Université Sidi Mohamed Ben Abdellah
Faculté des Sciences et Techniques Fès
Département Génie Electrique

Mémoire de Projet de fin d'étude

Préparé par

Oussama MIQUAS

Pour l'obtention du diplôme

Ingénieur d'Etat en

SYSTEMES ELECTRONIQUES & TELECOMMUNICATIONS

Intitulé

Développement d'une interface logicielle de traitement de la Parole : Application à la Langue Arabe

Encadré par :

Pr. A. MECHAQRANE

Pr. A.FARCHI (FST Settat)

Soutenu le 4 novembre 2014, devant le jury composé de :

Pr. A. MECHAQRANE	**: Encadrant**
Pr. A.FARCHI	**: Encadrant**
Pr. H. GHENNIOUI	**: Examinateur**
Pr. N. ES-SBAI	**: Examinateur**

ANNEE UNIVERSITAIRE 2014- 2015

Remerciements

Ce travail a été développé au sein du Laboratoire Télécommunication dirigé par Monsieur le Professeur **FARCHI ABDELMAJID** à la FST de SETTAT.

A cette occasion, j'adresse mes remerciements et ma profonde gratitude à mes encadrants **MECHEQUERANE ABDELLAH** professeur à la FST de FES et Monsieur **FARCHI ABDELMAJID** professeur à la FST de SETTAT pour leur aide précieuse, leurs conseils bienveillants, leurs directions, leurs compétences, leurs grandes expériences et leurs qualités humaines qui m'ont permis de mener à bien ce travail. Je leur exprime ma profonde reconnaissance.

Je tiens aussi à remercier mon co-encadrant Mr. SOUFYANE MOUNIR ainsi que Mr. KARIM TAHIRY pour leur soutien et leur assistance durant ce projet.

Je remercie également tous les membres de l'équipe du laboratoire qui m'ont aidé à effectuer ce travail dans les meilleures conditions.

Je tiens bien sûr à remercier amplement tout le cadre professoral de FST pour la formation de qualité qu'ils nous assurent. J'espère que le travail réalisé soit à la hauteur de leurs espérances ainsi qu'aux attentes de mon encadrant.

Je remercie également les membres du jury pour leur considération.

Dédicace

A celui qui m'a indique la bonne voie en me rappelant que la volonté fait toujours les grands Hommes...
A mon Père.

A celle qui a attendu avec patience les fruits de sa bonne éducation...
A ma Mère.

A tous mes frères, oncles et cousins.

A tous mes amis et tous ceux qui me sont chers…

Que Dieu vous garde.

Résumé

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en un signal électrique grâce au microphone. Dans un second temps, ce signal électrique résultant sera numérisé.

Notre interface graphique sert à manipuler les signaux vocaux numérisés. Un signal vocal brut et bruité doit être premièrement traité à l'aide des outils mathématiques nécessaires.

Les algorithmes de fenêtrage, de lissage de la puissance et de séparation sont implémentés dans notre interface afin de caractériser notre signal vocal (sa fréquence fondamentale, son énergie et son spectre) et pour le modéliser par la suite.

C'est une première étape dans le domaine de la reconnaissance automatique de la langue arabe. Nos résultats peuvent être exploités par l'équipe du laboratoire pour développer la recherche dans ce sens.

Abstract

The speech appears as a physical change in air pressure caused and issued by the articulatory system. Acoustic phonetics study this signal by transforming it into an electrical signal through the microphone as a first step. The resulting electrical signal is digitized.

Our graphical user interface is used to manipulate the digitized voice signals. A crude and noisy speech signal must be first treated using mathematical tools. The windowing algorithm, smoothing power and separation are implemented in our interface to be able to characterize our voice signal (its fundamental frequency, energy, and spectrum) and model it thereafter.

This is a first step in the field of automatic recognition of Arabic, and our results may be operated by the team from the lab to develop research in this direction.

Liste des Abréviations

FFT : Fast Fourier Transform

C_V : Consonne_Voyelle

ZCR : Zéro Crossing Rate

/b/ : “bae” ب

/m/ : “mim” م

/d/ : “dal” د

/t/ : “tae” ت

/k/ : “kaf” ك

/kh/ : “khae” خ

/a/ : “fatha” ◌َ

/i/ : “kasra” ◌ِ

/u/ : “damma” ◌ُ

Liste des figures

Sommaire

Introduction

La parole a été, depuis longtemps, l'outil de communication de l'être humain. La compréhension de cet outil nécessite, en premier lieu, une détection des mots prononcés par l'organe phonatoire puis une analyse de ces derniers. Dans le cadre de ce projet de fin d'étude, nous nous intéressons à la langue Arabe qui est parlée par environ 420 millions de personnes [http://fr.wikipedia.org/wiki/Arabes].

Les activités de recherche décrites dans ce rapport coïncident avec l'évolution du domaine de la reconnaissance automatique de la parole. Notre objectif est de développer une interface logicielle de traitement de la parole de la langue Arabe. L'originalité de ce projet de fin d'étude vient de la volonté d'aborder cette langue qui ne dispose pas de ressources nécessaires pour sa reconnaissance automatique. Nous nous intéressons plus particulièrement à la reconnaissance automatique de voyelles et consonnes des mots arabes isolés à savoir leurs caractéristiques afin de pouvoir les modéliser par la suite. L'intérêt de ce travail réside dans l'extraction du signal utile des mots enregistrés et de leur séparation à l'aide des deux algorithmes : Zéro Crossing Rate (ZCR) et l'algorithme de différence entre l'énergie de la consonne et celle de la voyelle. Il permet également de calculer les moments spectraux de ces consonnes et voyelles.

Ce rapport s'articule en quatre chapitres. Dans le premier chapitre, nous présenterons le signal vocal et ses caractéristiques anatomiques et phonatoires et nous décrirons ensuite les modes et les lieux d'articulation pour les différentes consonnes et voyelles.

Dans le deuxième chapitre, nous aborderons le principe de la reconnaissance automatique de la parole ainsi que les différentes étapes d'élaboration d'un système de reconnaissance.

