الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

# Mémoire de Master

Filière Télécommunication
Spécialité Réseaux et Télécommunication

Présenté par

TAHI Sofiane

&

BOUACALA Badreddine

# La reconnaissance des mots isolés En utilisant la méthode MFCC et l'algorithme DTW

Proposé par : Mr, Djebari Mustapha

Année Universitaire 2018-2019

**Nous tenons tout d'abord à remercier Dieu le tout puissant, miséricordieux qui nous a donné la force et la patience d'accomplir ce modeste travail.**

**Nous voudrions aussi présenter nos sincères remerciements à notre encadreur Mr, Djebari Mustapha, et nous voudrions également lui témoigner notre gratitude pour son soutien et ses conseils qui nous ont été très précieux afin de mener notre travail à bon port, Merci.**

**Nos vifs remerciements vont également au prestigieux membre du jury et à tout le staff du département de Génie Electrique de l'université de Blida 1 très spécialement nos professeurs durant tout ce cycle de 5ans.**

**Enfin, nous remercions chaleureusement nos chers parents pour leur soutien et leur patience, qui ont étaient toujours là dans les moments difficiles pour nous pousser à l'avant. Merci.**

**ملخص** تتيح عملية التعرف على الكلام للجهاز إمكانية فهم ومعالجة المعلومات المقدمة شفهيًا من قِبل مستخدم بشري، هذه العملية تتم عن طريق استخدام تقنيات مطابقة لمقارنة موجة صوتية بمجموعة من العينات المكونة من الصوت (وحدة الصوت الدنيا).

في هذا المشروع، سوف نقدم نظام التعرف التلقائي على الكلام على أساس التعرف على الكلمات المعزولة في حالة المفردات الصغيرة، وذألك باستخدام تقنينين الأولى تعمل على اساس معاملات السيب سترال في مقياس مال، والثاني هو فقط خوارزمية مقارنة ديناميكية.

**كلمات المفتاح** .MFCC. DTW كلمات معزولة.؛ صوت

**Résumé :** La reconnaissance de la parole permet à la machine de comprendre et de traiter les informations fournit oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore a un ensemble d'échantillons composé de phonème (unité sonore minimale).

Dans ce projet, nous allons présenter un système de reconnaissance automatique de parole basé sur la reconnaissance de mots isolés dans le cas de petit vocabulaire. En utilisant deux technique, l'une s'appuyant sur les coefficients cepstraux dans l'échelle de Mel (MFCC), et l'autre ce n'est qu'un algorithme de comparaison dynamique (DTW).

**Mots clés :** La parole ; MFCC ; DTW ; Mots isolés.

**Abstract:** Speech recognition allows the machine to understand and process information provided orally by a human user. It consists of using matching techniques to compare a sound wave to a set of samples composed of phoneme (minimum sound unit).
In this project, we will present an automatic speech recognition system based on the recognition of isolated words in the case of small vocabulary. Using two techniques, one based on the cepstral coefficients in the Mel scale (MFCC), and the other is only a dynamic comparison algorithm (DTW).

**Keywords:** The speech; MFCC; DTW; Isolated words.

## Listes des acronymes et abréviations

**DCT :** Discret Cosinus Transform

**DCT $^{-1}$ :** Transformée en Cosinus Inverse

**DTW :** Dynamic Time Warping

**FFT $^{-1}$ :** Transformée Fourier Inverse

**FFT :** Fast Fourier Transform

**HMM :** Hidden Markov Model (modèle de Markov caché)

**LPC:** Linear Prédictive Coding

**LFCC:** Linear Frequency Cepstral Coefficients

**LPCC:** Linear Predictive Coefficients Cepstral

**LSF:** Line Spectral Frequencies

**MSG:** Modulation SpectroGram

**MFCC:** Mel Frequency Cepstral Coefficients

**MATLAB:** Matrix Laboratory

**PLP:** Perceptual Linear Predictive

**SFS:** Speech Filing System

**SVM:** Support Vector Machine

**RAP :** Reconnaissance Automatique de la Parole

**TFD :** Transformée de Fourier Discrète

# Table des matières

**Chapitre 3 : Implémentation & Résultats**

## Liste des figures

## Liste des tableaux

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. Ces propriétés offrent une grande variété d'applications comme l'aide aux handicapés, la messagerie vocale, les services vocaux dans les téléphones portables, la production de documents écrits par dictée, etc.

Cependant le signal de parole est l'un des signaux les plus complexes à caractériser ce qui rend difficile la tâche d'un système RAP. Cette complexité du signal de parole provient de la combinaison de plusieurs facteurs, la redondance du signal acoustique, la grande variabilité inter et intra-locuteur, les effets de la coarticulation en parole continue et les conditions d'enregistrement. Pour surmonter ces difficultés, de nombreuses méthodes et modèles mathématiques ont été développés, parmi lesquels on peut citer : la comparaison dynamique, les réseaux de neurones, les machines à vecteurs supports SVM, les modèles de Markov stochastiques et en particulier les modèles de Markov cachés HMM. Ces méthodes et modèles travaillent à partir d'une information extraite du signal de parole considérée comme pertinente. Cette extraction est effectuée par une analyse acoustique qui conduit à rassembler cette information sous le terme de vecteur de paramètres acoustiques dont la dimension et la nature sont déterminants pour accéder à de bonnes performances des systèmes de RAP. Les différents types de paramètres acoustiques couramment cités dans la littérature sont les coefficients : LPC, LSF, MSG, LPCC, LFCC, PLP, MFCC, etc. Généralement, les coefficients MFCC sont les paramètres acoustiques (caractéristiques) les plus utilisés dans les systèmes RAP [1].

Cependant, des travaux de recherche ont permis d'étudier l'amélioration des performances des systèmes de RAP en combinant les coefficients MFCC avec d'autres LPCC, PLP, énergie, ondelettes [2] [3]. De plus d'autres travaux ont montré une amélioration de performances en intégrant les coefficients différentiels de premier ordre (appelés aussi delta $\Delta$) et deuxième ordre (appelés delta-delta $\Delta\Delta$) issus des coefficients 2 MFCC initiaux [4] [5]. Ces coefficients différentiels référés comme des paramètres dynamiques fournissent une information utile sur la trajectoire temporelle du signal de parole. Cette démarche a cependant pour effet de doubler ou tripler la dimension des vecteurs acoustiques. On peut penser qu'accroître le nombre de paramètres pertinents pourrait améliorer la précision de la reconnaissance. Dans les faits, cette idée se heurte à un problème connu sous le terme de "malédiction de la dimensionnalité". En effet, l'augmentation du nombre de paramètres se fait au prix d'un accroissement exponentiel du nombre d'échantillons constituant la base de données utilisée pour l'apprentissage du système de reconnaissance. En conséquence, lorsque la base de données est de taille finie et figée, les performances viennent même à se détériorer avec l'accroissement du nombre de paramètres. De plus, cette augmentation exige une quantité de ressources importante qui n'est pas en adéquation avec celle disponible dans les systèmes de RAP embarqués sur les téléphones (notamment reconnaissance des mots isolés ou connectés [7].

Enfin, notre projet consiste notamment à faire une reconnaissance des mots isolés (la prononciation des mots est séparée par des pauses de durée supérieure à quelques dixièmes de seconde (120 ms).) en combinant les coefficients MFCC avec la méthode de l'alignement globale et optimale (DTW) entre deux séquences, c'est-à-dire d'associer chaque élément de chaque séquence à au moins un élément de l'autre séquence en minimisant les coûts d'association entre les deux.

Dans ce travail de recherche, nous pouvons apporter une petite initiative au vaste domaine de la RAP, sous forme d'un programme Matlab qui peut également à l'aide de la MFCC et l'algorithme DTW reconnaitre des mots isolés. Pour atteindre notre objectif, notre étude s'articulera autour de trois chapitres :

- Le premier chapitre : il est destiné à la présentation de la parole humaine et de différents éléments qui gère sa production.
- Le deuxième chapitre : consacré à la décomposition théorique de l'algorithme DTW, et la méthode d'extraction des coefficients MFCC.
- Le troisième chapitre : il comportera la réalisation de l'interface applicable de notre programme, et notamment l'implémentation et les résultats de notre programme.
- Nous terminerons notre projet de fin d'étude, par une conclusion générale qui expliquera les avantages et les inconvénients de notre application, et ses différents domaines d'applications.

## 1.1 Introduction

La parole est un mécanisme de communication d'information entre les êtres humains. Pour cette raison, L'étude du mécanisme de la production de la parole humaine consiste donc à décrire, sur un plan physiologique, les mouvements des articulateurs vocaux qui produisent les sons et les mots.

Dans ce chapitre, nous présentons la physiologie anatomique du mécanisme de production de la parole humaine et le système de la reconnaissance automatique de la parole RAP.

## 1.2 Mécanisme et modélisation de la phonation

La phonation n'est pas une fonction physiologique vitale, elle ne possède pas de système organisé propre. La phonation se greffe sur l'appareil respiratoire pour emprunter ensuite, grâce au carrefour pharyngé, la voie digestive.

### 1.2.1 L'appareil phonatoire humain

L'appareil phonatoire ou appareil vocalique (**Fig. 1.1)** est l'ensemble des organes de la parole et des muscles qui les actionne. Ils permettent la production des phones, ou sons propres à la langue parlée. Voici les différents organes participant à la production de la parole [1] :

- **Les poumons :** fournit l'énergie nécessaire lorsque l'air est expiré par la trachée-artère, ainsi, il alimente en air les autres organes de la phonation.
- **Le larynx :** est l'ensemble des muscles et des cartilages mobiles qui entourent une cavité située dans la partie supérieure de la trachée.
- **Les cordes vocales :** sont deux lèvres symétriques placées en travers du larynx. Elles peuvent fermer complètement le larynx, l'air issu des poumons passe à travers les cordes vocales et se transforme en air phonatoire.
- **Le conduit vocal :** est considéré comme une succession de tubes aux cavités acoustiques de sections diverses, il prend naissance au niveau du larynx, juste après les cordes vocales, il se prolonge par des pharynx, puis il se subdivise en deux cavités :
    - Buccale : dont la variation est selon la position de la langue, les mâchoires et le palais.

- Nasale : sa forme est fixe, le couplage du conduit buccal et nasal produit un phénomène appelé nasalisation des sons produits.

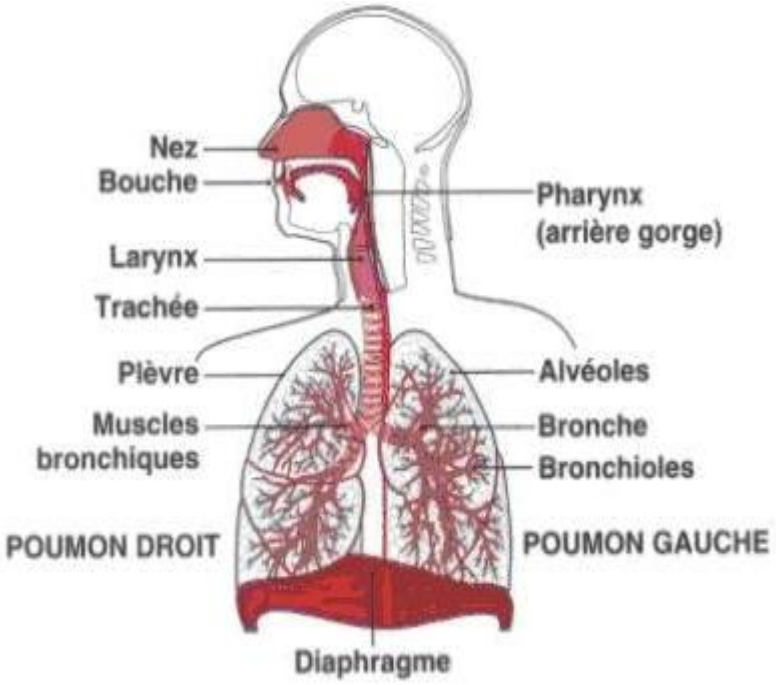



**Figure. 1.1 :** Appareil phonatoire humain [1]

### 1.2.2 Physiologie des organes de la phonation

Trois groupes d'organes assument les fonctions essentielles dans l'acte de parole, ou phonation **(Fig. 1.2)** [1] [2] :

- **Appareil respiratoire ou partie sub-glottique (diaphragme, poumons, trachée) :** qui fournit l'énergie nécessaire à la phonation en insufflant l'air vers la partie glottique.
- **Larynx ou partie glottique :** (ensemble de cartilages ligaments et muscles) contenant les cordes vocales (replis tendus horizontalement qui, sous l'effet des muscles, jouent un rôle de valve vis-à-vis de l'air des poumons libérant ainsi un flux d'air vers la partie supra-glottique.
- **Conduit vocal ou Partie supra-glottique :** formé des cavités orales (pharyngienne et buccale), à géométrie variable, en fonction des éléments articulateurs (langue, mâchoire inférieure, lèvres) et des cavités nasales,

à géométrie fixe, pouvant être couplées aux cavités orales par abaissement du voile du palais.

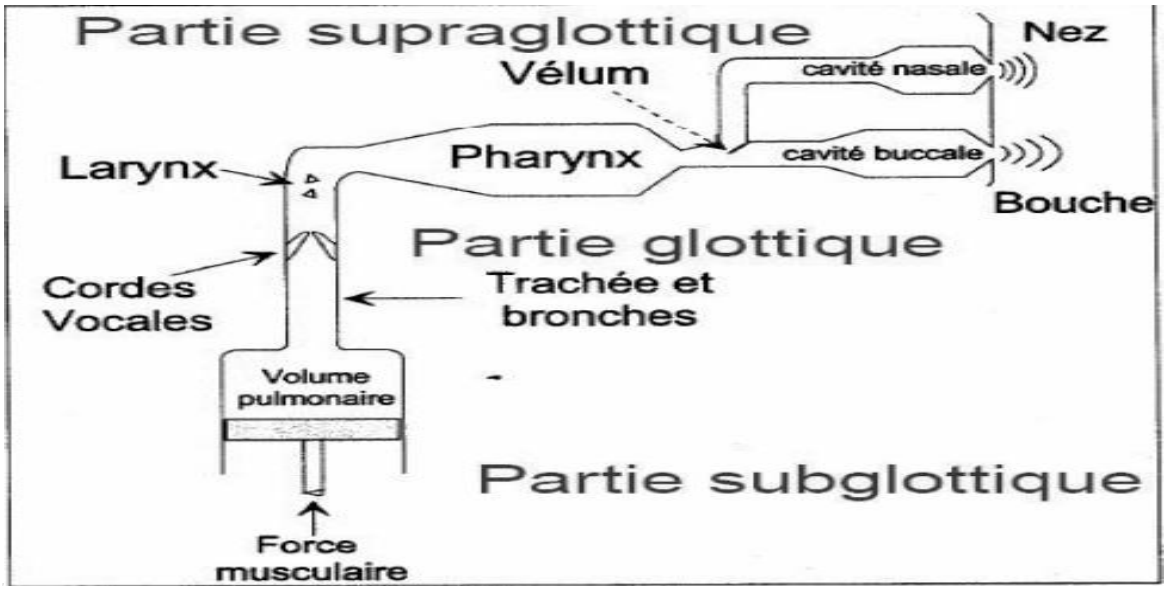



**Figure. 1.2 :** L'appareil phonatoire humain schématisé [3]

### 1.2.3 Principe de la production d'un son

Pour permettre la phonation, Le processus de production d'un son par l'appareil phonatoire peut se décomposer en trois phases : la création d'un écoulement d'air en provenance des poumons, la transformation de ce courant d'air en une énergie sonore par la vibration des cordes vocales (sons voisés) et/ou par la création de turbulences dues à un rétrécissement ou à une obstruction du conduit vocal, et enfin le filtrage de cette énergie sonore par les cavités supraglottiques [3].

Le principe de la production d'un son par l'appareil phonatoire humain repose sur le principe suivant (Fig. 1.3) :

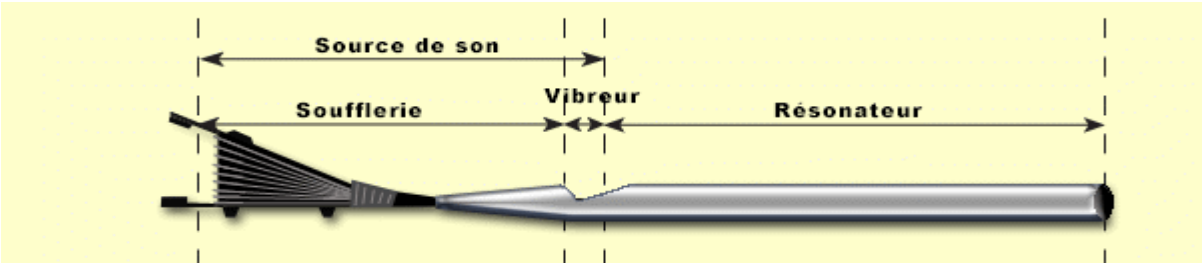



**Figure. 1.3 :** Principe de la génération d'un son de l'appareil phonatoire humain [4]

Une soufflerie (appareil respiratoire) créé une pression d'air engendrant au niveau d'un vibreur (cordes vocales) la création d'un son, qui est ensuite modulé par la partie résonatrice du système (les cavités bucco-nasales).

### 1.2.4 Bases fréquentielles de l'appareil phonatoire humain

La fréquence fondamentale de la voix est propre à chaque individu. Elle est en fonction de différents paramètres physiologiques tel que le volume et la masse de la glotte, la section de la trachée, sa longueur etc... [4]

| Catégorie | Homme | Femme | Enfant |
|---|---|---|---|
| La fréquence fondamentale $F_0$ (Hz) | 100 Hz | 200Hz | 300 à 400 Hz |

**Tableau 1.1 :** Différentes fréquence de l'appareil phonatoire humain

Contrairement aux instruments de musique, la fréquence fondamentale $F_0$ n'est pas prépondérante par rapport à ses harmoniques. L'énergie est répartie sur différentes fréquences supérieures de façon légèrement différente, d'un individu à l'autre pour des sons ou phonèmes équivalents.

### 1.2.5 Critère de classification

Les sons de la parole se forment lors du passage de l'air dans le chenal (tractus) vocal qui va des cordes vocales aux lèvres. La colonne d'air passe alors par un ensemble de cavités pouvant prendre une grande variété de configurations. Ce sont toutes ces modifications rapides qui déterminent la réalisation de sons très divers. Leur étude permet d'opérer un classement en voyelles et consonnes. En fonction des possibilités offertes par l'appareil vocal, chaque langue a adopté un ensemble particulier de sons distinctifs ou phonèmes qui sont les éléments sonores les plus brefs permettant de distinguer les différents mots [5].

