République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj – Bouira

Faculté des sciences et des sciences appliquées

Département de Génie Electrique

Mémoire de Master

Filière : Génie électrique

Option : Technologies des Télécommunications

Réalisé Par :
- GHERBI BAYA Amel
- YAMANI Yasmine

Thème :

Décodage Acoustique-Phonétique d'un signal de parole

Date de soutenance : 24/09/2017

Devant le jury composé de :

-ISSAOUNI Salim	MAA à l'université de Bouira	Président
-BENZIANE Mourad	MAA à l'université de Bouira	Rapporteur
- DIDICHE Mustapha	Algérie télécom Bouira	Co-rapporteur
- NOURINE Mourad	MCB à l'université de Bouira	Examinateur
- CHELBI Salim	MCB à l'université de Bouira	Examinateur

Remerciement

Tout d'abord, nous remercions le Dieu, notre créateur de nous avoir donné la force, la volonté et le courage afin d'accomplir ce travail modeste.

Nous adressons un grand remerciement à notre encadreur extérieur monsieur Didiche Mustapha qui a proposé le thème de ce mémoire, et qui nous a aidés tout au long de notre rédaction.

Nous remercions également monsieur Benziane Mourad pour le grand honneur qu'il nous a fait en nous encadrant lors de la poursuite de notre travail.

Nous remercions tous nos amis d'avoir toujours été là pour nous, particulièrement Khaled, Amine, Asma et Maachi pour leur présence et leurs encouragements.

Finalement, nous tenons à exprimer notre profonde gratitude à nos familles qui nous ont toujours soutenues et à tous ceux qui ont participé à la réalisation de ce mémoire. Ainsi que l'ensemble des enseignants qui ont contribué à notre formation.

Dédicace :

Je dédie ce modeste travail :

Au bon dieu qui m'a bénie de deux de ses plus belles créations, ces symboles de douceur, de générosité, de sacrifice mais surtout d'amour … Mes deux mamans qui ont tout fait et donné pour moi une de sang et l'autre de cœur.

Au père que la vie m'a offert, qui a fait de moi la personne que je suis aujourd'hui et qui m'a élevé et aimé comme l'une de ses filles.

A mes frères et sœurs :

Islam, Rayan, Yasmine, Melissa, Sheraz ; auxquels je souhaite la réussite dans leurs vies.

A mes grands-parents qui m'ont accordé leurs prières et leurs présences, que dieu me les gardent et leurs procurent une longue vie.

A mon cher binôme pour tout ce qu'elle a fait pour la réussite de ce projet.

A tous ceux qui m'ont aidé afin de réaliser ce travail.

A tous mes proches.

Amel

Dédicace :

Je dédie ce mémoire également :

A mes chers parents en témoignage d'estime, de respect et de vénération, que dieu vous garde pour moi.

A mon unique frère Malek et très chères sœurs :

Assia, Ryma, Lydia, ma belle sœur ; aux quels je souhaite un bel avenir.

A mes petites nièces :

Fareh et Anfel ; que dieu les bénissent.

A tous mes amis

A tous mes proches.

Yasmine

Table des matières

Chapitre III : Segmentation et paramétrisation des caractéristiques du signal de parole

Chapitre IV : Conception et mise en œuvre

Résumé

La parole comme un moyen de dialogue homme-machine efficace, a donné 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.

Dans ce cadre, le but de notre projet est d'implémenté un système de Décodage Acoustico-Phonétique (DAP) qui constitue une étape importante en reconnaissance de la parole continue. Cela a été fait en procédant à l'extraction des caractéristiques à savoir MFCC et LPC sous un environnement C++.

Le travail présenté dans ce mémoire propose de reprendre les toutes premières étapes de traitement des systèmes de reconnaissance de la parole, à savoir l'extraction des caractéristiques et la classification phonétique en faisant intervenir les réseaux de neurones qui rentre dans le cadre de l'étude de l'intelligence artificielle.

Mots clés : RAP, DAP, MFCC, LPC, MLP, FFT, NPC, LVQ.

Abstracts

Speech as an effective means of human-machine dialogue has given birth to several research projects in the field of Automatic Speech Recognition (RAP). A system of RAP is a system that has the ability to detect the speech and to analyze it for the purpose of transcribing this signal into a chain of words or phonemes representing what the person has spoken.

In this framework, the aim of our project is to implement an acoustic-phonetic decoding system (DAP) which is an important step in continuous speech recognition. This was done by extracting the features MFCC and LPC under a C ++ environment.

The work presented in this paper proposes to take up the very first stages of processing speech recognition systems, namely the extraction of the characteristics and the phonetic classification by involving neural networks which are considered as an important tool in machine learning.

Keywords: RAP, DAP, MFCC, LPC, MLP, FFT, NPC, LVQ.

Liste des figures

Liste des tableaux

Liste des abréviations

CELP: Code Exited Linear Prediction.

DAP: Décodage Acoustico-Phonétique.

DFT: Discret Fourier Transformer.

FFT: Faste Fourier Transformer.

HMM: Modèle de Markov Caché.

HRTF: Head-Related Transfer Functions.

I.A: Intelligence Artificielle.

IPA: Alphabet phonétique International.

LPC: Linear Predictif Coding.

LPCC: Linear Prediction Cepstral Coefficients.

LVQ: Linear Vector Quantization.

MFCC: Scal Mel frequency Cepstral Coefficient.

MLP: Multi Layer Perceptron.

NLP: Natural Language Processing.

NPC: Codage Neuro-Prédictif.

PLP: Perceptual Linear Prediction.

PPZ: Le Taux de Passage par Zéro.

RAP: Reconnaissance Automatique de Parole.

RELP: Residual Exited Linear Prediction.

TAL: Traitement Automatique des Langues.

TBF : Traitement par Bancs de Filtres.

Introduction générale

Le travail réalisé lors de ce mémoire s'inscrit dans le cadre général de la reconnaissance automatique de la parole. Les propriétés de la reconnaissance automatique de la parole (RAP) par les machines demeurent un défi pour les spécialistes. Ces propriétés offrent une grande variété d'applications comme l'aide aux handicapés, la messagerie vocale, les services vocaux dans les téléphones portables (numérotation automatique, reconnaissance de commandes vocales), la production de documents écrits par dictée, etc. Le plus important pour notre recherche est de faire retranscrire par la machine tous ce qui a été prononcé vocalement par l'être humain [2].

Aucun système, jusqu'à ce jour, n'a pu effectuer très efficacement la reconnaissance de la parole continue, en temps réel et dans un milieu bruité. Différents outils ont été développés pour permettre la reconnaissance de la parole continue. Notre intérêt porte sur la reconnaissance de la langue arabe, la reconnaissance globale est sans doute la plus adaptée dans le contexte actuel de notre problème .Les modèles mathématiques que nous avons choisi d'employer pour atteindre notre objectif sont l'extraction de caractéristiques à l'aide de méthodes temporelles comme le codage prédictif linéaire (LPC) ou de méthodes cepstrales comme le codage MFCC (Mel Frequency Cepstral Coding) [2].

Dans ce mémoire, nous nous intéressons à un décodage acoustico phonétique DAP qui consiste à transformé un signal vocale continu en une représentation symbolique discrète. Seulement cette reconnaissance peut être étudiée de deux façons l'une globale et l'autre analytique. Nous nous somme intéressé à l'étude analytique qui consiste à faire une extraction des caractéristiques dans le domaine fréquentielle (MFCC) et domaine temporelle (LPC) qui seront injectés dans un réseau de neurones de type MLP pour avoir des NPC.

Notre travail est organisé de la manière suivante :

Le premier chapitre concerne les méthodes de reconnaissance de la parole pour une machine et démontre toutes les contraintes et les problèmes posés en reconnaissance de la parole. Surtout dans un milieu fortement bruité.

Le deuxième chapitre consiste à décrire l'aspect physiologique des organes humains utilisés dans le cadre de la phonétique articulatoire et perceptive dont les propriétés physiques des sons en était traité.

Le troisième chapitre qui présente les différentes méthodes d'extractions des caractères des signaux et traite la segmentation et la paramétrisation via les coefficients MFCC et LPC.

La comparaison et discutions des résultats obtenus font l'objet du chapitre quatre qui met en application un logiciel d'analyse de parole sous un environnement c++. Cette interface est présentée par un oscillogramme, un spectrogramme et deux zones de texte qui affichent une transcription phonétique du signal de parole en caractère arabe et l'autre en caractère occidentaux, enfin un menu rempli de fonctionnalité. Les tests sont donnés pour le MLP pour l'apprentissage et la phase de généralisation qui est la reconnaissance.

En fin la conclusion a tirée de ce mémoire que les travaux doivent se poursuivre et les perspectives dans ce domaine restent encore très riches et très importants et doivent être prospectés.

Chapitre I :

La reconnaissance de la parole

I.1. Introduction

Le domaine de transfère de la parole à vue une révolution des cinq dernières décennies dans plusieurs domaines et un effort important doit être fournit dans cette étude vu la variété des langues dans le monde.

L'un des objectifs fondamental de l'amélioration de la communication Homme-Machine est le (T.A.L) une filière du traitement automatique de la langue naturelle et selon Shannon dans sa théorie de l'information**,** un message représenté comme une séquence de symboles discrets peut être quantifier en bits, l'information est d'une forme analogique continue « Speech Signal » ce qui est impossible de l'introduire directement dans la machine ; c'est pour cette raison qu'il faut faire des transformations (prétraitements) de numérisation afin que nous puissions l'exploiter sur machine.

I.2. Qu'est ce que la parole ?

L'information portée par le signal de parole peut être analysée de bien des façons. Nous en rappellerons simplement ici les aspects acoustiques.

La variation de la pression de l'air qui apparait physiquement causée et émise par le système articulatoire est une parole.

La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur).

De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre [1].

L'opération de numérisation, représenté dans la Figure I.1, demande successivement : un filtrage de garde, un échantillonnage, et une quantification.

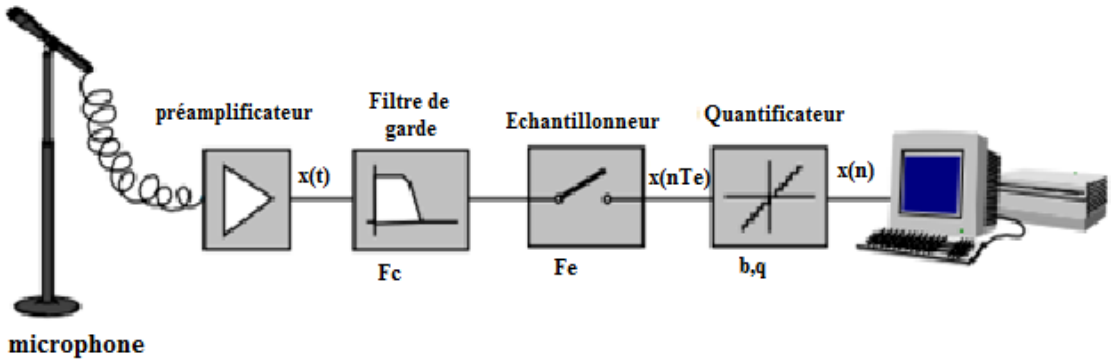


Figure I.1 : Enregistrement numérique d'un signal acoustique [1].

I.3. Les méthodes de reconnaissance de la parole

Le principe général d'un système de reconnaissance de la parole peut être décrit comme une suite de mots prononcés est convertie en un signal acoustique par l'appareil phonatoire [2]. Ensuite le signal acoustique est transformé en une séquence de vecteurs acoustiques ou d'observation (chaque vecteur est un ensemble de paramètre).Finalement le module de décodage consiste à associer à la séquence d'observation une séquence de mots reconnus [2].

Un système RAP transcrit la séquence d'observation en une séquence de mots en se basant sur le module d'analyse acoustique et celui du décodage .Le problème de la RAP est généralement abordé selon deux approches que l'on peut opposer du point de vue de la démarche : l'approche globale et l'approche analytique figure(I.2 [2].

La première considère un mot où une phrase en tant que forme globale à identifier en le comparant avec des références enregistrées.

La deuxième, utilisée pour la parole continue, cherche à analyser une phrase ou une chaine d'unités élémentaire en procédant à un décodage acoustico-phonétique exploité par des modules de niveau linguistiques [2].

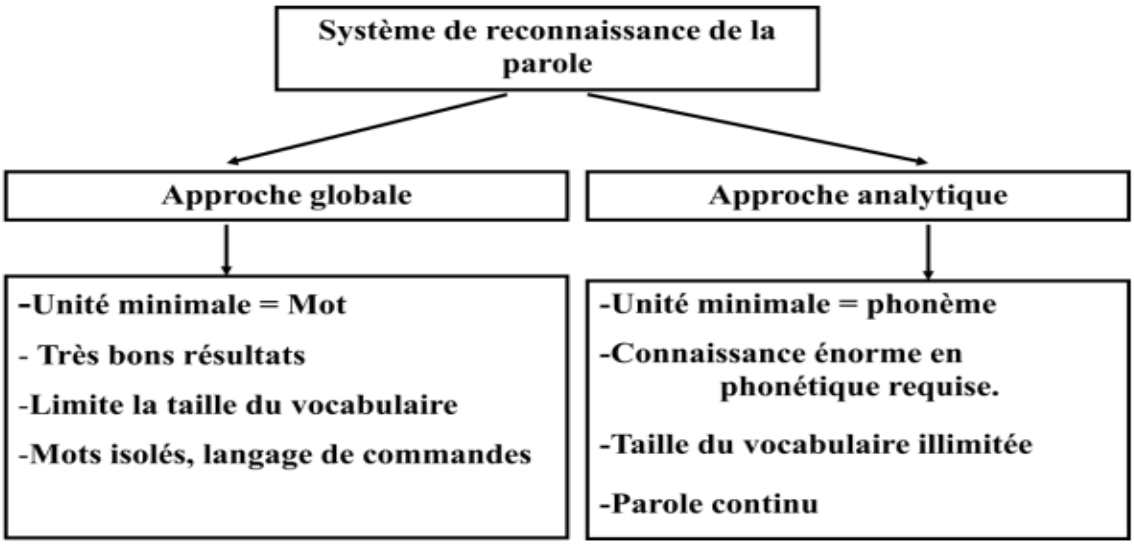


Figure I.2: les approches des systèmes de RAP [2].

I.3.1. La méthode globale

Cette méthode considère le plus souvent le mot comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans ce type de méthode, on compare globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique [3]. Ce type de méthode est utilisé dans les systèmes de reconnaissance de mots isolés, reconnaissance de parole dictée avec pauses entre les mots… et présente l'inconvénient de limiter la taille du dictionnaire.

Cette méthode a pour avantage d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes.

Généralement, on rencontre deux problèmes : le premier est relatif à la durée d'un mot qui est variable d'une prononciation à l'autre, et le deuxième aux déformations qui ne sont pas linéaires en fonction du temps. L'alignement temporel dynamique est la résolution de ces problèmes en appliquant un algorithme classique de la programmation dynamique.

I.3.2. La méthode analytique

C'est l'intervention du modèle phonétique du langage ou Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabe, demi syllabe, diphonie, phonème, phone homogène, etc.).

Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance dans cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification de ces unités en utilisant des méthodes de

reconnaissance des formes (classification statistique, réseau de neurones,...) ou des méthodes d'intelligence artificielle (systèmes experts par exemple) [3]. Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation et d'étiquetage phonétique. C'est pourquoi le DAP (Décodage Acoustico-Phonétique) [4], [5] est fondamental dans une telle approche.

Le processus de la reconnaissance de la parole dans une telle méthode peut être décomposé en deux opérations :

La segmentation : c'est la représentation du message (signal vocal) sous la forme d'une suite de segments de parole.

L'identification : c'est l'interprétation de segments trouvés en termes d'unités phonétiques.

Un système de reconnaissance de la parole peut être utilisé sous plusieurs modes :

- ✓ Dépendant du locuteur (mono locuteur) ;
- ✓ Multi locuteur ;
- ✓ Indépendant du locuteur.

I.4. Les problèmes de la reconnaissance de la parole

Il est bon d'en comprendre les différents niveaux de complexités et les différents facteurs qui en font un problème difficile, pour la reconnaissance automatique de la parole.

La parole est un phénomène à priori très simple à comprendre. Mais l'homme peut rencontrer des difficultés lorsqu'il essaie de suivre une conversation dans une langue qui lui est inconnue. De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance ainsi que par le milieu ambiant.

La grande variabilité de ces caractéristiques est l'obstacle principal dans la reconnaissance de parole. Nous allons maintenant voir les problèmes directement liés à la parole.

I.4.1. Variabilité interlocuteur

La variabilité interlocuteur est un phénomène majeur en reconnaissance de la parole. Comme

nous venons de le rappeler, un locuteur reste identifiable par le timbre de sa voix malgré une variabilité qui peut parfois être importante. La contrepartie de cette possibilité d'identification

à la voix d'un individu est l'obligation de donner aux différents sons de la parole une définition assez souple pour établir une classification phonétique commune à plusieurs personnes [6].

La cause principale des différences interlocuteurs est de nature physiologique. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence de base, le fondamental. Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents qui sont regroupés selon les classes que nous avons énoncées précédemment. Or le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le genre et l'âge. Ainsi, le conduit vocal féminin adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte. Le conduit vocal d'un enfant en bas âge est bien sûr inférieur en longueur à celui d'un adulte. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes [6].

La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux. Ces différences s'observeront d'autant plus facilement qu'une communauté de langue occupera un espace géographique très vaste, sans même tenir compte de l'éventuel rayonnement international de cette communauté et donc de la probabilité qu'a la langue d'être utilisée comme seconde ou, pire, troisième langue par un individu de langue maternelle étrangère. Là aussi, la définition phonétique tout autant qu'une définition stricte d'un vocabulaire ou d'une grammaire peuvent être mises à mal [6].

La variabilité interlocuteur telle qu'elle vient d'être présentée permet de comprendre aisément pourquoi les méthodes de reconnaissance des formes fondées sur la quantification de concordances entre une forme à analyser et un ensemble de définitions strictes plus ou moins formelles ne peuvent être appliquées, avec un succès limité, qu'à des applications où le nombre de définitions est restreint, limitant ainsi le nombre des possibles. D'une manière générale, la définition assez floue des différents phonèmes ou des différents mots d'une langue est la cause de nombreuses erreurs de classification dans les systèmes de décodage acoustico-phonétique(DAP). Mais la variabilité interlocuteur, malgré son importance

évidente, n'est pas encore la variabilité la plus importante car les différences au sein des classes phonétiques sont en nombre restreint. L'environnement du locuteur est porteur d'une variabilité beaucoup plus importante [6].

I.4.2. Variabilité intra locuteur

La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur.

Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie [6].

Il existe un autre type de variabilité intra-locuteur lié à la phase de production de parole ou de préparation à la production de parole. Cette variation est due aux phénomènes de coarticulation
[7]. Il est possible de voir la phase de production de la parole comme un compromis entre une minimisation de l'énergie consommée pour produire des sons et une maximisation des scores d'atteinte des cibles que sont les phonèmes tels qu'ils sont théoriquement définis par la phonétique.

Un locuteur adoptera donc un compromis qui est généralement partagé par une majorité de la communauté de langage à laquelle il appartient bien que ce compromis lui soit propre du fait de sa physionomie particulière. Ce compromis peut d'ailleurs être retrouvé à un plus haut niveau avec la notion d'idiolecte. Ce locuteur essaiera, lors d'une phase de production de parole, d'atteindre les buts qui lui sont fixés par les différents éléments de sa phrase tout en conservant un rythme naturel de production de la parole. Les cibles peuvent alors être modifiées du fait d'un certain contexte phonétique. Ce contexte peut être antérieur, lorsque le phonème provoquant une modification se trouve avant le phonème considéré, ou postérieur lorsque le phonème perturbateur se trouve après. La coarticulation peut enfin se produire à l'échelle d'un ou de plusieurs phonèmes adjacents, ce dernier cas étant cependant très rare [6].

La variabilité intra-locuteur est cependant beaucoup plus limitée que la variabilité inter-locuteur que nous allons étudier maintenant. Il est en effet possible, malgré les problèmes énoncés ci-avant, de mettre en œuvre des systèmes automatiques d'identification du locuteur, à la manière d'une personne reconnaissante une voix familière. Cette capacité est la preuve

qu'une certaine constance existe dans la phase de production de la parole par un même individu [6].

I.4.3. Variabilité due à l'environnement

La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution. Cette variation, considérée comme du bruit, sera étudiée ultérieurement. La variabilité liée à l'environnement peut, parfois, être considérée comme une variabilité intra-locuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières [6].

La variabilité environnementale due au locuteur peut tout d'abord être de nature physiologique. Ainsi, un système mécanique provoquant une déformation du conduit vocal provoquera immanquablement une variation dans le signal de parole produit. Ces contraintes physiques sont généralement rencontrées dans les systèmes de transport où une posture particulière, ou une accélération lors du déplacement, pourront provoquer une déformation [6].

Les moyens de transport peuvent également entraîner d'autres déformations du signal, d'origine psychologique. Le bruit ambiant peut ainsi provoquer une déformation du signal de parole en obligeant le locuteur à accentuer son effort vocal. Enfin, le stress et l'angoisse que certaines

personnes finissent par éprouver lors de longs voyages peuvent également être mis au rang des contraintes environnementales susceptibles de modifier le mode d'élocution [6].

I.4.4. Les différents types de bruits

Les différents bruits pouvant influés sur un message peuvent être divisés en deux grandes catégories : les bruits additifs et les bruits convolutionnels.

La distinction entre les deux peut être faite par le nombre d'agents agresseurs extérieurs à la transmission du message. Les bruits convolutionnels sont causés par la moindre qualité de la voie de communication, celle-ci ayant alors un rôle ambigu, du point de vue du message, de médium et d'agresseur, l'effet est plus dévastateur si le bruit n'est pas stationnaire et Les bruits additifs sont causés par des agents extérieurs et au trinôme source-voie-destinataire [2].

I.4.4.1. Les bruits additifs

Les bruits additifs sont dus à la multiplicité des systèmes de communication dans un même environnement. Plusieurs émetteurs et plusieurs receveurs pouvant être confinés dans un même espace, les messages de tous les émetteurs peuvent donc se trouver en concurrence sur une même voie sans que les récepteurs possèdent un mécanisme infaillible pour isoler le message qui leur est destiné. L'émetteur et le récepteur peuvent aussi se trouver en présence d'un ou de plusieurs équipements générant un bruit de fond de force variable [8].

I.4.4.2. Les bruits convolutionnels

Les bruits convolutionnels (ou multiplicatifs) sont dus à la distorsion induite par la voie de communication. Ils résultent de la mauvaise qualité d'un ou de plusieurs éléments de support du message ou, tout simplement, de son étroitesse en ban dépassant.

Les sociétés modernes utilisent de plus en plus de moyens de communication à longue distance tels que le téléphone, les moyens radiophoniques et, récemment, radiotéléphoniques. Ces moyens de communication à longue distance ont été élaborés à partir d'un compromis coût/efficacité. La parole, lorsqu'elle est transmise par un tel moyen, est forcément dégradée tout en gardant une grande intelligibilité [8].

Un des champs possibles d'application de la RAP sont les serveurs vocaux accessibles par les lignes téléphoniques. Mais la parole transmise par téléphone souffre de déformations variables induites par la qualité de la connexion

Une transmission peut ainsi souffrir de l'étroitesse de la bande passante, de la mauvaise qualité des microphones de certains terminaux téléphoniques, de bruits additifs stationnaires et de porteuses basses fréquences. La qualité de la transmission varie cependant très peu au cours d'une même communication. De manière plus générale, le bruit convolutionnel est présent dans toute application de RAP par l'intermédiaire du microphone utilisé pour la saisie de la voix. Un système de RAP mis au point avec certain microphone pourra voir ses capacités diminuer de manière conséquente lorsqu'un autre microphone sera employé [2].