Le troisième chapitre sera consacré à la présentation des différents modèles et algorithmes utilisés dans les systèmes de reconnaissance de la parole. Nous présenterons aussi deux outils indispensables dans ce domaine à savoir les logiciels Matlab et Praat.

Le dernier chapitre présentera les résultats de nos expérimentations. Une interface graphique développée sous Matlab englobe tous les algorithmes et les tâches effectuées au cours de ce projet de fin d'études.

Chapitre 1 : Phonétiques et caractéristiques de la parole

1.1 Introduction

La parole est selon la définition du Robert « la faculté de communiquer la pensée par un système de sons articulés émis par les organes de phonation » [1].

Jusqu'à ce jour, bien que de nombreuses recherches furent effectuées, il nous est impossible d'égaler ou même de cerner les fonctions du cerveau qui permettent la parole.

Par le fait de la position de la parole, au croisement de plusieurs disciplines, on lui distingue plusieurs niveaux de description [2]:

Le niveau phonétique, le niveau phonologique, le niveau acoustique, le niveau morphologique, le niveau syntaxique, le niveau sémantique et le niveau pragmatique.

Dans ce chapitre, nous présentons une brève description des caractéristiques anatomiques, phonatoires et phonémiques du signal vocal. Ensuite, nous présentons une classification phonémique pour bien définir les origines et les caractéristiques de chaque phonème. Finalement, nous présentons les différentes Caractéristiques fréquentielles du signal de parole.

1.2 Caractéristiques anatomiques, phonatoires et phonémiques du signal vocal

Etant donné que le conduit vocal agit comme un filtre acoustique, il donne ainsi au son les indices acoustiques qui distinguent les différents phonèmes. En effet, le signal vocal est représenté par un ensemble de sons qui peuvent être voisés ou non voisés. Par définition, le phonème est la plus petite unité phonique fonctionnelle [3]. Les phonèmes sont regroupés par classes selon leur mode, lieu de production, mode de voisement et de nasalité.

La réalisation des phonèmes est influencée par le phénomène de coarticulation lié à l’enchaînement d’une suite de sons.

Dans cette section, nous décrivons brièvement l’anatomie des organes de la parole, les modes phonatoires, les principales classes phonétiques ainsi que le phénomène de coarticulation.

1.3 Description de l'anatomie des organes de la parole

Pour analyser le signal de parole, il est intéressant de comprendre la façon dont il est produit par le système articulatoire. Le passage de l'air à travers le conduit vocal donne naissance à plusieurs variétés de sons. Les phonèmes diffèrent en fonction de leur lieu et leur mode d'articulation. La Figure 1, ci-dessous, présente les différents points d'articulation qui sont: les lèvres (labiales), les dents (dentales), les alvéoles (alvéolaires), le palais dur postérieur et antérieur (palatal), la voile du palais (vélaire), la luette (uvulaire), le pharynx (pharyngal), la glotte (glottal), dos (dorsal) et apex (apical).

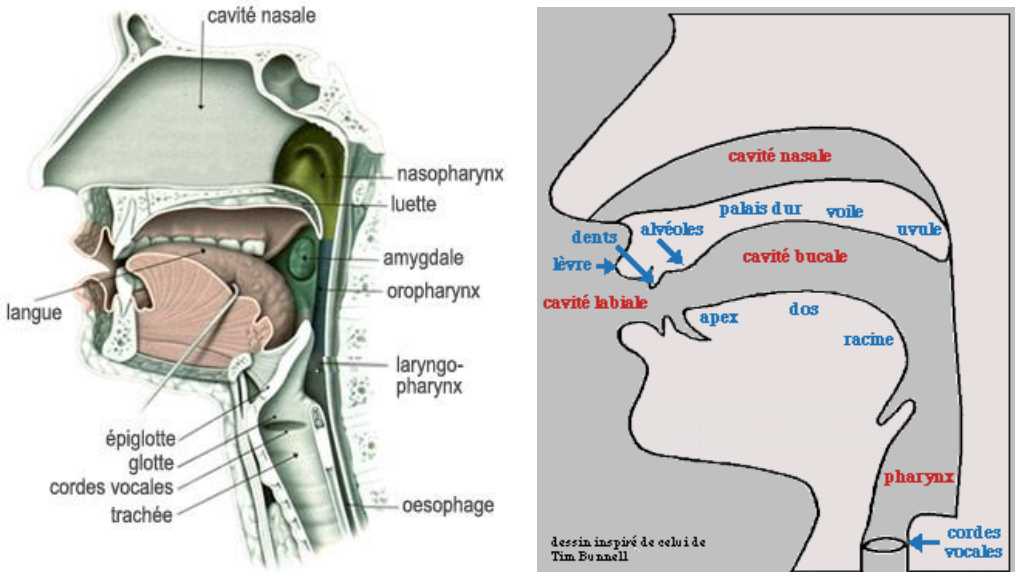

Figure 1 : Description détaillée de l'appareil vocal

1.3.1 Les modes phonatoires

Les principaux modes phonatoires sont : le voisement, l'aspiration, la voix soufflée et la laryngalisation. Le voisement caractérise les voyelles et certaines consonnes voisées contrairement aux consonnes sourdes qui sont caractérisées par le non voisement [4].

L'aspiration est une courte période non voisée [4]. On peut distinguer des consonnes aspirées et non-aspirées par rapport au délai d'établissement du voisement après le relâchement de l'occlusion ou VOT pour « Voice On set time ».

La voix soufflée se distingue lorsqu'une partie de l'air s'échappe sans être modulée et le son voisé est accompagné de souffle, la voix peut donc manquer d'amplitude.

La laryngalisation se distingue lorsque le son produit est caractérisé par une fréquence fondamentale basse.

1.3.2 L'appareil phonatoire

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. L'appareil respiratoire fournit l'énergie nécessaire à la production des sons, en poussant l'air à travers la trachée. Au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal (Figure 1).

