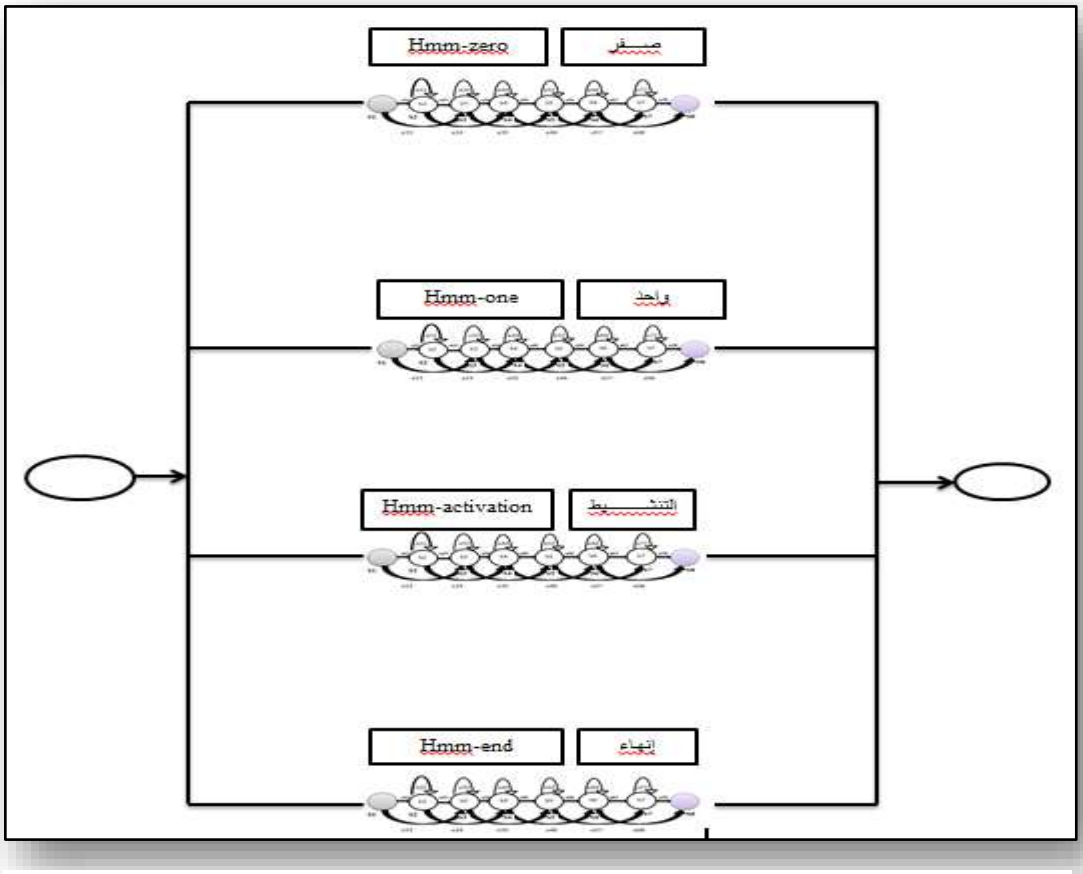


Figure III.11 : *Le réseau de connexion entre mots*

4.6 Reconnaissance :

Une fois les modèles HMMs entrainés obtenus, on peut procéder à l'étape de reconnaissance. La fonction utilisée est basé sur l'algorithme de Viterbi. Cette fonction est invoquée en HTk comme suit :

Syntaxe : HVite -Hmmdef.mmf -i reco.mlf -w reseau.slf dict.txt hmmlist.txt test.txt

Avec:

- **Hmmdef.mmf :** fichier texte contient la définition des HMMs.il est obtenu en copiant chaque définition des HMMs (de la dernière itération) dans un fichier unique. Chaque élément est séparé par une nouvelle ligne de l'autre élément, et une nouvelle ligne après le dernier élément.
- **Reco.mlf :** c'est le fichier de sortie (réseau de connexion obtenu).
- **Hmmlist :** fichier texte contenant les noms des 20 modèles à utiliser, les éléments sont séparés par une nouvelle ligne.
- **Test.txt :** le fichier script qui contient la liste des fichiers(MFCC) de teste.

La sortie est stockée dans un fichier **reco.mlf** qui contient le fichier de reconnaissance avec la vraisemblance de reconnaissance pour chaque mot. La figure (III.12) illustre l'utilisation de Hvite.