La parole enregistrée dans tous les corpus utilisés pour la recherche est en effet toujours bruitée puisque le microphone utilisé effectue toujours un filtrage linéaire. Enfin, certains milieux d'enregistrement sont de mauvaise qualité et peuvent provoquer des phénomènes de réverbération. C'est notamment le cas des pièces possédant de grandes surfaces faites d'un matériau dur ou lorsque le microphone utilisé pour l'enregistrement est placé assez loin du

locuteur. Pour résoudre ce problème, on utilise généralement un ensemble de microphones pour trouver le filtre inverse [2].

I.4.4.3. Les bruits physiologiques

D'autres bruits n'ont pas la généralité des bruits de type additif ou convolutionnel car ils sont spécifiques à l'être humain lors de sa phase de production de parole. Mails ils peuvent également être considérés dans le domaine de la RAP.

Le mal fonctionnement de la plupart des systèmes de RAP en milieu bruité a cause des contraintes posées par de tels environnements qui n'ont pas été prises en compte dès le départ.

La modification de la méthode de production de parole était une manière d'adaptation de l'homme aux conditions sonores rencontrées.

L'effet Lombard est l'un des phénomènes les plus remarquables de modification de production de la parole par l'homme. il modifie sa voix, et son effort vocal, en "haussant le ton" de manière à ce que la parole produite conserve un bon RSB par rapport à l'environnement grâce a l'emplacement de locuteur dans un environnement bruité [2].

Cette accentuation de la voix pose cependant un problème majeur aux systèmes de RAP car les spectres de tous les phonèmes peuvent être modifiés ce qui a pour effet de nettement amoindrir les taux de reconnaissance [8]. Certaines études montrent que l'homme arrive à avoir de meilleures capacités de compréhension dans le cas de la parole Lombard que pour la parole normale lorsqu'il lui est demandé de reconnaître des mots isolés ou de la parole continue masqués par du bruit.

I.4.5. La coarticulation

Il faut se souvenir que les articulations se succèdent très rapidement :

Une première articulation peut ne pas être achevée au commencement de la seconde et la représentation de la réalisation de chaque phonème peut varier considérablement en fonction de son voisinage [2].

Dans la production des phonèmes, il est plus facile et normal d'essayer de ne pas produire les sons de façon isolée, mais plutôt d'anticiper la prochaine articulation.

Cette anticipation, qui change légèrement la qualité des sons, sera permise à condition que la compréhension du message ne soit pas compromise [2].

Phénomènes de Coarticulation:

- **Assimilation** : phénomène par lequel un son tend, du fait de sa proximité par rapport à un autre, à devenir identique, ou à prendre certaines de ses caractéristiques (voisement ou dévoisement par exemple).
- **Dilation :** modification des caractéristiques d'un son due à l'anticipation d'un autre son qui ne lui est pas contigu.
- **Différenciation :** changement phonétique qui a pour but d'accentuer ou de créer une différence entre deux sons contigus.
- **Dissimilation :** changement phonétique qui a pour but d'accentuer ou de créer une différence entre deux sons voisins mais non contigus.
- **Interversion :** lorsque deux sons contigus changent de place dans la chaîne parlée. **Métathèse :** lorsque deux sons non contigus changent de place dans la chaîne parlée [2].

I.5. Conclusion

Dans ce chapitre nous avons parlé en général sur la reconnaissance de la parole et définit les différentes méthodes de la reconnaissance. Nous avons également présenté les différents problèmes rencontrés dans la reconnaissance automatique de la parole. Le signal acoustique de la parole présente une grande variabilité qui complique la tâche des systèmes RAP. Cette complexité provient de la combinaison de plusieurs facteurs, comme la redondance du signal acoustique, la grande variabilité intra et inter-locuteurs, les effets de la coarticulation en parole continue, ainsi que les conditions d'enregistrement. Pour surmonter ces problèmes, différentes approches sont envisagées pour la reconnaissance de la parole telles que les méthodes analytiques, globales et les méthodes statistiques

Chapitre II

La production de la parole

II. 1. Introduction

Le processus de production de la parole est un système dans lequel une ou plusieurs sources excitent un ensemble de cavités. La source sera soit générée au niveau des cordes vocales soit au niveau d'une constriction du conduit vocal [9].

Dans le premier cas, la source résulte d'une vibration quasi-périodique des cordes vocales et produit ainsi une onde de débit quasi-périodique.

Dans le second cas, la source sonore est soit un bruit de friction soit un bruit d'explosion qui peut apparaître s'il y a un fort rétrécissement dans le conduit vocal où si un brusque relâchement d'une occlusion du conduit vocal s'est produit [9].

L'ensemble de cavités situées après la glotte, dites les cavités supra glottiques, vont ainsi être excitées par la ou les sources et "filtrer" le son produit au niveau de ces sources. Ainsi, en changeant la forme de ces cavités, l'homme peut produire des sons différents. Les acteurs de cette mobilité du conduit vocal sont communément appelés les articulateurs [9].

II. 2. Le processus de production

On peut donc résumer le processus de production de la parole en trois étapes essentielles :

- La génération d'un flux d'air qui va être utilisé pour faire naître une source sonore (au niveau des cordes vocales ou au niveau d'une constriction du conduit vocal) C'est le rôle de la soufflerie.
- La génération d'une source sonore sous la forme d'une onde quasi-périodique résultant de la vibration des cordes vocales ou/et sous la forme d'un bruit résultant d'une constriction ou d'un brusque relâchement ou occlusion du conduit vocal : C'est le rôle de la source vocale.
- La mise en place des cavités supraglottiques (conduits nasal et vocal) pour obtenir le son désiré (c'est principalement le rôle des différents articulateurs du conduit vocale) [9].

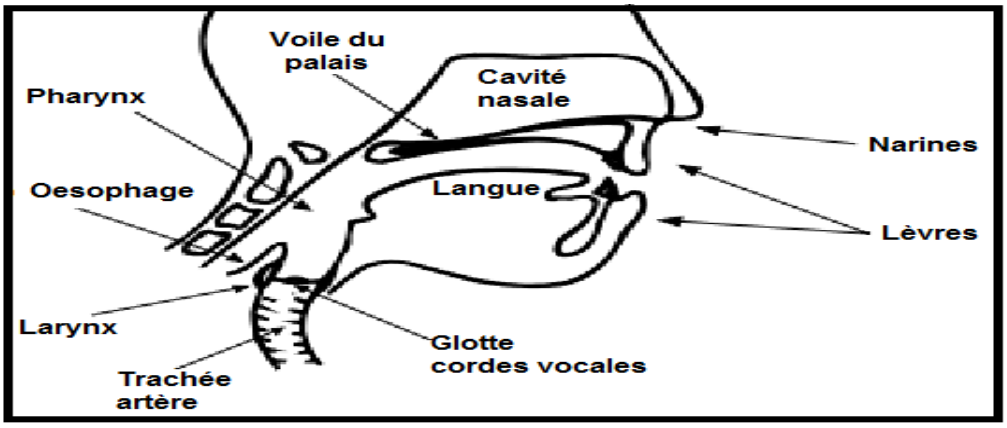


Figure II.1 : Coupe de l'appareil phonatoire humain [9].

II.3. Les différentes étapes de production de la parole

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure du tranché (Figure II.1). Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi

que pendant la phonation des sons non voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales [9].

Le larynx est d'abord complètement fermé, ce qui accroit la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer ; des impulsions périodiques de pression sont ainsi appliquée, au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation [9].

II. 4. Les organes de production de la parole

La parole est essentiellement produite par deux types de sources vocales. La première, plus sonore, est celle qui prend naissance au niveau du larynx suite à la vibration des cordes vocales. La seconde, moins sonore, prend naissance au niveau d'une constriction du conduit vocal ou lors d'un relâchement brusque d'une occlusion du conduit vocal. On parlera dans ce cas de sources de bruit [9].

II.4.1. Larynx

Le larynx est un organe situé dans le cou qui joue un rôle crucial dans la respiration et dans la production de parole. Le larynx (Figure II.2) est plus spécifiquement situé au niveau de la séparation entre la trachée artère et le tube digestif, juste sous la racine de la langue. Sa position varie avec le sexe et l'âge : il s'abaisse progressivement jusqu'à la puberté et il est sensiblement plus élevé chez la femme [9].

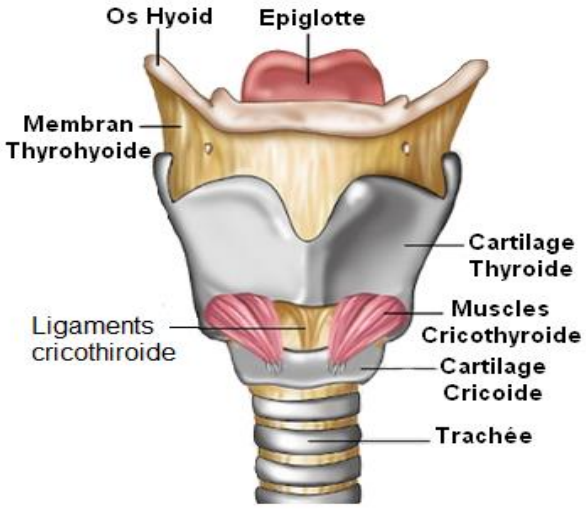


Figure II.2: schéma du larynx [35].

Il est constitué d'un ensemble de cartilages entourés de tissus mous. La partie la plus proéminente du larynx est formée de la thyroïde. La partie antérieure de cartilage est communément appelée la "pomme d'Adam". On trouve, juste au-dessus du larynx, un os en forme de 'U' appelé l'os hyoïde. Cet os relie le larynx à la mandibule par l'intermédiaire de muscles et de tendons qui joueront un rôle important pour élever le larynx pour la déglutition ou la production de parole [9].

La partie inférieure du larynx est constituée d'un ensemble de pièces circulaires, le cricoïde, sous lequel on trouve les anneaux de la trachée artère.

Le larynx assure ainsi trois fonctions essentielles :

- ✓ Le contrôle du flux d'air lors de la respiration ;
- ✓ La protection des voies respiratoires;
- ✓ La production d'une source sonore pour la parole [9].

Les muscles du larynx:

Les mouvements du larynx sont contrôlés par deux groupes de muscles. On distingue ainsi les muscles intrinsèques, qui contrôlent le mouvement des cordes vocales et des muscles à l'intérieur du larynx, et les muscles extrinsèques, qui contrôlent la position du larynx dans le cou.

La (Figure II.3) nous représente les muscles intrinsèques. Les cordes vocales sont ouvertes par une paires de muscles (les muscles cricoaryténoïdiens postérieur) qui sont situés entre la partie arrière du cricoïde et le cricoaryténoïdien [9].

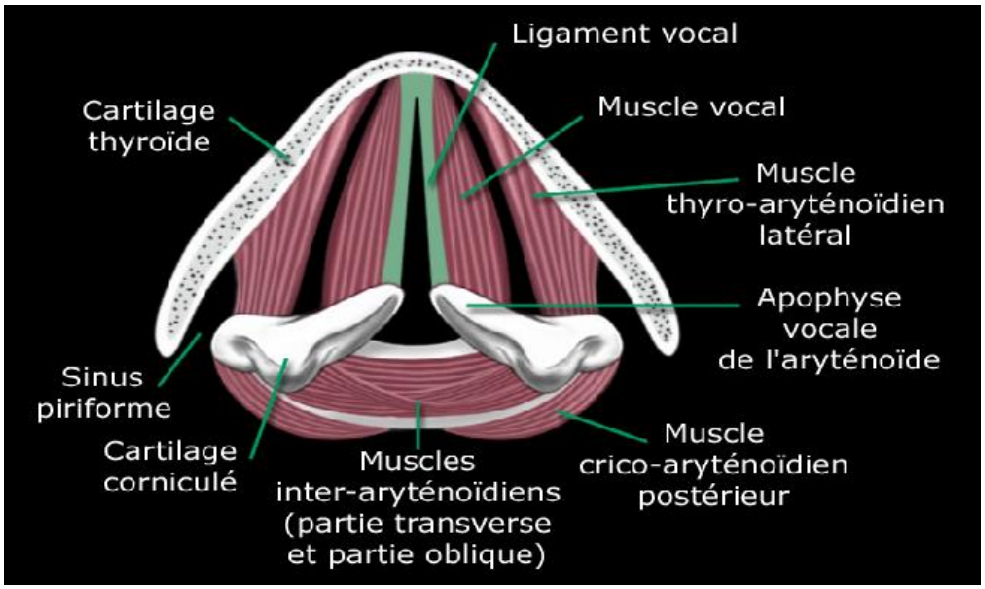


Figure II.3: Schéma des muscles intrinsèques du larynx [36].

Les cordes vocales :

Les cordes vocales situées au centre du larynx ont un rôle fondamental dans la production de la parole .Elles sont constituées de muscles recouverts d'un tissu assez fin couramment appelé la muqueuse. Sur la partie arrière de chaque corde vocale, on trouve une petite structure faite de cartilages : Les aryténoïdes. De nombreux muscles y sont rattachés qui permettent de les écarter pour assurer la respiration [9].

Durant la production de parole, les aryténoïdes sont rapprochés (voir Figure II.3). Sous la pression de l'air provenant des poumons, les cordes vocales s'ouvrent puis se referment rapidement. Ainsi, lorsqu'une pression soutenue de l'air d'expiration est maintenue, les cordes vocales vibrent et produisent un son qui sera par la suite modifié dans le conduit vocal pour donner lieu à un son voisé. Ce processus de vibration des cordes vocales est décrit un peu plus en détail ci-après [9].

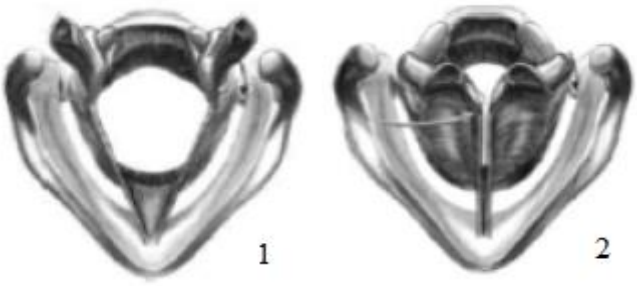


Figure II.4 : Les cordes vocales en position ouverte durant la respiration (1) et fermé pour la production de parole (2) [9].

Plusieurs muscles aident pour fermer et tendre les cordes vocales. Les cordes vocales sont-elles même constituées d'un muscle, le thyroaryténoide. Un autre muscle, l'interaryténoide, permet de rapprocher ces deux cartilages. Le muscle cricoaryténoïdiens latéral qui est lui aussi situé entre l'aryténoïde et le cartilage cricoïde sert à la fermeture du larynx [9].

Le muscle cricothyroide va du cartilage cricoïde jusqu'au cartilage thyroïde. Lorsqu'il se contracte, le cartilage cricoïde bascule en avant et tend les cordes vocales ce qui résultera à un élèvement de la voix, les muscles extrinsèques n'affectent pas le mouvement des cordes vocales mais élèvent ou abaissent le larynx dans sa globalité.

La (Figure II.**5)** donne une vue schématique d'une coupe verticale du larynx. Sur ce schéma, les cordes vocales sont ici clairement séparées, comme elles seraient durant la respiration. On peut également remarquer au-dessus des cordes vocales, des tissus ayant pour principal rôle d'éviter le passage de substances dans la trachée durant la déglutition : ce sont les fausses cordes vocales. Il est important de noter qu'elles ne jouent aucun rôle lors de la phonation. Le cartilage mou en forme grossière de langue qui se trouve au-dessus est appelé l'épiglotte et a également un rôle pour protéger l'accès de la trachée lors de la déglutition [9].

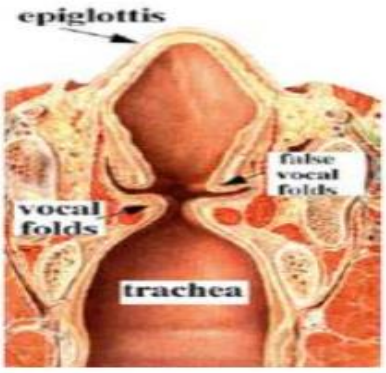


Figure II.5 : Vue longitudinale du larynx [9].

II.4.2. Les cavités supraglottiques

D'autres organes situés au-dessus des glottes (organes supraglottiques) interviennent également, même à un degré moindre, dans la production du son. On distingue, ainsi :

- **Le conduit vocal**

C'est un tube acoustique de section variable qui s'étend de la glotte jusqu'aux lèvres. Pour un adulte, le conduit vocal mesure environ 17 cm. Sa forme varie en fonction du mouvement des articulateurs qui sont les lèvres, la mâchoire, la langue et le velum. Ces articulateurs sont brièvement décrits ci-dessous [9].

- **Le conduit nasal**

Le conduit nasal est un passage auxiliaire pour la transmission du son. Il commence au niveau du velum et se termine aux niveaux des fosses nasales. Pour un homme adulte, cette cavité mesure environ 12 cm [9].

Le couplage acoustique entre les deux cavités est contrôlé par l'ouverture au niveau du velum (Figure II.1). On aura la production d'un son nasal, lorsque le velum -ou voile du palais- est largement ouvert. Et pour l production non – nasal lorsque le velum ferme le conduit nasal [9].

D'autre organes, dits articulateurs, joue également un rôle chacun en ce qui le concerne. Les articulateurs sont :

- **La langue**

La langue est une structure frontière, appartenant à la fois à la cavité buccale pour sa partie dite mobile et au glosso-pharynx pour sa partie dite fixe, qui appliquée contre le palais ou les dents constituent un organe vibratoire accessoire, intervenant dans la formation des consonnes. Elle a donc de l'importance pour la phonation.

On comprend que la langue est un articulateur fondamental puisque sa position est déterminante dans le conduit vocal [9].

- **La mâchoire**

La mâchoire possède un nombre de degrés de liberté plus faible et étant un corps rigide ne peut pas se déformer comme la langue. Néanmoins, la mâchoire peut non seulement s'ouvrir et se fermer, mais peut également s'avancer ou effectuer des mouvements de rotation.

Son rôle dans la parole n'est cependant pas primordial dans la mesure où il est

possible en bloquant la mâchoire de parler de façon très intelligible [9].

- **Les lèvres**

Les lèvres sont situées à l'extrémité du conduit vocal et comme pour la langue, elles possèdent une grande mobilité en raison des nombreux muscles impliqués dans leur contrôle. Les commissures sont les points de jonction des lèvres supérieure et inférieure ces dernières jouent un grand rôle dans la diplomatie (pour le sourire, bien sûr...).

Au point de vue acoustique, c'est l'espace intérolabial qui est important. On peut observer différents mouvements importants pour la phonation dont :

- ✓ L'occlusion (les lèvres sont fermées)
- ✓ La protrusion (les lèvres sont avancées vers l'avant)
- ✓ L'élévation et l'abaissement de la lèvre inférieure
- ✓ L'étirement, l'abaissement ou l'élévation des commissures [9].

II.5. Les phonèmes et phonétique

II. 5.1. Notion de phonétique

Les Phonèmes est un ensemble d'unités linguistiques élémentaires qui ont la propriété de changer le sens d'un mot et qui sont formées par un nombre finis d'éléments sonores distinctifs de la parole**.**

Un phonème est donc la plus petite unité phonique fonctionnelle, c'est-à-dire distinctive. Il n'est pas défini sur un plan acoustique, articulatoire, ou perceptuel, mais bien sur le plan fonctionnel. Les phonèmes n'ont pas d'existence indépendante : Ils constituent un ensemble structuré dans lequel chaque élément est intentionnellement différent de tous les autres, la différence étant à chaque fois porteuse de sens. La liste des phonèmes pour la plupart des langues européennes a été établie dès la fin du 19è siècle [9].

Les phonèmes peuvent ainsi être vus comme les éléments de base pour le codage de L'information linguistique.

Cependant, ces phonèmes peuvent se regrouper en classes dont les éléments partagent des caractéristiques communes. On parlera ici de "traits distinctifs" [9].

Trais distinctifs

Un trait distinctif est l'expression d'une similarité au niveau articulatoire, acoustique ou

perceptif des sons concernés [9].

Par exemple, pour les voyelles on distinguera 4 traits distinctifs :

- La nasalité : la voyelle a été prononcée à l'aide du conduit vocal et du conduit nasal suite à l'ouverture du velum.
- Le degré d'ouverture du conduit vocal
- La position de la constriction principale du conduit vocal, cette constriction étant réalisée entre la langue et le palais.
- La protrusion des lèvres.

De même, les consonnes seront classées à l'aide de 3 traits distinctifs :

- Le voisement : la consonne a été prononcée avec une vibration des cordes vocales
- Le mode d'articulation (on distinguera les modes occlusif, fricatif, nasal, glissant ou liquide).
- Lieu d'articulation est souvent l'appellation de la position de la constriction principale du conduit, qui contrairement aux voyelles n'est pas nécessairement réalisé avec le corps de la langue.

En fait, les phonèmes (qui peuvent donc être décrits suivant leurs traits distinctifs) sont des éléments abstraits associés à des sons élémentaires. Bien entendu, les phonèmes ne sont pas identiques pour chaque langue et le /a/ du français n'est pas totalement équivalent au /a/ de l'anglais. Ainsi, est née l'idée de définir un alphabet phonétique international (alphabet IPA) qui permettrait de décrire les sons et les prononciations de ces sons de manière compacte et universelle.

Il existe d'autres façons d'organiser les sons, par exemple en opposant les sons sonnants (voyelles), les consonnes nasales, les liquides ou les glissantes » aux sons obstruant « occlusives, fricatives ».

Les voyelles

Les voyelles sont typiquement produites en faisant vibrer ses cordes vocales. Le son de telle ou telle voyelle est alors obtenu en changeant la forme du conduit vocal à l'aide des différents articulateurs. Dans un mode d'articulation normal, la forme du conduit vocal est maintenue relativement stable pendant quasiment toute la durée de la voyelle [9].

Les consonnes

Comme pour les voyelles, les consonnes vont pouvoir être regroupées en traits distinctifs. Contrairement aux voyelles, elles ne sont pas exclusivement voisées (même si les voyelles prononcées en voix chuchotée sont, dans ce cas également, non voisées) et ne sont pas nécessairement réalisées avec une configuration stable du conduit vocal [9].

- **Les consonnes voisées**

Elles sont produites avec une vibration des cordes vocales. Lorsqu'en plus du voisement, une source de bruit est présente due à une constriction du conduit vocal, on pourra parler de consonnes à excitation mixte [9].

- **Les fricatives**

Sont produites par un flux d'air turbulent prenant naissance au niveau d'une constriction du conduit vocal. On distingue plusieurs fricatives suivant le lieu de cette constriction principale :

- Les labiodentales, pour une constriction réalisée entre les dents et les lèvres.
- Les dentales, pour une constriction au niveau des dents.
- Les alvéolaires, pour une constriction juste derrière les dents.

En fait, suivant les langues, en regardant plusieurs langues, on s'aperçoit que quasiment tous les points d'articulations du conduit vocal peuvent être utilisés pour réaliser des fricatives [9].

C'est d'ailleurs l'une des difficultés de l'apprentissage des langues étrangères car il n'est pas aisé d'apprendre à réaliser des sons qui demandent de positionner la langue à des endroits inhabituels [9].

- **Les plosives**

Elles sont caractérisées par une dynamique importante du conduit vocal. Elles sont réalisées en fermant le conduit vocal en un endroit. L'air provenant des poumons crée alors une pression derrière cette occlusion qui est ensuite soudainement relâchée suite au mouvement rapide des articulateurs ayant réalisé cette occlusion. De même, que pour les fricatives, l'un des traits distinctifs entre les plosives est le lieu d'articulation. Pour les plosives, on aura ainsi:

- Les labiales, pour une occlusion réalisée au niveau des lèvres.