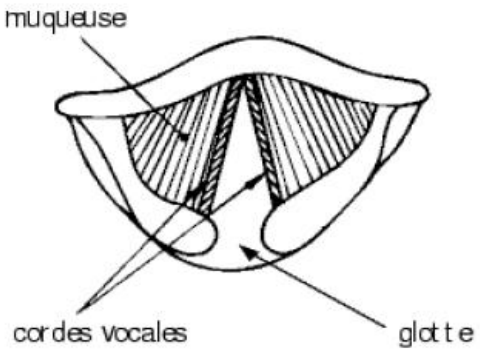

Figure 2 : Le larynx vu du dessus

- **le type de phonation** (le mode de vibration des cordes vocales) : son voisé ou non voisé, chuchoté, crié, soufflé, etc.

- **le lieu d'articulation** (la région de rétrécissement maximal du canal buccal) suivant la position du larynx, du voile du palais, de la langue, des mâchoires, des dents et des lèvres. Plus généralement, une configuration géométrique donnée de ces éléments fixe certaines résonances (les formants) qui permettent de distinguer les sons voisés entre eux.

- **le mode d'articulation** (pour les sons non voisés) : occlusif ou plosif (le passage de l'air est fermé et le son résulte de son ouverture subite).

- **le caractère nasal** suivant la position du voile du palais qui enclenche ou non la résonance de la cavité buccale.

De nombreux facteurs de variabilité dans le contrôle articulatoire du conduit vocal sont donc responsables du caractère unique et reconnaissable d'un locuteur particulier.

1.4 Classification des sons selon la source: voisés ou non voisés

Selon la nature de la source d'excitation à la sortie du larynx, un signal de parole est tantôt périodique, tantôt aléatoire. Ceci amène à la classification des sons en deux types : voisés ou non voisés [4][6].

Les sons voisés tels que les voyelles par exemple, sont produits par le passage de l'air des poumons à travers la trachée, qui met en vibration les cordes vocales [7].

Pour la parole voisée, l'excitation possède un caractère périodique et des propriétés particulières dues à la forme de l'onde de débit glottique.

Les sons voisés sont généralement quasi-périodiques. Ce type de sons représente la majorité du temps de phonation, et est caractérisé en général par une énergie élevée en basse fréquence avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante à l'information linguistique [6].

Par contre, les sons non voisés présentent une structure apériodique, les cordes vocales sont écartées et n'entrent pas en vibration. L'énergie de ce type de sons est concentrée dans les hautes fréquences et correspond au bruit [7].

Au niveau du spectrogramme, comme le montre la Figure 3, ci-dessous, les parties voisées du signal apparaissant sous la forme de successions de pics spectraux denses en énergie sur lesquels on a superposé les courbes des trois premiers formants (F1, F2 et F3), dont les fréquences centrales ne sont pas forcément des multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière.

La figure suivante est une représentation des trois premiers formants (F1, F2, F3) superposés sur le spectrogramme du signal : **أخذ إجازةً** « Il a pris des vacances » où il y a une alternance de sons voisés et non voisés. Dans le cas voisé, une structure formantique est présentée.

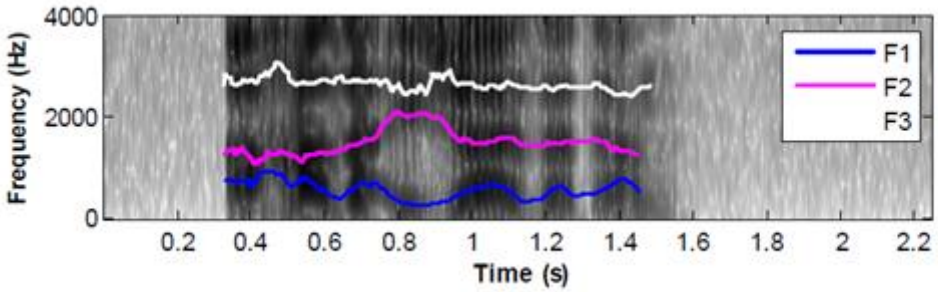

Figure 3 : Représentation des trois premiers formants F1, F2, F3.

1.5 Classification phonémique

Les langues du monde font un usage spécifique des possibilités physiologiques et anatomiques des organes articulateurs et n'exploitent chacune, qu'une partie des mouvements et positions articulatoires que l'homme peut produire. Ces mouvements articulatoires servent à produire des classes de voyelles et de consonnes distinctes [5].

La division de l'ensemble de ces sons, ou phonèmes, en classes distinctes, est à l'origine de la constitution d'alphabets phonétiques qui caractérisent les différentes langues. On distingue généralement trois classes principales: les voyelles, les consonnes et les semi-voyelles.

1.5.1 Les voyelles

Les voyelles peuvent être spécifiées à l'aide de quatre traits qui sont: la nasalité, le degré d'ouverture du conduit vocal (aperture vocal), la position de la constriction principale du conduit vocal (antérieure/postérieure) et la protrusion des lèvres (arrondissement) [4].

Les voyelles nasales diffèrent des voyelles orales dans le fait que le voile du palais est abaissé pour leur articulation, ce qui met en parallèle les cavités nasale et buccale [6]. On définit et on c1asse les voyelles, comme les consonnes, selon les critères de mode et de point (ou de zone) d'articulation. « Trois positions extrêmes serviront à définir les voyelles principales. En schématisant, elles occupent les sommets d'un triangle qu'on appelle le « triangle vocalique » [8]. (Voir Figure 4)