-Les critères permettant de classifier les sons d'une langue sont les suivants :

- Le mode articulatoire a trait à la qualité du passage de l'air dans le canal buccal. La réalisation des voyelles implique un passage libre de l'air le long du canal buccal. Le degré d'ouverture de la cavité buccale permet de distinguer quatre types de voyelles (les voyelles ouvertes, mi-ouvertes, mi-fermées et fermées). Pour les consonnes, deux modes articulatoires sont à distinguer. Le passage de l'air est totalement bloqué ou obstrué lors de la production des consonnes occlusives. Le passage est rétréci suffisamment pour permettre l'émission d'un bruit continu lors de la réalisation des consonnes constrictives ou fricatives.

- La résonance orale ou antirésonance nasale est fonction de la fermeture ou de l'ouverture de l'accès vers les fosses nasales. Lors de la production de voyelles ou de consonnes nasales, le voile du palais est abaissé et permet le passage de l'air à la fois par le canal buccal et par les fosses nasales, ce qui confère aux sons une coloration particulière. Les voyelles et les consonnes produites sont alors dites " nasales ". Lorsque le voile du palais est relevé et bien accolé à la paroi pharyngale, l'air ne passe que par la cavité buccale, donnant naissances aux sons vocaliques et consonantiques dits oraux.

- Le rôle des cordes vocales détermine le caractère sourd ou sonore des différentes articulations. Lorsque les cordes vocales vibrent, les sons seront dits voisés ou sonores par opposition aux sons non voisés ou sourds. La réalisation des voyelles implique la mise en vibration des cordes vocales. Pour les consonnes, l'absence ou la présence de ces vibrations détermine leur caractère sourd ou sonore.

- Le lieu d'articulation se situe nécessairement dans la partie supérieure du canal buccal dans une zone allant de la lèvre supérieure jusqu'à la paroi pharyngale. C'est le point duquel l'articulateur se rapprochera ou avec lequel il entrera en contact.

- L'articulateur est constitué par la région inférieure du canal buccal. Il s'agit de la lèvre inférieure et des différentes parties de la langue. La réalisation de toute articulation implique un rapprochement plus ou moins grand ou un contact franc entre l'articulateur et le lieu d'articulation.

### 1.2.6 Classification des phonèmes

En générales la langue est composée de plus d'une trentaine de phonème repartis en plusieurs classes [6].

#### 1.2.6.1 Les voyelles

Sont caractérisées acoustiquement par la présence des zones de fréquence où les harmoniques sont particulièrement intenses. Elles sont générées par une vibration laryngienne des cordes vocales. On distingue deux types de voyelles :

- Les voyelles nasales : dont le conduit nasal est couplé à la cavité buccale et l'émission se fait à la fois par les narines et par la bouche (exemple : blanc, bon, lin, brun).
- Les voyelles orales : sont des sons voisés, chacune d'elles correspond à une configuration particulière du conduit vocal, sans intervention de la cavité nasale qui est alors isolée par fermeture du voile du palais (exemple : plat, lait, bol, blé, roue).

#### 1.2.6.2 Les consonnes

Ce sont des sons résultants d'une fermeture partielle (constriction) ou totale (occlusion) du conduit vocal lors du passage de l'air phonatoire, elles peuvent être voisées ou non voisées, nasales ou orales. Les consonnes sont classées selon les trois principaux types suivants :

- Fricatives (constrictives) : dans cette classe sont regroupés les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences et peut

être voisée [v], [z] (exemple : verre, Asie) ou sourde [f], [s], [š] (exemple : fer, assis, chou).

- Plosives (occlusives) : un son plosive est produit par une occlusion momentanée du conduit vocal, en un point donné suivi par une ouverture brusque, et peut être sonore (exemple : basse, doux, goût) ou sourde (exemple : passe, toux, cou).
- Nasales : dans ce cas le conduit vocal est fermé et l'air s'écoule par la cavité nasale. On distingue deux consonnes nasales, toutes les deux voisées :
  [m] : dont le lieu d'articulation est labial (exemple : masse).
  [n] : dont le lieu d'articulation est dentaire (exemple : nous, signal).

#### 1.2.6.3 Les semi-voyelles ou semi-consonnes

Sont des phonèmes intermédiaires entre les voyelles et les consonnes. Quand on les prononce, on entend le timbre d'une voyelle auquel s'ajoute le frottement d'une consonne spirante. Leur fréquence d'emploi est liée à la vitesse du débit de la parole, plus celui-ci est rapide, plus il y aura de semi-voyelle (exemple : hier, huit, oui).

#### 1.2.6.4 Les sonantes

Se caractérisent par une structure de formants et elles ne possèdent que peu ou pas de bruit, plusieurs sous-classes existent :

- Les vibrantes : il s'avère qu'il en existe une seule qui est le [r] (exemple : rue) qui est produite par une vibration de la langue.
- Les liquides : il en existe une seule qui est produite par une obstruction partielle du conduit buccal et un écoulement latéral [l] (exemple : lent).

Voici les différents modes d'articulation des phonèmes qui sont classifiés dans la (Fig. 1.4)

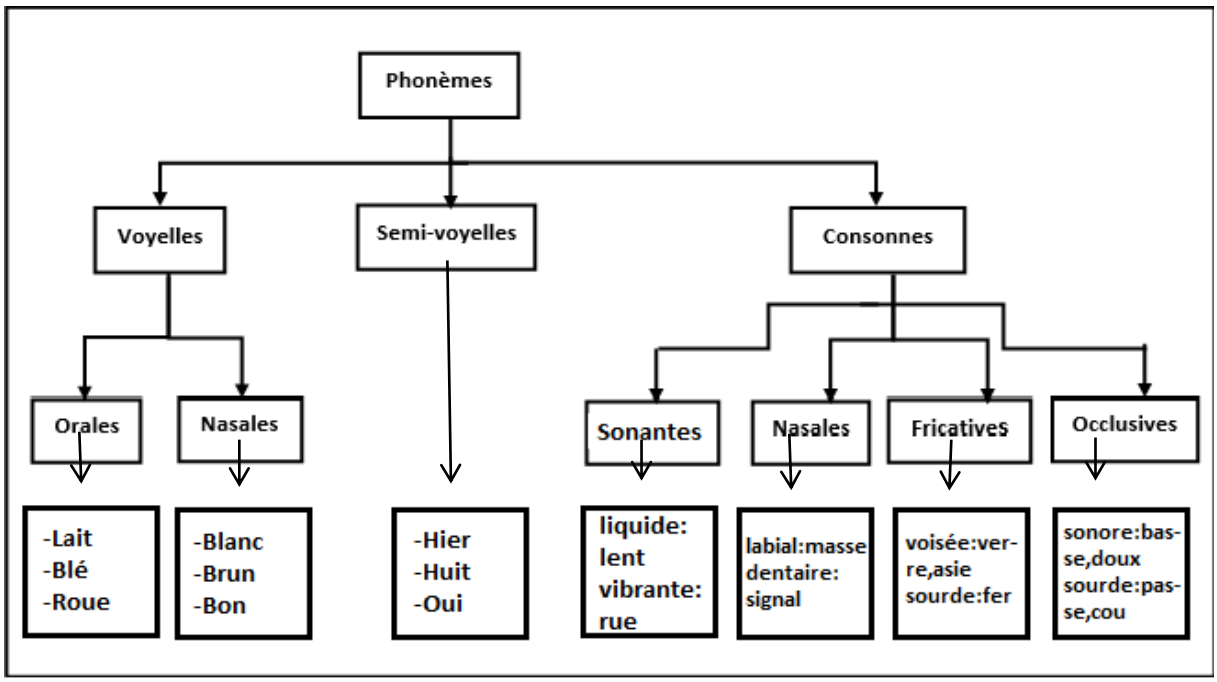



**Figure. 1.4 :** Classification des phonèmes

## 1.3 Problèmes de variabilité de la parole

La capacité qu'ont les êtres humains à pouvoir communiquer au travers du langage est un processus à la fois simple et extrêmement complexe. Simple, car nous n'éprouvons pas de difficulté à produire, ni à comprendre les messages linguistiques dans des conditions normales. La rapidité de ces processus en est une illustration convaincante. Simple donc, mais également extrêmement complexe dès lors que l'on tente d'expliquer la nature des mécanismes qui entrent en jeu dans ces processus. Une des multiples raisons de cette complexité est qu'il existe une relation très indirecte entre notre volonté d'exprimer un message et les diverses formes que peut prendre la réalisation de cette volonté. Autrement dit, d'une part l'être l'humain peut produire des réalisations différentes pour un même message linguistique, d'autre part il est capable à partir de ces réalisations d'identifier le message linguistique unique dont ces réalisations sont la représentation. La variabilité de la parole est au centre de cette problématique [7].

Nous allons maintenant voir les problèmes directement liés à la parole. Ceux-ci sont relatifs à la différence innée de la prononciation vis- à-vis d'un ou plusieurs locuteurs.

### 1.3.1 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. Il existe aussi un autre type de variabilité intra-locuteur lié à la phrase de production de parole ou de préparation à la production de parole. Cette variation est due aux phénomènes de coarticulation.

### 1.3.2 Variabilité inter-locuteur

La variabilité inter-locuteur est un phénomène majeur en reconnaissance de la parole. Elle concerne les différences du signal vocal des locuteurs différents. Ces différences sont liées à l'âge, l'accent régional, le sexe, etc.

### 1.3.3 Variabilité due à l'environnement :

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. 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 est considérée comme du bruit. La variabilité environnementale due au locuteur peut tout d'abord être de nature physiologique. Ainsi, des systèmes mécaniques. 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. 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 [7].

## 1.4 Reconnaissance Automatique de la Parole

La Reconnaissance Automatique de la Parole (RAP) est un système qui a la capacité de détecter la parole et de l'analyser dans le but de générer une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. Cette analyse se fonde sur l'extraction des paramètres descriptifs de la parole. Cependant le signal parole ne contient pas seulement des informations sur le texte parlé mais aussi des informations sur le locuteur, la langue, les émotions dont leur extraction n'est pas l'objectif de la RAP. On s'intéresse dans un premier temps à une étape primordiale de la RAP, en l'occurrence la sélection des paramètres descriptifs pertinents pour la tâche de reconnaissance des mots isolés. Avant d'aborder cette étape, nous allons décrire dans ce chapitre les principes généraux et les problèmes de la RAP, ainsi que les différentes étapes constituant un tel système.

### 1.4.1 Approches de la reconnaissance de la parole

Le principe général d'un système de RAP peut être décrit par la figure (**Figure. 1.5)**

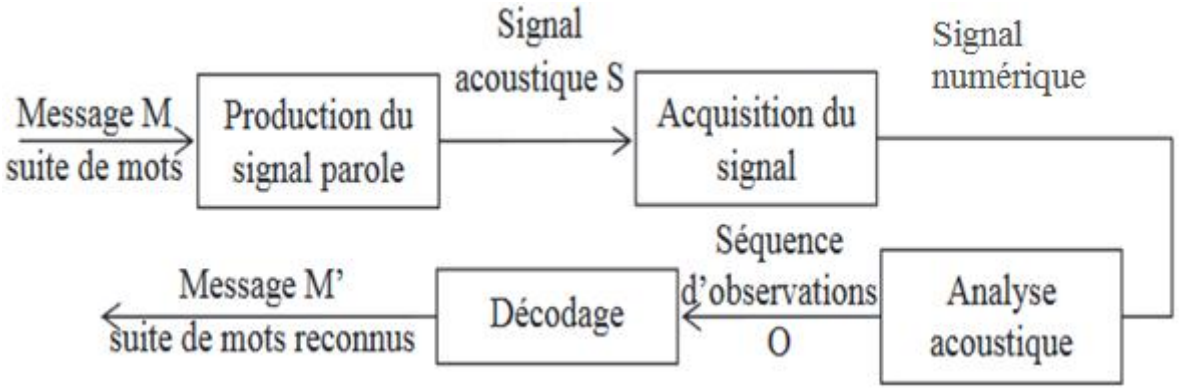



**Figure. 1.5 :** Principe de la reconnaissance de la parole

La suite de mots prononcés M est convertie en un signal acoustique S par l'appareil phonatoire. Ensuite le signal acoustique est transformé en une séquence de vecteurs acoustiques ou d'observations O (chaque vecteur est un ensemble de paramètres acoustiques). Finalement le module de décodage consiste à associer à la séquence d'observations O une séquence de mots reconnus M'.

Un système RAP transcrit la séquence d'observations O en une séquence de mots M en se basant sur le module d'analyse acoustique et celui de décodage. Cependant, les systèmes RAP en majorité utilisent des méthodes statistiques à base de modèles de Markov. Ces méthodes sont hybrides (globale et analytique).

#### 1.4.1.1 Approche Globale

L'approche globale considère l'énoncé entier comme une seule unité indépendamment de la langue. Elle consiste ainsi à abstraire totalement les phénomènes linguistiques et ne retenir que l'aspect acoustique de la parole. Cette approche est destinée généralement pour la reconnaissance des mots isolés séparés par au moins 200 ms (voir **Figure. 1.6)** ou enchaînés, appartenant à des vocabulaires réduits.

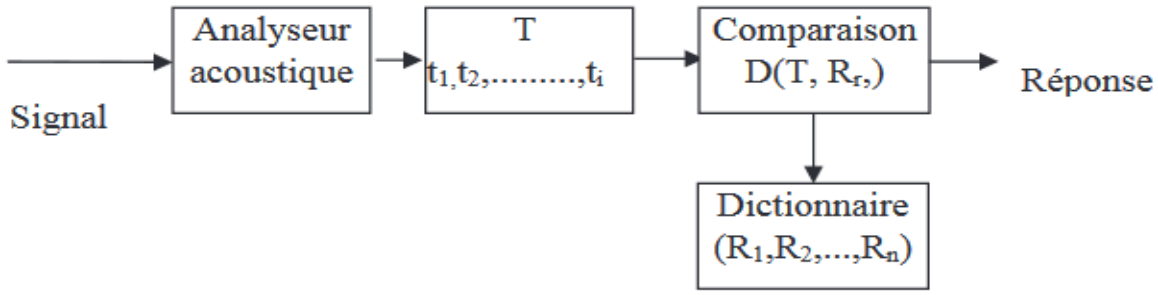



**Figure. 1.6 :** Reconnaissance de mots isolés

Dans les systèmes de reconnaissance globale, une phase d'apprentissage est nécessaire, pendant laquelle l'utilisateur prononce la liste des mots du vocabulaire de son application. Pour chacun des mots prononcés, une analyse acoustique est effectuée permettant d'extraire les informations pertinentes sous forme de vecteurs de paramètres acoustiques. Le résultat est stocké ensuite en mémoire. Donc, les méthodes globales mettent en jeu une ou plusieurs images de références acoustiques (R1, ....., Rn), a priori pour chaque mot.

Lors de la phase de reconnaissance, lorsque l'utilisateur prononce un mot T, la même analyse est effectuée : l'image acoustique du mot à reconnaître est alors comparée à toutes celles des mots de référence du vocabulaire.

Le mot ressemblant le plus au mot prononcé est alors reconnu. 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. Ces problèmes peuvent être résolus en appliquant un algorithme classique de la programmation dynamique appelé alignement temporel dynamique (Dynamic Time Warping DTW). Ce type d'algorithme permet sous certaines conditions d'obtenir une solution optimale à un problème de minimisation d'un certain critère d'erreur sans devoir considérer toutes les solutions possibles. Dans le cas de la RAP, cet algorithme consiste à chercher le meilleur

alignement temporel qui minimise la distance entre la représentation d'un mot de référence et la représentation d'un mot inconnu. Dans le cas de grand vocabulaire ou de la parole naturelle continue, cette approche devient insuffisante et il est alors nécessaire d'adopter une nouvelle approche.

#### 1.4.1.2 Approche analytique

L'approche analytique cherche à trouver des solutions au problème de la reconnaissance de la parole continue ainsi qu'au problème du traitement de grands vocabulaires en contexte multi-locuteurs puisqu'on ne mémorise qu'un nombre restreint d'éléments, indépendant de la taille vocabulaire. Cette approche consiste à segmenter le signal vocal en constituants élémentaires (mot, phonème, syllabe…etc.), puis à identifier ces derniers, et enfin à reconstituer la phrase prononcée par étapes successives en exploitant des modules d'ordre linguistique. Le processus de la reconnaissance de la parole dans une telle méthode peut être décomposé en deux opérations :

- Représentation du message (signal vocal) sous la forme d'une suite de segments de parole, c'est la segmentation.
- Interprétation des segments trouvés en termes d'unités phonétiques, c'est l'identification.

#### 1.4.1.3 Approche statique

L'approche statistique se fonde sur une formalisation statistique simple issue de la théorie de l'information permettant de décomposer le problème de la reconnaissance de la parole continue. Cette approche est construite sur le principe de fonctionnement des méthodes globales (avec phase d'apprentissage et de reconnaissance) mais avec l'exploitation des niveaux linguistiques. Ainsi une analyse acoustique est nécessaire pour convertir tout signal vocal en une suite de vecteurs acoustiques. Ces vecteurs sont considérés comme des observations dans la phase d'apprentissage des modèles statistiques et dans la phase de reconnaissance qui effectue une classification de chaque observation (par un index d'état dans le cas de la modélisation Markovienne) [8] [9].

### 1.4.2 Difficultés de la reconnaissance de la parole

Le signal de parole est l'un des signaux les plus complexes à caractériser et analyser car, c'est un sujet à une grande variabilité. Cette complexité est liée à la production du signal de parole, ainsi qu'à l'aspect technologique. Le signal de parole varie non seulement avec les sons prononcés, mais également avec le locuteur, l'âge, les émotions, la santé, l'environnement. De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance (les appareils d'acquisition et de transmission), ainsi que par le milieu ambiant.

#### 1.4.2.1 La redondance

Le signal vocal présente un caractère redondant. Il contient plusieurs types d'information : les sons, la syntaxe et la sémantique de la phrase, l'identité du locuteur et son état émotionnel. Bien que cette redondance assure une certaine résistance du message au bruit.

#### 1.4.2.2 La variabilité

Le signal vocal de deux prononciations à contenu phonétique égal est différent pour un même locuteur (variabilité intra-locuteur) ou pour des locuteurs différents (variabilité interlocuteur).