- Les dentales, pour une occlusion au niveau des dents.
- Les vélo-palatales, pour une occlusion au niveau du palais.

En plus du lieu d'articulation, les plosives peuvent également être voisées ou non voisées [9].

Ainsi, une dentale voisée qui est prononcée avec le même lieu d'articulation.

- **Les consonnes nasales**

Elles sont en général voisées et sont produites en effectuant une occlusion complète du conduit vocal et en ouvrant le vélum permettant au conduit nasal d'être l'unique résonateur. Comme pour les autres consonnes, on aura, suivant le lieu d'articulation :

- Les labiales, pour une occlusion du conduit vocal réalisée au niveau des lèvres.
- Les dentales, pour une occlusion du conduit vocal au niveau des dents.
- Les vélo-palatales, pour une occlusion du conduit vocal au niveau du palais [9].

- **Les glissantes et les liquides**

Cette classe de consonnes regroupe des sons qui ressemblent aux voyelles. Les liquides sont d'ailleurs parfois appelés semi consonnes ou semi-voyelles. Les glissantes et les liquides, en général, voisées et non nasales [9].

- Les glissantes, comme leur nom l'indique, sont des sons en mouvement et précédent toujours une voyelle « ou un son vocalique ».
- Les liquides « ou semi-voyelles » sont des sons tenus, très similaires aux voyelles mais en général avec une constriction plus conséquente et avec l'apex de la langue plus relevé.

L'alphabet arabe comporte 28 consonnes. On peut les classer selon plusieurs critères: des consonnes articulées avec une vibration des cordes vocales et des consonnes qui n'engendrent pas une vibration des cordes vocales, le franchissement de l'air à travers le conduit vocal donne naissance à d'autres variétés de sons. Les consonnes ne peuvent être décrites, contrairement aux voyelles, par des valeurs formantiques statiques car elles sont généralement perçues par des fréquences variables rapidement qui, au fait, reflètent les mouvements articulatoires générés lors de leur production. Le tableau ci-dessous montre une classification des consonnes arabes selon leurs modes d'articulation.

Nature		Trans A.P.I	Voisement	Mode articulation	Code maison	Trans arabe
VOYELLES	Courte	œ	+	Vocalique	Oe	أَوَوَوَوَ
		u	+	Vocalique	U	أُ
		i	+	Vocalique	I	إ
	Longue	œ :	+	Vocalique	Oe :	آ
		u :	+	Vocalique	u :	أو
		i :	+	Vocalique	i:	إي
CONSONNES		b	+	Plosive	B	ب
		d	+	Plosive	D	د
		d,	+	Plosive	d.	ض
		t	-	Plosive	T	ت
		t,	-	Plosive	t.	ط
		k	-	Plosive	K	ك
		q	-	Plosive	K-	ق
		ð	+	Fricative	Z-	ذ
		z.	+	Fricative	Z.	ظ
		z	+	Fricative	Z	ز
		ʒ	+	Fricative	ch -	ج
		ç	+	Fricative	V	ع
		ɣ	+	Fricative	v-	غ
		f	-	Fricative	F	ف
		θ	-	Fricative	Gh	ث
		s	-	Fricative	S	س
		ş	-	Fricative	s.	ص
		ʃ	-	Fricative	s-	ش
		ħ	-	Fricative	Ch	ح
		x	-	Fricative	ch-	خ
		h	-	Fricative	&	ه
		m	+	Nasale	M	م
		n	+	Nasale	N	ن
		w	+	Sonnante centrale	W	و
		j	+	Sonnante centrale	j	ي
		l	+	Sonnante latérale	L	ل
		ļ	+	Sonnante latérale	l.	ل
		r	+	Vibrantes	R	ر
		ŗ	+	Vibrantes	R.	ر

Tableau II.1 : Classification des phonèmes arabes selon le mode d'articulation et la transcription API.

II. 6. L'appareille auditif

Le système auditif humain possède des capacités remarquables. Sa sensibilité s'étend sur une plage de fréquences audibles allant de 20Hz à 20kHz et couvre un ensemble de pressions acoustiques de 20µPa à 20Pa ce qui représente une incroyable dynamique 1 de 120dB. En outre, les capacités de discrimination aussi bien en fréquence qu'en intensité sont excellentes. Cette sensibilité est étonnante au vu du faible nombre de cellules sensorielles dont dispose la cochlée (3500 cellules ciliées internes) en comparaison avec les millions de cellules photosensibles de l'œil [10].

Cette qualité « se paie » en retour au niveau de la complexité de fonctionnement de la chaîne de traitement que constitue l'oreille. L'appareil auditif, dont la structure est représentée dans la Figure II.6 est divisée en trois éléments de nature et de fonction différentes mais complémentaires :

— l'oreille externe ;

— l'oreille moyenne ;

— l'oreille interne.

Nous allons indiquer de manière simplifiée le fonctionnement global de l'appareil auditif [10].

Captées par l'oreille externe, les vibrations acoustiques sont transmises par l'oreille moyenne au milieu liquidien de la cochlée, organe de l'audition de l'oreille interne. Au sein de la cochlée, les vibrations provoquent la mise en mouvement des liquides et des différentes membranes qui la constituent. Ces mouvements provoquent à leur tour l'inclinaison des stéréo-cils des cellules ciliées déclenchant ainsi l'activation des fibres nerveuses. Ces dernières transmettent alors un message électrique vers le cortex cérébral [10].

Dans l'oreille interne, ces mécanismes vibratoires ne sont pas seulement le fait de réactions mécaniques passives. En effet, des mécanismes actifs complexes viennent modifier et enrichir les propriétés mécaniques de la cochlée [10].

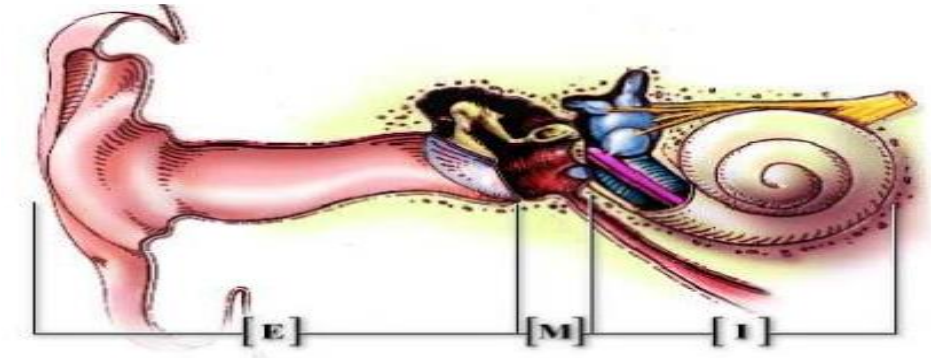


Figure II.6: Schéma de l'appareil auditif humain comprenant l'oreille externe [E], l'oreille moyenne [M] et l'oreille interne [I]. (Image INSERM reproduit de [10]).

II.6.1. L'oreille externe

L'oreille externe est le premier maillon de la chaîne que constitue l'appareil auditif. Elle reçoit les vibrations acoustiques aériennes et les transmet à l'oreille moyenne. L'oreille externe est composée du pavillon et du conduit auditif externe (Figure II.6).

L'oreille externe a une fonction d'antenne acoustique pour l'audition. Le pavillon amplifie de quelques décibels les fréquences avoisinant 5kHz et le conduit auditif externe amplifie d'une dizaine de décibels celles autour de 2, 5kHz. L'effet total du corps et de l'oreille externe engendre globalement une amplification qui varie entre 5 et 20dB sur la plage des fréquences de 2 à 7kHz [12, 13].

Chez l'homme, le pavillon n'étant pas mobile, son intérêt pour la localisation des sources est moindre que chez la plupart des autres mammifères. Néanmoins, le cerveau parvient à compenser, en partie, cet inconvénient. En effet, les ondes sonores transmises à l'oreille moyenne sont altérées par la combinaison entre les effets de diffraction acoustique sur le corps ou sur la tête et les effets de la géométrie de l'oreille externe. Pour chaque oreille, l'altération est fonction de la position de la source sonore par rapport au pavillon et par rapport à l'axe du conduit auditif externe. Elle concerne principalement les hautes fréquences. L'oreille externe joue donc un rôle dans la localisation des sources sonores en transmettant au reste du système auditif différentes informations spectrales et temporelles témoignant de l'origine spatiale de la source et qui sont interprétées par le cortex cérébral .Plus précisément, pour une position particulière de la source sonore, la vibration acoustique parvient, à chacune des oreilles, affectée par une certaine fonction de transfert [14].Ces fonctions de transfert varient selon les positions relatives de la source et de l'auditeur et correspondent aux

fonctions de transfert de tête ou Head-Related Transfer Functions (HRTF) utilisées en traitement du signal dans le contexte de la spatialisation des sons [15, 16].

Cette faculté à localiser les sources sonores est relativement importante et intervient de manière capitale lorsque l'environnement est particulièrement bruyant en permettant à l'auditeur de focaliser son attention sur le locuteur. Cette aptitude que l'on désigne par le terme d'effet cocktail-party assure une bonne compréhension de la parole dans des situations à faible rapport signal à bruit [10].

En résumé, le rôle auditif de l'oreille externe est donc moindre que celui des autres organes de l'oreille. Celle-ci assure principalement une fonction de protection du tympan, d'amplification des sons (dans la zone 2 à 7kHz) et permet la localisation des sources sonores [10].

II. 6.2 L'oreille moyenne

L'oreille moyenne est l'élément essentiel de la transmission sonore. Comme le montre le schéma II.7, elle est composée principalement par la membrane tympanique et la chaîne des osselets ou chaîne ossiculaire [10].

Sa fonction principale est de transmettre les vibrations aériennes vers l'oreille interne. Afin de parvenir à ce résultat, l'oreille moyenne réalise l'adaptation d'impédance nécessaire entre le milieu aérien et le milieu liquidien de l'oreille interne. Sans cette adaptation d'impédance, l'énergie acoustique serait réfléchie dans sa quasi-totalité puisqu'il s'agit d'une transmission d'un milieu de basse impédance vers un milieu de haute impédance [13, 17].

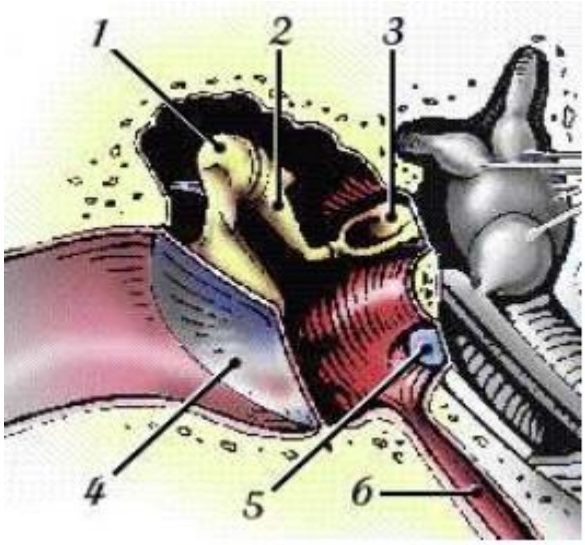


Figure II.7: Schéma de l'oreille moyenne (image INSERM reproduit de [11]).

La chaîne ossiculaire comprend le marteau **(1),** l'enclume **(2)** et l'étrier **(3).** Le tympan **(4)** sépare le conduit auditif externe de la cavité de l'oreille moyenne qui est en relation avec la cavité buccale par la trompe d'Eustache**(6)**. La fenêtre ovale, sur laquelle s'applique la platine de l'étrier**(3)**, et la fenêtre ronde**(5)** séparent l'oreille moyenne et l'oreille interne.

L'adaptation est réalisée en partie par un effet de levier mais surtout grâce au rapport de surface entre la surface du tympan et celle de la platine de l'étrier qui s'appuie sur la fenêtre ovale de la cochlée [18].

Ce rôle d'adaptation d'impédance est particulièrement efficace sur la plage des fréquences de la parole et permet à environ 46% de l'énergie d'être transmis [18]. Dans [19], Moore cite une autre propriété probable du système ossiculaire : la protection aux bruits internes (comme par exemple les bruits de mastication). Sans l'organisation particulière du système ossiculaire, ceux-ci seraient transmis à la cochlée par conduction osseuse [10].

❖ **Le réflexe stapédien**

La deuxième fonction de l'oreille moyenne est de protéger, par réflexe, l'oreille interne aux sons de fort niveau, en particulier pour les basses fréquences. Ce réflexe se déclenche normalement en 150ms pour des sons de plus de 80dB HL (c'est à dire 80dB au dessus du seuil auditif 2) et en 25 à 35ms pour des niveaux plus élevés. Il met en jeu le muscle stapédien qui, en se contractant, augmente la rigidité de la chaîne tympan-ossiculaire et réduit ainsi la transmission sonore vers l'oreille interne [17, 11].

Cette protection est néanmoins très limitée:

- en niveau : de 10 à 20dB jusqu'à 2000Hz
- en durée : au bout de quelques minutes, les muscles se relâchent après épuisement.
- en fréquence : aux sons en dessous de 1 à 2kHz.
- en nature : il ne protège pas des bruits impulsifs.

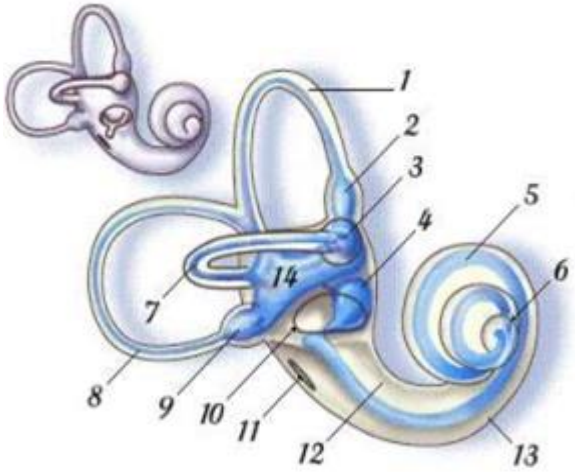
Ce réflexe ossiculaire interviendrait également pour un locuteur ou un chanteur dans la réduction de la perception de sa propre voix [10].

II. 6.3. L'oreille interne

L'oreille interne est l'organe principal de l'audition, car c'est le responsable de la transduction du signal acoustique en message nerveux. Son anatomie et sa physiologie très complexes sont à l'origine des capacités auditives très performantes dont disposent les mammifères [10].

- **La cochlée**

  L'oreille interne regroupe deux organes sensoriels distincts : le vestibule, organe de l'équilibre que nous ne décrirons pas ici et la cochlée, organe de l'audition.



(1) Canal antérieure, (2) Ampoule (du meme canal), (3) Ampoule (canal horizontal), (4) Saccule, (5) Canal cochléaire, (6) Hélicotréme, (7) Canal latérale, (8) Canal postérieure, (9) Ampoule, (10) Fenêtre ovale, (11) Fenêtre ronde, (12) Rampe vestibulaire, (13) Rampe tympanique, (14) Utricule.

Figure II.8 : Schéma de l'oreille interne (image INSERM reproduit de [11]).

La cochlée est constituée d'un ensemble de trois tubes enroulés en spirale sur deux tours et demi (chez l'homme) et remplis de liquides (Figures II.8 et II.9) :

- les rampes tympanique et vestibulaire remplies par la périlymphe,
- le canal cochléaire rempli par l'endolymphe.

Les vibrations mettent en mouvement le tympan et la chaîne des osselets. L'étrier, plaqué sur la fenêtre ovale (Figures II.7 et II.8), transfère la vibration au compartiment périlymphatique de la rampe vestibulaire selon le sens de circulation indiqué par les flèches de la Figure II.9.

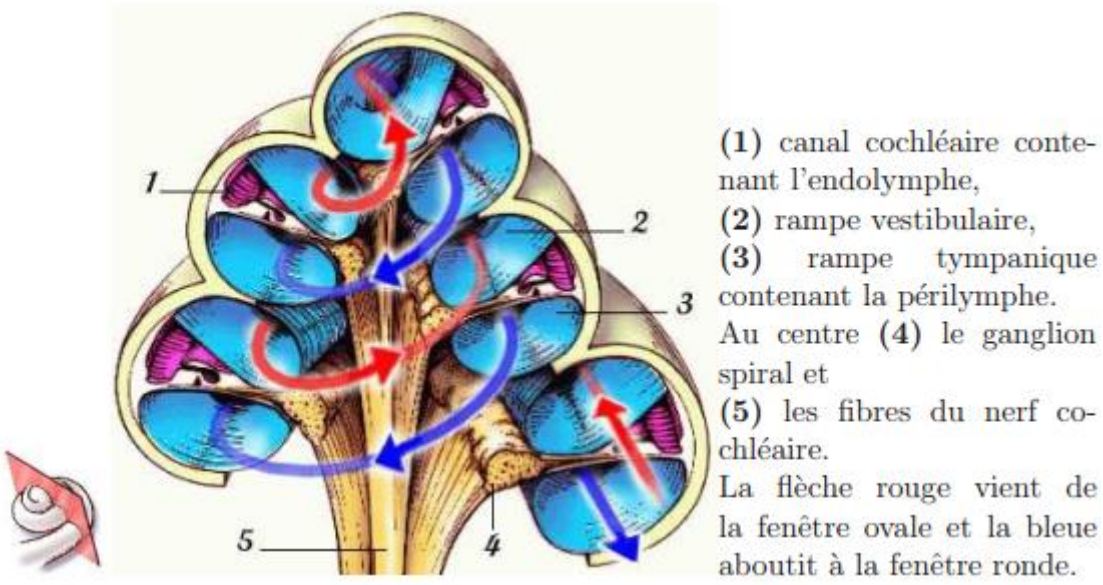


Figure II.9 : Section axiale de la cochlée (image INSERM reproduit de [11])

C'est dans le canal cochléaire (indiqué par 1 sur le schéma II.9) que se trouve l'organe sensoriel essentiel de l'audition : l'organe de Corti.

II.7. Traitement du signal vocal

L'information contenue dans le signal de parole peut être analysée de bien des façons. Si nous observons la forme que produit la parole selon l'audiogramme présenté par la Figure II.10.

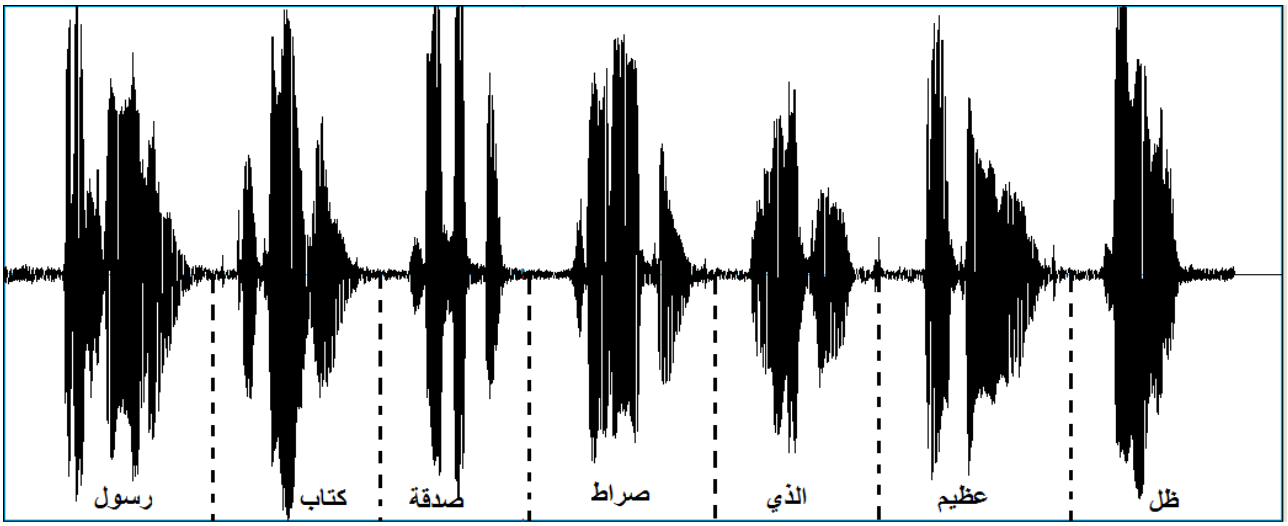


Figure II.10: Audiogramme d'un signal vocal enregistré par Praat.

Nous remarquons une forme apériodique avec des amplitudes variantes ou pseudopériodique. Ainsi, aux cotés droit et gauche du signal principal nous distinguons des petites courbes non identifiées, ce que nous appelons le bruit. Il y a plusieurs travaux sur le

sujet de reconnaissance de parole/ non parole basés sur le bruit (Speech/-NON Speech) [19]. En plus, chaque individu possède sa propre information vocale qui le caractérise. Et cette information peut être extraite à partir des signaux sortant du résonateur. Les traits acoustiques du signal de parole sont directement liés à sa production dans l'appareil phonatoire. Tout d'abord, nous avons l'énergie du son [20] ; celle-ci est liée à la pression de l'air en amont du larynx. Puis nous avons la fréquence fondamentale F_o [21] ; cette fréquence correspond à la fréquence du cycle d'ouverture/fermeture des cordes vocales.

Enfin, nous avons le spectre du signal de parole [22] ; celui-ci résulte du filtrage dynamique du signal en provenance du larynx par le conduit vocal qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses (Figure II.11).

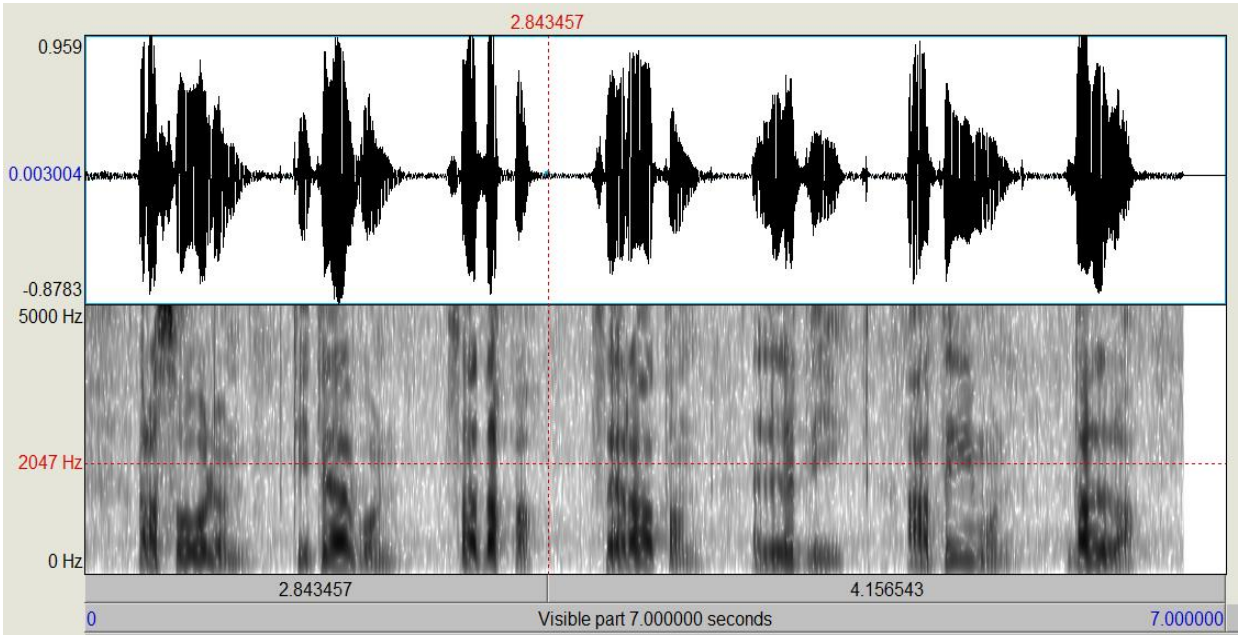


Figure II.11 : Un signal vocal et le spectrogramme associé.