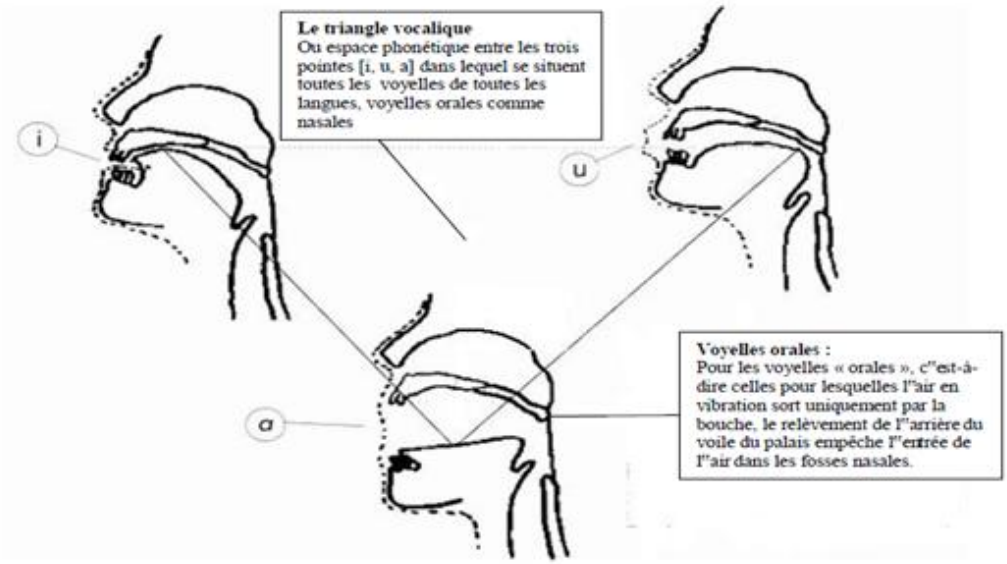

Figure 4 : Présentation des voyelles principales dans le triangle vocalique [10]

[i] : voyelle orale fermée antérieure rétractée.

[a] : voyelle orale ouverte

[u] : voyelle orale fermée postérieure arrondie

La phonétique classique classe les voyelles d'après la position de la langue dans la cavité buccale. Pour les caractériser, elle prend également en compte la position et la forme des lèvres. Ces derniers peuvent être plus au moins écartées, étirées, arrondies et protruites à des degrés divers. En effet les lèvres peuvent [10]:

- soit se projeter en avant et s'arrondir, et la voyelle est une voyelle arrondie ou labialisée, comme /y, oe, o, u/.
- soit s'étirer ou rester en position neutre : la voyelle est alors une voyelle non labialisée ou non arrondie, comme /i, e, ɛ, a/.

Il faut noter qu'il existe une certaine corrélation entre la hauteur de la langue et la labilité [5].

On distingue aussi, selon les mouvements horizontaux de la langue dans la bouche, les voyelles antérieures (en avant), les voyelles centrales (au milieu), et les voyelles postérieures (en arrière), et, selon l'écartement entre la langue et le lieu d'articulation appelé aperture (on parle donc de degré d'aperture des voyelles), les voyelles fermées et ouvertes. Toutes les voyelles de toutes les langues sont toujours situées dans le triangle vocalique, orales comme nasales. (Voir Figure 5)

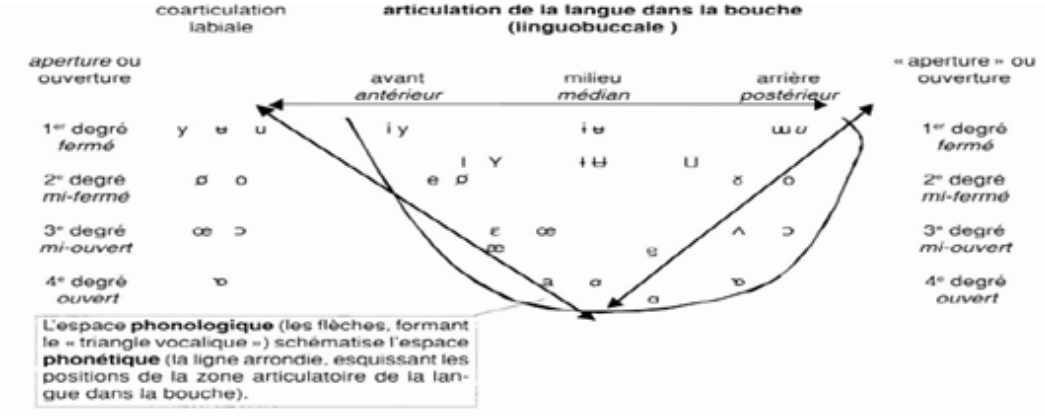

Figure 5 : Triangle vocalique : Les voyelles selon l'alphabet phonétique international (API)

Le triangle vocalique constitue un moyen de repérage facile pour placer les voyelles.

On distinguera des niveaux d'ouverture (de 2 à plus de 5) et des niveaux de profondeur (de 2 à 7: 4 en Français). Notons que l'indication antérieure / postérieure n'a pas de sens pour la voyelle ouverte étant donné que la langue est abaissée et est alors centrée par rapport à la profondeur [8].

L'aperture vocalique désigne la distance verticale qui sépare le sommet du dôme de la langue et le palais. Elle renvoie au degré de courbure convexe de la surface de la langue et à sa hauteur relative dans la cavité buccale. Sur cette base on peut distinguer des voyelles hautes comme /i/ et /u/, mi-hautes comme /e/ et /o/, mi-basses comme /ɛ/ et /ɔ/ et basses comme /a/ et /ɑ/. Il s'agit donc d'un classement qui repose sur une dimension verticale et non sur l'aire au lieu d'articulation comme pour les consonnes.

La plupart des voyelles utilisées dans les langues sont sonores, c'est-à-dire qu'elles sont prononcées en faisant vibrer les cordes vocales, mais des voyelles sourdes, sans vibration des cordes vocales, sont utilisées dans certaines langues comme le Cheyenne et le japonais.

Le chuchotement utilise aussi, par définition, des voyelles sourdes.

La voyelle constitue le pivot d'une syllabe. Tandis qu'une voyelle peut constituer à elle seule une syllabe cela ne peut être le cas d'une consonne. Celle-ci doit être nécessairement associée à une voyelle.

1.5.2 Les consonnes

Comme leur nom l'indique « consonna : „dont le son se joint avec „ », elles précèdent ou suivent un élément vocalique [5]. Lorsque le conduit vocal au passage de l'air se rétrécit par endroits, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, de leur voisement et leur nasalisation [5]. Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général: l'organe articulatoire, la source sonore, l'intensité, l'aspiration, la palatalisation, et la direction du mouvement de l'air [6].