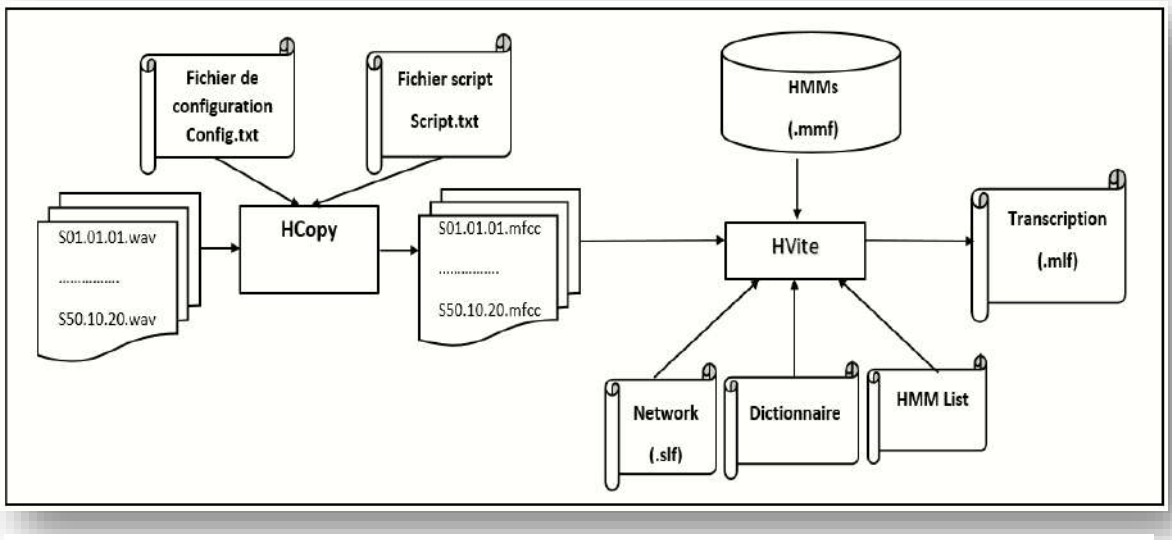


Figure III.12 : *processus du codage et reconnaissance*

La figure (III.13) illustre un extrait du fichier de résultat de transcription.

```
#!MLF!#
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.01.rec"
0 8600000 zero -3843.732666
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.02.rec"
0 9400000 one -4443.527832
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.03.rec"
0 9300000 two -4532.809082
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.04.rec"
0 8900000 three -4518.205078
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.05.rec"
0 9400000 four -4573.066895
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.06.rec"
0 9300000 five -4883.871582
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.07.rec"
0 9500000 six -4643.278809
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.08.rec"
0 9200000 seven -4411.139160
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.09.rec"
0 9700000 eight -4516.868652
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.10.rec"
0 9000000 nine -4585.495117
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.11.rec"
0 11700000 activation -5641.246582
.
"D:/PFE2019-2020/htk/bin.win32/manipE8/mfcc/s01.03.12.rec"
```

Figure III.13 : *Exemple d'un fichier reco*

5. Analyse des résultats des tests sur la RAP :

L'étude consiste à effectuer une série d'expérimentations en faisant varier un certain nombre de paramètres afin d'observer l'évolution du taux de reconnaissance. Les étapes de l'expérimentation sont les suivantes :

- étude de l'effet de la topologie et du nombre d'états par modèle HMM
- étude de l'effet de la méthode d'initialisation pour les deux topologies.
- étude de l'effet du nombre de gaussiennes par état pour les deux topologies.
- Conclusions.

6. Résultats et interprétations :

6.1 *L'effet de la topologie et du nombre d'états :*

On a utilisé 2 topologies dans nos expériences comme illustré dans cette figure :