Chacun de ces traits acoustiques est lui-même intimement lié à une autre grandeur perceptuelle, à savoir l'intensité, le rythme, et le timbre. Le spectrogramme est la représentation temps-fréquence qui permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné. L'ensemble des spectres constituant le spectrogramme sont calculé par la transformé de Fourier que nous allons voir plus en détails par la suite.

II. 7.1. Intensité d'un signal vocal

L'intensité d'un son, appelée aussi volume, permet de distinguer un son fort d'un son faible. Elle correspond à l'amplitude de l'onde. L'amplitude est donnée par l'écart maximal

de la grandeur qui caractérise l'onde. Pour le son, onde de compression, cette grandeur est la pression. L'amplitude sera donc donnée par l'écart entre la pression la plus forte et la plus faible exercée par l'onde acoustique. Lorsque l'amplitude de l'onde est grande, l'intensité est grande et donc le son est plus fort. L'intensité du son se mesure en décibels (dB). On distingue différentes façons de mesurer l'amplitude d'un son :

- La puissance acoustique : La puissance acoustique est associée à une notion physique. Il s'agit de l'énergie transportée par l'onde sonore par unité de temps et de surface. Elle s'exprime en Watt par mètre carré ($W.m^{-2}$).
- Addition de sons : L'échelle des décibels est une échelle dite logarithmique, ce qui signifie qu'un doublement de la pression sonore implique une augmentation de l'indice d'environ 3 : avec 3 dB de plus, l'intensité est en fait doublée [23].

II. 7.2. La fréquence fondamentale

La fréquence fondamentale $\mathbf{F_o}$(ou pitch [période des sons voisés) joue un rôle important dans la parole [24]. C'est elle qui véhicule une grande partie de l'information prosodique. L'intensité de la voix et les durées successives des syllabes complètent ces informations. D'une manière générale, la prosodie, qui peut être considérée comme l'effet des différentes variations de la fréquence fondamentale $\mathbf{F_o}$, de l'intensité et de la durée, peut faire ressortir bien des caractéristiques du locuteur, comme son genre, ses origines géographiques et culturelles, ses émotions, etc. mais participe aussi à la caractérisation de la langue elle-même, par la manière dont elle est utilisée pour différencier les divers éléments syntaxiques comme les énoncés (interrogatifs, exclamatifs ou déclaratifs), l'importance de certains mots, ou bien même pour caractériser les différences lexicales entre les mots. $\mathbf{F_o}$moyen-homme 100 à 150 Hz, F_o moyen-femme 200 Hz à 300 Hz et $\mathbf{F_o}$moyen enfant 350 à 400 Hz [2].

II.7.3. Les formants

Le voisement que nous percevons est différent de celui produit à la source par les cordes vocales. Ce que nous entendons est le fruit d'un phénomène de filtrage sur une onde complexe. La capacité d'amplifier ou, au contraire, d'atténuer certaines fréquences est la propriété de tout résonateur. Dans le cas précis de la parole, le stimulus est fourni par l'onde périodique complexe provenant du mouvement des cordes vocales et ce sont les cavités supra-glottiques qui assurent la fonction de résonateur. La forme, la dimension ainsi que la matière qui compose ces cavités sont autant de particularités qui détermineront les fréquences qui seront mises en évidences et celles qui seront atténuées. Les cavités supra-glottiques ont la

capacité de neutraliser certaines harmoniques et d'en mettre d'autres en évidence par un simple changement de configuration. Lorsque l'on prononce, sur une note constante ou à une hauteur de voix constante, des voyelles aussi différentes que" oe i u ", c'est le procédé d'atténuation et de renforcement qui entre en jeu et qui est responsable de l'apparition du timbre propre à chacune des voyelles. Donc le conduit vocal possède une fonction de transfert (filtre) appliquée à une source produit un phonème, on appelle formant les maxima locaux (pics) de cette fonction de transfert notés **F1**, **F2**, **F3**… et correspondent aux zones de renforcement maximal. Néanmoins, du point de vue perceptif, seuls quelques formants jouent un

rôle central au niveau de la parole. En théorie on décrit souvent les voyelles grâce à leurs 2 premiers formant **F1**, **F2** a travers un triangle acoustique [2].

II.7.4. Le rythme

Le rythme est la durée des silences et des phones. Il est difficile de les en extraire car un mot prononcé d'une façon naturelle, sans aucun traitement, donne un mélange de phones chevauchés entre eux et un silence d'intensité non nulle (le bruit).

II.7.5. Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix. Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez ; ce sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son. Les éléments physiques du timbre comprennent :

- la répartition des fréquences dans le spectre sonore,
- les relations entre les parties du spectre, harmoniques ou non,
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence),
- l'évolution dynamique globale du son,
- l'évolution dynamique de chacun des éléments les uns par rapport aux autres.

II. 8. Conclusion

Dans ce chapitre on à cité les organes productives de la parole puis défini l'aspect perceptive et les différentes propriétés physiques du son. Comme on le souligne souvent, il n'y a pas dans le corps humain un organe responsable de la parole : la parole se fait par la collaboration de différents organes (qui servent normalement à respirer, mastiquer etc.). Le son est fondamentalement une vibration de l'air. Cet air est fourni par les poumons ; la

vibration est fournie par les plis vocaux ; enfin, des mouvements compliqués de la langue (principalement) et des lèvres donnent au son une « couleur » particulière.

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 trois grandes classes d'organes: les poumons, le larynx, et les cavités supraglottiques.

Chapitre III

Segmentation et Paramétrisation des caractéristiques du signal de parole

III.1. Introduction

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 paramétrisation sur ce signal. L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes [8], un tel système prend un signal en entrée et retourne un vecteur de paramètre (appelé indifféremment vecteur acoustique ou encore vecteur d'observation).Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

Donc la numérisation consiste à faire passer le signal par trois étapes essentielles et qui sont : l'échantillonnage, la quantification et le codage.

Une fois la numérisation finalisée, le signal doit subir une préaccentuation afin de relever les hautes fréquences, puis segmenté en trames. Chaque trame est constituée d'un nombre N fixe d'échantillons de parole.

En générale, N est fixe de telle manière que chaque trame corresponde à environ 20 ms de parole (durée pendant laquelle la parole peut être considérée comme stationnaire).

Enfin, un fenêtrage de Hamming est effectué, afin de limiter les effets du phénomène de Gibbs. Toutes cette mise en forme du signal, permet de calculer les coefficients à intervalle temporel régulier qui doivent représenter au mieux le signal à modéliser, et extraire le maximum d'information nécessaire à la reconnaissance de la parole.

Ces coefficients jouent un rôle capital dans les approches utilisées pour la reconnaissance de la parole. En effet, ces paramètres qui modélisent le signal seront fournis au système de reconnaissance.

Dans notre travail, étant donné que nous nous intéressons au milieu bruité, nous nous sommes limités à deux approches les plus utilisées en littérature :

- Paramétrisation basée sur un modèle de production de la parole (LPC).
- Paramétrisation basé sur une analyse dans le domaine Cepstral (MFCC).

Donc, notre chapitre sera consacré à traiter le signal de la parole par les étapes suivantes :

- Mise en forme du signal.
- Détermination des coefficients MFCC.
- Détermination des coefficients LPC.

III.2. Approches et techniques d'extraction de caractéristiques

Dans cette partie nous allons détailler les approches d'extraction de caractéristique qu'ils existent et les techniques utiliser dans notre recherche :

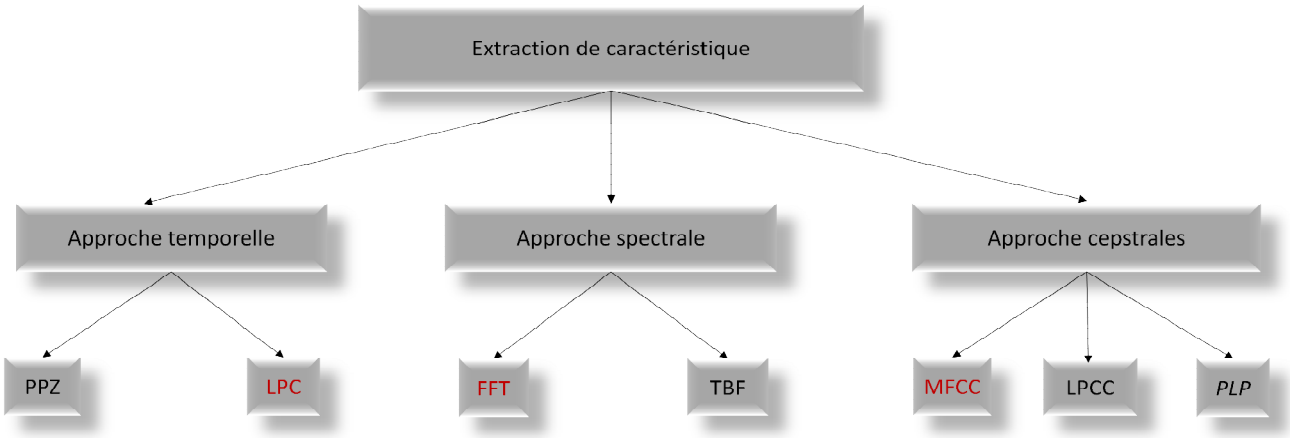


Figure III.1: Schéma présentant les différentes méthodes d'extraction de caractéristique [30].

III.2.1. Approche temporelle

Cette approche étudier le signal de parole de manière à observer la forme temporelle du signale. On peut déduire un certain nombre de caractéristiques à partir de cette forme temporelle qui pourront être utilisées pour le traitement de la parole. Il est, par exemple, assez claire de distinguer les partie voisées, dans lesquelles on peut observer une forme d'onde quasi-périodique, des parties non voisées dans lesquelles un signal aléatoire de faible amplitude est observé [9].

Le signal de parole est un signal quasi-stationnaire. Cependant, sur un horizon de temps supérieur, il est clair que les caractéristiques du signal évoluent significativement en fonction des sons prononcés comme illustré sur la figure ci-dessous [9].

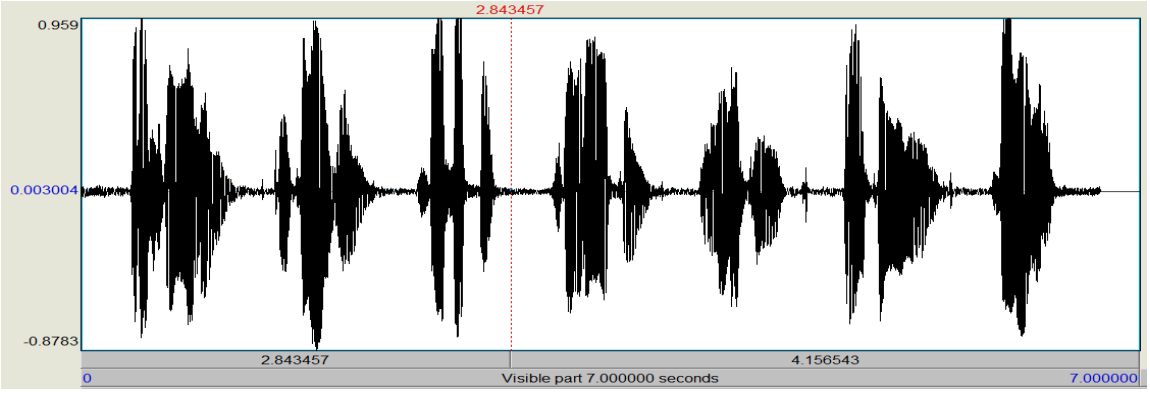


Figure III.2: Représentation temporelle(Audiogramme) de signaux de parole.

Les méthodes de type temporel sont basées sur l'analyse des caractéristiques temporelles du signal vocal telles que : l'énergie, le taux de passage par zéro, le calcul de la fréquence fondamental etc. Différentes techniques permettent l'analyse de l'aspect temporel du signal vocal afin de permettre de déduire ses paramètres, parmi ces méthodes nous trouvons [25] :

- ❖ Le taux de passage par zéro (PPZ),
- ❖ L'analyse par prédiction linéaire (LPC).

➢ **Codage prédictif linéaire (LPC) :**

Le codage prédictif linéaire (LPC, Linear Predictive Coding) est une méthode de codage et de représentation de la parole [26]. Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire. Il s'agit donc de prédire le signal à un instant n à partir des p échantillons précédents .La parole n'étant cependant pas un processus parfaitement linéaire, la moyenne mobile que constitue la somme pondérée du signal sur p pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme e(n).

Le codage par prédiction linéaire consiste donc à déterminer les coefficients a_i qui minimisent l'erreur e (n), ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage.

La méthode du codage par prédiction linéaire est tout autant utilisée en RAP qu'en compression pour le transfert de la voix par téléphone ou radio. Elle n'est cependant pas parfaite puisque l'erreur de prédiction peut être importante sans qu'il soit possible, par cette méthode, de la corriger.

La méthode RELP, Residual Excited Linear Prediction, permet de réduire une partie de cette erreur. Le principe consiste à comparer, lors de la prédiction linéaire, le signal obtenu avec le signal original. L'erreur, obtenue par soustraction, représente la partie du signal original que le prédicteur n'arrive pas à modéliser. Dans la méthode RELP, l'erreur résiduelle

est passée dans un filtre passe-bas permettant de conserver l'erreur effectuée dans la seule bande fréquentielle allant de 0 à1000 hertz. La sortie du filtre est alors codée et passée au receveur qui peut alors reconstruire un signal à partir de la prédiction et de l'erreur observée [6].

Pour pallier le problème de l'erreur résiduelle, d'autres méthodes fondées sur la prédiction linéaire ont été développées. Ainsi la méthode CELP, Code Excited Linear Prediction, permet d'effectuer une compression de la parole par codage d'une trame vis-à-vis de références stockées dans un corpus. Ainsi, une trame de parole sera codée selon une combinaison linéaire de certaines trames du corpus et c'est cette combinaison linéaire qui sera considérée à la place de la trame dans les traitements ultérieurs. Cette méthode de codage de la parole est surtout employée pour la compression et la transmission de la parole à de faibles débits [27].

III.2.2. Approche fréquentielle ou spectrale

La deuxième approche pour caractériser et représenter le signal de parole est d'utiliser une représentation spectrale [7].

Ces méthodes sont fondées sur une décomposition fréquentielle du signal sans connaissance a priori de sa structure fine. Il s'agit donc de transformer le signal original de la représentation temporelle à une représentation fréquentielle par la transformé de Fourier représente sous la formule (III.1) [7].

$$F(w) = \int_{-\infty}^{+\infty} f(t) e^{-iwt}\, dt \qquad \text{(III.1)}$$

Ou j2 = -1 et f(t) est la fonction temporelle.

- **Transformé de Fourier Rapide (TFR) :**

La Transformée de Fourier Rapide (notée par la suite FFT) est simplement une TFD calculée selon un algorithme permettant de réduire le nombre d'opérations et, en particulier, le nombre de multiplications à effectuer. Il faut noter cependant, que la réduction du nombre d'opérations arithmétiques à effectuer, n'est pas synonyme de réduction du temps d'exécution. Tout dépend de l'architecture du processeur qui exécute le traitement [26].

III.2.3. Approche cepstrale

Pour séparer les deux informations présentes dans le signal de parole que sont la fréquence fondamentale et la transformation, supposée linéaire, effectuée par le conduit vocal, il est nécessaire d'effectuer une déconvolution a posteriori du signal pour connaître la

contribution des cordes vocales et du conduit vocal lors de la génération du signal qui a, par la suite, été observé en entrée du système. Cette déconvolution peut être effectuée grâce au cepstre. Il est à noter que le nom même decepstre est défini à partir du mot spectre (cepstre = $(\text{spec})^{-1}$ tre). De même, la représentation temps-fréquence associée n'est plus qualifiée de fréquentielle mais de quéfrentielle.

Le cepstre est une méthode qui se fonde sur la transformée de Fourier mais qui, grâce à une méthode efficace, permet d'isoler la fréquence initiale du fondamental de la transformation qui a été opérée par le conduit. Comme pour le calcul du spectrogramme, le signal est pré accentué puis convolué avec une fenêtre mobile. Une première transformée de Fourier est alors calculée pour obtenir un spectre du signal, comme pour un spectrogramme. Ces coefficients sont ensuite transformés par logarithme module. La convolution étant un opérateur multiplicatif, ce passage par les logarithmes permet de passer les coefficients dans un espace additif. Une transformée de Fourier inverse permet alors d'obtenir un cepstre dont un coefficient représente le fondamental, les autres coefficients permettant d'obtenir le spectre de la convolution effectuée sur le fondamental. Cette méthode de calcul des cepstres est élémentaire [27], il existe également des méthodes itératives effectuant un lissage, ce qui permet d'obtenir des cepstres de meilleure qualité.

Une extension possible des cepstres est leur passage dans un espace fréquentiel non linéaire proche de l'audition humaine. Il est ainsi possible de modifier la procédure de calcul précédente pour que les coefficients obtenus soient répartis selon une échelle Mel. Une telle procédure, permet d'obtenir des coefficients cepstraux à échelle Mel, Mel Frequency Cepstral Coefficients, MFCC. Ces coefficients ont été très utilisés en RAP du fait des bons résultats qu'ils ont permis d'obtenir. [28] avait d'ailleurs comparé cette méthode à d'autres du même ordre avec des conclusions qui, déjà, laissaient entrevoir la qualité des informations extraites par la méthode MFCC. Parmi les méthodes auxquelles les MFCC avaient été comparés se trouvaient des méthodes fondées sur la prédiction linéaire.

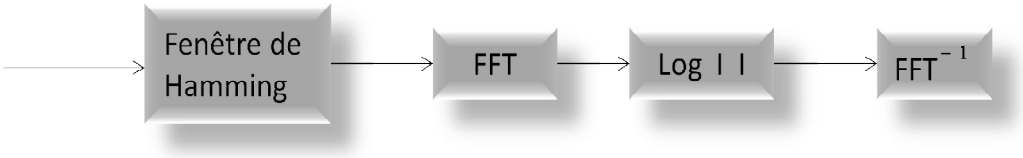


Figure III.3: Analyse cepstrale sur une fenêtre temporelle [30].

- **Coefficients cepstraux (MFCC) :**

Les coefficients cepstraux (MFCC) ont été très utilisés en RAP du fait des bons résultats qu'ils ont permis d'obtenir. Lorsque le spectre d'amplitude résulte d'une FFT sur le

signal de parole prétraité, lissé par une suite de filtres triangulaires répartis selon l'échelle Mel, les coefficients sont appelés Mel Frequency Cepstral Coefficients (MFCC). L'échelle non linéaire de Mel est donnée par la formule suivante [16]:

$$\text{Mel} = \left(\frac{1000}{\log 2}\right)\log\left(1+\frac{\text{F}}{1000}\right)$$

Afin de réduire l'information, une suite de filtres (triangulaires, rectangulaires...) est appliquée dans le domaine spectral selon l'échelle précédemment décrite. Les coefficients obtenus sont alors synonymes d'énergie dans des bandes de fréquence. La Figure III.4 donne un exemple de répartition d'une suite de filtres selon l'échelle Mel, couramment utilisée [25].

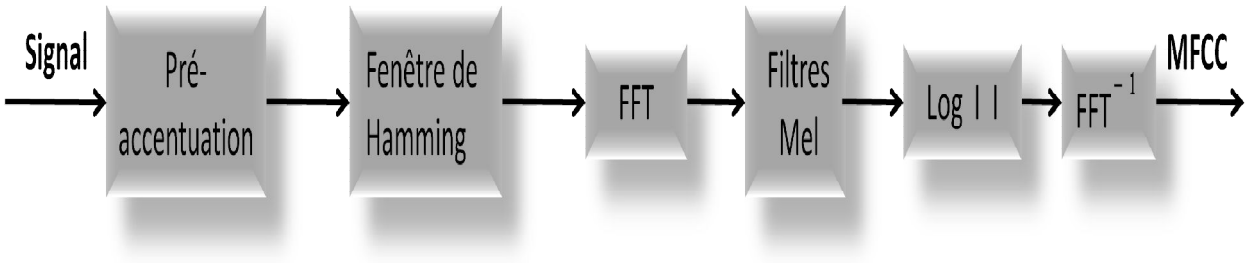


Figure III.4: Calcul des coefficients cepstraux MFCC [30].

III.3. Numérisation du signal

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 qui consiste en trois étapes: l'échantillonnage, la quantification et le codage.

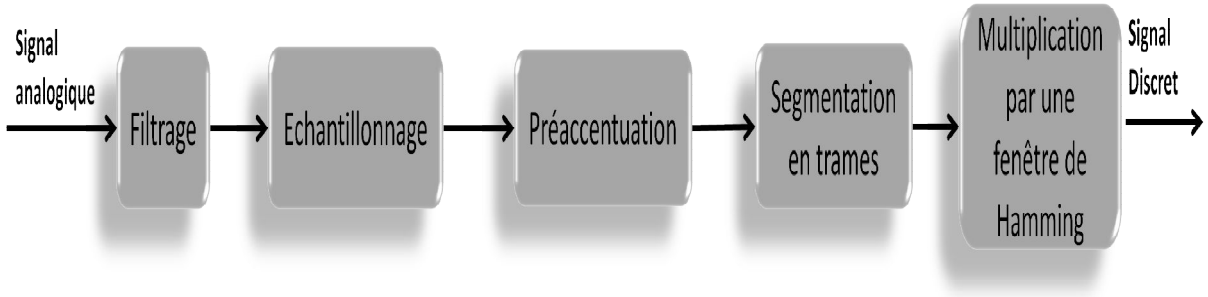


Figure III.5 : Mise en forme du signal de parole [30].

III.3.1. L'échantillonnage

L'échantillonnage consiste à transformer une fonction S(t) à valeurs continues en une fonction $\hat{S}(t)$ discrète constituée par la suite des valeurs S(t) aux instants d'échantillonnage

t = kT avec k un entier naturel (Figure III.6).Le choix de la fréquence d'échantillonnage n'est pas aléatoire car une petite fréquence nous donne une présentation pauvre du signal. Par contre une très grande fréquence nous donne des mêmes valeurs, redondance, de certains échantillons voisins donc il faut prélever suffisamment de valeurs pour ne pas perdre l'information contenue dans S(t). Le théorème suivant traite cette problématique :

Théorème (de Shannon) : La fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieur à 2 fois la fréquence haute du spectre du signal analogique.

$$F_{ech} = 2 \times F_{max}$$

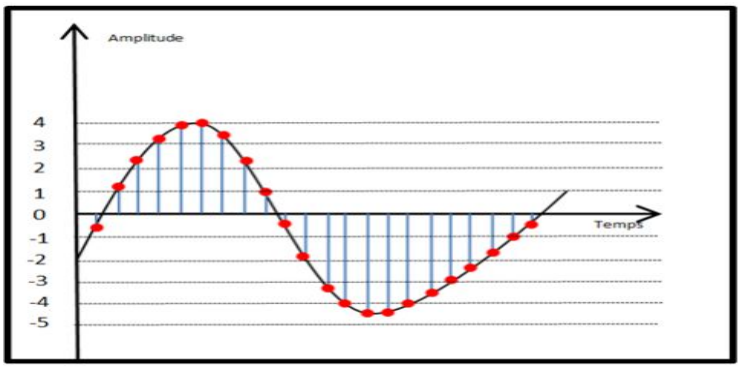


Figure III.6 : Un signal échantillonné.

Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit F_e= 8000 Hz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6000 à 16000 Hz. Par contre pour le signal audio (parole et musique), on exige une bonne représentation du signal jusque 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48kHz. Pour les applications multimédias, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz [29].