Les consonnes forment donc une classe très hétérogène que l'on peut décomposer en trois sous-classes principales ayant des caractéristiques distinctes: les fricatives, les occlusives et les sonantes [4].

- Les consonnes fricatives ou constrictives prennent naissance dans la production d'un bruit de friction qui résulte d'une turbulence aérodynamique en un ou plusieurs points du conduit vocal en raison de la présence d'un fort resserrement (ou constriction) dans le flot d'air expiratoire [4]. Ce sont essentiellement les lèvres et la langue qui, selon leur position et leur tension musculaire particulière, conditionnent le type de friction réalisée. Cette friction est réalisée au niveau d'un lieu d'articulation qui peut être le palais [ʃ, ʒ], les dents [s, z], ou les lèvres [f, v], elle produit un bruit en haute fréquence. On parle généralement de fricatives non voisées et des fricatives voisées. Les fricatives non voisées sont caractérisées par le passage d'un écoulement d'air à travers la glotte ouverte, tandis que dans le cas des fricatives voisées, la source vocale est active, il y aura donc une vibration des cordes vocales incomplète, c'est-à-dire que les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète et combinée avec un bruit de friction : c'est une combinaison de composantes d'excitation périodique et turbulente [6].
- Les consonnes occlusives se caractérisent principalement par un silence dû à la fermeture complète du conduit vocal appelée « occlusion » en un lieu bien défini. Une forte pression est créée en amont de cette occlusion qui peut être au niveau du palais [k, g], des dents [t, d], ou des lèvres [p, b]), puis relâchée brusquement. La période d'occlusion est appelée la phase de tenue. Les occlusives peuvent être généralement soit des occlusives voisées (sonores) ou non voisées (sourdes). Les occlusives voisées sont en général plus brèves au niveau du silence que les occlusives sourdes. Dans le cas des occlusives voisées

par exemple [b, d, g], le silence n'est pas total dans la mesure où la vibration des cordes vocales pendant la tenue articulatoire se traduit par la « barre de voisement » qui est une faible énergie en très basse fréquence au niveau du spectrogramme. Les consonnes occlusives non voisées, par exemple [t, k], sont caractérisées par un silence pendant la tenue articulatoire suivi par une explosion au moment de relâchement. Conventionnellement, on mesure « le délai d'établissement du voisement » (VOT : Voice Onset Time) à partir de la barre d'explosion. « La mesure est comptée négativement si le voisement précède la barre d'explosion, et positivement dans le cas contraire ». Les consonnes voisées par exemple [b, d, g] sont à VOT négatif et les consonnes non voisées [t, k] sont à VOT positif. Les transitions formantiques des occlusives sont caractérisées par des déflexions fréquentielles rapides des formants que l'on observe au passage d'une consonne à une voyelle et réciproquement et cela est dû à la diminution du degré de constriction qui suit la rupture de l'occlusion et les mouvements des organes articulatoires vers une nouvelle cible [6][4].

- Les consonnes sonantes possèdent comme les voyelles, une structure formantique, mais au contact des consonnes sourdes, les sonantes qui sont intersèquement sonores perdent leur voisement. Cette classe est en fait constituée du regroupement de deux sous-classes que sont les liquides et les nasales :
    - Les liquides [l, r] sont assez difficiles à classer. L'articulation de [l], qui une latérale, ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal. Le son de [r], quant à lui, admet plusieurs réalisations fortes différentes qui peuvent être soit trille c'est-à-dire répétitive au niveau du spectrogramme, soit tape avec la tenue d'un silence de durée très limitée [6][4].

Généralement, au niveau du spectrogramme, on observe des formants vocaliques pour les liquides.

- Les nasales [m, n] font intervenir les cavités nasales par abaissement du voile du palais. Les spectrogrammes de nasales présentent de nombreuses similitudes avec ceux des voyelles; on y découvre, le plus souvent, un premier formant fort en basse fréquence [10].

1.5.3 Les semi-voyelles

Les semi-voyelles [j, μ, w] combinent certaines caractéristiques des voyelles et des consonnes, on les appelle parfois semi-consonnes ou glides [8]. D'ailleurs on peut les classer selon la sous-classe sonantes des consonnes. Comme les voyelles, leur position centrale est

assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes [6]. Les fréquences initiales des formants d'une semi-voyelle sont proches de celles de la voyelle correspondante.

1.6 Phénomène de coarticulation

La parole est produite par une chaîne de gestes articulatoires qui se réalisent dans le temps. Cependant, les sons de la parole sont caractérisés par une importante variabilité selon leur entourage phonétique [5]. Cette variabilité a été en partie attribuée au fait que les mouvements accomplis par les articulateurs dans la production de la parole se chevauchent sur l'axe temporel, par exemple, en prononçant la syllabe CV, les gestes articulatoires associés à la consonne initiale et à la voyelle qui la suit sont partiellement superposés [11]. Ces phénomènes de recouvrement temporel sont généralement désignés sous le terme de coarticulation. En effet un phonème ne se prononce pas de la même manière, tout dépend s'il est prononcé seul ou dans une chaîne parlée (nommé « allophone ») et tout dépend de son entourage phonétique.

Cependant, la modification de la configuration du conduit vocal pour passer d'un phonème à un autre se fait de façon progressive et les deux sons subissent une distorsion. Il faut noter aussi 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 [5][10].

Il est reconnu que les phonèmes s'influencent aussi bien à droite par rétention de l'articulation qu'à gauche par anticipation. Par exemple, une consonne sonore suivie d'une consonne sourde a tendance à se dévoiser. Ainsi le mot « médecin », après la chute du « e » devient « médcin » où le « d » perd sa marque de sonorité; d'où l'effet du phénomène de coarticulation. Le phénomène de coarticulation est primordial à l'intelligibilité de la phrase. En effet, si on synthétise une phrase en mettant bout à bout les phonèmes composant cette phrase, mais sans aucune contrainte de coarticulation entre phonèmes, on obtient une phrase devient incompréhensible [5].