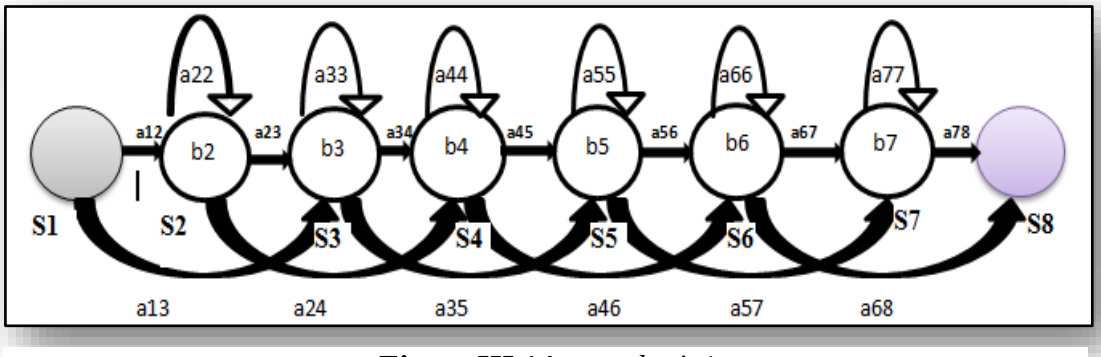


Figure III.14 : *topologie1*

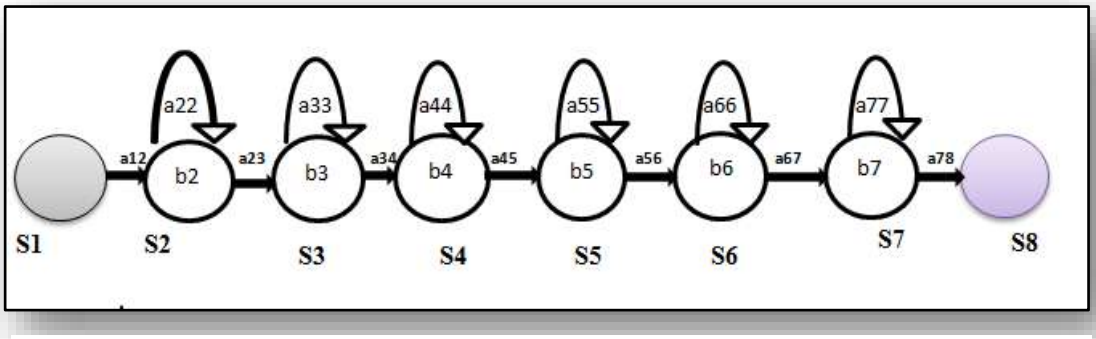


Figure III.15 : *topologie2*

Les résultats sont donnés dans le tableau suivant :

Nombre d'états	Taux de reconnaissance (%) pour la topologie 1	Taux de reconnaissance (%) pour la topologie 2
5 états	68.4764	68.8266
6 états	80.7856	79.8849
7 états	79.7723	87.1654
8 états	81.1859	91.2309
9 états	76.2197	92.7446

Tableau III.1 : *le taux de reconnaissance obtenu pour des nombres d'états différents et pour les deux topologies*

Interprétation :

- Pour les deux topologies, la tendance générale est l'amélioration du taux de reconnaissance en augmentant le nombre d'états. Cette amélioration tend à disparaitre lorsque le nombre d'états est important car dans ce cas le nombre de paramètres du HMM

devient important et on aura besoin de plus d'exemples d'apprentissage (un corpus plus volumineux).

- L'amélioration mentionnée plus haut est plus apparente pour la topologie 2 car elle introduit une certaine contrainte sur la durée minimale de chaque unité phonétique du mot prononcé.
- Les modèles ayant entre 7 et 9 états donnent de bons résultats étant donné qu'ils représentent bien le nombre d'unités phonétiques de chaque mot du vocabulaire.

6.2 Étude de l'effet de la méthode d'initialisation sur les performances de reconnaissance :

Les résultats sont donnés dans le tableau suivant :

Nombre d'états	Taux de reconnaissance (%) pour la topologie 1	Taux de reconnaissance (%) pour la topologie 2
5 états	68.4764	68.8266
6 états	80.7856	79.8849
7 états	79.7723	87.1654
8 états	81.1859	91.2434
9 états	76.2197	92.7946

Tableau III.2 : *Taux de reconnaissance de l'initialisation de chaque mot avec sa classe d'exemples*

Nombre d'états	Taux de reconnaissance (%) pour la topologie 1	Taux de reconnaissance (%) pour la topologie 2
5 états	68.4764	68.8266
6 états	80.7856	79.8849
7 états	79.7723	87.1654
8 états	81.1859	91.2309
9 états	76.2197	92.7446

Tableau III.3 : *taux de reconnaissance de l'initialisation globale*

Interprétation :

On remarque une légère amélioration en passant de l'initialisation globale vers l'initialisation de chaque mot avec les exemples qui lui correspondent, ce qui est logique car l'initialisation globale (appelée aussi initialisation plate) donnent des modèles HMMs initiaux identiques alors que la deuxième méthode d'initialisation donne des HMMs initiaux différents. Chaque HMM démarre d'un point initial proche de la forme qu'il représente permettant à l'apprentissage d'atteindre un meilleur optimum.