III.3.2 Quantification

Cette étape consiste à approximer les valeurs réelles des échantillons selon une échelle de n niveaux appelée échelle de quantification. Il y a donc 2^n valeurs possibles comprises entre -2n-1 et 2n-1 pour les échantillons quantifiés (Figure III.7). L'erreur systématique

que l'on commet en assimilant les valeurs réelles de l'écart au niveau du quantifiant le plus proche est appelé bruit de quantification.

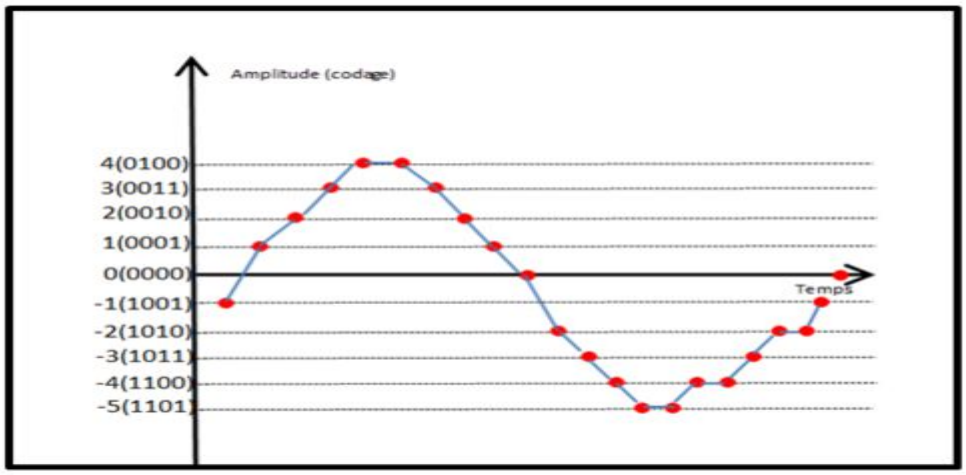


Figure III.7 : Un signal quantifie.

III.3.3. Codage

C'est la représentation binaire des valeurs quantifiées qui permet le traitement du signal sur machine (Figure III.7).

III.4. Préaccentuation

Le signal échantillonné S_e est ensuite préaccentué afin de relever les hautes fréquences qui sont moins énergétiques que les basses fréquences. Cette étape consiste à faire passer le signal $S_n$dans un filtre numérique à réponse impulsionnelle finie de premier ordre donné comme suit :

$$H(Z)=\mathbf{1}-\alpha Z^{-1} \qquad \text{Avec} \quad 0\leq \alpha \leq 1 \qquad \text{(III.2)}$$

Ainsi, le signal préaccentué S_a est lié au signal se par la formule suivante :

$$S_a\,(\boldsymbol{n})=S_e\,(\boldsymbol{n})-\propto S_e\,(\boldsymbol{n}-\mathbf{1}) \qquad \text{(III.3)}$$

III.5. Segmentation

Le processus de reconnaissance de la voix passe par une phase très importante : c'est la segmentation, tel qu'aucun système n'utilise cette phase, car elle prépare le signal de parole pour les traitements ultérieurs. Cette phase possédé une grande influence sur la qualité des caractéristiques à obtenir et par conséquent, le taux de classification à obtenir. Le but de cette phase est l'extraction des segments de base à traiter selon l'unité de base de

traitement, à savoir : mot, syllabe ou phonème … etc. ce processus est très influencé par le bruit intégré dans le signal enregistré [30].

Dans ce qui suit nous allons détailler les méthodes de segmentation que nous allons citée ultérieurement.

III.5.1. Segmentation en voisées/ non voisées

Les sons voisés sont produits par la vibration des cordes vocale. Les voyelles sont intrinsèquement voisées, tandis que les consonnes peuvent l'être ou non. On peut donc considérer qu'un mot est constitué d'une suite de segments voisés, de segments non voisés et de silences brefs. Cependant toute suite de ces trois segments de base ne correspond pas à un mot, du bruit peut être constitué par des sons voisés. Un des paramètres de voisement est le pitch [27].

III.5.2. Segmentation en phonèmes

La segmentation d'un signal de parole en phones consiste à délimiter sur le continuum acoustique de ce signal une séquence de segments caractérisés par des étiquettes appartenant à un ensemble discret et fini d'éléments, qui est l'alphabet phonétique de la langue. La segmentation phonétique de la parole est une tâche difficile car le signal de parole n'est pas clairement composé de segments discrets bien délimités [31].

D'un côté, nous constatons que l'élocution d'un énoncé se caractérise par un mouvement continu des organes de la parole et par l'absence d'un quelconque positionnement statique de ces organes. Le passage d'une cible articulatoire d'un phone, à une autre cible articulatoire d'un autre phone, se fait de manière continue, avec un chevauchement entre les deux configurations articulatoires, ce qui donne naissance au phénomène de coarticulation.

D'un autre côté, sur la base de notre perception de la parole, nous pouvons affirmer que ce signal se compose d'une série d'éléments sonores distincts. En effet, l'examen du spectrogramme d'un signal de parole permet de distinguer des zones spectralement homogènes .Ce fait révèle, à un certain degré, la nature segmentale de la parole. Le paradoxe entre la perception des segments de parole et la variabilité acoustique de cette dernière démontre que la segmentation est un problème fondamentalement complexe. Même si les frontières entre certains phones semblent relativement claires, il n'y a pas de transitions franches entre beaucoup de phones [31].

III.5.3. Segmentation en syllabe

La syllabe est une unité structurante de la langue. Généralement, la structure d'une syllabe se décompose souvent en 3 parties : l'attaque (une ou plusieurs consonnes -facultatif), le noyau (une voyelle ou une diphtongue - obligatoire) et la coda (une ou plusieurs consonnes - facultatif). A cause de la caractéristique facultative des consonnes surs l'attaque et sur la coda, il y a parfois des ambiguïtés de segmentation d'une phrase en syllabes [32].

V. Berment, dans le cadre de sa thèse, a construit un outil nommé «Sylla » permettant de mettre au point rapidement des « modèles syllabiques » pour une langue peu dotée. Il a appliqué cet outil pour construire des modèles grammaticaux des syllabes des langues d'Asie du Sud-est : laotien, birman, thaï et khmer. L'outil et la méthode de construction d'un modèle syllabique permet de créer rapidement un « reconnaisseur syllabique » : pour une chaîne de caractères en entrée, le reconnaisseur teste si la chaîne peut constituer une syllabe dans la langue considérée [32].

Pour la segmentation en syllabes, un segmenteur syllabique sera construit en employant un algorithme de programmation dynamique, à l'aide d'un modèle syllabique, qui segmente une phrase de texte en optimisant le critère de « plus longue chaîne d'abord »(LongestMatching), ou le critère de « plus petit nombre de syllabes » (Maximal Matching) [32].

III.5.4. Segmentation en mots

La segmentation d'un message parlé en ses constituants élémentaires est un sujet difficile. Pour l'éviter, de nombreux projets de la RAP se sont intéressés à la reconnaissance de mots prononcés isolement. La reconnaissance des mots isolés ou tous les mots prononcés sont supposés être séparés par des silences de durée supérieure à quelques dixièmes de secondes, se fait essentiellement par l'approche globale [33].

III.5.5. Segmentation en locuteurs et tour de parole

La segmentation selon le locuteur est née relativement récemment pour répondre au besoin créé par le nombre toujours croissant de documents multimédia devant être archivés et accédés. Les tours de parole et l'identité des locuteurs constituent une intéressante clé d'accèsà ces documents. Le but de la segmentation selon le locuteur est donc de segmenter en tour de parole (un tour de parole est un segment contenant une intervention d'un locuteur) un document audio contenant N locuteurs et d'associer chaque tour de parole au locuteur

l'ayantprononcé. En général, aucune information apriori n'est disponible, sur le nombre de locuteur sous leurs identités [34].

La segmentation en macro classes acoustiques est nécessaire pour supprimer les parties du document ne contenant pas de parole (comme la musique, les silences…) ou pour réaliser des traitements spécifiques à des conditions acoustiques données (genre des locuteurs, parole téléphonique, parole au-dessus de la musique…). Le processus de segmentation acoustique proposé en trois niveaux: parole/non parole, parole propre/parole avec musique/parole téléphonique et homme/femme. La classification est réalisée suivant un procédé hiérarchique en trois étapes [34]:

- Le premier niveau de segmentation correspond à une séparation "parole/non parole". Le procédé est basé sur une modélisation statistique des deux classes. Il consiste en une discrimination trame à trame suivie d'un ensemble de règles morphologiques. Ces dernières permettent de définir des contraintes sur les segments, comme leur durée minimale;
- La deuxième étape de segmentation consiste à répartir les zones étiquetées "parole" en trois classes : "parole propre", "parole et musique" et "parole téléphonique". Cette étape repose sur un décodage de type Viterbi associé à un HMM ergodique.
- La dernière étape est dédiée à la séparation "homme/femme". Un procédé de même type que pour l'étape précédente est employée, avec des états dépendant de la classe acoustique et du genre (une classe "parole dégradée" est ajoutée, pour augmenter la robustesse du procédé).

III.6. Fenêtrage

Si nous définissons w(n) comme fenêtre où 0 < n < N -1 et N représente le nombre d'échantillons dans chacune des trames, alors le résultat du fenêtrage est le signalx_a, donné par la formule (III.5).

$$X_a = x(n)\, w(n), \qquad 0 < n < N - 1 \qquad \text{(III.4)}$$

Les fenêtres les plus utilisées sont :

- **Fenêtre de Hamming**:

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) & 0 \le n \le N - 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{(III.5)}$$

- **Fenêtre rectangulaire** :

$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & \text{sinon} \end{cases} \quad \text{(III.6)}$$

- **Fenêtre triangulaire** :

$$w(n) = \begin{cases} \frac{2n}{N-1} & \text{si } 0 \le n \le \frac{N-1}{2} \\ \frac{2(N-n-1)}{N-1} & \text{si } \frac{N-1}{2} < n \le N-1 \\ 0 & \text{sinon} \end{cases} \quad \text{(III.7)}$$

- **Fenêtre de Hanning** :

$$w(n) = \begin{cases} 0.5 - 0.5\cos\frac{2\pi n}{N-1} & \text{si } 0 \le n \le N-1 \\ 0 & \text{sinon} \end{cases} \quad \text{(III.8)}$$

- **Fenêtre de Blackman** :

$$w(n) = \begin{cases} 0.42 - 0.5\cos\frac{2\pi n}{N-1} + 0.08\cos\frac{4\pi n}{N-1} & \text{si } 0 \le n \le N-1 \\ 0 & \text{sinon} \end{cases} \quad \text{(III.9)}$$

La figure III.8 illustre la forme que prennent les fonctions définies ci-dessus :

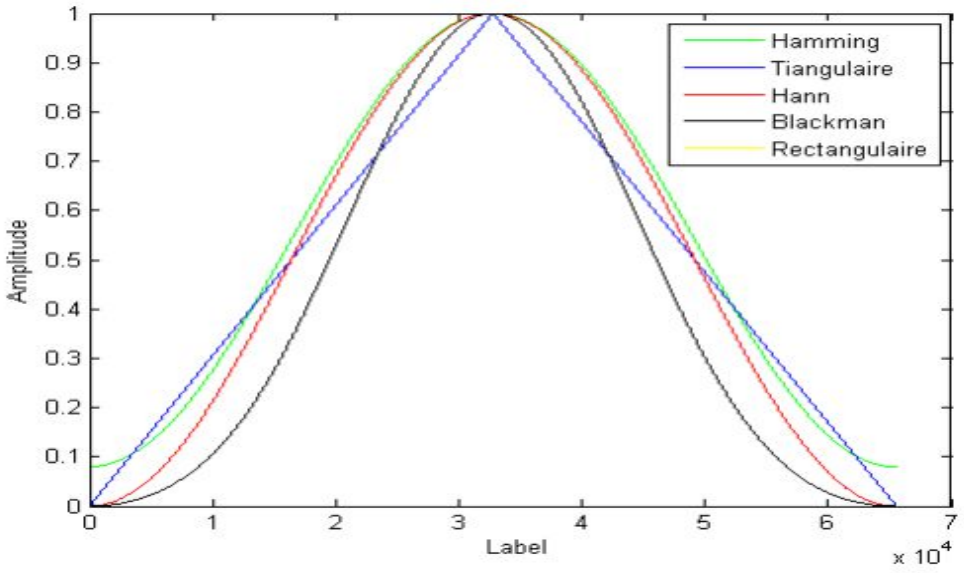


Figure III.8: Les fonctions de fenêtrage [2].

III.7. Paramétrisation via les MFCC

Dans le cadre des applications de reconnaissance de la parole, seule l'estimation de l'enveloppe spectrale est nécessaire. L'extraction de coefficient MFCC est basée sur l'analyse par banc de filtre qui consiste à filtrer le signal par un ensemble de filtres passe-bande [2].

L'énergie en sortie de chaque filtre est attribuée à sa fréquence centrale.

Pour simuler le fonctionnement du système auditif humain, les fréquences centrales sont réparties uniformément en une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large [2].

Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole.

Les échelles perceptives les plus utilisées sont l'échelle Mel ou l'échelle Bark. Du point de vue performance des systèmes de reconnaissance de la parole, ces deux échelles sont quasiment identiques. La relation utilisée pour concrétiser ceci est : [2]

$$M(f) = B \log_{10}(1 + \frac{f}{c}) \qquad \text{(III.10)}$$

Avec : B=2595, C=700, et f : la fréquence.

Mel	500	600	700	800	900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
Fré	510	630	770	920	1080	1270	1480	1720	2000	2320	2700	3150	3700	4400	5300	7700

Tableau III.1 : Transformation des échelles [2].

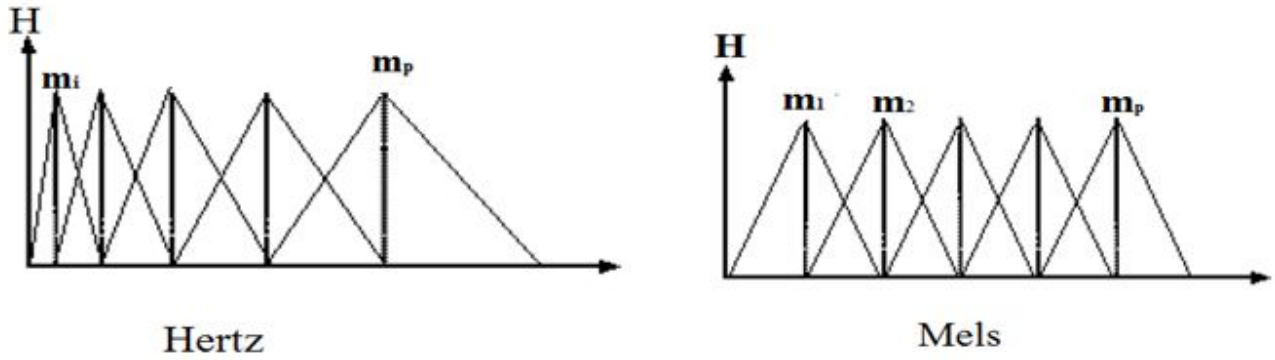


Figure III.9 : Répartition des filtres triangulaires sur les échelles [2].

Lorsque ce band de filtre est en place, il est possible de calculer les coefficients MFCCs. Le nombre de filtre utilisé dans une telle analyse est choisi de manière empirique : zwiker propose 24 filtres. De la même manière, on choisit empiriquement le type des filtres [2].

Après la mise en forme du signal de parole (commune à la plupart des méthodes d'analyse), une transformée de Fourier Discrète (DFT : Discret Fourier Transforme), en particulier FFT (Transformée de Fourier Rapide : Faste Fourier Transforme), est appliquée pour passer dans le domaine fréquentiel et pour extraire le spectre du signal [2].

Ensuite le filtrage est effectué en multipliant le spectre obtenu par les gabarits des filtres. Ces filtres sont en général triangulaires.

Le traitement décrit ci-dessus permet d'obtenir une estimation de l'enveloppe spectrale (densité spectrale lissée). Il est possible d'utiliser les sorties du banc de filtres comme entrée pour le système de reconnaissance.

Cependant, d'autres coefficients dérivés des sorties d'un banc de filtres, sont plus discriminants, plus robustes au bruit ambiant et moins corrélés entre eux. Il s'agit des coefficients cepstraux dérivés des sorties du banc de filtres répartis linéairement sur échelle Mel, ce sont les coefficients « MFFC ».

Le Cepstre est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale.

Ceci à une interprétation du point de vue de la déconvolution homomorphique : alors que le filtrage linéaire permet de séparer des composantes combinées linéairement, dans le cas de composantes combinées de façon non linéaire (multiplication ou convolution), les méthodes homomorphique permettent de se ramener au cas linéaire [2].

L'algorithme peut être décrit comme suit :

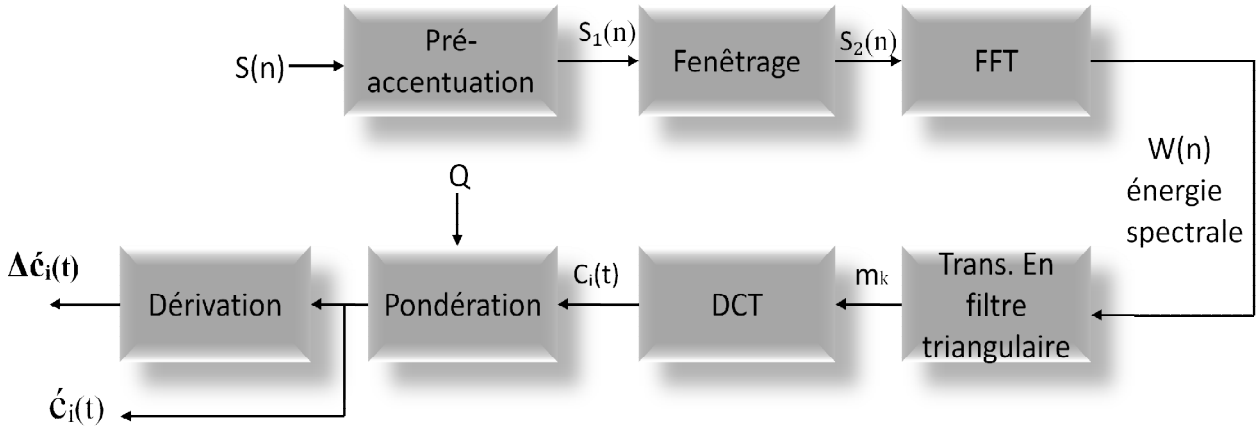


Figure III.10: Processus d'extraction des MFCCs [30].

III.7.1. Calcul de la transformée de Fourier (FFT)

Au cours de cette étape chacune des trames, de N valeurs, est convertie du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapide pour le calcul de la transformée de Fourier discret (DFT) et est définie par la formule. Les valeurs obtenues sont appelées le spectre [2].

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{2j\pi}{N}kn} \qquad \text{(III.11)}$$

En général, les valeurs X[k] sont des nombres complexes et nous nous utilisons que les valeurs absolues (énergie de la fréquence) [2].

III.7.2. Transformation en échelle Mels et Filtres triangulaires

C'est une chaîne des nombres dont la longueur dépend du nombre de FFT. Dans cette étape, l'énergie spectrale sera calculée [2] :

$$Wn = |S(fn)|^2 \qquad \text{(III.12)}$$

La chaîne des coefficients m_k (k = 1,2, …, K) de K filtres obtenue par la somme accumulée après avoir transformé $w_n$en échelle Mels et après avoir passé à travers de chaque filtre de bande.

L'ensemble des filtres de bande comprend usuellement plus de 20 filtres triangulaires.

III.7.3. Transformation en Cosinus Discret: DCT (Discret Cosinus Transforme)

Ensuite, les valeurs logarithmiques des $m_k$seront transformées en domaine temporel en utilisant la transformation en Cosinus Discret :

$$C_i = \sqrt{2/k}\sum_{j=1}^{K} \ln(\text{(mj)}\cos(\frac{\pi i}{k}(j-0.5)) \quad i=1, 2...K \qquad \text{(III.13)}$$

Normalement, on n'utilise que des premières valeurs des **ci**. Dans quelques systèmes de reconnaissance de la parole les 12 coefficients de MFCCs plus un coefficient normalisé d'énergie des trames sont choisies pour les caractéristiques typiques de la parole. On a donc 13 coefficients.

Nous allons vérifier si cette combinaison de 13 coefficients plus leurs variations en 1ére ordre, en 2éme sont les meilleures pour la reconnaissance [2].

III.7.4. Pondération

Comme mentionné ci-dessus, la chaîne de ces valeurs est pondérée par une fonction de fenêtrage présentée comme l'expression suivante [2] :

$$\hat{C}_i = \left[1 + \frac{Q}{2}\sin\left(\frac{\pi i}{Q}\right)\right] C_i \qquad 1 \le i \le Q \qquad \text{(III.14)}$$

III.7.5. Dérivation

Pour augmenter la qualité de la reconnaissance, on a encore implanté une relation entre les trames consécutives (relation temporelle) de signal. Ce sont les variations en **1ér** ordre et en **2éme** ordre des coefficients [2] :

$$\Delta\hat{C}_i = \frac{\sum_{\theta=1}^{\Theta} \theta(\hat{C}_{t+\theta} - \hat{C}_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta_2} \qquad \text{(III.15)}$$

Où **Θ**est la longueur de la fenêtre pour calculer Δet qui est usuellement choisi par 2 ou 3.

En général d'après ce qu'on vient de voir l'extraction via MFCCs s'effectue en étapes suivantes :

- **Signal découpé en trame de N échantillon :** Le Signal échantillonné doit subir une préaccentuation puis un fenêtrage.
- On fait passer un filtre de Hemming. On obtient des trames pondérées
- Celles-ci subissent une transformé de Fourier discret DFT.
- Puis les spectres de signal obtenus par DFT sont multipliés par des filtres triangulaires B(m).
- Ces derniers doivent passer par une étape ou on fait sortir des logarithmes d'énergie appelés aussi des logarithmes de spectre d'amplitudes. **E(m)= log |. |**
- Enfin les coefficients cepstraux des MFCC, peuvent être obtenus par une transformé de cosinus discrète de E(m). Appelé C(n) ou iDCT.

Les coefficients MFCC [4] sont un type de coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole.

Le codage MFCC utilise une échelle fréquentielle non-linéaire. L'échelle Mel est définie par :

$$B(f) = 2595\log_{10}(1 + \frac{f}{700}) \qquad \text{(III.16)}$$

f est la fréquence en Hz, B(f) est la fréquence en Mel de f.