En revanche, à cause du phénomène de coarticulation, il est souvent difficile de segmenter le signal de parole en unités discrètes pour l'analyse, ce qui serait très utile dans les applications automatiques de reconnaissance de la parole. Dans certains cas, une telle division est facile, surtout lorsque l'excitation change brusquement avec le début ou la fin de la vibration des

cordes vocales (c'est-à-dire une transition voisée-non voisée), ou au moment de l'ouverture et fermeture du conduit vocal (par exemple, la fermeture des lèvres) [12].

1.7 Caractéristiques fréquentielles du signal de parole

L'analyse dans le domaine fréquentiel désigne l'analyse des fonctions mathématiques ou des signaux selon la fréquence, plutôt qu'une fonction de temps. Une fonction donnée ou un signal peuvent être convertis entre les domaines temporels et fréquentiels à l'aide d'un opérateur mathématique appelé une transformation. Un exemple est la transformée de Fourier, qui décompose une fonction en une somme d'un nombre (potentiellement infini) d'ondes sinusoïdales. Le «spectre» est la représentation dans le domaine fréquentiel du signal. La transformée de Fourier inverse convertit la fonction du domaine fréquentiel à une fonction dans le domaine temporel. Un analyseur de spectre est l'outil couramment utilisé pour visualiser le monde réel des signaux dans le domaine fréquentiel.

On affiche généralement la façon dont le spectre de parole change au fil du temps sous la forme d'un spectrogramme comme le montre la Figure 5. Le but de l'estimation spectrale est de décrire la distribution fréquentielle de la puissance contenue dans un signal.

Le signal de parole est caractérisé comme suit [13] :

- La bande passante du signal est d'environ 10 kHz.
- Le signal voisé est périodique avec une fréquence fondamentale entre 80 Hz et 350 Hz.
- La distribution spectrale présente des pics d'énergie
- L'enveloppe du spectre de puissance du signal montre une atténuation quand la fréquence s'élève (-6 dB par octave).

1.7.1 La bande passante

La bande passante du signal de parole est beaucoup plus élevée que 4 kHz. Pour les fricatives, il y a une quantité importante d'énergie en haute fréquence.

Une bande passante de 4 kHz qui est celle du téléphone, contient toutes les informations nécessaires pour comprendre la voix humaine sauf certaines fricatives.

1.7.2 La fréquence fondamentale

La vitesse à laquelle les cordes vocales s'ouvrent et se referment lors du processus de phonation, produit une vibration d'une hauteur variable appelée fréquence fondamentale dont la valeur est étroitement liée à la taille de l'appareil phonatoire de la personne. Cette fréquence est quasi stationnaire pour un signal de type voisé, elle varie de [14]:

- De 80 à 200 Hz pour une voix masculine,
- De 250 à 450 Hz pour une voix féminine,
- De 200 à 600 Hz d'une voix d'enfant.

Deux sons de même intensité et de même hauteur se distinguent par le timbre, qui est déterminé par les harmoniques de la fondamentale [14]. Un intérêt majeur pour cette fréquence se trouve dans les applications de la synthèse de parole.

1.7.3 Les formants

Le spectre du signal vocal résultant de l'action des sources de sons sur le conduit vocal présente des maximums et des minimums qui correspondent aux résonances et aux antirésonances du conduit vocal appelés formant et anti-formants. Du point de vue perceptif, seuls les trois premiers formants jouent un rôle essentiel pour caractériser le spectre vocal [14]. On peut caractériser toute voyelle en n'utilisant que ses trois premiers formants. En général, la fréquence du premier formant varie de 200 à 900 Hz, celle du second de 500 à 2500 Hz et le troisième se situe entre 1500 et 3500 Hz. Des formants d'ordre supérieur existent même si leur rôle sur le plan perceptif est limité, ils contribuent à caractériser la voix.

1.7.4 La représentation de la parole dans les domaines temps et fréquence

Une représentation de l'évolution temporelle du signal vocal ou audiogramme est donnée dans la Figure 6. Cependant, pour avoir plus d'informations sur la fréquence fondamentale et les formants, on utilise, généralement, une représentation 3-D (amplitude/fréquence/temps) appelée spectrogramme.

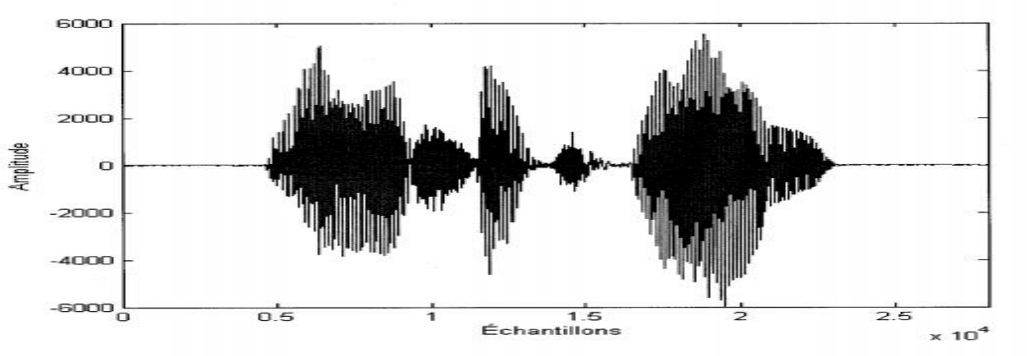

Figure 6 : Représentation numérisée d'un signal vocal

1.7.5 Le spectrogramme

Le spectrogramme est une représentation tridimensionnelle, où le temps est représenté sur l'axe X, la fréquence sur l'axe Y et le niveau de chaque fréquence sur l'axe Z (symbolisé par le niveau du gris). Pour l'obtenir, on effectue sur le signal une FFT (Fast Fourier Transform) à fenêtre glissante.