6.3 Etude de l'effet du nombre de gaussiennes par état :

Jusqu'à présent, les modèles utilisés comportaient une seule gaussienne par état. Nous allons maintenant faire varier le nombre de gaussiennes afin d'observer l'effet de ce paramètre sur les performances de reconnaissance.

Les résultats pour deux gaussiennes sont donnés dans le tableau suivant :

Nombre d'états	**2 gaussiennes**	
	Taux de reconnaissance pour la topologie 1	Taux de reconnaissance pour la topologie 2
5 états	81.8864	86.8276
6 états	84.8887	90.3803
7 états	85.6392	93.983
8 états	86.7150	95.1464
9 états	88.2787	95.7843

Tableau III.4 : *utilisation de 2 gaussiennes par état pour les deux topologies*

Les résultats pour trois gaussiennes sont donnés dans le tableau suivant :

Nombre d'états	**3 gaussiennes**	
	Taux de reconnaissance pour la topologie 1	Taux de reconnaissance pour la topologie 2
5 états	81.7738	87.4656
6 états	85.8019	91.7688
7 états	85.4141	97.0831
8 états	86.6775	95.0463
9 états	88.4163	95.4841

Tableau III.5 : *utilisation de 3 gaussiennes par état pour les deux topologies*

Interpretations:

- L'augmentation du nombre de gaussiennes par état améliore la performance de reconnaissance, ce qui est logique car plus on utilise de gaussiennes plus on modélise la distribution des vecteurs spectraux d'une manière plus précise.
- Il convient cependant de rappeler que l'augmentation du nombre de gaussien conduit à une rapide augmentation du nombre de paramètres des modèles. Rappelons que chaque gaussienne est représentée par un vecteur moyenne (39 valeurs) et une matrice de covariance (39 valeurs) ; ce qui fait qu'on sera rattrapé par le problème de sous-apprentissage (Under training) dont la résolution nécessite l'augmentation du nombre d'exemples d'apprentissage. Ce problème a été soulevé plus haut lors de l'étude de l'effet du nombre d'états.

7. Conclusion :

Ce chapitre expose les résultats de reconnaissance automatique de la parole (RAP) sur une base de données de mots arabes isolés. L'objectif était de mettre la lumière sur l'importance de l'effet du nombre d'états choisi, la topologie des modèles HMM, la méthode d'initialisation et le nombre de gaussiennes sur les performances de reconnaissance.

La conclusion globale pouvant être tirée se résume dans le fait que, bien qu'ayant un caractère empirique, la conception des modèles HMMs pour la reconnaissance doit tenir compte d'un certain nombre de paramètres pour obtenir de bons résultats. Le premier paramètre est le nombre d'unités phonétiques de chaque mot. Le deuxième paramètre est la topologie adoptée pour les modèles HMMs ; on a vu dans cette étude que la deuxième topologie était supérieure à la première en termes de performances de reconnaissance. Le troisième paramètre est le nombre de gaussiennes par état, ce paramètre permet de pousser les performances de reconnaissance à des niveaux plus que satisfaisants.

Conclusion générale

Dans ce travail nous avons abordé un domaine en cours d'expansion durant cette dernière décennie : c'est la reconnaissance automatique de la parole et particulièrement en arabe. Après une lecture bibliographique conséquente relative à ce domaine, nous avons fait le choix de travailler avec les chaines de Markov cachées (HMM) qui représentent un outil très robuste. Nous avons opté ensuite pour l'outil HTK afin de manipuler les HMMs. Plus particulièrement, nous avons étudié un système de reconnaissance d'un vocabulaire des mots arabes isolés. Ce système se compose d'une base d'apprentissage et d'une base de test.

Notre étude avait pour objectif d'étudier l'effet des paramètres de la modélisation Markovienne sur les performances d'un système de reconnaissance RAP. Les paramètres ayant fait l'objet de l'étude sont : le nombre d'états des modèles, la topologie choisie pour les modèles, les méthodes d'initialisation des modèles ainsi que le nombre de Gaussiennes par état.