En effet, lorsque la même personne prononce deux fois le même énoncé, on constate des variations sensibles sur le signal vocal causées par :

- L'état physique, par exemple, la fatigue ou le rhume.
- Les conditions psychologiques, comme le stress.
- Les émotions du locuteur.
- Le rythme lié à la durée des phonèmes (façon dont s'exprime le locuteur) et l'amplitude (voix normale, voix chuchotée, voix criée).

Cependant la variabilité interlocuteur est a priori la plus importante. Elle s'explique par :

- Les différences physiologiques entre locuteurs.
- Les habitudes acquises en fonction du milieu social et géographique comme les accents régionaux.

Cette variabilité rend très difficile la définition d'invariants et complique la tâche de reconnaissance. Ainsi, il faut pouvoir séparer ce qui caractérise les phonèmes, de l'aspect particulier à chaque locuteur.

#### 1.4.2.3 La continuité et la coarticulation

La production d'un son est fortement influencée par le son qui le précède et qui le suit en raison de l'anticipation du geste articulatoire. La localisation correcte d'un segment de parole isolé de son contexte est parfois impossible. Évidement la reconnaissance des mots isolés bien séparés par un silence est plus facile que la reconnaissance des mots connectés. En effet, dans ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés.

#### 1.4.2.4 Condition d'enregistrement

L'enregistrement du signal de parole dans de mauvaises conditions rend difficile l'extraction des informations pertinentes indispensables pour la reconnaissance des mots contenus dans ce signal. En effet, les perturbations apportées par le microphone (selon le type, la distance, l'orientation) et l'environnement (bruit, réverbération) compliquent beaucoup le problème de la reconnaissance [10].

### 1.4.3 Applications de la reconnaissance de la parole

Les applications de la RAP sont nombreuses. Elles existent là où la parole peut remplacer ou compléter une interface existante pour communiquer avec une machine (tab.1.2) [11].

| Domaine | Application |
|---|---|
| **Industrie** | -Commande de machines, conduite de processus, routage, Programmation de machine-outil à commande numérique.<br>-Entrée de données dans les systèmes de Conception Assistée par Ordinateur |
| **Télématique** | -Demande de renseignement, réservation, consultation de bases de données.<br>Numérotation téléphonique automatique etc. |
| **Bureautique** | -Commande de fonctions, entrée de données, machine à écrire automatique. |
| **Aviation** | Commande d'appareillages, contrôle aérien automatique. |
| **Sécurité et la justice** | -Empreinte vocale pour accès en zone règlementée (lieu, fichier)<br>-Identification des suspects. |

| | |
|---|---|
| **Enseignement et la formation** | -Formation des pilotes, programmation, enseignement assisté par ordinateur (langues, …) |
| **Aide au médecin** | -Diagnostic assisté par ordinateur, choix de médicaments, comptes rendus.<br>-Commande d'appareillages divers (chirurgie…).<br>-Repérage des indices physiologiques (zézaiement, bégaiement, …) et psychologiques (émotivité, timidité, agressivité, …) |

**Tableau 1.2 :** Différentes applications de la RAP

### 1.4.4 Les avantages et la complexité de la reconnaissance de la parole

Les avantages que l'on attend de la reconnaissance de la parole sont multiples. Elle libère complètement l'usage de la vue et des mains (contrairement à l'écran et au clavier), et laisse l'utilisateur libre de ses mouvements. La vitesse de transmission des informations est supérieure, dans la RAP à celle que permet l'usage du clavier. Enfin tout le monde ou presque sait parler, alors que peu de gens sont à l'abri des fautes de frappe et d'orthographe. Ces avantages sont à l'origine d'une grande variété d'applications comme :

- L'aide aux handicapés.
- La messagerie.
- La commande de machines ou de robots.
- Le contrôle de qualité et la saisie des données.
- L'accès à distance : téléphone, internet.
- La dictée vocale.

Toutes ces applications bénéficient de l'évolution technologique qui se traduit par l'apparition de composants intégrés spécialisés (en traitement du signal pour la programmation dynamique) et du développement des techniques et des méthodes algorithmiques de plus en plus performantes.

Enfin l'insertion d'un système RAP dans son environnement réel d'utilisation dépend de son contexte d'application et de ses conditions d'utilisation ce qui peut vite rendre le dispositif très complexe.

Un système RAP peut être décrit selon quatre grands axes graduant cette complexité :

- La dépendance du locuteur (système optimisé pour un locuteur bien particulier) ou l'indépendance du locuteur (système pouvant reconnaître n'importe quel locuteur).
- Le mode d'élocution : mots isolés, mots connectés, mots-clés, parole continue lue ou parole continue spontanée.
- La complexité du langage autorisé : taille du vocabulaire et difficulté de la grammaire.
- La robustesse aux conditions d'enregistrement : systèmes nécessitant de la parole de bonne qualité ou fonctionnement en milieu bruité.

## 1.5 Système de reconnaissance

On peut classer les différents systèmes de reconnaissance suivant leurs performances techniques, à savoir la reconnaissance de mots isolés par rapport à parole continue, systèmes mono locuteurs par rapport à systèmes multi locuteur, etc.

### 1.5.1 Reconnaissance des mots

#### 1.5.1.1 Reconnaissance des mots isolés

La technique de la reconnaissance des mots isolés est basée sur l'approche globale ou les mots à reconnaître sont comparés à un dictionnaire de références acoustiques conçu lors de la phase d'apprentissage. On rencontre ce type de reconnaissance aussi bien dans les systèmes de pilotage d'application les plus simples, correspondant à des situations connues à l'avance (jeu de commandes d'un système d'exploitation, réponses à des questions ou à des choix proposés par un serveur vocal, etc.) [12].

#### 1.5.1.2 Reconnaissance de parole continue

Cette technique qui apporte un confort d'utilisation indéniable est beaucoup plus complexe que la précédente en raison de phénomène de coarticulation ou de liaisons entre des mots contigus.

#### 1.5.1.3 Reconnaissance de mots enchainés

La méthode de la reconnaissance de la parole continue ne peut pas être appliquée aux mots enchaînés. En effet une simple analyse de l'enveloppe du signal ne peut donc pas détecter

les frontières entre les mots. Pour pallier à ce problème, une segmentation des mots s'impose [13].

### 1.5.2 Reconnaissance de locuteur

On distingue deux systèmes de la reconnaissance.

#### 1.5.2.1 Système mono-locuteur

En raison de la variabilité importante du signal de parole entre locuteurs différents, de nombreux systèmes ne peuvent fonctionner qu'avec un seul locuteur. Les plus simples se contentent de stocker et de rapprocher les différentes prononciations d'un même mot. Ce qui suppose de la part de l'utilisateur un entrainement préalable du système. D'autres possèdent déjà une représentation standard des unités phonétiques, réalisées par le constructeur, complétées par une phase d'apprentissage durant laquelle on améliore le modèle en fonction des caractéristiques de la voix de l'utilisateur.

#### 1.5.2.2 Système multi-locuteur

Le système multi-locuteur est plus précisément adapté aux applications visant un large public. Leur mise au point requiert cependant de la part du constructeur un travail important, en effet, une liste de mots est présentée à un grand nombre de personnes, c'est la phase d'apprentissage dans le but est de créer un dictionnaire de référence [14].

## 1.6 Conclusion

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 interlocuteurs, 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. Actuellement la majorité des systèmes RAP sont construits selon la méthode statistique en utilisant plusieurs modèles comme le modèle de Markov cachés HMM. Ou Les coefficients MFCC (Mel Frequency Cepstral Coefficients) ou aussi les paramètres LPCC (Linear Prediction Cepstral Coefficients). Ainsi, dans ce chapitre, nous avons décrit brièvement le principe de fonctionnement des systèmes RAP.

## 2.1 Introduction

La voix est un signal d'information infinie. Une analyse directe et la synthèse du signal vocal complexe sont dues à une trop grande quantité d'informations contenues dans le signal. Par conséquent, les processus de signaux numériques tels que l'extraction de caractéristiques et la correspondance de caractéristiques sont introduits pour représenter le signal vocal. Plusieurs méthodes, telles que le codage prédictif par couche (LPC), le modèle de Markov caché (HMM), le réseau de neurones artificiels (ANN), etc., sont évaluées dans le but d'identifier une méthode simple et efficace pour le signal vocal. Le processus d'extraction et d'appariement est mis en œuvre juste après que le signal de pré-traitement ou de filtrage est exécuté. Les coefficients de fréquence de Mel MFCC, est utilisée comme technique d'extraction. L'alignement de séquence non linéaire connu sous le nom de Dynamic Time Warping (DTW) a été utilisé comme technique d'appariement de caractéristiques. Comme il est évident que le signal vocal a tendance à avoir une cadence temporelle différente, l'alignement est important pour produire de meilleures performances.

Ce chapitre présente la viabilité de MFCC pour extraire des fonctionnalités et de DTW pour comparer les modèles de test.

## 2.2 Prétraitement du signal de la parole

Avant d'aborder l'analyse acoustique, il est recommandé de faire subir au signal vocal un prétraitement pour lui donner une représentation moins redondante, tout en permettant une extraction assez précise des paramètres pertinents qui caractérisent le signal de la parole. Le prétraitement se présente en plusieurs étapes qui sont schématisées dans la figure 2.1 [15].

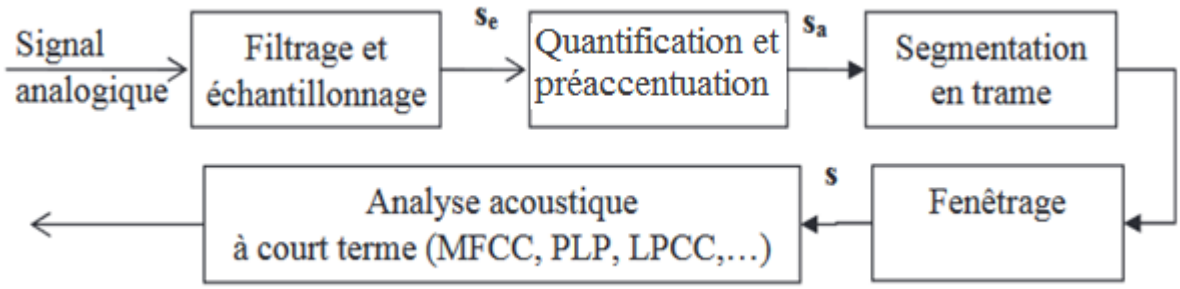



**Figure 2.1 :** Etapes de prétraitement du signal de parole

### 2.2.1 Le préamplificateur

La parole apparaît physiquement comme une variation de la pression de l'air, 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 [15].

### 2.2.2 Filtre de garde

Afin de réduire le coût du traitement numérique d'une façon notable, on doit limiter le spectre en utilisant un filtre dont la fréquence de coupure fc est choisie en fonction de la fréquence d'échantillonnage.

### 2.2.3 Echantillonnage

Le signal de la parole étant analogique, il s'avère nécessaire de le numériser (échantillonnage + quantification) avant tout traitement. Cette opération consiste en l'échantillonnage du signal qui est présenté dans la figure 2.2 [16].

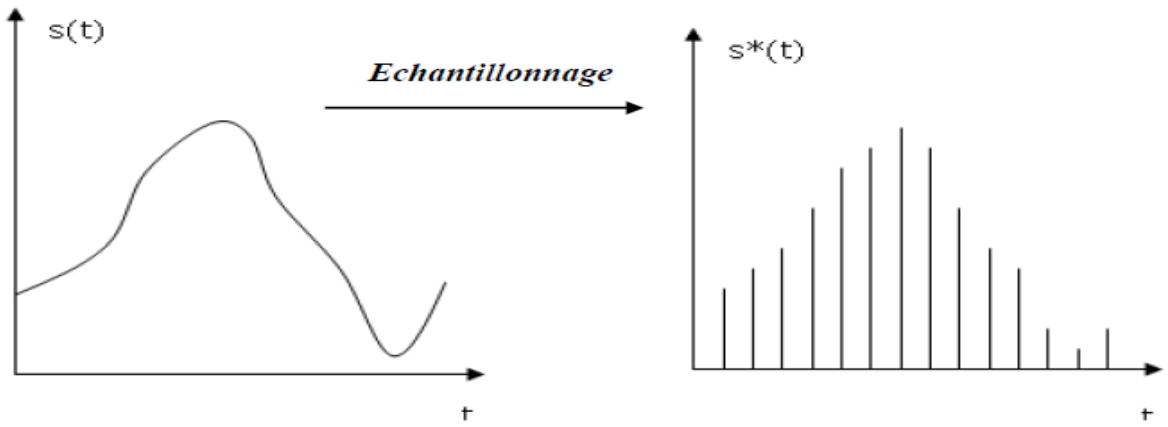



**Figure 2.2 :** Echantillonnage

D'après Shannon, la perte d'information entre le signal analogique et le signal discret correspondant est nulle si et seulement si :

$$f_e \geq 2 \times f_{max} \qquad (2.1)$$

fe : la fréquence d'échantillonnage.

fmax : la fréquence maximale que contient le signal à traiter.

### 2.2.4 La quantification

La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore à chaque pas de l'échantillonnage. Le choix de la fréquence d'échantillonnage est aussi déterminant pour la définition de la bande passante représentée dans le signal numérisé qui est présenté dans la figure 2.3.

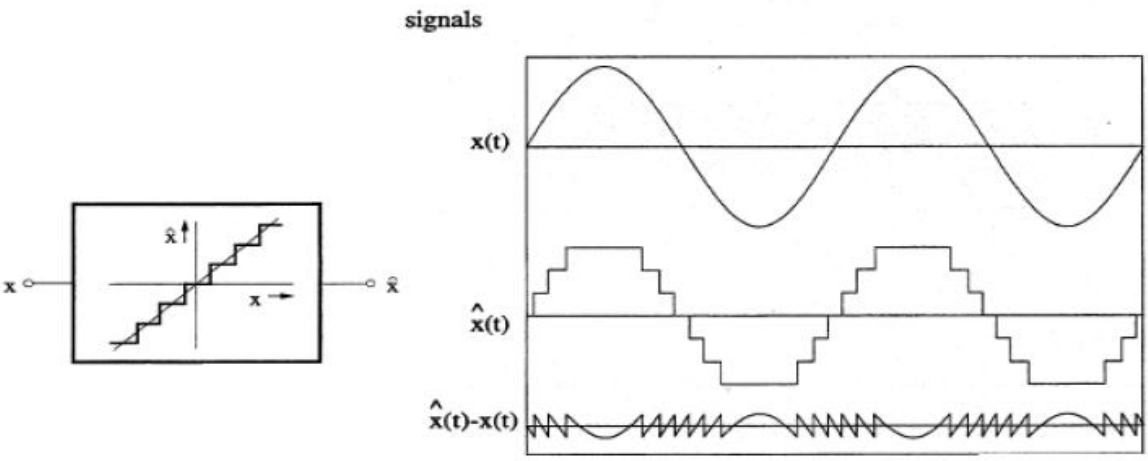



**Figure 2.3 :** La quantification

### 2.2.5 Préaccentuation

En général, le signal vocal se caractérise par une perte de 6 dB/octave, due à l'influence de la source d'excitation et au rayonnement des lèvres. Une perte de 6 dB/octave veut dire que les hautes fréquences ont une énergie plus faible que celle des basses fréquences. Pour pallier à cet inconvénient la préaccentuation permet d'égaliser les sons aigus avec les sons graves (voir figure (2.4)).

L'opération consiste à faire passer le signal à travers un filtre de transmittance :

$$H(z) = 1 - az^{-1} \qquad (2.2)$$

Le facteur de préaccentuation est pris entre 0.9 et 1 (souvent 0.95). Comme conséquence, la préaccentuation introduit une légère distorsion spectrale [17].

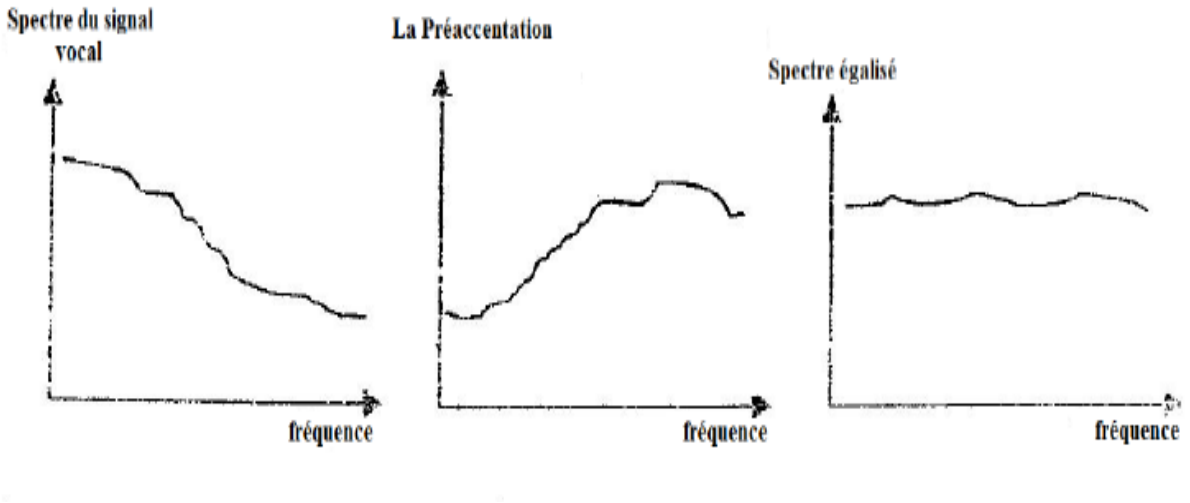



**Figure 2.4 :** la préaccentuation

Des études comparatives entre différents types de paramètres acoustiques ont été effectuées, pour déterminer ceux qui représentent mieux le signal vocal. Une étude comparative classique a été effectuée entre plusieurs représentations du signal vocal : cepstre en sortie d'un banc de filtres en échelle Mel (MFCC) ou en échelle linéaire (LFCC), coefficients de prédiction linéaire (LPC) ou de réflexion (RC), cepstre calculé à partir des coefficients auto-régressifs (LPCC). Ces représentations ont été appliquées dans un système de reconnaissance fondé sur un alignement DTW entre un mot de test et des mots de référence. Dans cette étude, les MFCC donnent les meilleurs résultats, ce qui montre plus généralement l'intérêt d'un pré-traitement par banc de filtres, d'une échelle fréquentielle non linéaire, et de la représentation cepstrale [18].

## 2.3 Mel Frequency cepstral coefficients (MFCC)

Le point principal à comprendre à propos de la parole est que les sons générés par un humain sont filtrés par la forme de l'appareil vocal (la langue, les dents, etc.). Si nous pouvons déterminer la forme avec précision, cela devrait nous donner une représentation précise du phonème produit.