On distingue deux types de spectrogrammes, les spectrogrammes à bandes larges (voir Figure 7) et les spectrogrammes à bandes étroites (voir Figure 7). Les premiers sont obtenus avec des fenêtres de faible durée. Ils mettent en évidence l'enveloppe spectrale (les formants du signal), les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont obtenus avec des fenêtres de l'ordre de 30 à 40 ms, ils offrent une bonne résolution au niveau fréquentiel, les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales.

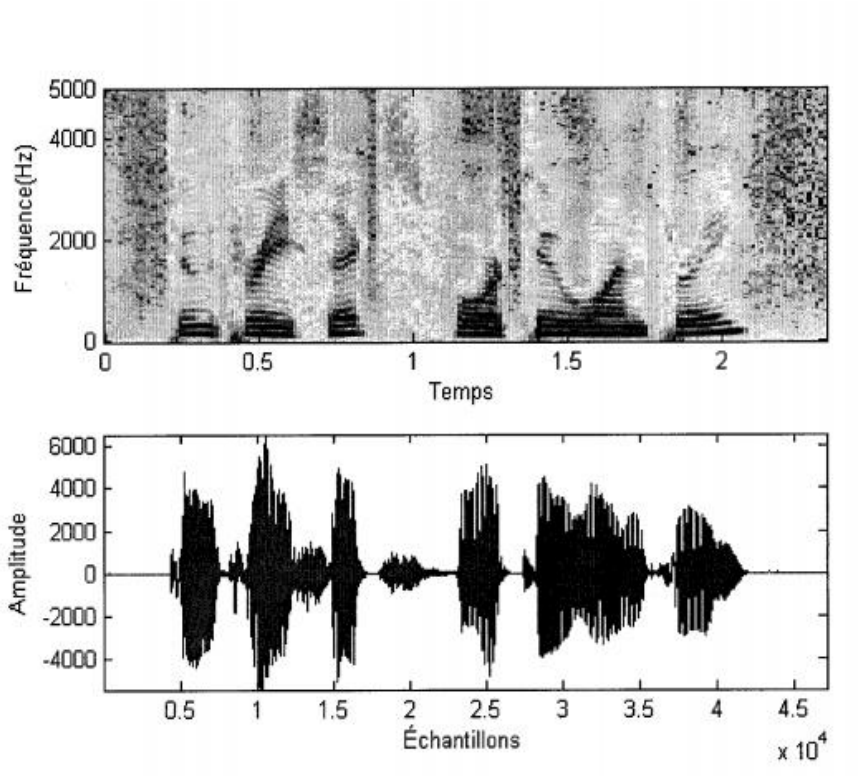

Figure 7 : Spectrogramme à bande large (128 échantillons)

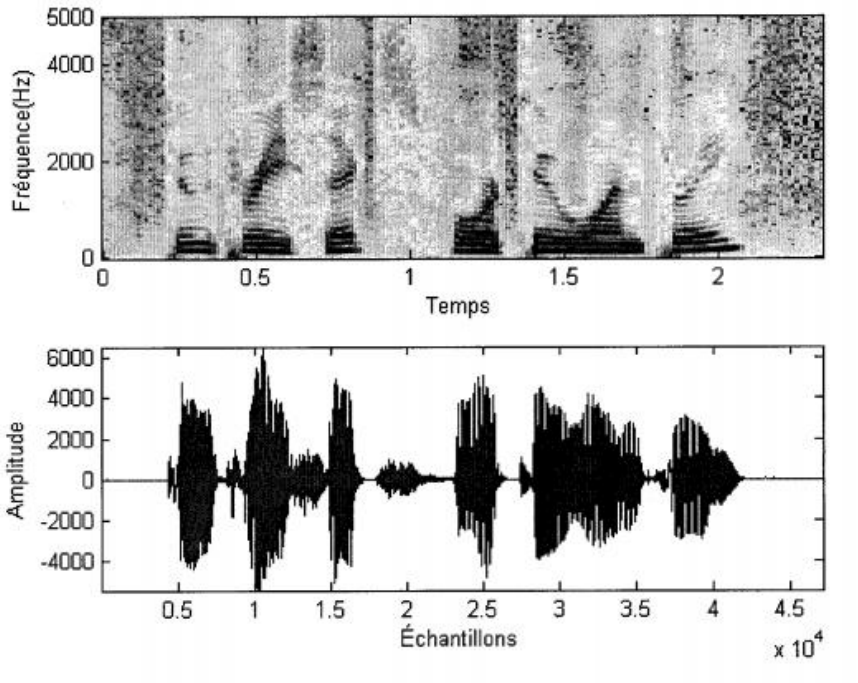

Figure 8 : Spectrogramme à bande étroite (512 échantillons)

1.8 Conclusion

Dans ce chapitre, nous avons passé en revue le mécanisme de la production de la parole, le principe de son audition ainsi que les caractéristiques générales du signal vocal.

Principalement, on peut classer le signal vocal en deux catégories : les sons voisés, résultants de la vibration des cordes vocales, et les sons non voisés qui ne nécessitent par l'intervention du larynx.

Nous avons présenté, dans ce chapitre, une brève description des caractéristiques anatomiques, phonatoires et phonémiques du signal vocal. Ensuite, nous avons présenté une classification phonémique pour bien définir les origines et les caractéristiques de chaque phonème du point de vue articulatoire et acoustique. Finalement, nous avons présenté les caractéristiques spectrales principales du signal vocal.

Chapitre 2 : Phonétique de la Langue Arabe et état de l'art

2.1 Introduction

La mise en œuvre d'un algorithme de reconnaissance de la parole fait appel à différentes sources de connaissance, correspondant à différents niveaux de représentation de la parole : le niveau acoustique, phonétique, lexical, syntaxique et sémantique, et pragmatique.