Soit un signal discret {x[n]} avec $0 \leq n \leq N-1$, N est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète S[k]est obtenue par (III.17):

$$S[k] = \sum_{n=0}^{N-1} X[n] e^{\frac{-j2\pi nK}{N}} \quad \text{avec} \quad 0 \leq K < N \qquad \text{(III.17)}$$

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

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \leq m \leq M+1 \qquad \text{(III.18)}$$

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

Les points f [m]discrets correspondants sont calculés par l'équation :

$$f[m] = \frac{N}{f_s} B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad \text{(III.19)}$$

Où B^{-1}est la transformée de Mel-fréquence en fréquence :

$$B^{-1} = 700 * \left(10^{b/2595} - 1\right) \qquad \text{(III.20)}$$

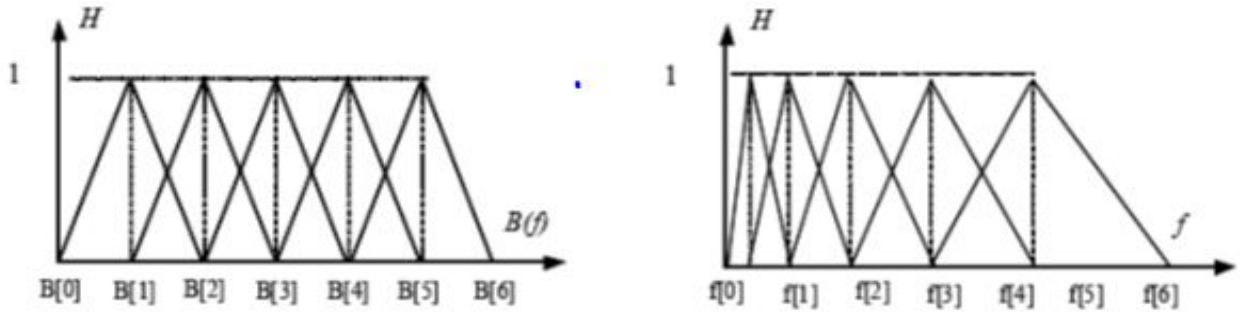


Figure III.11 : Les filtres triangulaires passe-bande en Mel-Fréquence (**B(f)**) et en fréquence **(f)) [2].**

Le coefficient $\mathbf{H_m}$ **[k]** de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & \text{si } K \le f[m-1] \\ \frac{K - f[m-1]}{f[m] - f[m-I]} & \text{si } f[m-I] \le K \le f[m] \\ \frac{f[m+I] - K}{f[m+I] - f[m]} & \text{si } f[m] \le K \le f[m+I] \\ 0 & \text{si } K \ge f[m+I] \end{cases} \qquad \text{(III.21)}$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme d'énergie sera :

$$E[m] = \log\left[\sum_{K=0}^{N-1} |S[k]|^2 H_m[k]\right] \qquad 0 \le m \le M \qquad \text{(III.22)}$$

Les coefficients cepstraux de Mel-fréquence (MFCCs) peuvent être obtenus par une transformée de cosinus discrète de E[m] :

$$C[n] = \sum_{m=0}^{M-1} \cos\left(\frac{\pi n(m+\frac{1}{2})}{M}\right) \qquad 0 \le n \le M \qquad \text{(III.23)}$$

En résumé:

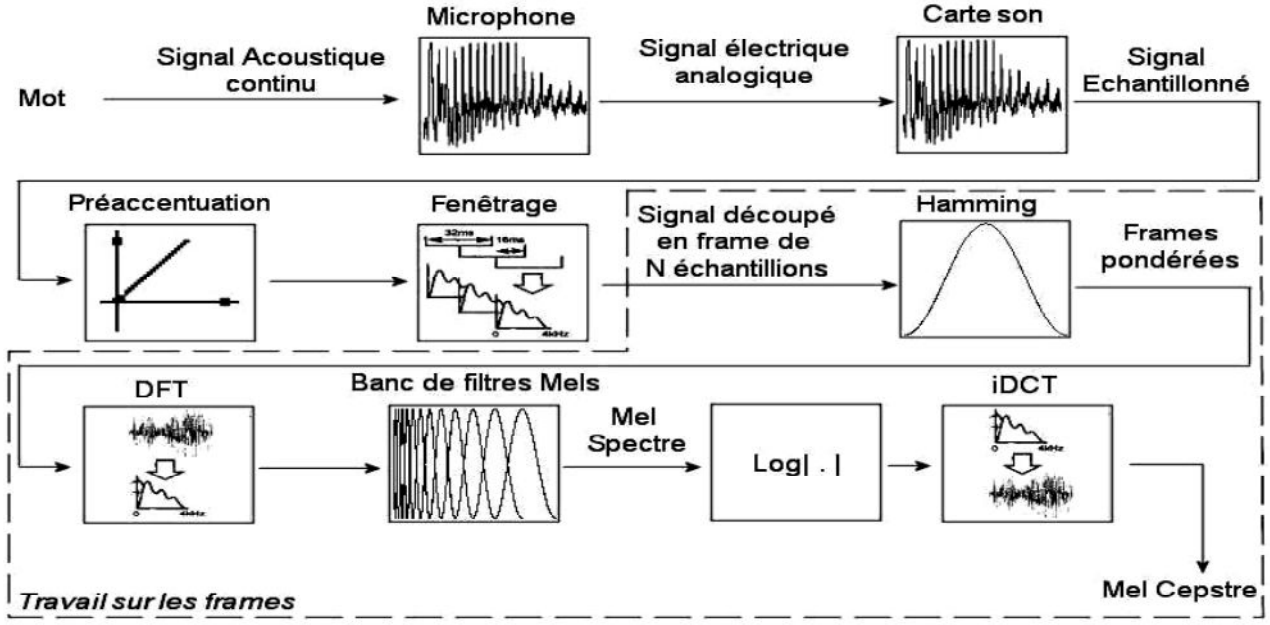


Figure III.12 : Étapes du calcul des coefficients MFCC [2].

III.8. Paramétrisation basée sur un modèle de production de la parole (LPC)

Les méthodes dites de codage prédictif linéaire LPC ont été largement utilisée pour l'analyse de la parole. Elles font référence à un modèle du système de phonation, que l'on représente en général comme un tuyau sonore à section variable. L'analyse LPC est utilisée essentiellement en codage et en synthèse de la parole. La prédiction linéaire est une technique issue l'analyse de la production de la parole permettant d'obtenir des coefficients de prédiction linéaire (linear prédiction coefficient LPC) [2].

Dans ce cadre d'analyse, le signal de parole x est considéré comme la conséquence de l'excitation du conduit vocal par un signal provenant des cordes vocales. La prédiction s'appuie sur le fait que les échantillons de parole adjacents sont fortement corrélés et que par conséquent, l'échantillon Sn peut être estimé en fonction des p échantillons précédents [2].

Par prédiction linéaire, on obtient donc une estimation du signal :

$$\hat{S}_n = \sum_{i=1}^{p} a_i \, x_{n-i} \qquad \text{(III.24)}$$

Où les $\mathbf{a_i}$sont des coefficients constants sur une fenêtre d'analyse. La définition devient exacte si on inclut un terme d'excitation :

$$x_n = \sum_{i=1}^{p} a_i \, x_{n-i} + \, Ge_n \qquad \text{(III.25)}$$

Où **e** est le signal d'excitation et **G** gain de l'excitation. La transformée en Z de cette égalité donne :

$$GE(Z) = \left(1 - \sum_{i=1}^{p} a_i Z^{-i}\right) X(Z) \qquad \text{(III.26)}$$

D'où :

$$H(Z) = \frac{X(Z)}{E(Z)} = \frac{G}{\left(1 - \sum_{i=1}^{p} a_i Z^{-i}\right)} = \frac{G}{A(Z)} \qquad \text{(III.27)}$$

Cette équation peut être interprétée comme suit :

Le signal **x** est le résultat de l'excitation du filtre tout pôle $\boldsymbol{H}\ (\boldsymbol{Z}) = \frac{G}{A(Z)}$ par le signal d'excitation **e.**

Les coefficients **ai** sont les coefficients qui minimisent l'erreur quadratique moyenne :

$$E_n = \sum_m (G - e_{n+m})^2 = \left[\sum_m \left(x_{n+m} - \sum_{i=1}^{p} a_i x_{n+m-i}\right]^2 \right. \qquad \text{(III.28)}$$

Calcul des coefficients a_i : La prédiction de $S_n$est

$$\hat{S} = \sum_{k=1}^{p} a_k \, S_{n-k} \qquad \text{(III.29)}$$

Et l'erreur est donc :

$$\varepsilon_n = S_N - \hat{S}_n = S_n - \sum_{k=1}^{p} a_k \, S_{n-k} \qquad \text{(III.30)}$$

Fonction d'énergie associée (minimisation de l'erreur au sens des moindres carrés est :

$$E = \sum_{n=1}^{N} (S_n - \hat{S}_n)^2 \qquad \text{(III.31)}$$

Où **N** est le nombre total d'échantillons. En minimisant par rapport aux coefficients $\mathbf{a_i}$ on obtient :

$$\forall i = 1 \cdots, p \frac{\partial E}{\partial a_i} = 0 \quad \text{(III.32)}$$

C'est-à-dire

$$\forall i = 1 \cdots, p \quad \sum_{n=1}^{N} S_{n-i} S_n = \sum_{n=1}^{N} \sum_{k=1}^{p} a_k \, s_{n-i} \, s_{n-k} \quad \text{(III.33)}$$

Ou encore :

$$\forall i = 1 \cdots, p \qquad \sum_{k=0}^{p} a_k \left(\sum_{n=1}^{N} s_{n-i} \, s_{n-k}\right) = 0 \quad \text{(III.34)}$$

En posant $\mathbf{a_0=1}$, et sachant que la fonction d'auto corrélation du signal **S** est :

$$\varphi_s(i) = \sum_{n=1}^{N} s_{n-i} \, s_n \qquad i = 1 \cdots, p \quad \text{(III.35)}$$

On note:

$$\phi(i,k) = \sum_{n=1}^{N} s_{n-i} = \varphi_s(i-k) \quad \text{(III.36)}$$

(Par prolongement périodique du signal échantillonné) et on obtient :

$$\forall i = 1, \ldots, p \qquad \sum_{k=1}^{p} a_k \varphi_s(i-k) = \varphi_s(i) \quad \text{(III.37)}$$

Finalement :

$$\begin{Bmatrix} \varphi s(0)\varphi s(-1) & \cdots & \varphi s(1-p) \\ \varphi s(1)\varphi s(0) & \cdots & \varphi s(2-p) \\ \vdots & \vdots & \vdots \\ \varphi s(p-1)\varphi s(p-2) & \cdots & \varphi s(0) \end{Bmatrix} \begin{Bmatrix} a1 \\ a2 \\ \vdots \\ ap \end{Bmatrix} = \begin{Bmatrix} \varphi s(1) \\ \varphi s(2) \\ \vdots \\ \varphi s(p \end{Bmatrix} \quad \text{(III.38)}$$

On obtient donc les coefficients (a_k) $1 \leq k \leq p$ par résolution du système linéaire ci-dessus. Pour la mesure de la prédiction linéaire on utilise la structure Toeplitz de la matrice de covariance ϕ de façon à résoudre ce système en O (p^2) opération [2].

R_s: est la matrice d'auto covariance du signal s, c'est une matrice de Toeplitz.

Cette équation mène au système d'équation appelée équation normales :

$$R_s a = [\sigma^2, 0, \ldots \ldots \ldots]^T \quad \text{(III.39)}$$

Ces équations sont résolues par les algorithmes de LEVINSON-DURBIN ou de SCHUR.L'ordre P est un élément important .Atal et al proposent un ordre P = 12 pour modéliser les caractéristiques du conduit vocal.

Le codage LPC est très largement utilisé en traitement de la parole, notamment en transmission. Ce modèle est le plus facile et rapide à mettre en œuvre dans les systèmes temps-réel. Cependant, il possède certaines limitations [2] :

- Ces modèles linéaires qui ne tiennent pas compte des phénomènes non-linéaires présents lors de la production de la parole.

- Il présente également une grande variance pour les différentes classes en reconnaissances.
- Il est optimisé pour la tâche de reconnaissance.

Les étapes d'analyse spectrale du module de paramétrisation du signal sont :

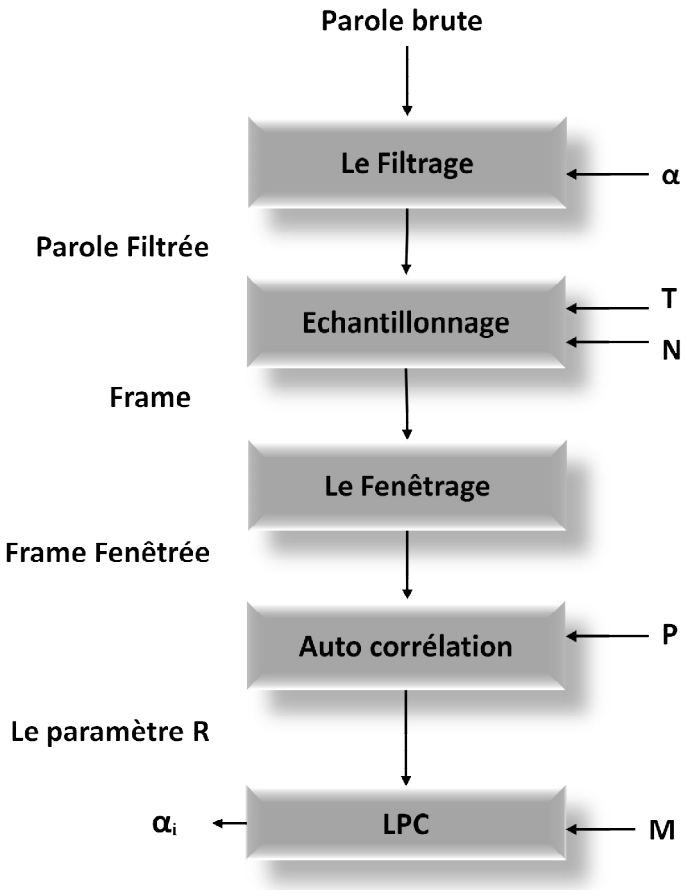


Figure III.13: Étapes de calcul des coefficients des LPC (a_i) [2].

III.9. Paramétrisation via NPC

Le codage neuro-prédictif est une extension du codage LPC, donc une méthode de codage temporelle mais contrairement à la méthode LPC, le codeur NPC extrait les caractéristiques non linéaires d'un phonème. Il est basé sur un MLP à une couche cachée suivi d'une couche de sortie à 1 neurone appelé cellule de prédiction. L'étape d'apprentissage consiste à prédire un échantillon (extrait du signal acoustique d'un phonème) à partir des n échantillons précédents grâce à un MLP.

III.10. Modélisation selon LVQ

Le classifieur à prototypes LVQ (Linear Vector Quantization) a été appliqué avec succès dans différents domaines comme la reconnaissance de l'écriture ou bien la reconnaissance de la parole. La procédure d'apprentissage consiste en l'ajustement des prototypes (représentants) afin de décrire les frontières optimales des classes.

L'algorithme LVQ qui est parfois considéré comme une variante des «cartes topologiques», est un algorithme de classification supervisée dont l'efficacité est remarquable.

III.11. Conclusion

Dans ce chapitre nous avons parlé dans un premier lieu des approches et techniques d'extraction de caractéristiques d'un signal ensuite nous avons présenté les différentes étapes de segmentation et les techniques d'extraction de caractéristique du signal de parole, le résultat de cette opération est utilisé par les méthodes de la phase suivante.

Chapitre IV

Conception et mise en œuvre

IV.1. Introduction

L'objectif de ce projet est le « Décodage acoustico phonétique du signal de parole », notre contribution s'effectuera essentiellement pour la langue arabe et par conséquent cette langue connait une série de consonnes complexes dites 'emphatiques' et certaines plosives difficiles à séparer. Leur reconnaissance consiste en l'occurrence un réel défie pour tout systèmes de Décodage Acoustico Phonétique (DAP).

Et comme tous ceux qui connaissent un peu les mécanismes internes de traitement de l'information linguistique par les ordinateurs savent que, pour ces machines, tous les systèmes d'écriture se ramènent finalement à des codes numériques et qu'il n'y a donc strictement aucune différence de ce point de vue entre langue arabe et n'importe quelle autre langue, notamment alphabétique.

Alors le problème n'est donc pas dans les machines, mais réside dans les hommes, ou plus précisément dans les exigences du dialogue homme-machines. Pour cela, nous avons décidé de construire une petite base de données pour la langue arabe. Notre système reçoit en entrée le signal de parole et renvoie comme résultat un ensemble phonétique. Autour des modules de reconnaissance, nous avons développé des procédures pour l'analyse phonétique et l'affichage graphique ainsi que des modules d'évaluations des performances du système. Mais cela n'était pas une chose facile, nous nous sommes heurtés à de nombreux problèmes :

- Le problème de prononciation de certains phonèmes par les locuteurs ;
- Le bruit qui résulte de la mauvaise qualité du matérielles et de l'environnement ;
- L'étiquetage nécessite des connaissances en phonétique.

Nous avons aussi intégrer un système d'apprentissage et de décodage acoustico-phonétique basé sur un corpus composé de fichiers de format wav qui doivent être soigneusement étiquetés. Nous avons eu le libre choix des méthodes afin de résoudre le problème du décodage nous avons utilisé les coefficients MFCC et LPC qui on très réputé par les bon résultats qui ont permet d'obtenir en DAP et leurs robustesse aux bruits.

La base de données est constituée d'enregistrement de prononciations de phonèmes et de mots produits par 18 locuteurs algériens, ayant prononcé isolément en contexte

consonantique emphatique ou non emphatique. L'enregistrement a été effectué en utilisant le logiciel d'analyse Praat chaque locuteur prononce les 14 consonnes arabes avec harak't el fatha, damma et kasra sous forme de fichiers sous format (wav, 16 bits, 16kHz, mono). Les logiciels utilisés: Praat pour calculer les formants, la durée temporelle, pitch, intensité, pulses et le spectrogramme et pour l'enregistrement sous format wav. On a aussi la possibilité d'adjoindre notre logiciel qui est d'abord un analyseur de parole entièrement développé sous un environnement C++ qui donne en présence les possibilités suivantes :

1) Affichage du contenu de la base de données.
2) Extraction des paramètres en utilisant les méthodes LPC, MFCC.
3) Taux de reconnaissance des voyelles et des consonnes.

IV.2. Interface du système

Notre système est démarré par l'interface suivante qui schématisé dans la figure ci-dessous.

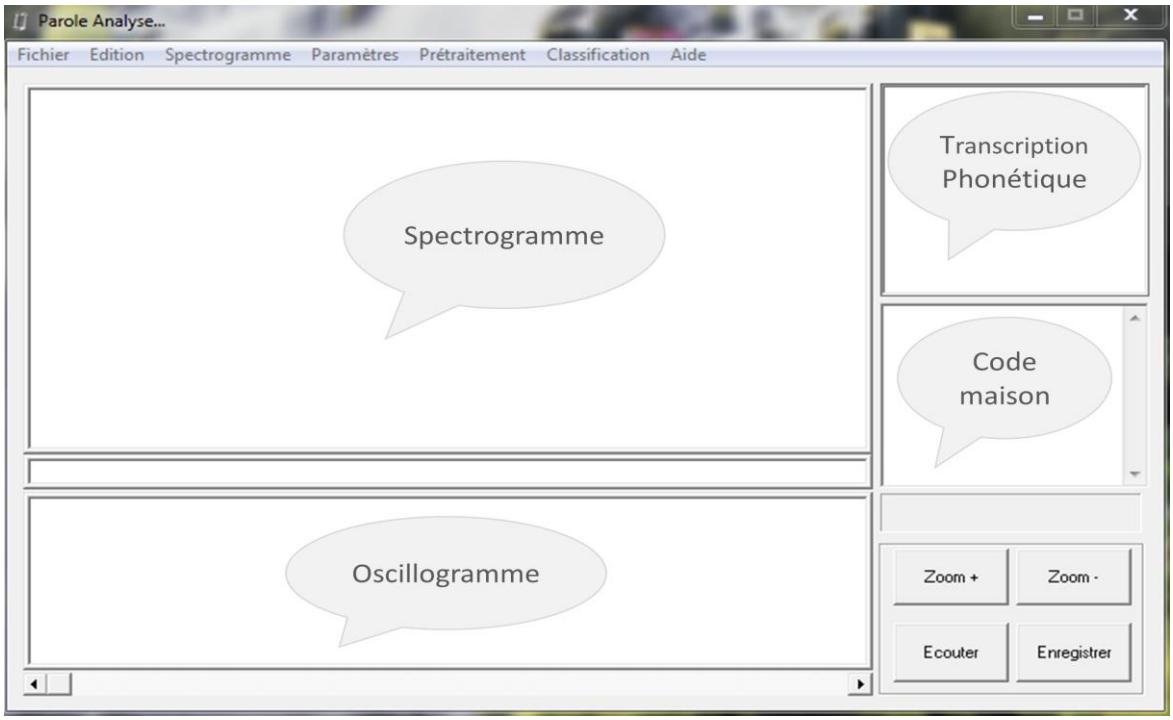


Figure IV.1 : L'interface de démarrage de notre système.

- Un oscillogramme qui permet une visualisation temporelle du signal de parole.
- Un spectrogramme permettant une visualisation fréquentielle de la parole.
- Deux zones de textes qui affichent une transcription phonétique du signal de parole une fois le mécanisme de reconnaissance enclenché, l'une en caractères arabes et l'autre en code maison en caractères occidentaux.
- Un menu rempli de fonctionnalités.

IV.3. Extractions des caractéristiques

IV.3.1. Acquisition du signal

Acquérir de la parole, avec un temps limité à 4 secondes. Le signal est stocké sur fichier disque contenant les échantillons de taille 16 bits et la fréquence d'échantillonnage fixé à 16KHz.

IV.3.2 Module acoustique

Ce module se charge d'extraire les paramètres acoustiques à partir du signal temporel, il permet en particulier de :

- Calculer et afficher un spectrogramme à bande large avec une fenêtre de Hamming.
- Spectrogramme à bande étroite pour séparer les harmoniques de la fréquence fondamentale.

Spectrogramme à bande étroite

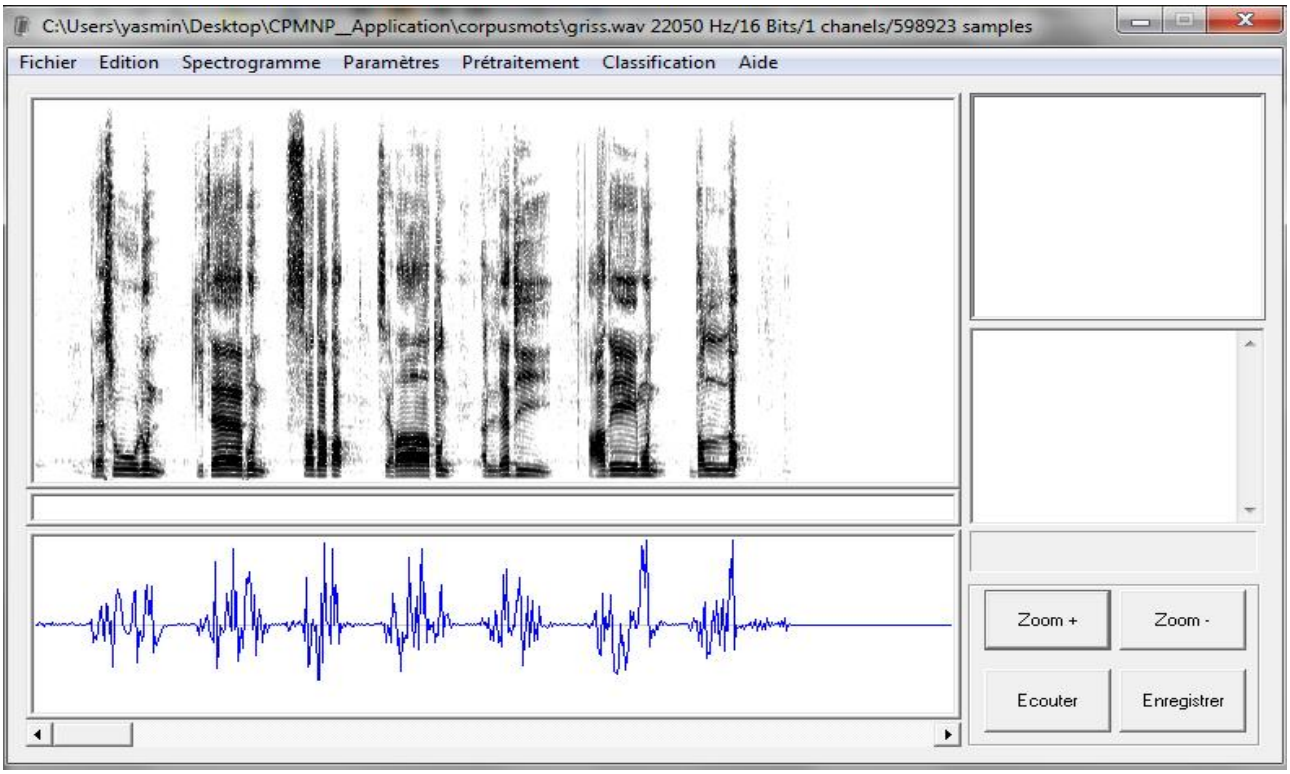


Figure IV.2 : Bande étroite.