La forme du conduit vocal se manifeste dans l'enveloppe du spectre de puissance à court terme, et le rôle des MFCC est de représenter avec précision cette enveloppe.

L'étude des coefficients MFCC du signal permet d'extraire des caractéristiques de celui-ci autour de la FFT et de la DCT, convertis sur une échelle de Mel. Il s'agit de la méthode la plus utilisée pour représenter un signal en reconnaissance de la parole, car elle est très robuste. Son principal avantage est que les coefficients obtenus sont décorrélés.

La première étape de tout système de reconnaissance automatique de la parole consiste à extraire des caractéristiques, c'est-à-dire à identifier les composants du signal audio qui sont utiles pour identifier le contenu linguistique et éliminer tous les autres éléments contenant des informations telles que le bruit de fond, les émotions, etc [19].

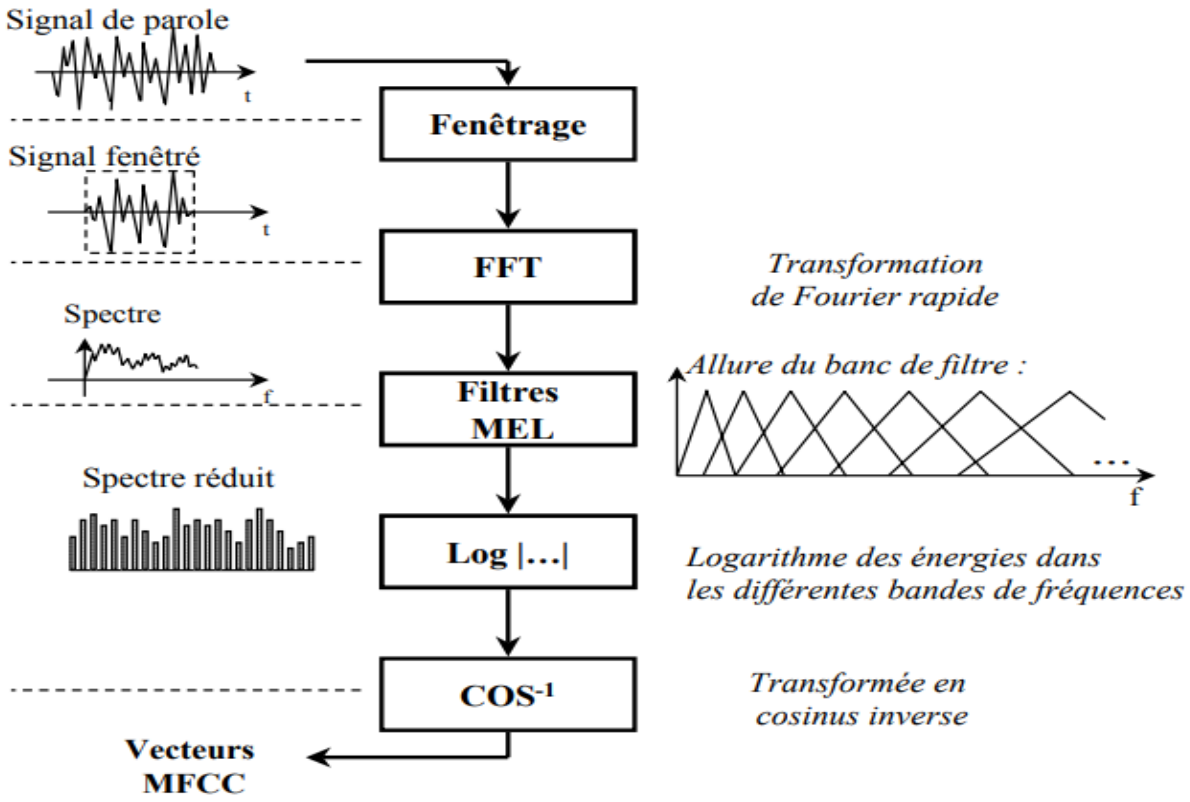



**Figure 2.5 :** schéma synoptique des étapes d'extraction des coefficients MFCC

### 2.3.1 Extraction des coefficients MFCC (Mel Frequency Cepstral Coefficients)

Le calcul des coefficients MFCC est réalisé de la manière suivante :

#### 2.3.1.1 Cadrez le signal en trames courtes (Fenêtrage)

Un signal audio change constamment, donc pour simplifier les choses, supposons que sur des échelles de temps courtes, le signal audio ne change pas beaucoup (lorsque nous disons qu'il ne change pas, nous entendons statistiquement, c'est-à-dire statistiquement stationnaire), évidemment, les échantillons changent constamment. Même des échelles de temps courtes. C'est pourquoi nous encadrons le signal dans des trames de 20 à 40ms [20].

❖ **Fenêtre rectangulaire :**

Elle est définie par :

$$f(nT)=\begin{cases}1 \xrightarrow{Si} |nT| < T'. \\ 0 \xrightarrow{Si} ailleurs\end{cases} \qquad (2.3)$$

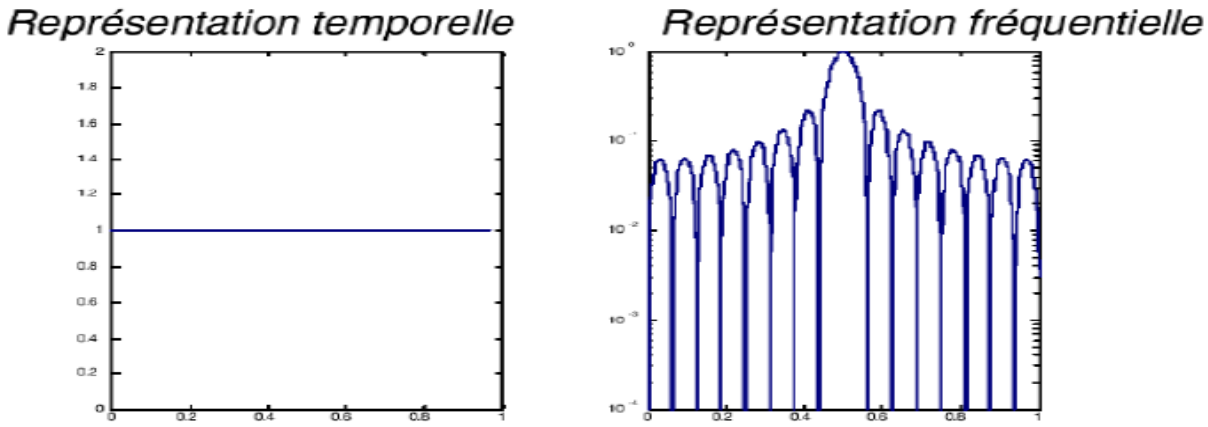



**Figure 2.6 :** Fenêtre rectangulaire et son spectre

❖ **Fenêtre de Hanning :**

Elle est définie par :

$$f(nT)=\begin{cases}0.5\left(1-\cos\left(\frac{n\pi T}{T'}\right)\right) \xrightarrow{Si} |nT| < T' \\ 0 \xrightarrow{Si} ailleurs\end{cases} \qquad (2.4)$$

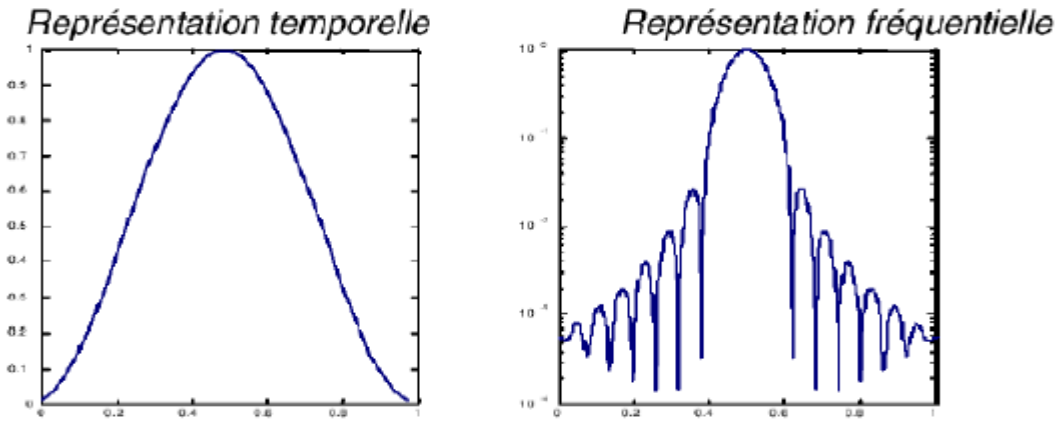



**Figure 2.7 :** Fenêtre de Hanning et son spectre

❖ **Fenêtre de Hamming :**

Elle est définie par :

$$f(nT)=\begin{cases} 0.54-0.46.\cos\left(\frac{\pi nT}{T'}\right) \xrightarrow{Si} |nT| < T'. \\ 0 \xrightarrow{Si} ailleurs \end{cases} \qquad (2.5)$$

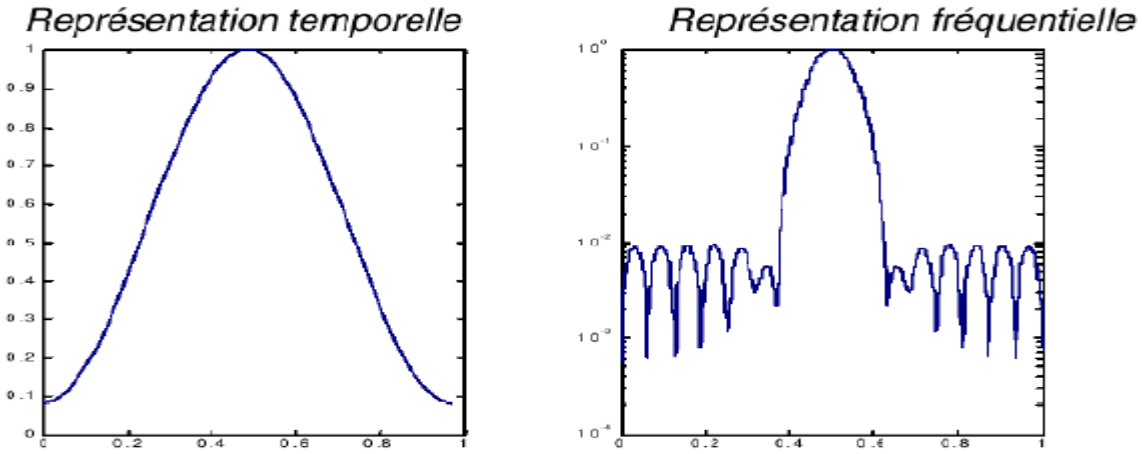



**Figure 2.8 :** Fenêtre de Hamming et son spectre

Parmi ces fenêtres, la fenêtre de Hamming est la plus convenable à la parole, car elle entraîne un minimum de distorsion spectrale du signal de parole, par rapport aux autres fenêtres. (Atténuation du rapport du lobe principal au lobe secondaire est égale à - 41dB, c'est-à-dire que la concentration d'énergie dans le lobe principal est égale à 99.96%).

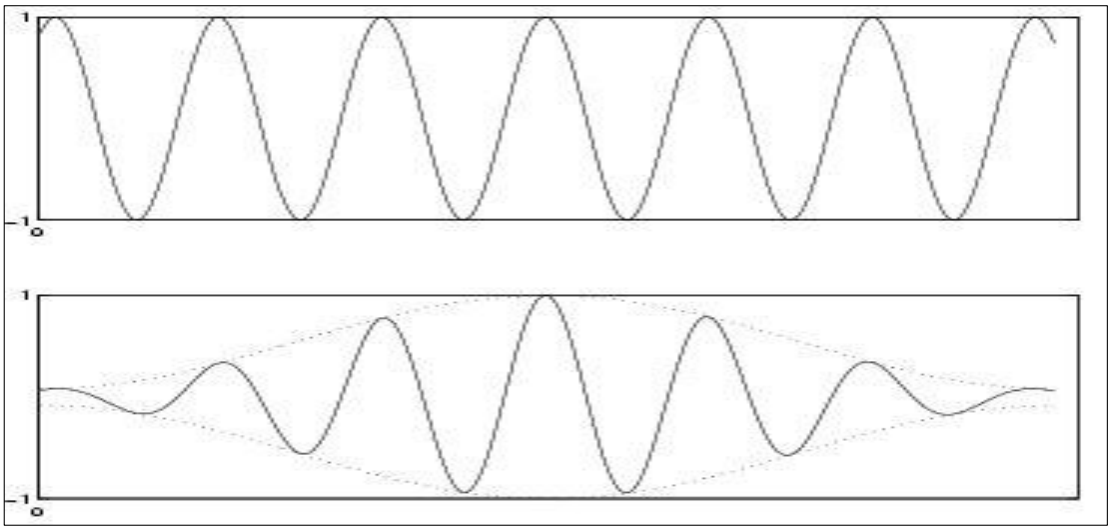


**Figure 2.9 :** Représentation d'un signal sinusoïdal pondéré par la fenêtre de Hamming

#### 2.3.1.2 FFT (Fast Fourier Transform)

La transformée de Fourier rapide (acronyme anglais : FFT) est un algorithme de calcul de la transformée de Fourier discrète (TFD). Ainsi, pour le temps de calcul de l'algorithme rapide peut être 100 fois plus petit que le calcul utilisant la formule de définition de la TFD.
Cet algorithme est couramment utilisé en traitement numérique du signal pour transformer des données du domaine temporel au domaine fréquentiel [19].

#### 2.3.1.3 Calculez l'estimation par périodogramme du spectre de puissance

Cette étape consiste à calculer le spectre de puissance de chaque image. Ceci est motivé par la cochlée humaine (un organe dans l'oreille) qui vibre à différents endroits en fonction de la fréquence des sons entrants. Selon l'emplacement de la cochlée qui vibre (qui bouge de petits poils), différents nerfs se déclenchent pour informer le cerveau que certaines fréquences sont présentes. Notre estimation de périodogramme effectue un travail similaire pour nous en identifiant les fréquences présentes dans la trame [19].

#### 2.3.1.4 Appliquez le banc de filtres mel au spectre de puissance

L'estimation spectrale par périodogramme contient encore beaucoup d'informations non requises pour la reconnaissance vocale automatique. En particulier, la cochlée ne peut pas discerner la différence entre deux fréquences rapprochées. Cet effet devient plus prononcé lorsque les fréquences augmentent. Pour cette raison, nous prenons des groupes de groupes de périodogrammes et les résumons pour avoir une idée de la quantité d'énergie disponible dans diverses régions de fréquence. Ceci est effectué par notre banque de filtres Mel : le premier filtre est très étroit et donne une indication de la quantité d'énergie disponible près de 0 Hertz. À mesure que les fréquences augmentent, nos filtres s'élargissent à mesure que nous nous inquiétons moins des variations. Nous ne sommes intéressés que par la quantité d'énergie produite à chaque endroit. L'échelle Mel nous dit exactement comment espacer nos bancs de filtres et quelle largeur leur donner [19].

#### 2.3.1.5 Prenez le logarithme de toutes les énergies de banque de filtres

Une fois que nous avons les énergies du banc de filtres, nous en prenons le logarithme. Ceci est également motivé par l'audition humaine : nous n'entendons pas le volume sur une échelle linéaire. Généralement, pour doubler le volume perçu d'un son, il faut y mettre 8 fois plus d'énergie. Cela signifie que les fortes variations d'énergie peuvent ne pas sembler très différentes si le son est fort au départ. Cette opération de compression rend nos fonctionnalités plus proches de ce que les humains entendent réellement [19].

**Pourquoi le logarithme et non une racine de cube ?** Le logarithme nous permet d'utiliser la soustraction de la moyenne cepstrale, qui est une technique de normalisation de canal.

#### 2.3.1.6 Prenez le DCT des énergies de la banque de filtres log

La dernière étape consiste à calculer la DCT des énergies de la banque de filtres de journaux. Cela s'explique principalement par 2 raisons. Parce que nos bancs de filtres se chevauchent tous, les énergies des bancs de filtres sont très corrélées les unes aux autres. La DCT décorrèle les énergies, ce qui signifie que les matrices de covariance diagonale peuvent être utilisées pour modéliser les caractéristiques [19].

Mais notez que seuls 12 des 26 coefficients DCT sont conservés. En effet, les coefficients DCT les plus élevés représentent des changements rapides dans l'énergie du banc de filtres et il s'avère que ces changements rapides dégradent les performances RAP. Nous obtenons donc une légère amélioration en les abandonnant.

### 2.3.2 L'échelle de Mel

L'échelle Mel relie la fréquence perçue, ou hauteur d'un ton pur, à la fréquence réelle mesurée. Les humains sont beaucoup mieux à même de discerner les petits changements de hauteur dans les basses fréquences que dans les hautes fréquences. En incorporant cette échelle, nos fonctionnalités correspondent plus étroitement à ce que les humains entendent [21].

La formule de conversion de la fréquence en échelle de Mel est la suivante :

$$B(f) = 2595\log\left(1 + \frac{f}{700}\right) \qquad (2.6)$$

Où :

f : est la fréquence en Hz.

B(f) : est la fréquence Mel-échelle de f.

Pour revenir de Mels à la fréquence :

$$B^{-1}(b) = 700 * \left(10^{b/2595} - 1\right) \qquad (2.7)$$

### 2.3.2.1 Calcul des coefficients dans l'échelle MEL

Soit un signal discret {s[n]} avec 0 < n < N-1, N est le nombre d'échantillons d'une fenêtre analysée, Fs est la fréquence d'échantillonnage, la transformée de Fourier discrète S[k] est obtenu :

$$S[k]=\sum_{n=0}^{N-1} s[n]e^{-j2\pi nk/N} \qquad avec \quad 0\le k<N \qquad (2.8)$$

Le spectre du signal est multiplié avec des filtres triangulaires (figure. 2.10) dont les bandes passantes sont équivalentes en domaine mel-fréquence. Les points frontières B[m] des filtres en Mel-fréquence sont calculés ainsi :

$$B[m]=B(f_1)+m\frac{B(f_h)-B(f_1)}{M+1} \qquad 0\le m\le M+1 \qquad (2.9)$$

Avec :

M : le nombre de filtres.

$f_h$ : la fréquence la plus haute.