Les résultats exposés dans le chapitre précédent montrent l'importance de ces paramètres vu que le taux de reconnaissance changeait significativement lorsqu'on faisait varier ces paramètres. Il a été particulièrement relevé que plus le système contient de gaussiennes et plus la reconnaissance est meilleure, ce qui montre l'importance capitale de bien modéliser la distribution des vecteurs spectraux. Il faut néanmoins souligner qu'une étude plus poussée est susceptible d'être menée sur des aspects qui n'ont pas été traités dans ce travail, à savoir :

- Le nombre optimum de gaussiennes à utiliser.
- Construire des HMMs discrets et comparer leurs performances avec les HMMs à distributions continues.

.

Références bibliographiques

[1] René Boite, "Traitement de la parole " collection électricité, presses polytechniques et universitaires Romandes, PPUR press polytechniques : 2000.

[2] Khaled Zaabi, "implémentation d'une méthode de reconnaissance de la parole sur le processeur de traitement numérique du signal, thèse, université du Québec, Montréal, le 10 juin : 2004.

[3] Bachir Salah dine, "réalisation d'une application en vue de la reconnaissance de la parole transcodée Speex et G729, thèse, université Mohammed Boudiaf, M'sila, juin : 2016.

[4] Anatomie de l'appareil auditif sleep area remedees, S Annie, Woldrop.

[5] Claude Gimenes.Fr/signal/théorie-du signal/-Viii-échantillonnage-des-signaux.

[6] Benammar Riadh, " Traitement de la parole arabe par les HMMs",25 septembre : 2012.

[7] H.Sakoe and S.Chiba, "A dynamic programming approach to continuous recognition", Budapest Hungary: 1971.

[8] Chandarli Bouchera et Labiod Asma, " Etude de l'effet de l'étage d'entré sur des performances d'un système de reconnaissance de la parole de vingt(20) mots isolés prononcé en arabe", mémoire de master, Université Dr Yahia Fares Médéa : 2019.

[9] Collectif Evasy – évaluation des systèmes de synthèse de parole, compagne d'évaluation, EVALDA/Technolangue : 2003.

[10] Reynolds, D Speaker, "identification and verification using Gaussian mixture speaker models speech communication»:1995.

[11] Reynolds, D.A, Quartieri, T.F ET DUNN, R.B, "speaker verification using adapted Gaussian mixture models, digital signal processing", page 41: 2000.

[12] Khelfi Amel, "fenêtrage multiple pour la reconnaissance robuste de la parole", thèse de master, Université Dr Yahia Fares Médéa : 2016-2017.

[13] Jean- Paul Haston Loria Inria, " la reconnaissance automatique de la parole", université de France Heneri Poincaré institue, Nancy : 2005.

[14] Sean .A.Fulop, "Speech Spectrum analysis", Springer Heidelberg Dordrecht London New York Springer-Verlag Berlin Heidelberg: 2011.

[15] Asbai Nassim, "Identification et Authentification de locuteurs par les techniques de fusion des paramètres et modèles dans un environnement réel" : 29-06-2015.

[16] BENDAHMANE Abderrahmane, "Cours de traitement de la parole", Université Mohammed Boudiaf d'Oran : 2013-2014.

[17] Steve Young Gunnar Evermann Mark Gales Thomas Hain Dan Kershaw Gareth Moore Julian Odell Dave Ollason Dan Povey Valtcho Valtchev Phil Woodland, "The HTK book", Microsoft corporation :1995.

[18] Ferrah Dalila, "Reconnaissance automatique de la parole arabe par les HMMs A L'aide de HTK toolkit ", Université SAAD Dahleb, Blida : 2013-2014.

[19] Meberbeche Kamel, Modèles de Markov cachés : "application à la reconnaissance automatique de la parole", Université Mouloud – Mammeri, Tizi Ouzou.

[20] Esma Fatıma Bilgin Tas¸demir, "A Large Vocabulary Online Handwriting Recognition System for Turkish", These, Sabancı University June: 2018.

[21] Lawrence-R-Rabine ET Ronald W.Schafer, "Introduction to digital speech processing", Taylor and Francis Inc: 2007.

[22] Nicolas Moreau, Basic tutorial HTK (v.3.1) : Un Système de Reconnaissance e Automatique de la Parole, Ceci est la traduction française par Antoine Chemardinde : 02-02-2002.