Spectrogramme à bande large

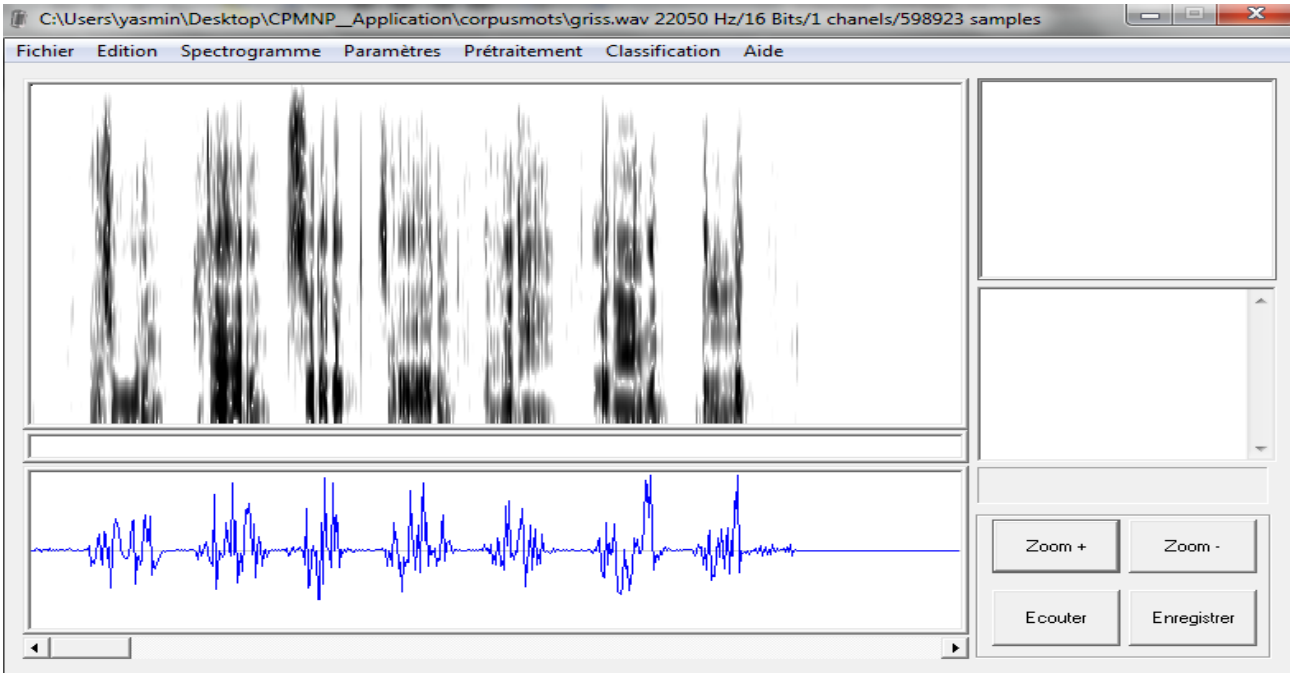


Figure IV.3 : bande large.

- **Spectrogramme lissé cepstralement**

Les algorithmes utilisent le processus vectoriel, ce qui permet d'obtenir un temps de calcul d'environ 1 seconde pour un spectrogramme de 4 secondes de parole.

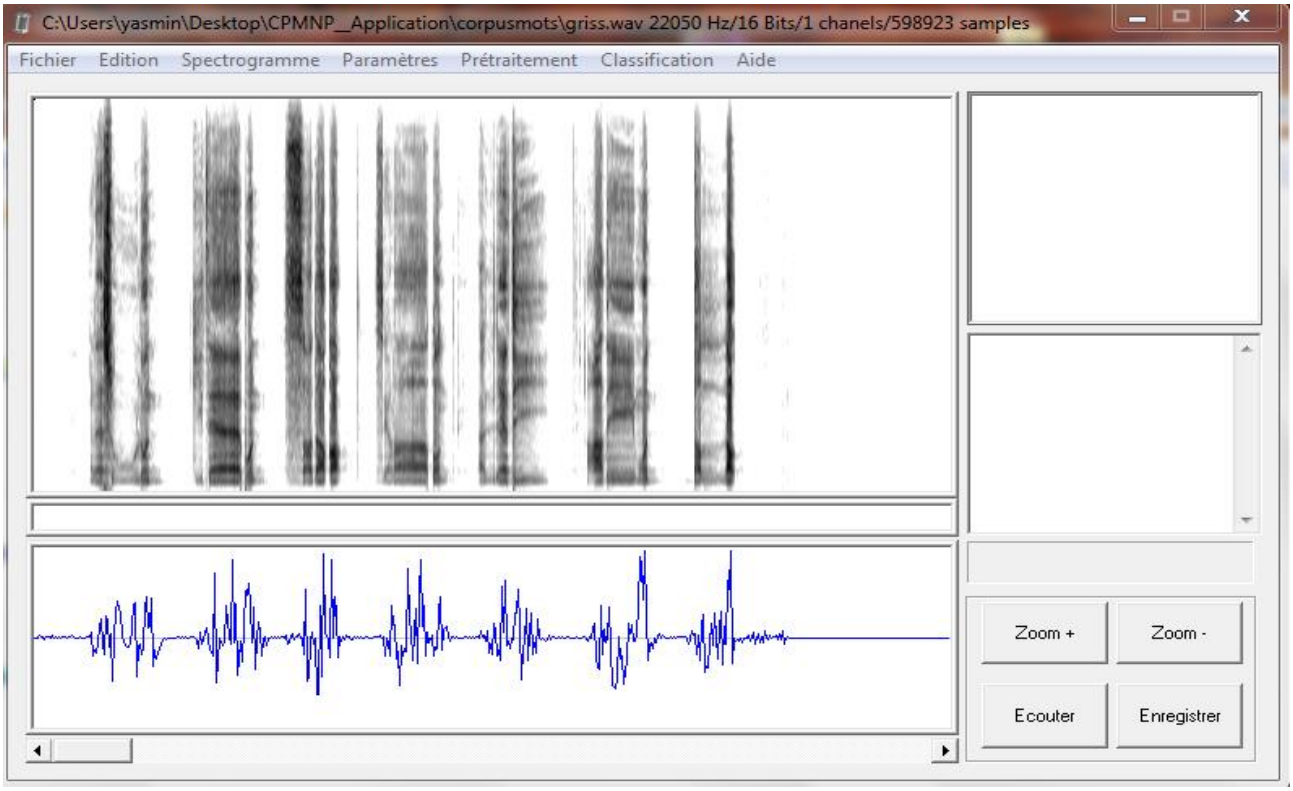


Figure IV.4 : Lissage cepstrale du spectrogramme.

- Calculer la fréquence fondamentale ou pitch (F_0), qui correspond aux vibrations des cordes vocales. La méthode utilisée est celle de l'auto corrélation.
- Calculer l'énergie, les formants.

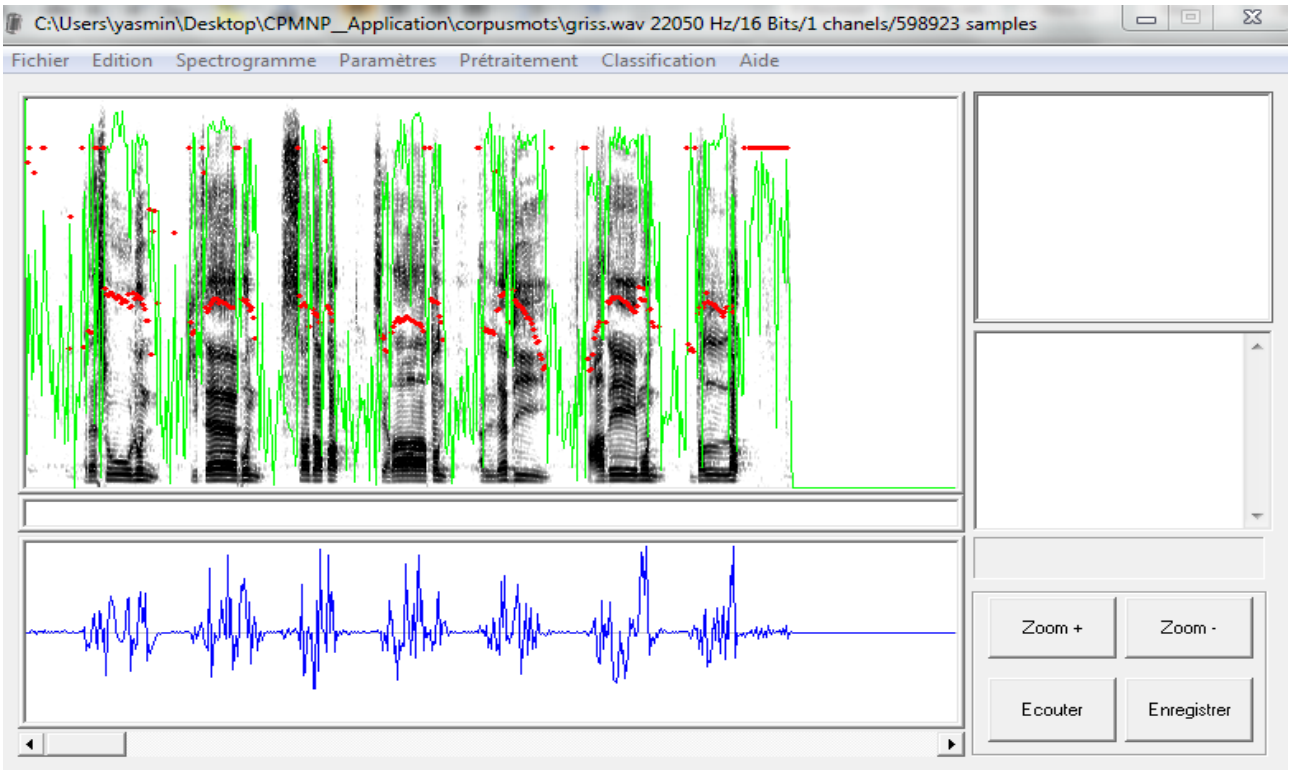


Figure IV.5 : pitch (F0 en rouge et le taux de voisement en vert).

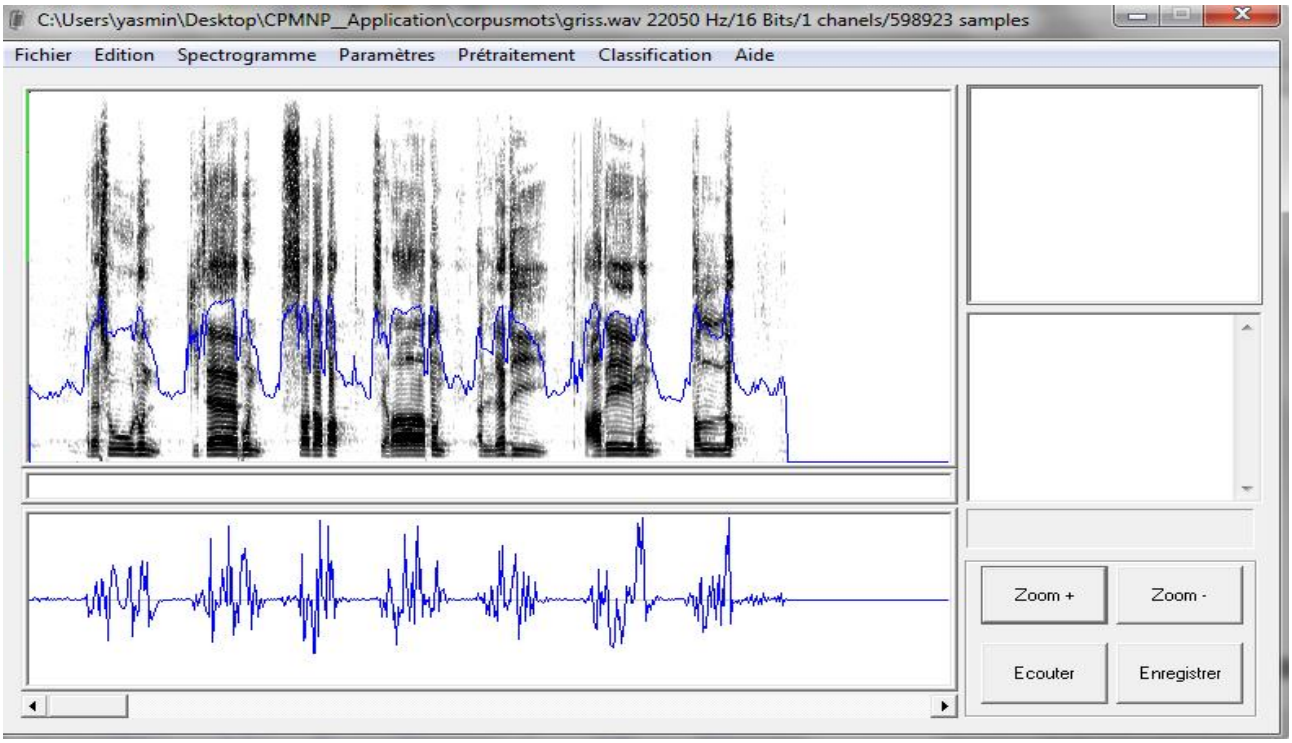


Figure IV.6 : l'énergie.

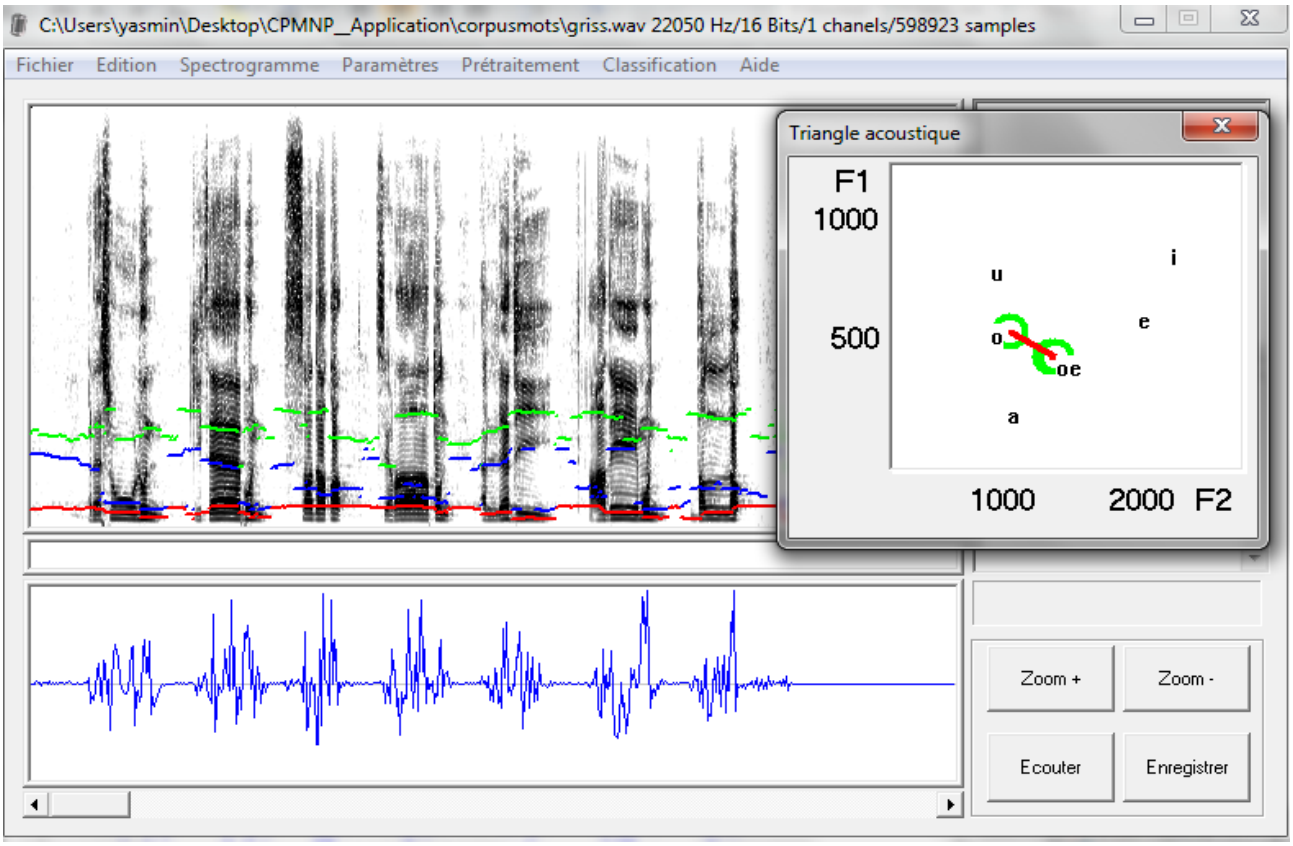


Figure IV .7 : les formantsF_1(rouge) ,F_2 (bleu), F_3(vert) et leur transition.

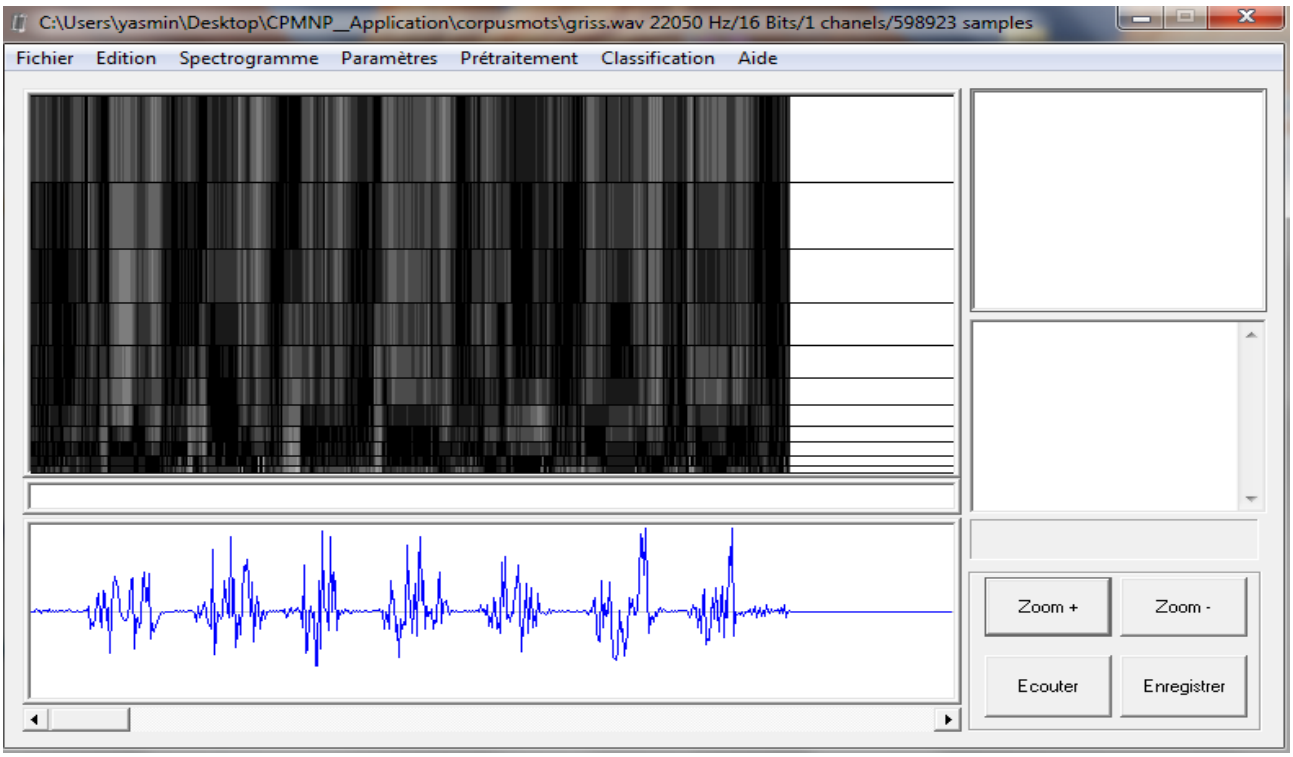


Figure IV.8 : Analyse du spectrogramme par bandes de fréquences linéaires ou sur échelle de Mel.

IV.3.3. Le module de décodage

- **La segmentation**

Il consiste à segmenter le signal de parole en grande classes phonétiques en utilisant des algorithmes non contextuels et reposant sur des critères simples. Nous avons retenu deux grandes méthodes : le temps, classe (voyelles, consonnes : fricative, plosives).

- **Le calcule d'indice**

L'extraction des indices phonétiques pertinents est une étape très importante dans le processus de décodage phonétique. Nous avons développé une procédure pour chaque indice phonétique, ces indices sont :

- La durée d'un segment ;
- Le degré de voisement ;
- Les valeurs des formants ;
- Les transitions formantiques ;
- Le centre de gravité énergétique.

IV.3.4. Le module d'étiquetage

C'est à ce niveau que se fait le décodage proprement dit. En partant des segments fournis par le module de segmentation, le module tente de trouver les bons phonèmes prononcés en utilisant les indices extraits lors de l'étape précédente.

Une étiquette manuelle permet d'affecter des étiquettes phonétiques à des segments de parole à partir de la représentation spectrographique de la phrase. L'étiquetage est possible que sur une interface graphique et il permet de :

- Insérer une étiquette phonétique
- Effacer une étiquette
- Changer l'étiquette d'un segment
- Déplacer la limite d'un segment
- Calculer et afficher un spectrogramme
- Ecouter un morceau de signal

Le résultat de l'étiquetage manuel est sauvegardé dans un fichier qu'il est ensuite possible de lire pour le consulter ou le modifier. Il servira en particulier à évaluer les performances du système tant au niveau segmentation qu'au niveau reconnaissance.

Notre base de données se divise en deux parties :

- Une première partie « base de voyelles » où on a demandé à une vingtaine de locuteurs prononcer isolément les six voyelles établis précédemment oe, u, i, a, o, é.
- Une deuxième partie « base de mots » où on a demandé à une dizaine de locuteurs de prononcer 7 mots :

رسول, كتاب, صدقة, صراط, الذي, عظيم, ظل

Remarque : l'acquisition s'est faite sur une vingtaine de locuteurs mais beaucoup d'échantillons ont été rejetés à cause des problèmes cités précédemment.

La parole est un phénomène extrêmement variable, il faut beaucoup d'expérience pour pouvoir se rapprocher de l'étiquetage exact d'un échantillon de parole.

Comme nous l'avons déjà dit, le passage d'un phonème vers un autre n'est pas direct, mais passe par une phase de transition appelé momentum où le premier phonème commence peu à peu à perdre ses caractéristiques à cause des mouvements articulatoires qui commencent à virer vers la position cible du deuxième phonème.