$f_1$ : 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 :

$$f[m]=\left[\frac{N}{F_S}\right]B^{-1}\left[B(f_1)+m\frac{B(f_h)-B(f_1)}{M+1}\right] \qquad (2.10)$$

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

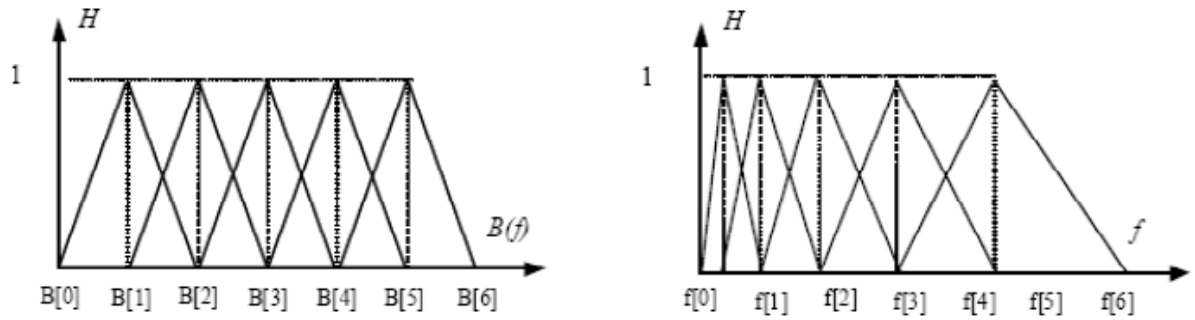



**Figure 2.10 :** Les filtres triangulaires passe-bande en Mel-fréquence et en fréquence [22].

#### 2.3.2.2 Analyse Cepstrale

Les coefficients en sortie des bancs de filtres peuvent être utilisés pour mesurer des différences entre deux trames. Ils présentent cependant des inconvénients comme de dépendre de l'énergie du signal ou de l'excitation. La transformation cepstrale permet d'obtenir une information normalisée. L'analyse cepstrale est une méthode basée sur le modèle de production de la parole. Le signal de la parole peut être représenté par la convolution de la source (cordes vocales) et du filtre (canal buccal) dans le domaine temporel comme suit :

$$s(t) = e(t) \otimes h(t) \qquad (2.11)$$

On passe dans le domaine fréquentiel pour obtenir l'enveloppe spectrale qui permet de faire apparaître les différences de fréquences. La convolution devient donc une multiplication :

$$S(f) = E(f) \cdot H(f) \qquad (2.12)$$

On souhaite séparer la source du filtre pour récupérer l'enveloppe spectrale du signal. Pour cela, on utilise la fonction log :

$$\log(|S(f)|) = \log(|E(f)|) + \log(|H(f)|) \qquad (2.13)$$

On applique ensuite la transformée inverse pour obtenir les coefficients temporels appelés coefficients cepstraux

### 2.3.3 Les coefficients Cepstraux

C'est l'étape finale, on transforme les données dans l'échelle des Mels–fréquentielle donc vers l'échelle des temps. Le résultat de cette étape sera les MFCC proprement dit. Il suffit d'effectuer l'inverse de la transformée de Fourier **FFT$^{-1}$** ou la transformée en cosinus inverse **DCT$^{-1}$**, ce qui revient au même puisque la transformée en Cosinus inverse donne la partie réelle de la transformée de Fourier [23] [24],

$$s(n) \xrightarrow{\text{FFT}} S(f) \xrightarrow{Log|\ |} Log|S(f)| \xrightarrow{\text{FFT}^{-1}} \text{cepstre}$$

**Figure 2.11 :** Différentes étapes de l'analyse cepstrale.

Les coefficients cepstraux sont donnés par :

$$c(n) = \frac{1}{N}\sum_{j=0}^{N-1} Log(|S(j)|)\, e^{\frac{2ijn}{N}} \qquad pour\ n = 0, 1, \dots, N-1 \qquad (2.14)$$

## 2.4 Dynamic Time Warping (DTW)

Il s'agit d'une méthode apparue dans les années 80 dans le domaine du traitement de la parole et encore utilisée dans des systèmes de reconnaissance vocale disposant de ressources matérielles limitées. Dans les systèmes de reconnaissance basés sur la DTW, chaque mot du lexique est représenté par une réalisation de référence. Le processus de reconnaissance consiste à évaluer la distance d'une observation à chacune des références.

L'alignement temporel de différents énoncés est le principal problème de la mesure de distance en reconnaissance vocale. Un petit changement entraîne une identification incorrecte. DTW est une méthode efficace pour résoudre le problème de l'alignement temporel.

L'algorithme vise à aligner deux séquences de vecteurs caractéristiques en déformant l'axe temporel de manière répétitive jusqu'à ce qu'une correspondance optimale soit trouvée entre les deux séquences. Cet algorithme effectue une cartographie linéaire par axe de l'axe des temps pour aligner les signaux [25].

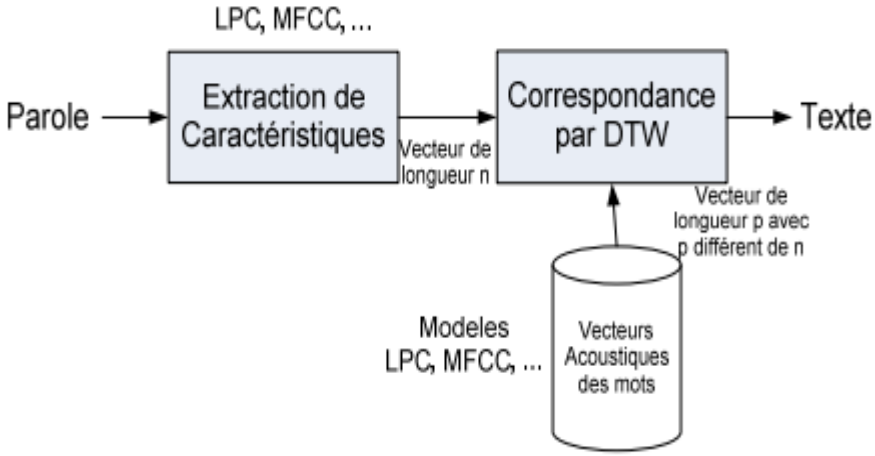



**Figure 2.12 :** Schéma d'un système de reconnaissance basé sur la comparaison dynamique (DTW) [27].

La DTW est reconnu récursivement par :

$$D(A_i, B_j) = \delta(a_i, b_j) + \min \begin{cases} D(A_{i-1}, B_{j-1}), \\ D(A_i, B_{j-1}), \\ D(A_{i-1}, B_j) \end{cases} \qquad (2.15)$$

Ou :

A : séquence stocker ou enregister et représente la sous séquence {a1 ,….,ai}

B : séquence test et représente la sous séquence {b1,….,bi}

i : amplitudes de la séquence A

j : amplitudes de la séquence B

D : la distance minimale entre les séquences A et B.

On peut voir l'alignement entre deux signaux temporels en utilisant la DTW dans la figure suivante :

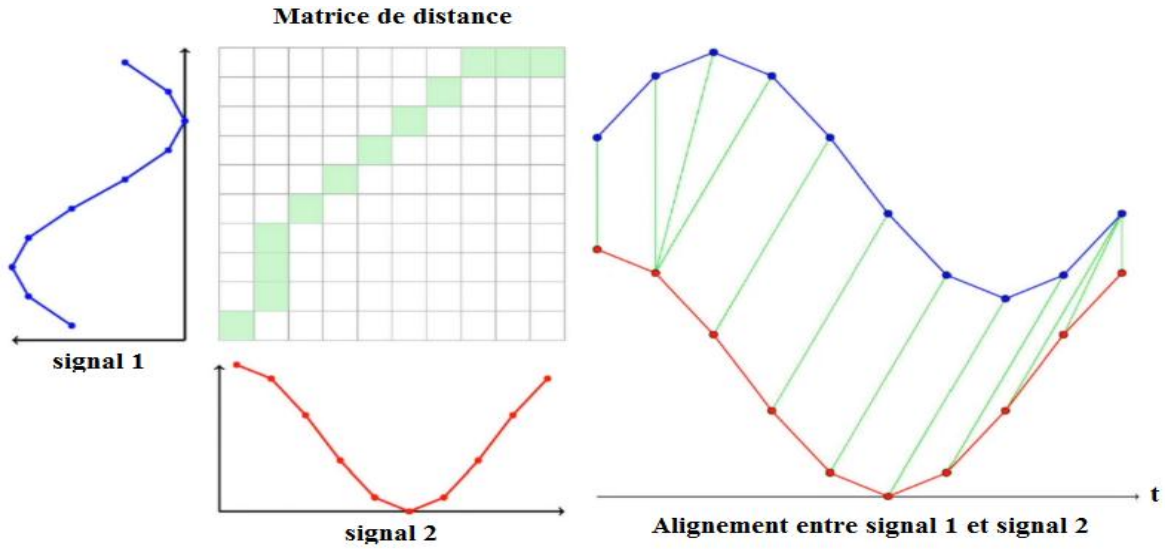



**Figure 2.13 :** l'alignement entre deux signaux temporels en utilisant la DTW

Pour bien comprendre le fonctionnement de l'algorithme DTW on donne un petit exemple qui vise à extraire d'une façon analytique le meilleur chemin d'alignement entre deux séquence temporelles A et B.

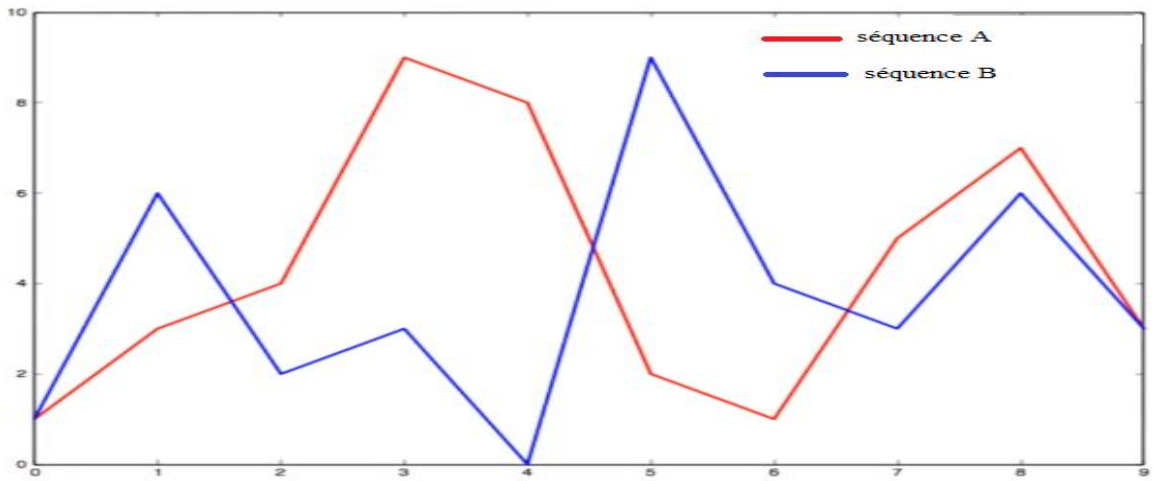



**Figure 2.14 :** deux séquence temporelles A et B

Tout d'abord on calcule la matrice de distance entre les deux séquences en se basant sur la formule analytique dite auparavant (2.15) :

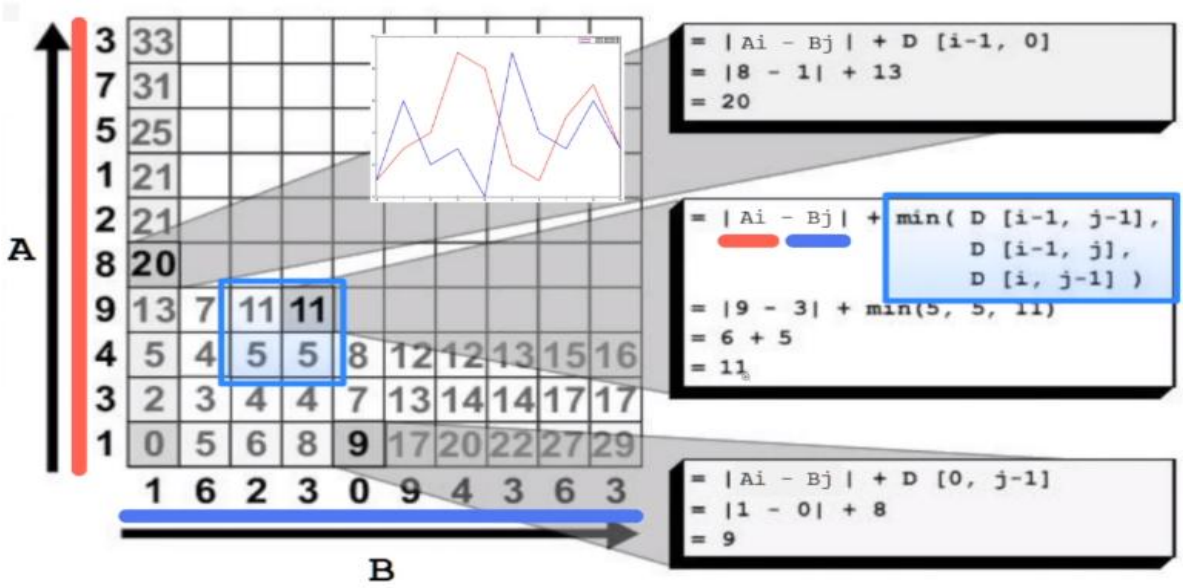



**Figure 2.15 :** calcule de la matrice de distance entre les deux séquences A et B

Et la phase finale c'est de faire la somme des paramètres de la matrice représenté dans la figure :

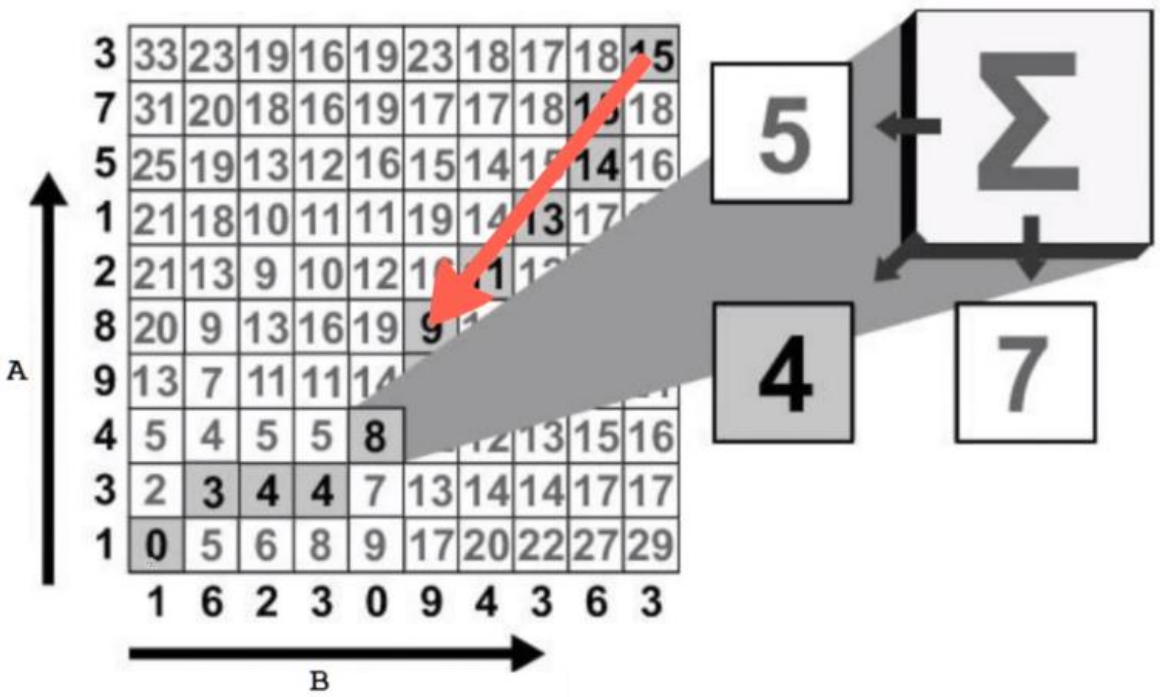



**Figure 2.16 :** calcule de la distance minimale entre les deux séquences

D= 15+15+14+13+11+9+6+8+4+4+3+0

D= 102

La distance retenue D est celle correspondant à l'alignement de coût minimal. Rapide dans des tâches à petit vocabulaire, cette technique a un certain nombre d'inconvénients importants qui limitent son champ d'application. D'une part, la modélisation des mots par une instance est très peu robuste à l'ensemble des variabilités acoustiques. Cette faiblesse peut être partiellement limitée par l'utilisation de plusieurs références par mot, par un choix plus fin des références ou encore par l'usage de distances spectrales robustes, cette technique est plus adaptée à un contexte d'utilisation mono-locuteur en environnement peu bruité.
D'autre part, la complexité des modèles et du décodage sont proportionnels à la taille du lexique, ce qui exclut l'utilisation de la DTW dans des systèmes grand vocabulaire. Enfin, bien que diverses extensions à la reconnaissance de la parole continue aient été expérimentées, cette méthode ne permet, dans sa version standard, que la reconnaissance de mots isolés [26] [27].

## 2.5 Conclusion

Ce chapitre constitue à connaitre les méthodes ou les techniques utilisé pour la reconnaissance de mots isolés MFCC et DTW.
L'extraction des caractéristiques a été réalisée à l'aide des coefficients de fréquence de Mel Melquency (MFCC) et la correspondance des caractéristiques a été réalisée à l'aide de la technique de Dynamic Time Warping (DTW).
Les fonctionnalités extraites ont été stockées dans un fichier.wav à l'aide de l'algorithme MFCC. Une mesure de distorsion basée sur la réduction de la distance euclidienne a été utilisée pour faire correspondre le signal vocal inconnu à la base de données de signaux vocaux.

## 3.1 Introduction

L'objectif de notre travail est de trouver les meilleurs paramètres caractéristiques du signal vocal pour avoir un Taux de Reconnaissance plus élevé. Nous allons présenter en premier lieu le logiciel d'acquisition du signal vocal ainsi que les différentes étapes de traitement à savoir les choix des paramètres d'acquisition ainsi que la segmentation manuelle, pour avoir les meilleurs séquences test qui nous aiderons à faire bien marcher notre programme. Par la suite, nous parlerons des algorithmes utilisés pour l'extraction des paramètres MFCC avec une variante basée sur la FFT, et la DCT. Ensuite nous calculons la distance minimale ou l'alignement le plus proche entre les signaux test et les signaux enregistrer en utilisant la DTW.

## 3.2 Architecture du système de reconnaissance

L'architecture du système de reconnaissance comprend les modules suivants (figure 3.1).

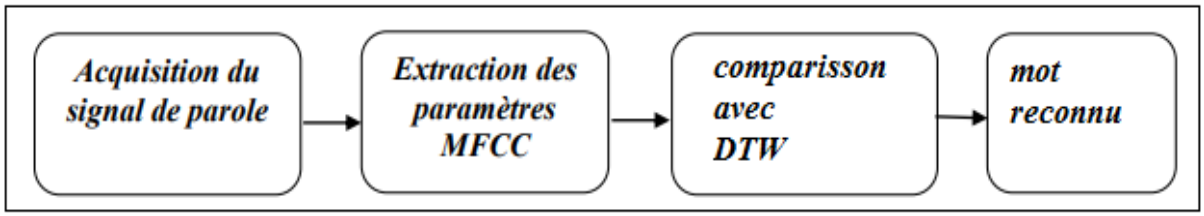



**Figure 3.1 :** Schéma global du système de reconnaissance des mots isolés

Nous avons utilisé pour notre travail un micro-ordinateur muni d'une horloge de 2.30 GHz, 8.00 Go de mémoire vive (RAM), équipé d'une carte son et d'un microphone pour l'enregistrement du signal de la parole.

La programmation a été faite sous **MATLAB** version 7.5.0 (R2009a) qui est un environnement de calcul technique conçu pour le calcul numérique et la visualisation à haute performance.

Le MATLAB offre des familles d'applications réunies sauvegardées dans des boîtes à outils, donc, il s'avère être le logiciel le plus adéquat pour nos applications de par les avantages qu'il nous a offerts en termes de traitement du signal et l'extraction des paramètres que nécessite notre étude ou par la facilité qu'elles engendrent pour les applications des MFCC et DTW en vue de la reconnaissance des mots isolés.

## 3.3 Les étapes de traitement

Notre travail se compose de deux parties essentielles :

- L'extraction des paramètres MFCC.
- L'alignement temporelle DTW.

La première partie comporte les étapes suivantes :

- Le prétraitement de signal vocal.
- L'extraction des paramètres caractérise le signal acquis tels que les coefficients cepstraux.

Dans la deuxième partie, les étapes effectuées sont :

- La comparaison ou l'alignement temporelle entre le signal test et le signal enregistré
- L'apprentissage et la reconnaissance.

### 3.3.1 Acquisition des données

Le signal vocal est recueilli par des capteurs transformant ce dernier en un signal électrique analogique. Il est échantillonné afin de faciliter son traitement. L'acquisition des données consiste à enregistrer les mots isolés du corpus choisi, en utilisant le logiciel Speech Filing System (SFS) (figure 3.2).

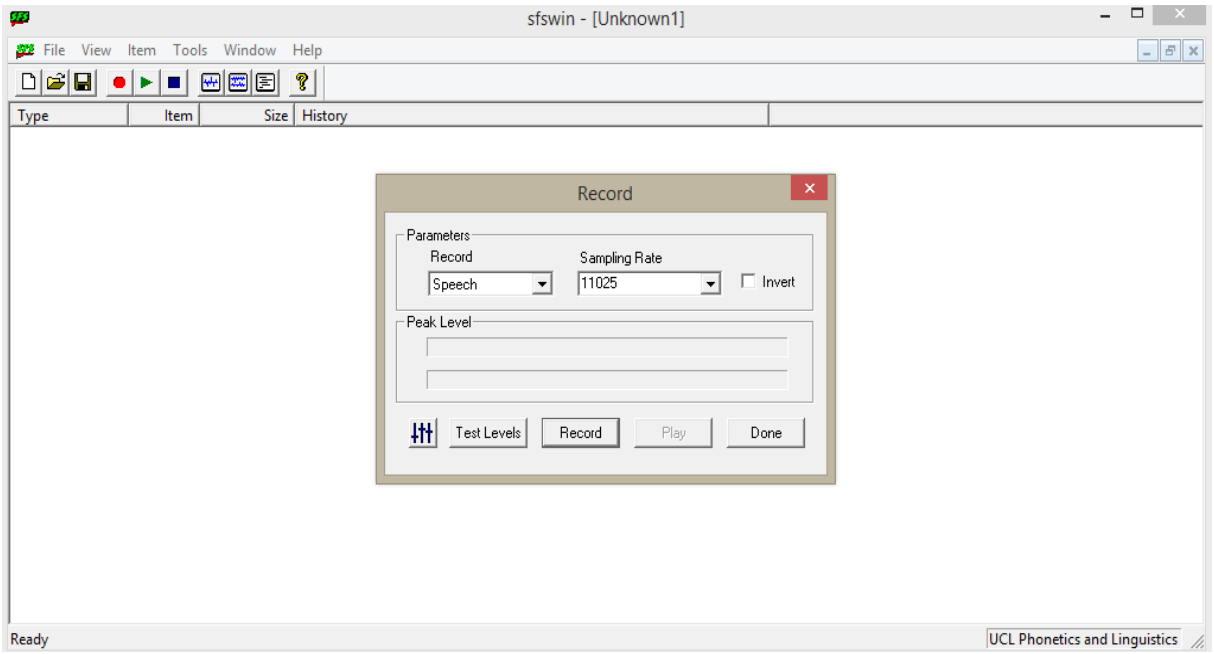



**Figure 3.2 :** Fenêtre principale du logiciel SFS

Nous avons choisi la fréquence d'échantillonnage de 11025 Hz, les échantillons ont été codés sur 16 bits par échantillon. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole.

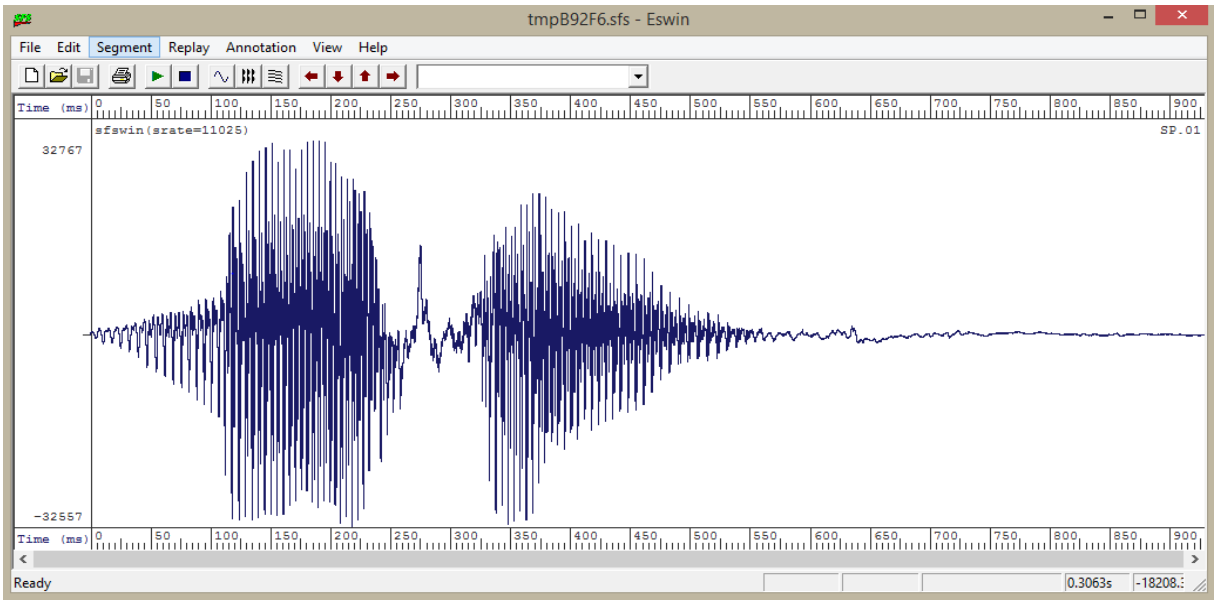



**Figure 3.3 :** Signal vocal du mot isolé [maison]

### 3.3.2 Segmentation des mots isolés

La phase de segmentation joue un rôle très important dans les systèmes de RAP. Chaque segment représente un mot isolé. Pour notre travail, la segmentation a été effectuée manuellement, car elle nous offre la possibilité de limiter la portion du signal à analyser. La segmentation des 4 mots isolés est sauvegardée dans un fichier Speech Filing System (SFS), après l'avoir convertie dans un fichier (*.wav), ensuite transférée vers l'espace de travail du MATLAB (fig 3.4).

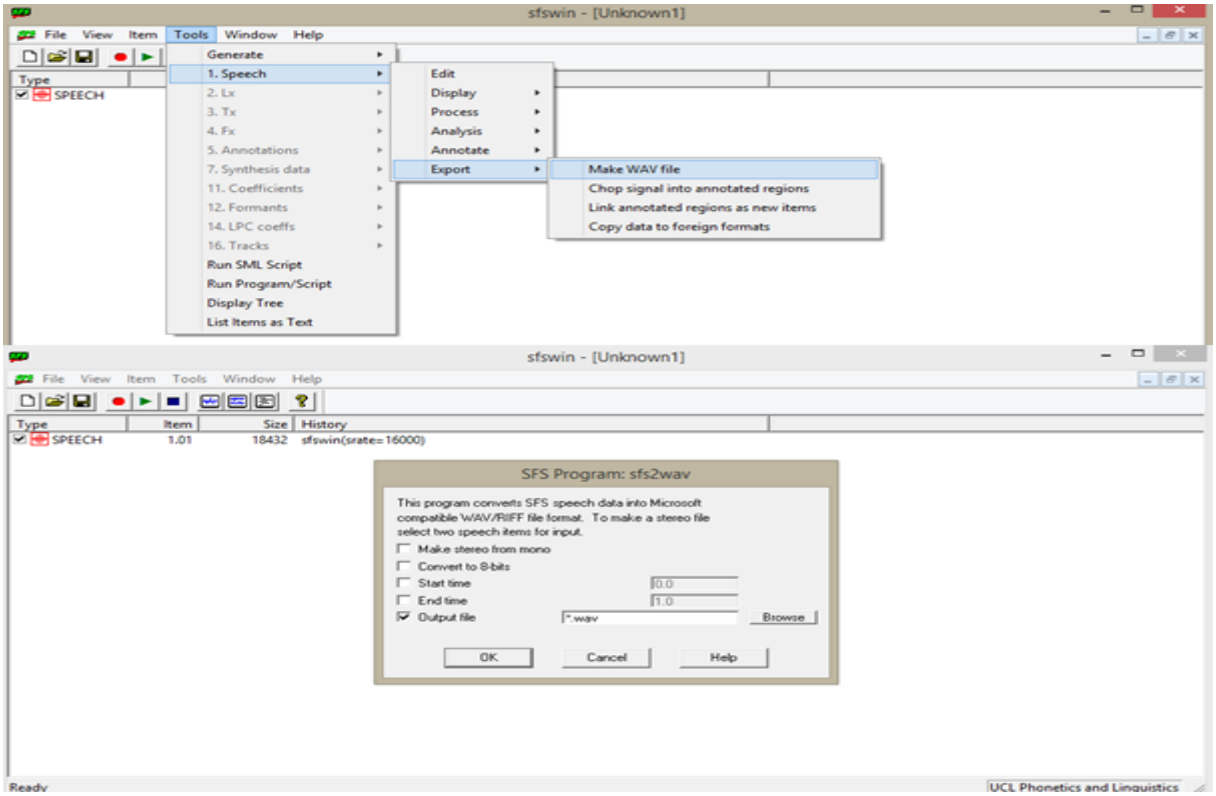



**Figure 3.4 :** Interface du logiciel SFSWAV

### 3.3.3 Prétraitement du signal vocal

Pour extraire les paramètres pertinents du signal vocal, une préaccentuation est effectuée sur le signal échantillonné (filtré), ensuite segmenté (ou fenêtré) en trames, est appliquée toutes les 10 ms sur des fenêtres de 20 ms, en utilisant une fenêtre temporelle. Notre choix s'est porté sur la fenêtre de Hamming. Cette dernière est effectuée, afin de réduire les effets de bords, avec un recouvrement de moitié pour éviter toute perte d'informations du signal à étudier, ainsi une meilleure représentation du signal est donnée. La largeur de celle-ci est de 256 échantillons.

### 3.3.4 Calcul des paramètres MFCC

La Reconnaissance Automatique de la Parole (RAP) repose sur l'extraction des paramètres du signal acquis. La méthode d'analyse que nous avons choisie est l'analyse cepstrale (chapitre 2), en prenant les paramètres de MFCC, le calcul de ces derniers est schématisé par un organigramme (fig 3.5).

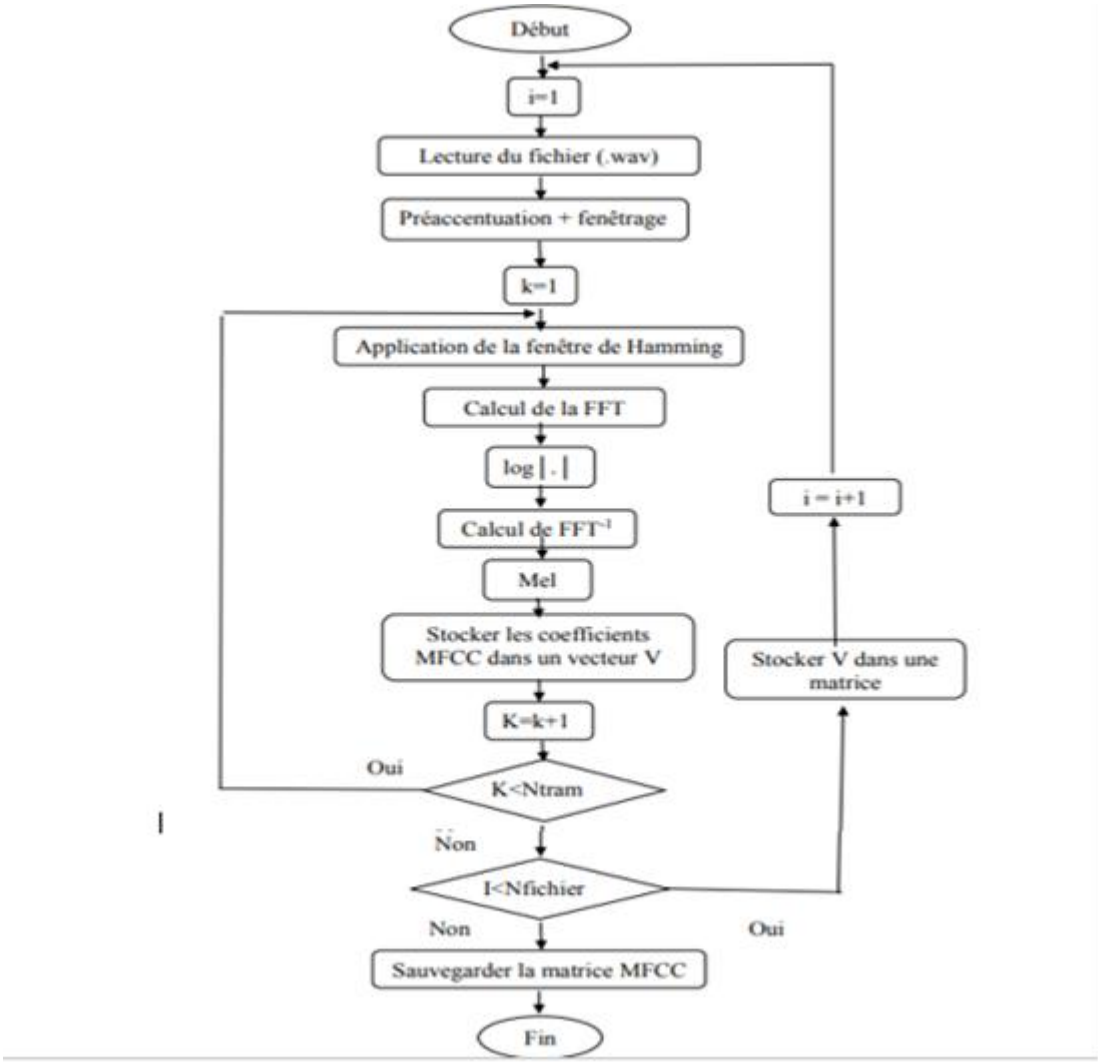



**Figure 3.5 :** Organigramme de l'extraction des paramètres MFCC

Avec : k : numéro de la trame.

Ntram : nombre de la trame.

i : numéro de fichier à traiter.

Nfichier : nombre de fichiers.

V : vecteur du paramètre MFCC

Logiciel Speech Filing System (SFS) nous permet aussi de faire l'analyse mathématique du signal enregistrer et il nous affiche les courbes selon nos besoins. Dans notre cas on fait subir le signal a un banc de filtre triangulaire pour extraire les coefficients MFCC. Les figure (3.6) et (3.7) explique comment procéder pour faire subir le signal dans la figure (3.3) a un banc de filtre et l'analyser avec la méthode MFCC.