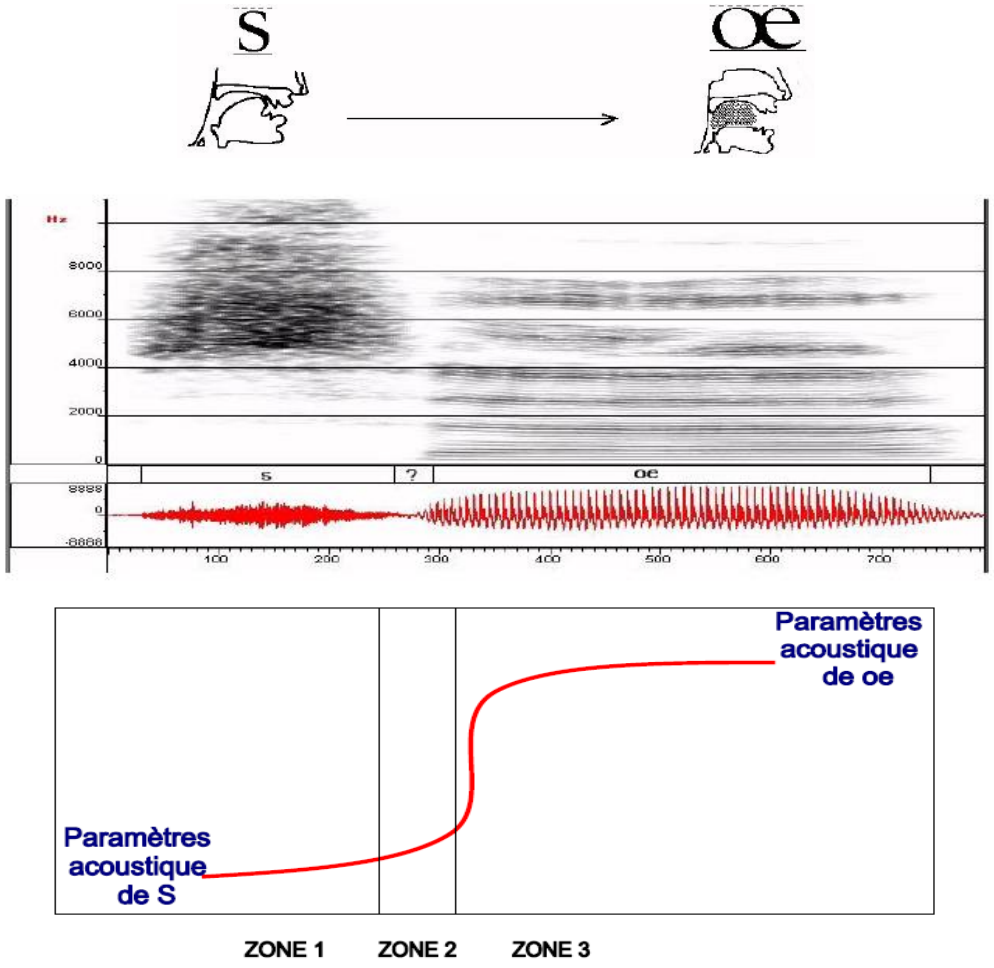


Figure IV.9 : Le changement des paramètres acoustiques de la transition phonétique /s/->/œ/ en fonction d'un changement linéaire des paramètres articulatoires.

Bien qu'il est difficile de déterminer un début et une fin de phonème, des phonèmes comme la fricative /s/ et la voyelle /oe/ montrent une certaine stabilité des caractéristiques acoustiques au centre du phonème, zone où les positions articulatoires cibles ont été atteintes, mais ce n'est pas toujours le cas, le meilleur exemple est les plosives, la stabilité que l'on remarque chez les voyelles et les fricatives sont due à la durée de ces phonèmes qui sont vraiment plus importante que les plosives, ainsi un locuteur peut rallonger cette durée à sa guise, mais pour les plosives le mode d'élocution continu est impossible, leur production passe par des étapes essentielles à leurs perception, et pour la plupart la perception ne se fait qu'à la fin de la plosive et au début de la voyelle qui la suit, contrairement aux fricatives et aux voyelles où le phonème est perçu rien qu'en entendant les premières vingtaines de millisecondes du phonème donc ils sont perçus tout au long de leurs productions.

IV.3.5. Module du prétraitement

Afin d'utiliser les fichiers wav. Qui vont nous servir comme base d'apprentissage où base de test, on va d'abord leurs faires subir une série de prétraitements qui vise à préparer les entrées de notre réseau MLP. Après le chargement d'un fichier (22 kHz, 16bits, mono).

Ce signal est fenêtré par une fenêtre glissante de 10ms de type Hamming et qui avance avec un pas de 5ms, jusque là, on a découpé le signal en fenêtres de 10ms, pour chaque fenêtre on calcule les coefficients (LPC/ MFCC) sous forme de vecteurs acoustiques.

On consulte ensuite la liste des étiquettes, si notre fenêtre appartient à un bloc étiqueté, on ajoute aux vecteurs acoustiques, une information sur l'identifiant du phonème (étiquette) sinon inconnu. On va ensuite insérer ce vecteur dans la liste des vecteurs destinés à l'apprentissage où la liste des vecteurs destinés au test du réseau MLP.

Ces vecteurs seront considérer comme vecteur code pour le LVQ en utilisant un classifieur basé sur l'algorithme des plus proches voisins.

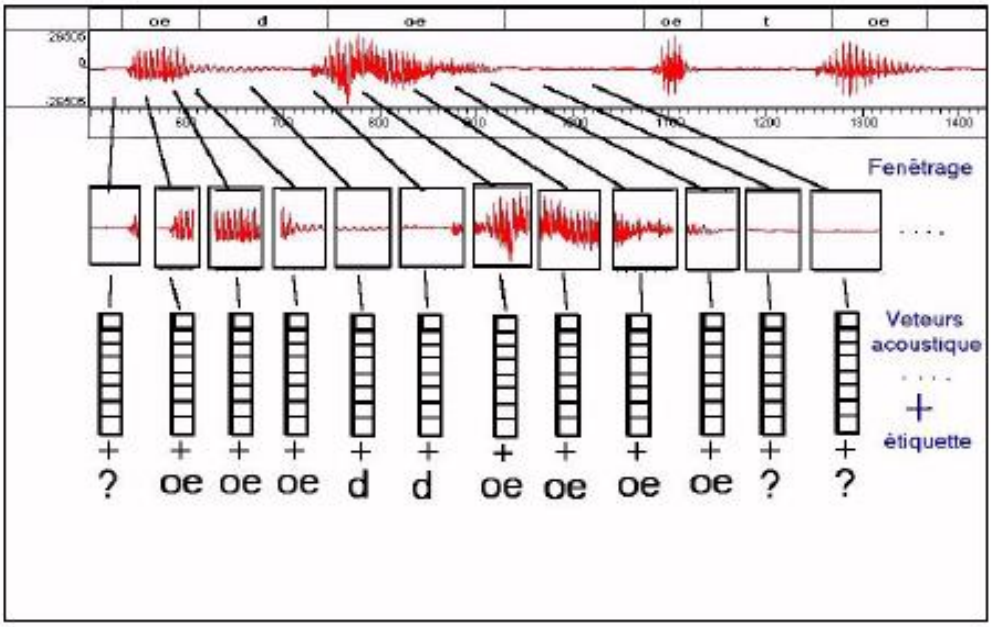


Figure IV.10 : L'extraction des vecteurs acoustiques à partir d'un signal avec étiquettes.

IV.4. Mise en œuvre de l'application

IV.4.1. Apprentissage

On va présenter les vecteurs acoustiques à notre réseau selon leur ordre dans la liste.

Un vecteur acoustique n'ayant pas d'étiquette appartenant à un phonème.

Ces entrées seront intégrées dans l'algorithme de la rétropropagation du gradient.

Ainsi le fichier en entier est présenté au réseau et on fait de même avec les autres fichiers.

IV.4.2. Adaptation de l'étiquetage

A. Base des voyelles :

Les voyelles sont des phonèmes assez stables acoustiquement surtout dans leur partie centrale, donc leur étiquetage n'était pas difficile. Pour cette base, on a choisi d'étiqueter le signal en segments de durée assez importante qui couvre presque toute la partie centrale de la voyelle, donc tous les vecteurs acoustiques se trouvant à l'intérieur de ces segments mettront le réseau en mode apprentissage.

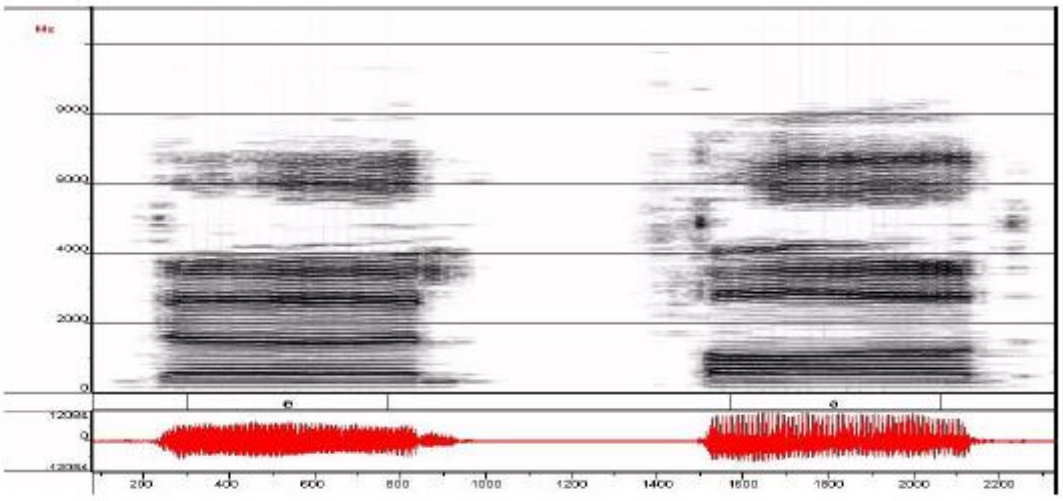


Figure IV.11 : Un exemple d'étiquetage pour les voyelles /e/, /a/.

B. Base de mots:

La base de mots qui comporte des plosives qui sont des phonèmes acoustiquement non stables. Pour l'étiquetage de cette base, nous avons commencé par les voyelles et les fricatives qui présentent plus de stabilité, nous avons choisi des segments dans leur partie centrale comme pour la base des voyelles, mais cette fois de plus petite durée afin d'accélérer l'apprentissage en minimisant le nombre des vecteurs destinés à l'apprentissage, et surtout afin de nous assurer que le réseau arrive bien à reconnaître les autres segments non étiquetés (non appris) donc arrive à généraliser.

Pour les plosives nous avons essayé de régler le moment de l'apprentissage selon notre perception, nous avons remarqué que la perception d'une plosive est plus puissante juste après la fin du burst, au début de la voyelle qui le suit, donc nous avons décidé de mettre l'étiquette à cet endroit, zone qui est non stable car les caractéristiques acoustiques de la plosive et de la voyelle qui la suit se mélangent dans cette zone, mais dans cette position le réseau aura sans doute récolter (mémoriser) assez d'information sur la plosive pour la reconnaître.

IV.4.3. Généralisation

La détection d'un phonème passe par des étapes, après avoir chargé le fichier wav et extrait les vecteurs acoustiques, ils sont propagés dans le réseau. On va se débrouiller pour ignorer les vecteurs à énergie faible en définissant un seuil minimal d'énergie (exemple 20dB), la prise de décision se fait à chaque entrée, donc on aura autant de sorties que d'entrées.

On représente les segments ignorés par un /_ /, ainsi on remarque que la plupart des vecteurs ont été bien affectés mais il existe des parasites dus au bruit, à la coarticulation…, donc une étape de lissage est nécessaire pour obtenir le résultat.

Ensuite, on prend soin de regrouper les segments similaires en notant leurs durées, pour des segments de durée inférieur à un seuil (exemple 50 ms), ce segment est ignoré (supposé bruit), pour des segments dépassant un seuil (exemple 150 ms) on rajoutera la marque / : / ce qui signifie pour une voyelle qu'elle est longue, /oe:/, /u:/, /i:/, pour une consonne, cela signifie plutôt qu'elle est renforcée.

Notons qu'un mot peut avoir plusieurs transcriptions phonétiques. Il est nécessaire de prendre en compte toutes les variantes de prononciations possibles de ce mot afin de bien le reconnaître.

IV.4.4. Test et résultats en reconnaissance phonétique:

Afin de tester les performances de notre système, nous allons l'appliquer sur notre base de données. Pour cela, nous allons extraire une suite de vecteurs de 16 coefficients (MFCC/ NPC) de nos fichiers wav, et lancer le mécanisme (apprentissage/ test) comme évoqué précédemment.

Base des voyelles avec un réseau MLP

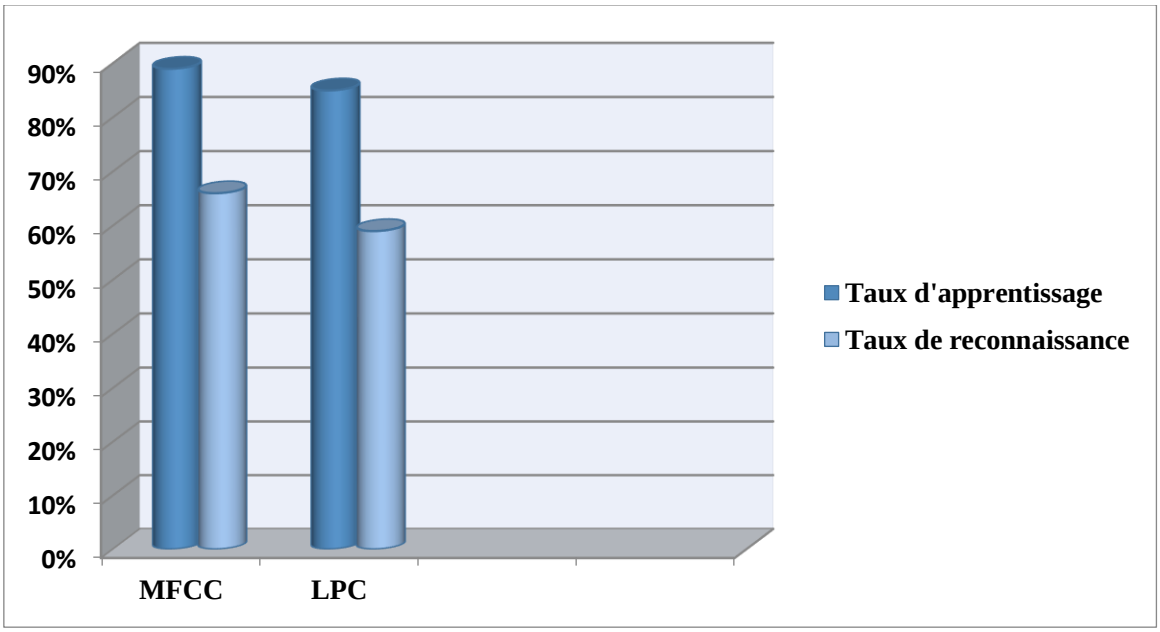


Figure IV.12 : Test et résultats en reconnaissance phonétique.

Pour ce test, on a utilisé un apprentissage sur un MLP simple, après 1000 itérations en utilisant 1é coefficients LPC, et 16 coefficients MFCC. Les résultats montrent une supériorité

des coefficients MFCC que se soit en taux d'apprentissage (89% MFCC, 65% LPC) ou en taux de reconnaissance (65% MFCC, 58% LPC). Le codage MFCC étant une méthode fréquentielle arrive à mieux reconnaitre les voyelles que le codage LPC. Notre corpus étant bruité, les coefficients MFCC montrent une meilleure résistance aux bruits que les LPC.

Base des mots avec le réseau MLP

ب	د ض	ت ط	ك	ق	ذ ظ	ع	س ص	م	ن	ل	ر
13%	**44%**	**53%**	**31%**	**12%**	**83%**	**0%**	**99%**	**70%**	**96%**	**96%**	**65%**

Tableau IV.1 : Taux de reconnaissance sur les consonnes de la base des mots avec un codage MFCC.

ب	د ض	ت ط	ك	ق	ذ ظ	ع	س ص	م	ن	ل	ر
60%	**70%**	**80%**	**60%**	**66%**	**92%**	**23%**	**99%**	**71%**	**61%**	**63%**	**67%**

Tableau IV.2 : Taux de reconnaissance sur les consonnes de la base des mots avec un codage LPC.

ب	د ض	ت ط	ك	ق	ذ ظ	ع	س ص	م	ن	ل	ر
74%	**72%**	**64%**	**73%**	**50%**	**99%**	**80%**	**88%**	**69%**	**70%**	**67%**	**69%**

Tableau IV.3 : Taux de reconnaissance sur les consonnes de la base des mots avec un codage LVQ-NPC.

On remarque une avance considérable se fait remarquer en taux de reconnaissance pour le codage LPC. En ce qui touche aux plosives, on dirait que le codage LPC arrive à mieux gérer les phonèmes rapides dans le temps.

On peut remarquer dans le codage LVQ-NPC certainement d'autres améliorations dans les scores de reconnaissances par rapport a ce qui a été cité plus haut. Mais les occlusifs restes encore loin de nos aspirations.

Conclusion générale et perspectives

Le travail réalisé lors de ce mémoire s'inscrit dans le cadre général de la reconnaissance automatique de la parole en langue arabe, le système de DAP est constitué essentiellement pour assurer la transcription d'un signal acoustique de parole continue en unités phonétiques.

L'étude théorique c'est basée essentiellement sur les caractéristiques acoustiques nécessaires au système de reconnaissance et quelques notions de base sur la phonétique de la langue arabe. Le système qu'on a utilisé dans notre travail suit principalement les étapes suivantes : la segmentation du signal de parole en classes phonétiques, l'extraction des indices phonétiques utilisés en reconnaissance et l'étiquetage des segments manuellement qui nécessite beaucoup de connaissance en phonétique. Après la phase d'analyse du signal de parole on trouve plusieurs informations, qui entrainent une grande variabilité du signal et qui nous ont permis d'extraire les caractéristiques appelées coefficients LPC et MFCC dans le réseau MLP.

La présence des phonèmes non-linéaires lors de la production de la parole limitent la validité du codage LPC. Cela nous a poussé à amélioré le système de DAP en utilisant un classifieur à prototype LVQ et un modèle neuronale pour obtenir en fin LVQ-NPC.

Compte tenu des résultats satisfaisants obtenus, le système reste toujours ouvert à d'autres améliorations à savoir l'enrichissement de la base données pour augmenter le taux de reconnaissance du modèle neuronal utilisé et utiliser d'autres algorithmes d'apprentissage.

Références Bibliographiques

[1] THIERRY, Dutoit. "Introduction au traitement automatique de la parole ". 1ér édition : Faculté polytechnique de Mons, 2000 ; P 6-7.

[2] DIDICHE,Mustapha Kamel Abderrahmane. "Modélisation neuro-prédictive pour la classification phonétique ". 139P.Thése présenté en vue de l'obtention du diplôme de doctorat en science en génie électriqueUniversité Mohammed Khider – Biskra : soutenu le 11/12/2014.

[3] BENDAHMANE, Abderrahmane. "Reconnaissance de la parole par distance DTW exemple d'application pour la reconnaissance de chiffres isolé dans la langue arabe". Laboratoire SIMPA, Oran, Algérie.

[4] LOUNI, Abderrahmane et BENYETTOU, Abdelkader. "Un codage neuro-prédictif pour l'extraction des traits distinctifs appliqué à la reconnaissance des phonèmes arabe". April, 2008, Oran, Algérie.

[5] BENYETTOU, Abdelkader. " Analyse et paramétrisation des phonèmes arabe en vue de la reconnaissance automatique de la parole". 1986, université d'Oran.

[6] LAURENT, Buniet. "Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques. Interface homme-machine [cs.HC] ". Université Henri Poincaré - Nancy I, 1997.

[7] J.-P. Zerling. "Articulation et coarticulation dans les groupes occlusive-voyelle en français". Thèse de doctorat de 3ème cycle, Université de Nancy 2, Nancy (France), 1979.

[8] LAURENT, Buniet." Traitement automatique de la parole en milieu bruité : étude de Modèles connexionnistes statiques et dynamiques". 2000.

[9] BERBECHE, Kamel." Modèle de Markov Cachés : Application à la reconnaissance automatique de la parole". Mémoire de magistère en Électronique option Télédétection, Université Mouloud Mammeri de Tizi Ouzou (Algérie).

[10] FILLON,Thomas." Traitement numérique du signal acoustique pour une aide aux malentendants "domain other. Télécom Paris Tech, 2004.

[11] R. Pujol : Promenade autour de la cochlée.
http://www.iurc.montp.inserm.fr/cric/audition/index.htm, 2004.

[12] E.A.G. Shaw : Handbook of Sensory Physiology, volume 5/1, chapter The External Ear, pages pages 455–490. Springer, 1974.

[13] James O. Pickles. "An Introduction to the Physiology of Hearing". Academic **Press**, 1982.

[14] M.C. Botte, G. Canevet, L. Demany et C. Sorin. "Psychoacoustique et Perception Auditive". INSERM/SFA/CNET, 1988.

[15] Jens Blauert." Spatial Hearing". MIT Press, Cambrige MA, 1983.

[16] Jean-Marc Jot, Véronique Larcher et Olivier Warusfel." Digital signal processing issues in the context of binaural and transaural stereophony". In Proceedings of the 98th Convention of the Audio Engineering Society, Paris, France, 1995.

[17] J.M. Aran, A. Dancer, J.M. Dolmazon et R. Pujol." Physiologie de la Cochlée". INSERM/SFA, 1988.

[18] A. Robier.Otologie. http://www.med.univ-tours.fr/enseign/orl/Otol/index.html, 2004. Site ORL de l'université de Médecine de Tours.

[19] J.Ramírez and al. "Speech/non-speech discrimination based on contextual information integrated bispectrum lrt". August 2006. VOL. 13, NO. 8.

[20] Van Den Heuvel H, Rietveld T, and Cranen B. "Methodological aspects of segment and speaker-related variability". A study of segmentai durations in dutch. 1994. No 22, pp 389-406.

[21] Atal B. S. Text-independent speaker recognition. April 1972. Paper presented at the Program of the 83rd meeting of the Acoustical Society of America , Buffalo, NY.

[22] T.Matsui and S.Furui."Text-independent speaker recognition using vocal tract and pitch information". 1990. pp 137-140.

[23] A.Malraux. http://andremalrauxtpeson.e-monsite.com/pages/ la-physique-du-son/l-intensite-du-son.html.

[24] Nguyen Quoc Cuong."Reconnaissance de la parole en langue Vietnamienne".2000.

[25] Halima, A."Un système neuro-expert pour la reconnaissance de la parole Neural Expert System for Speech Recognition".Mémoire pour l'obtention d'un Doctorat d'Etat en Informatique, 2005.

[26] Lotfi .A. "Un Systeme Hybride Ag/Pmc Pour La Reconnaissance De La Parole Arabe". Mémoire pour L'obtention Du Diplôme De Magister en Informatique, Université Badji Mokhtar Annaba, Année 2005.

[27] Calliope (ouvrage collectif)."La parole et son traitement automatique". Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pp, 1989.

[28] S. B. Davis et P. Mermelstein."Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no 4, pp 357-366, 1980.

[29] BERNARD, Gosselin."Représentation de l'information et quantification des signaux". Faculté Polytechniques de Mons, 2000. Belgique.

[30] MAAMRA,OumElhana et SETTOU,Trablesse."Proposition d'un modèle de déscripteur structurel pour la voix arabe,Application saisie des notes".73P.Thése de master Academique, UNIVERSITÉ HAMA LAKHDAR D'EL-OUED,2015.

[31] Samir.N." Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance". Thèse Docteur de l'Université de Rennes 1en Informatique, Année 2004.

[32] LÊ Viet.B." Reconnaissance automatique de la parole pour des langues peu dotes". Thèse Docteur de L'université Joseph Fourier - Grenoble 1 en Informatique, juin 2006.

[33] Oualid.D."Reconnaissance Automatique De La Parole Arabe Par Cmu Sphinx 4".Mémoire pour L'obtention Du Diplôme De Magister en électronique, Université Ferhat Abbas -Sétif 1-, Année 2013.

[34] Daniel.M, Sylvain.M, Corinne.F, Laurent.B, Jean-François.B."Segmentation selon le locuteur : les activités du Consortium ELISA dans le cadre de Nist RT03". Avignon Cedex 9-France, Année 2004.

[35] http://infirmier-freedom.blogspot.com/2015/05/anatomie-de-larynx.html.

[36] http://www.md.ucl.ac.be/didac/med2308/2.htm.