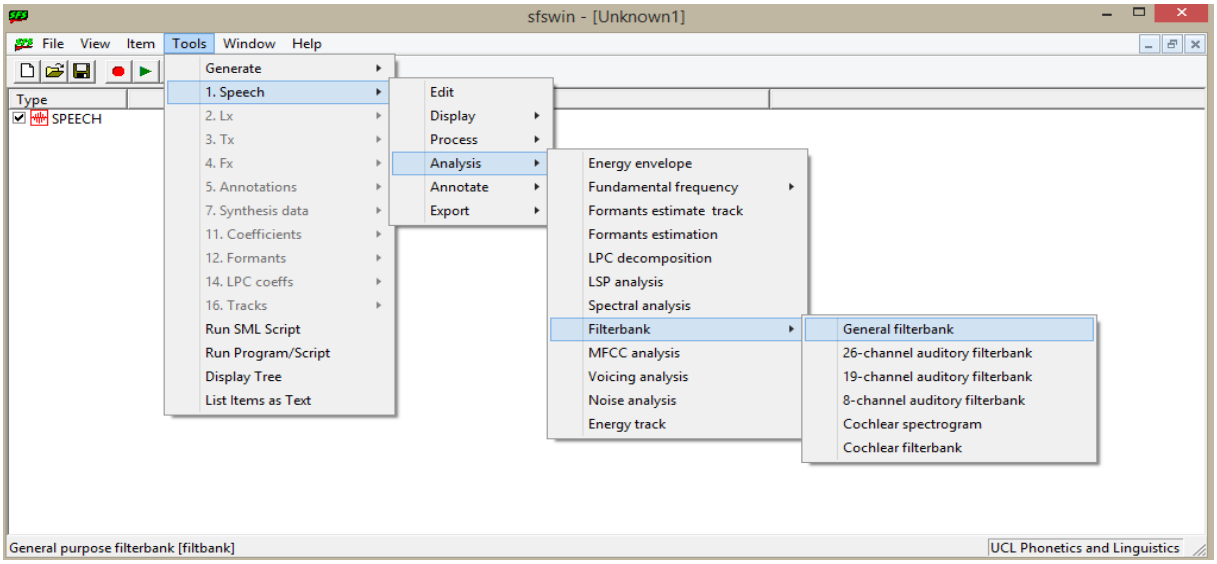



**Figure 3.6 :** Analyser le signal avec un banc de filtre triangulaire

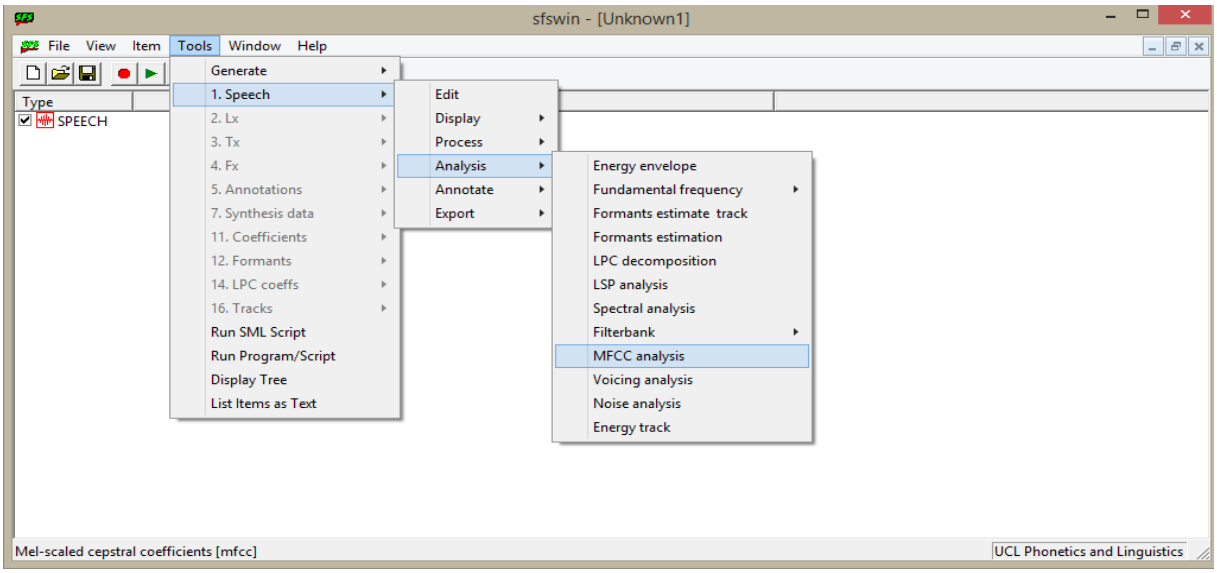



**Figure 3.7 :** analyser le signal avec la méthode MFCC

Cette analyse nous permet notamment de tracer les coefficients MFCC présente dans notre signal enregistrer et aussi le banc de filtre triangulaire.

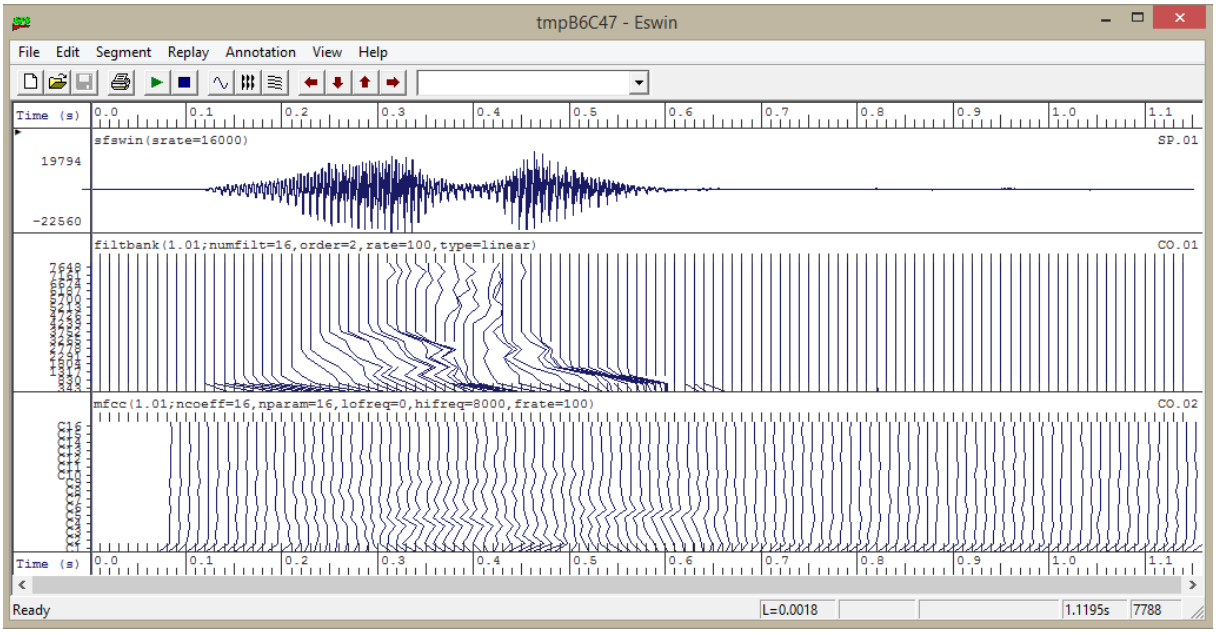



**Figure 3.8 :** graphes du signal du mot « Maison » le banc de filtres triangulaires et les coefficients MFCC

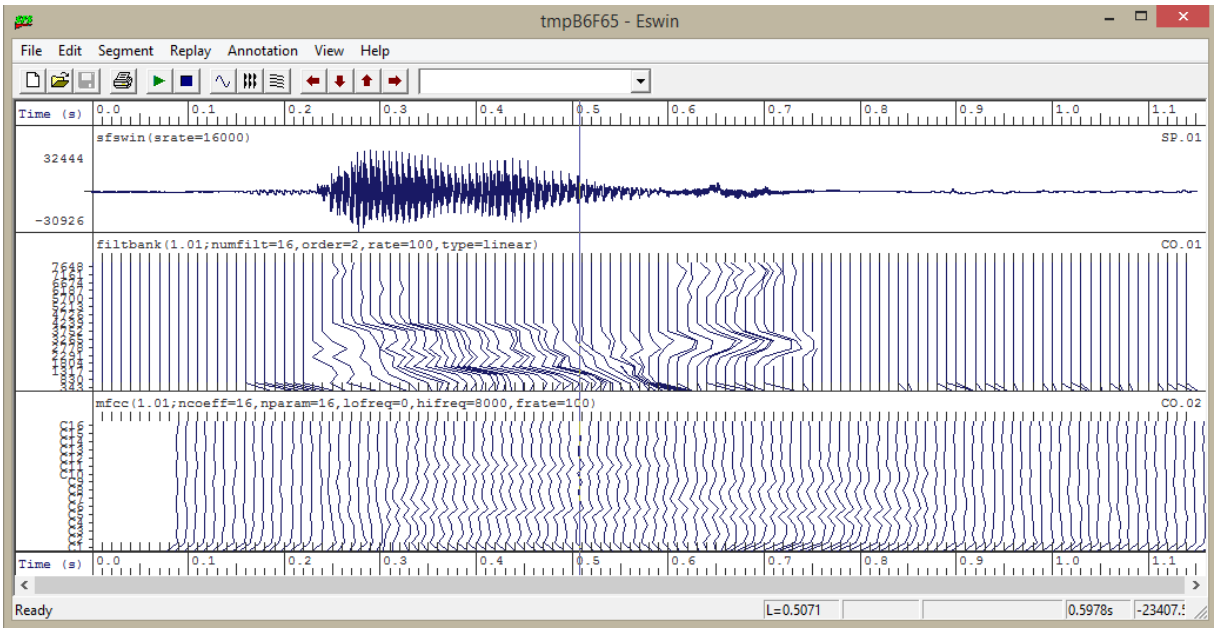



**Figure 3.9 :** graphes du signal du mot « Garage » le banc de filtres triangulaires et les coefficients MFCC

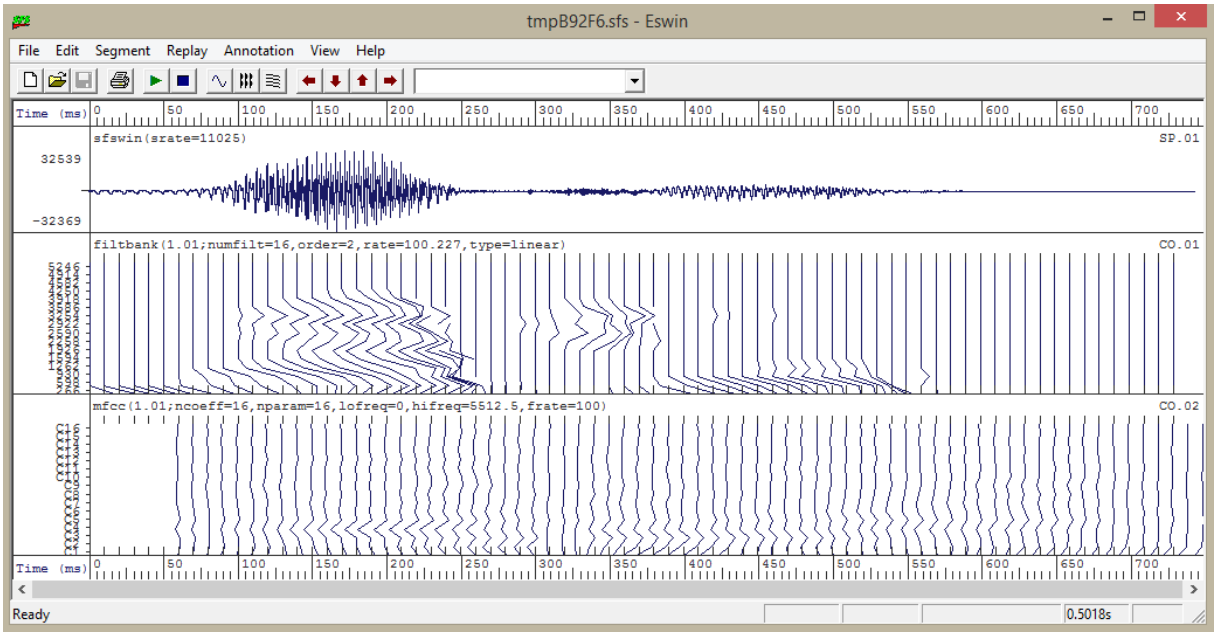



**Figure 3.10 :** graphes du signal du mot « Voiture » le banc de filtres triangulaires et les coefficients MFCC

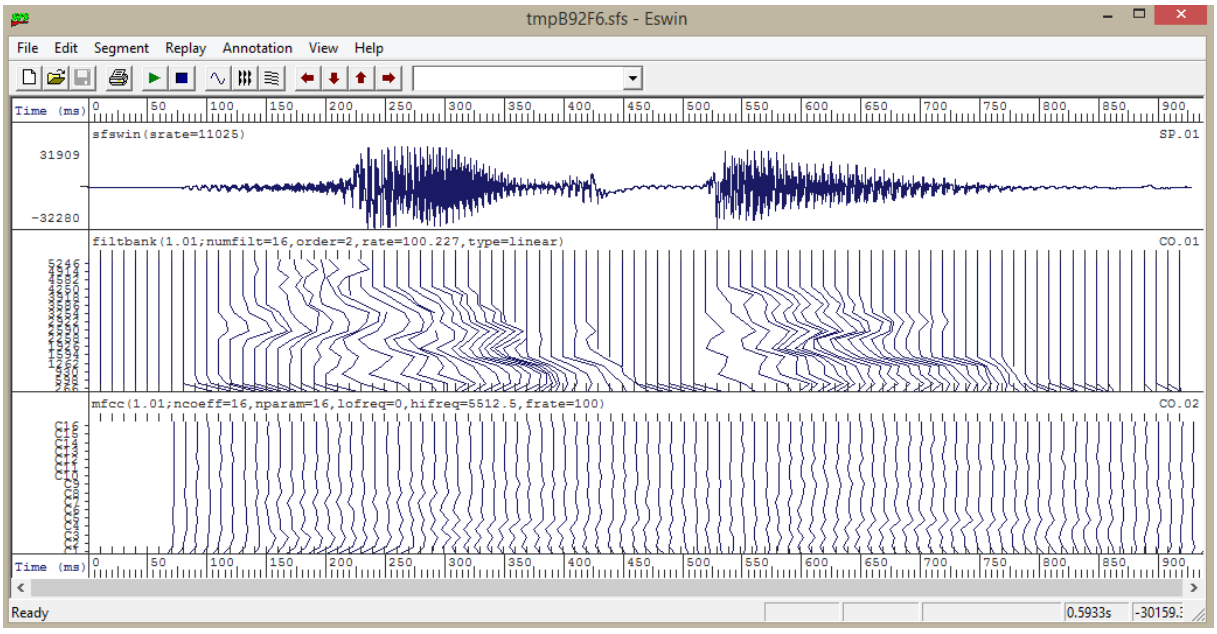
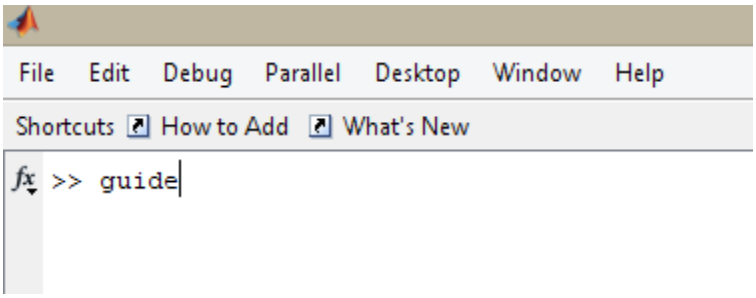



**Figure 3.11 :** graphes du signal du mot « Jardin » le banc de filtres triangulaires et les coefficients MFCC

## 3.4 Interfaces de l'application

Nous avons créé une interface d'application, qui représente et mis en valeur notre programme Matlab. Pour construire cette interface on écrit le mot « guide » dans l'espace de travail de Matlab :




Ensuite pour concevoir l'application on se bases sur les paramètres suivants :

- Un bouton de simulation (puch Button).
- Texte statique représenté par chaque enregistrement dans notre programme Matlab.
- Deux axes pour tracer les signaux tests et enregistrés respectivement

La figure (3.12) représente notre guide pendant notre travail :

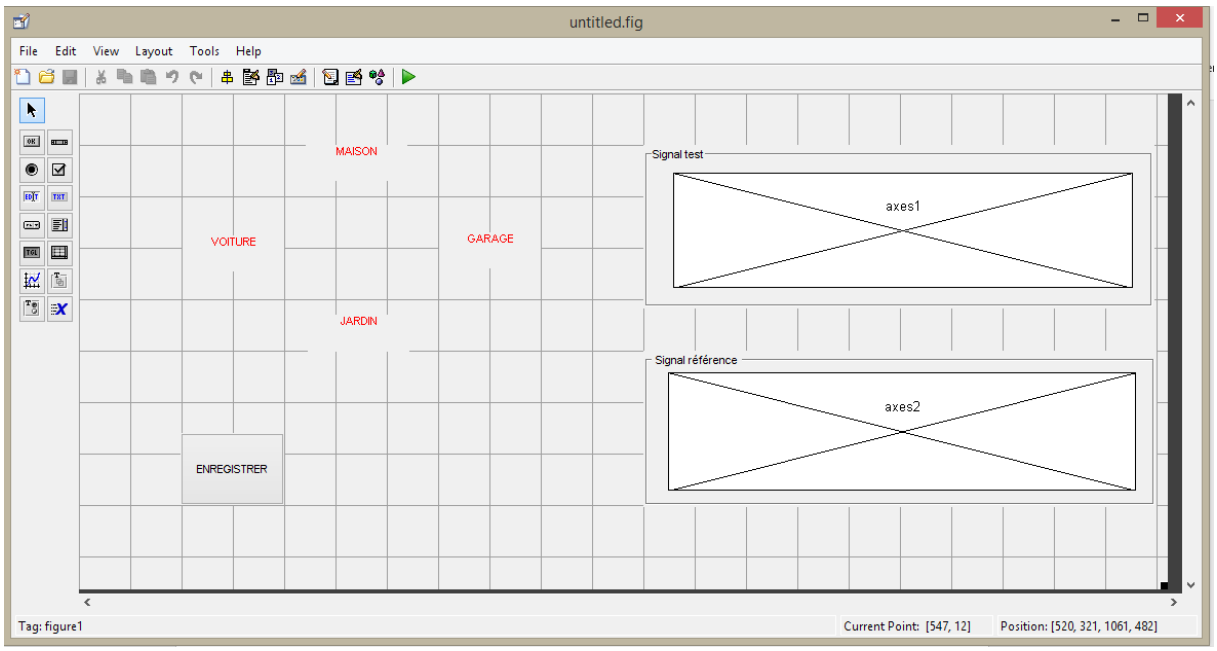



**Figure 3.12 :** Création de l'interface

## 3.5 Résultats expérimentaux

Finalement après avoir conçu notre interface on l'enregistre dans le dossier comportant nos codes Matlab, et on fait la simulation (Run). Notre application s'affiche comme suit :

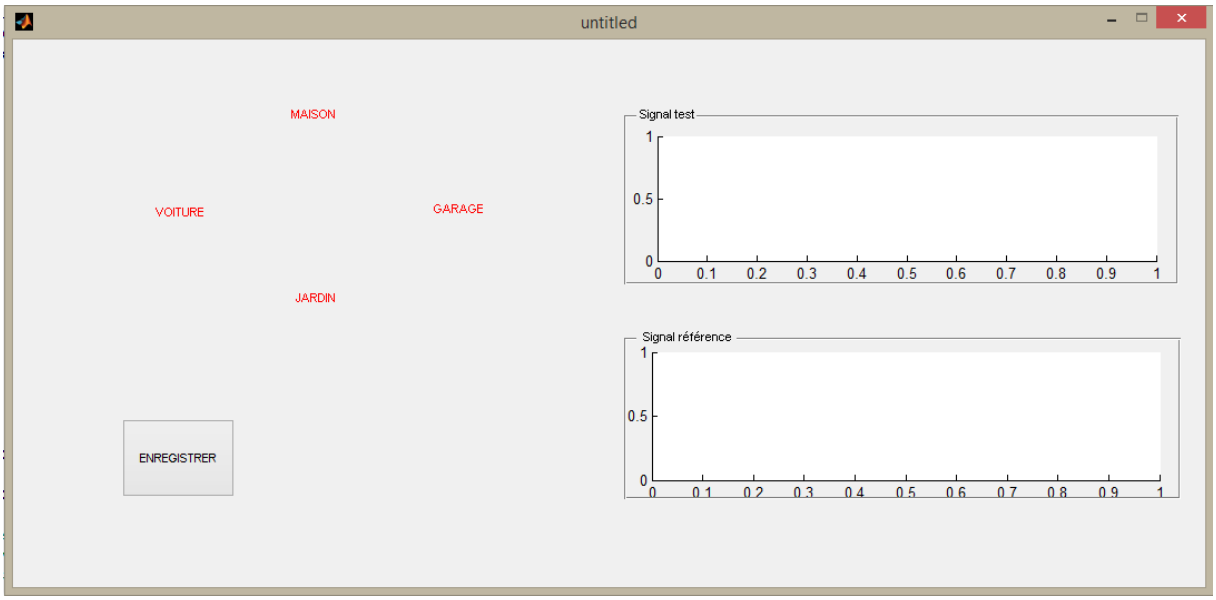



**Figure 3.13 :** L'interface du programme Matlab

On appuis sur le bouton « Enregistrer » et on prononce un des mots isolés déjà enregistrer auparavant :

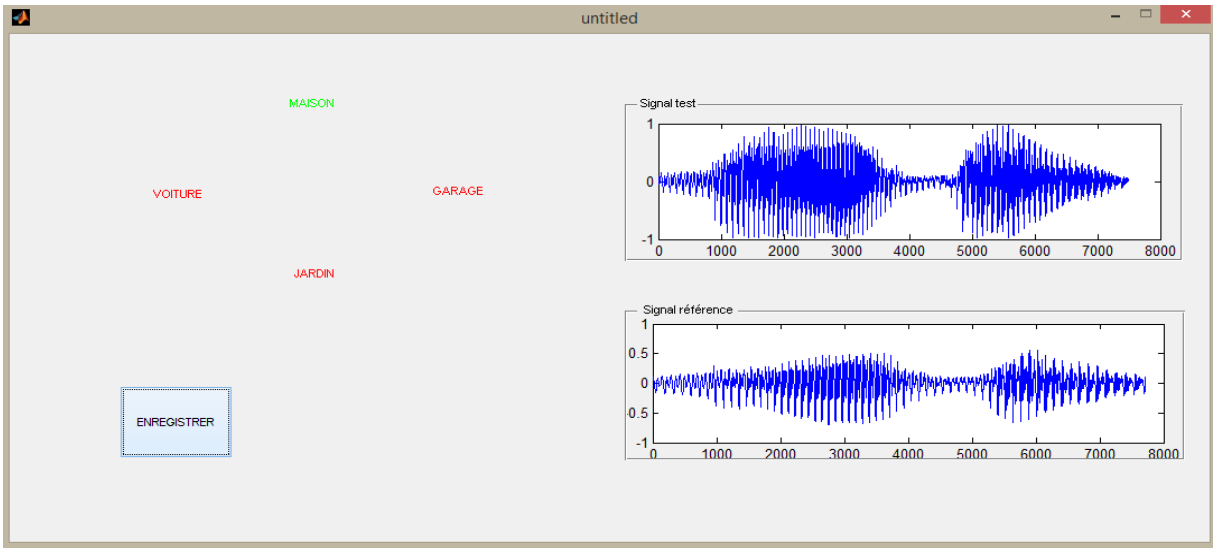



**Figure 3.14 :** Résultat après la prononciation du mot « Maison »

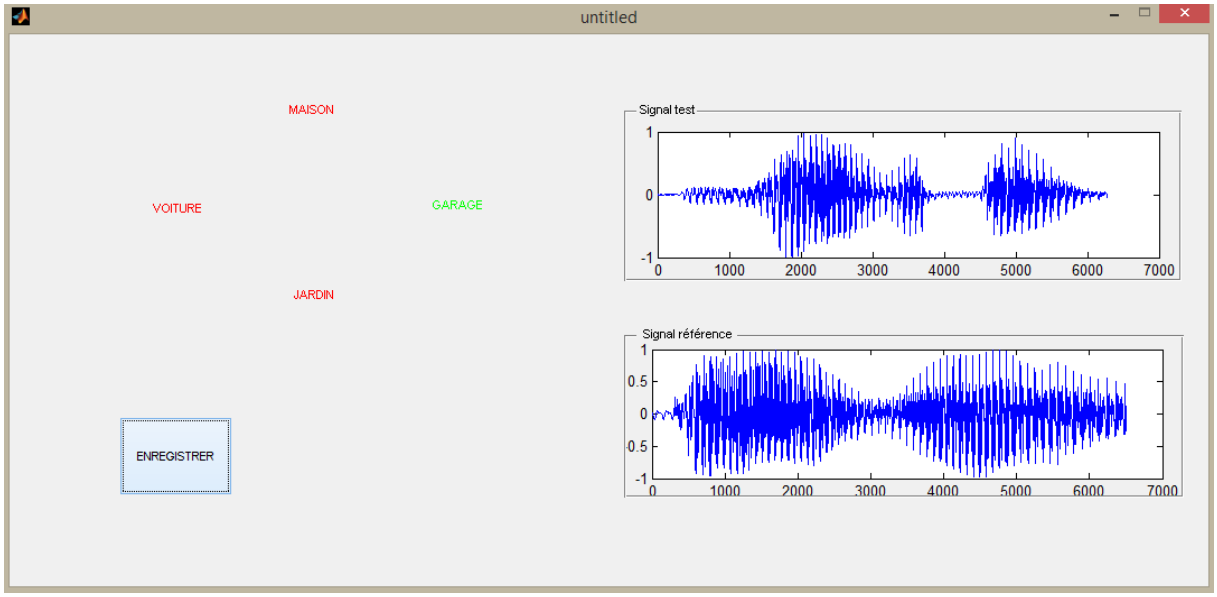



**Figure 3.15 :** Résultat après la prononciation du mot « Garage »

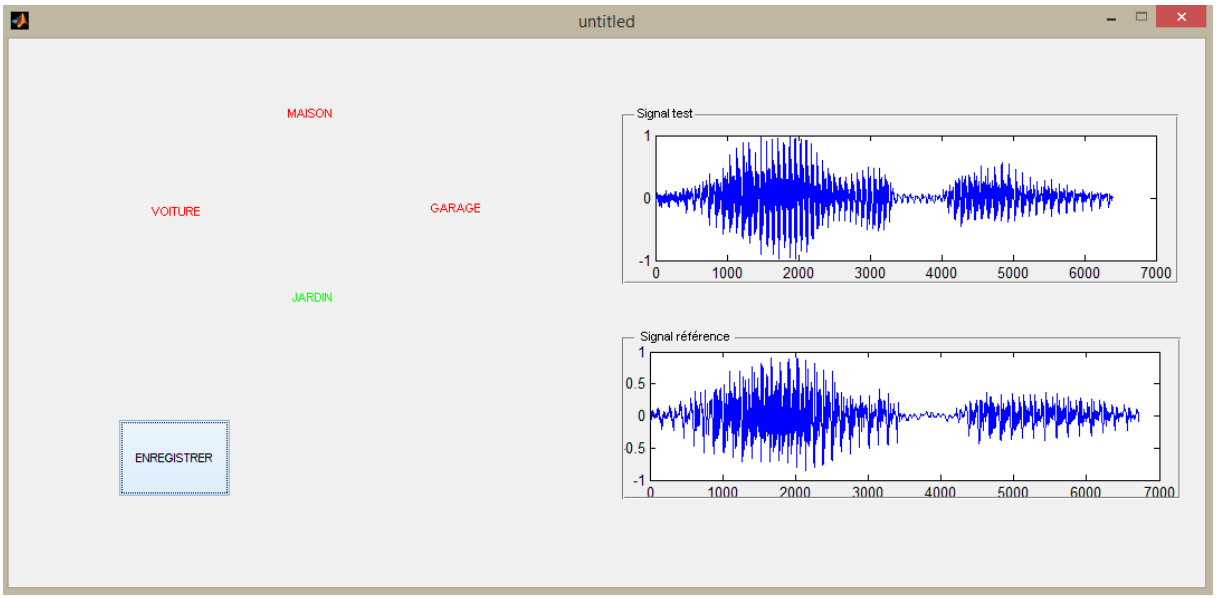



**Figure 3.16 :** Résultat après la prononciation du mot « Jardin »

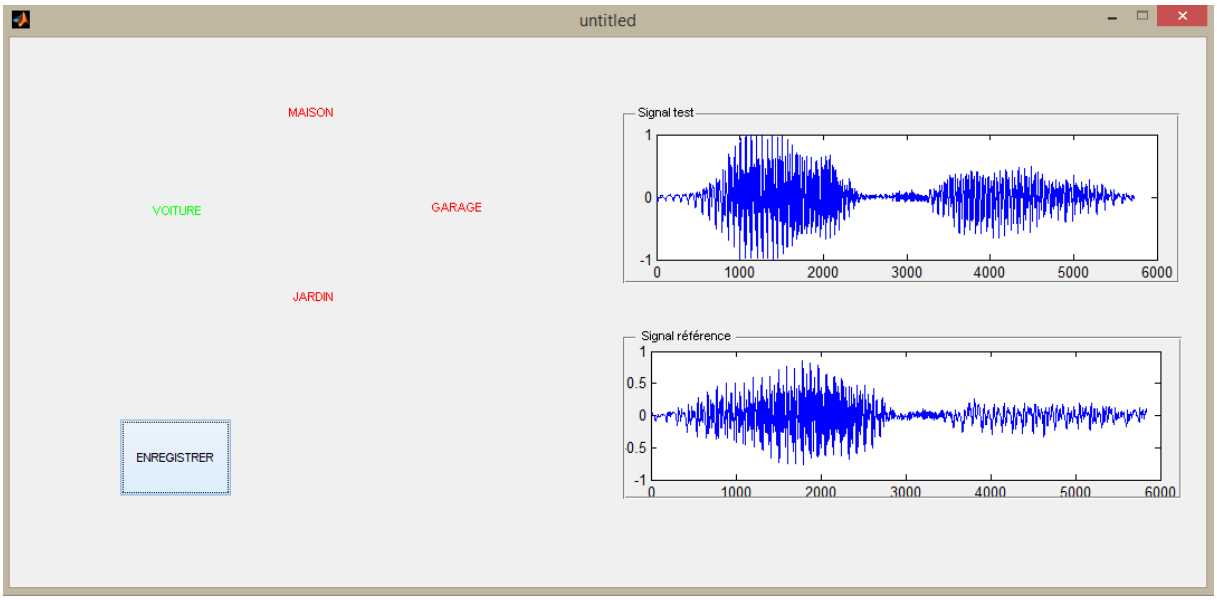



**Figure 3.17 :** Résultat après la prononciation du mot « Voiture »

**Commentaire :**

Les deux méthodes qu'on a appliqués dans notre programme nous ont permet d'avoir des résultats acceptables, vu le potentiel instrumental réduit qu'on a utilisé.

## 3.6 Conclusion

L'extraction des caractéristiques se fait par la méthode MFCC présentée précédemment. Le résultat de cette étape est un ensemble de cepstre d'apprentissage ou de test représentés par des matrices de taille, et par la suite en fais appel à la DTW qui a le rôle de trouver le meilleur alignement entre les cepstres de test et d'enregistrement. Cela nous permet de faire une identification méthodique et souple, dans le but de l'appliquer dans notre vie et nos différentes applications.

Les résultats expérimentaux ont été analysés à l'aide de MATLAB et il est prouvé que les résultats sont efficaces. Ce processus peut être étendu pour un nombre n des mots isolés. Le projet montre que le DTW est la meilleure technique d'appariement de caractéristiques non linéaire en identification de la parole, avec des taux d'erreur minimes et une vitesse de calcul élevée. DTW recevra la plus haute importance pour la reconnaissance vocale dans la reconnaissance des mots isolés.

Le principal intérêt des coefficients MFCC est d'extraire des informations pertinentes et en nombres limités, c'est donc la technique la plus efficaces et la plus utilisés dans les systèmes RAP.

L'algorithme DTW est un très bon outil capable de comparer deux spectres audios ayant des durées différentes, un débit, une intensité de la voie différente et cela de façon optimale en cherchant le meilleur chemin pour passer d'un spectre a l'autre. Néanmoins d'autre méthodes existent, comme les méthodes de Markov cachés (HMM) par exemple bien plus puissant que l'algorithme DTW mais bien plus compliquer.

Nous avons donc vu dans ce projet une approche globale de la reconnaissance de la parole. Mais le problème majeur étant que c'est un système mono locuteur se qui fait que chaque utilisateur doit créer son propre dictionnaire dans la phase d'apprentissage. Malgré ça cette méthode peut être parfaitement utilisées dans des appareils d'utilisations courante comme les téléphones portables (utilisation type appel numérotation automatique), les consoles automobiles ou pourquoi pas dans le domaine de la domotique et même pour la commande vocale pour les handicapés.

Pour une utilisation plus poussée type dictée vocale ou standard téléphonique automatique nécessitant une reconnaissance de la parole continue, d'autre systèmes doivent être utilisés qui ont tendance à reconnaitre instantanément n'importe quelle voix. On utilise dans ce cas on utilise une reconnaissante automatique qui ne considère pas la parole comme une suite de mots isolés. L'ordinateur reconnait ici les phonèmes d'un langage donné, déterminés par les rapports qu'ils l'entretiennent avec les autres sons de ce langage. Cette reconnaissance se fait en fonction de la contraintes phonétiques et linguistiques.

**[01]** R. Boite et M. Kunt, Traitement de la parole, Lausanne, Ed. Presses Polytechniques Romandes, 1987.

**[02]** M. Khalida et H. Hassina, Application des MFCCs à la reconnaissance des phonèmes arabes, PFE, Institut d'électronique, Université de Blida (Algérie), 2007.

**[03]** J. P.HATON,N. CARBONNEL,D. FOHR,J. F.MARI,A. KRIOUILLE,Interaction between stochastic modeling andnowledge-based techniques in acoustic-phonetic decoding of speech, IEEE ICASSP'87, pp. 868-871, Dallas, 1987.

**[04]** J. Cantineau, Cours de phonétique arabe, librairie, C. Klinchesiek, Ecole des langues orientales, Paris (France), 1960.

**[05]** CHAH. - Critères de Classification sur des Données Hétérogènes,Revue de Statistique Appliquée, volume XXXIII, N° 2, 1985.

**[06]** CHAH. - Calcul des Partitions Optimales d'un Critère d'Adéquation à une Préordonnance, Publication de l'ISUP, Vol. XXIX, Fasicule 1, 1984.

**[07]** BENZEGHIBA,M.,MORI,R.,DEROO,O.,DUPONT,S.,ERBES,T.,JOUVET,D.,FISSORE,L.,LAFACE,P.,MERTINS,A.,RIS,C.,ROSE,R.,TYAGI,V.etWELLEKENS,C.(2007).Automatic speech recognition and variability: a review. Speech Communication, vol. 49, pp. 763-786, 2007.

**[08]** B. H. Jueng, L. R. Rabiner and J. G. Wilpon, On the use of band pass liftering in speech recognition, IEEE. Transactions on Acoustic and Speech Signal Processing, 35(7): 947- 954, 1987.

**[09]** J. P. Haton, J. M. Pierrel, G. Perennou, J. Caelen et J. L Gauvain, Reconnaissance automatique de la parole, Ed. Dunod, Paris (France), 1991

**[10]** L.R. Rabiner and B.H. Juan, Fundamentals of speech recognition. Englewood Cliffs, N.J., USA: Prentice Hall, 1993.

**[11]** R. Battault, La reconnaissance vocale, techniques utilisées, applications actuelles et futures, Thèse d'Ingénieur en C.N.A.M, Université de Paris11 (France), 1998.

**[12]** Mémoire Master 2005 Rachidi julien.

**[13]** C.AZARA, H.AISSAOUI « Expériences de reconnaissance automatique de la parole à base de module de Markov cachés discrets », thèse d'ingénieurs, institut d'informatique USTHB, Alger (1999).

**[14]** I. Tamanna, Interpolation of linear prediction coefficients for speech coding, Thèse de Doctorat of Electrical Engineering, MC Gill University, Montreal, 2000.

**[15]** J.P. Haton, C. cerisara, D. Fohr, Y. Laprie, and K. Smaili, Reconnaissance automatique de la parole: du signal à son interprétation. Paris: Dunod, 2006.

**[16]** C. Barras, "Reconnaissance de la parole continue : Adaptation au locuteur et contrôle temporel dans les modèles de Markov Cachés," Université de Paris VI, Paris, Thèse de Doctorat 1996.

**[17]** L.R. Rabiner and B.H. Juan, Fundamentals of speech recognition. Englewood Cliffs, N.J., USA: Prentice Hall, 1993.

**[18]** H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," Acoustic Society Am, vol. 87, no. 4, pp. 1738-1752, April 1990.

**[19]** Mémoire d'ingéniorat 2007 «Application des MFCCs à la reconaissance des phonèmes arabes», Université Saad Dahleb Blida.

**[20]** F.J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," in Proc. IEEE, vol. 66 (1), Jan 1978, pp. 51-83.

**[21]** J. S. BRIDLE, M. D. BROWN, R. M. CHAMBERLAIN, An Algorithmfor Speech Recognition,ICASSP'82, pp. 899-902, Paris, May1982.

**[22]** CALLIOPE « La parole et son traitement automatique », collection technique et scientifique des télécommunications, Edition Masson (1989).

**[23]** L.R. Rabiner and R.W. Schafer, "Introduction to digital speech processing," vol. 1, no. 1, pp. 1-194, January 2007.

**[24]** B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the use of band pass liftering in speech recognition," in Proc. ICASSP, vol. 11, Tokyo, Japan, April 1986, pp. 765-768.

**[25]** Kovacs-Vajna ZM (2000) A fingerprint verification system based on triangular matching and dynamic time warping. IEEE Trans Pattern Anal Mach Intell 22(11):1266–1276

**[26]** Gollmer K, Posten C (1995) Detection of distorted pattern using dynamic time warping algorithm and application for supervision of bioprocesses. On-line fault detection and supervision in chemical process industries

**[27]** Keogh E, Pazzani M (2000) Scaling up dynamic time warping for data mining applications. In: 6th ACM SIGKDD international conference on knowledge discovery and data mining, Boston