Le développement d'un système de décodage phonétique de l'Arabe nous impose de bien étudier la composante phonétique de la langue afin de dégager les caractéristiques. En effet, le rôle de cette étape consiste, à partir d'un signal sortant d'un micro, à fournir une description phonétique d'un énoncé. Au niveau phonétique, ce signal est décrit par une séquence de phonèmes. L'utilisation des phonèmes comme unités de base permet de décrire un vocabulaire théoriquement illimité. En effet, l'ajout d'un mot au vocabulaire se réalise simplement en donnant sa transcription phonétique, ce qui ne nécessite pas de définition d'un nouveau modèle.

2.2 La langue Arabe et ses variantes

L'arabe est une langue parlée par plus de 200 millions de personnes [15]. Elle est langue officielle d'au moins 22 pays. C'est aussi la langue de référence pour plus d'un milliard de musulmans.

Le développement de la langue arabe a été associé à la naissance et la diffusion de l'islam. La langue arabe s'est imposée, depuis l'époque arabo-musulmane, comme langue religieuse mais plus encore comme langue de l'administration, de la culture et de la pensée, des dictionnaires, des traités des sciences et des techniques. L'arabe peut être considéré comme un terme générique rassemblant plusieurs variétés :

- l'arabe classique : la langue du Coran, parlée au VIIe siècle ;
- l'Arabe Standard Moderne (ASM) : une forme un peu différenciée de l'arabe classique, et qui constitue la langue écrite de tous les pays arabophones. L'ASM reste le langage de la presse, de la littérature et de la correspondance formelle, alors que l'arabe classique appartient au domaine religieux.
- les dialectes arabes : malgré l'existence d'une langue commune, chaque pays a développé son propre dialecte.

1. les dialectes arabes, parlés dans la Péninsule Arabique : dialectes du Golfe, dialecte du najd, yéménite ;

2. les dialectes maghrébins : algérien, marocain, tunisien, hassaniya de Mauritanie.

3. les dialectes proche-orientaux : égyptien, soudanais, syro-libano-palestinien, irakien (nord et sud).

L'arabe est un ensemble complexe dans lequel s'étendent des variétés écrites et orales répondant à un spectre très varié d'usages sociaux. Mais au-delà de cette variété, les sociétés arabes ont une conscience aiguë d'appartenir à une communauté linguistique homogène, d'où l'importance de l'ASM qui forme un terrain commun pour cette large population. L'ASM est la langue des médias officiels, de la communication écrite et de tout type de communication non spontanée. Elle se distingue des dialectes arabes par son système grammatical partagé avec l'arabe classique.

2.3 Phonétique de la langue Arabe

La finalité première des moteurs de Reconnaissance Automatique de la Parole est la transcription intégrale du signal de parole en mots. Cependant, outre le mot, il existe de nombreuses autres unités possibles pour la segmentation de la parole, en l'occurrence [16]:

- Les phones, ou unités sous phonèmes, qui, fusionnées entre elles, permettent d'obtenir des unités plus longues;
- Les phonèmes, ou l'unité la plus courte qu'un être humain est capable d'identifier dans la parole continue;
- Les allophones, ou les différentes réalisations sonores possibles d'un phonème;
- Les diphones, demi-syllabes, et syllabes qui permettent d'incorporer les phénomènes co-articulatoires;
- Les morphèmes, ou les plus petites unités porteuses de sens qu'il soit possible d'isoler dans un énoncé;

Le phonème est considéré comme étant la plus petite unité distinctive, c'est à dire permettant de faire la distinction entre deux mots. On identifie donc les phonèmes d'une langue par appariement des mots de sens différents mais ne différant que par un seul son. Ces mots sont

appelés paires minimales, à l'exemple des mots poulet et boulet, mots de sens différents dont la forme sonore ne se distingue que par le /p/ et le /b/. Cependant, chaque son d'une langue n'est pas forcément considéré comme un phonème. En effet, si on prend l'exemple des différentes prononciations du son /r/ dans le mot rat, qui peut être roulé ou non, on aboutit à deux sonorités différentes malgré un sens du mot identique. Ces deux sonorités correspondent à un même phonème mais sont appelées allophones.

La langue arabe standard a été étudiée par les phonéticiens arabes il y'a bien des siècles. Son système phonétique est constitué essentiellement par 34 phonèmes qui se composent de 6 voyelles et 28 consonnes. L'analyse qui suit est basée sur les propriétés spectrales des phonèmes de l'Arabe.

2.3.1 Classification articulatoire des voyelles

En ce qui concerne les voyelles de l'arabe, La durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales. Elles sont représentées dans le tableau suivant :

Courtes	Longues
ُ ِ َ	ا ي و

Figure 9 : Classification des voyelles de la langue arabe

Les voyelles se caractérisent principalement par la présence de zones de fréquences où les harmoniques (formants), sont particulièrement intenses. Ces formants apparaissent sur le spectrogramme sous forme de bandes noires plus ou moins parallèles à l'axe des temps. Le premier formant F1 est le pic spectral ayant la fréquence la plus basse, le deuxième formant F2 est le pic suivant, etc.

On ne tient pas compte du pic de très basse fréquence (formant glottal), vers 200-300 Hz, qui peut apparaître pour les voyelles ouvertes (type a) d'intensité faible ainsi que du formant supplémentaire qu'on observe occasionnellement dans la zone 800-1200 Hz.

Il existe trois voyelles principales, toutes orales, se trouvant aux extrémités du triangle vocalique : /a/, /i/ et /u/ (appelées fatha pour /a/, kasra pour /i/ et dhamma pour /u/). Elles sont caractérisées par deux classes de localisation : antérieure étirée (/i/) et postérieure arrondie (/u/), et deux degrés d'aperture: fermé (/i/ et /u/) et ouvert (/a/).

En principe [17], une voyelle longue est deux fois plus longue qu'une voyelle brève, mais cette durée est en fait variable, notamment en fonction :

- de la position de la syllabe dans le mot (en finale, la longue est abrégée)
- de l'accent (une longue non accentuée est abrégée).

Dans les parlers du Maroc, la distinction brèves/longues est très atténuée. Certaines voyelles brèves sont réalisées comme des longues et vice-versa. Cette caractéristique influe parfois sur la prononciation de l'arabe standard.

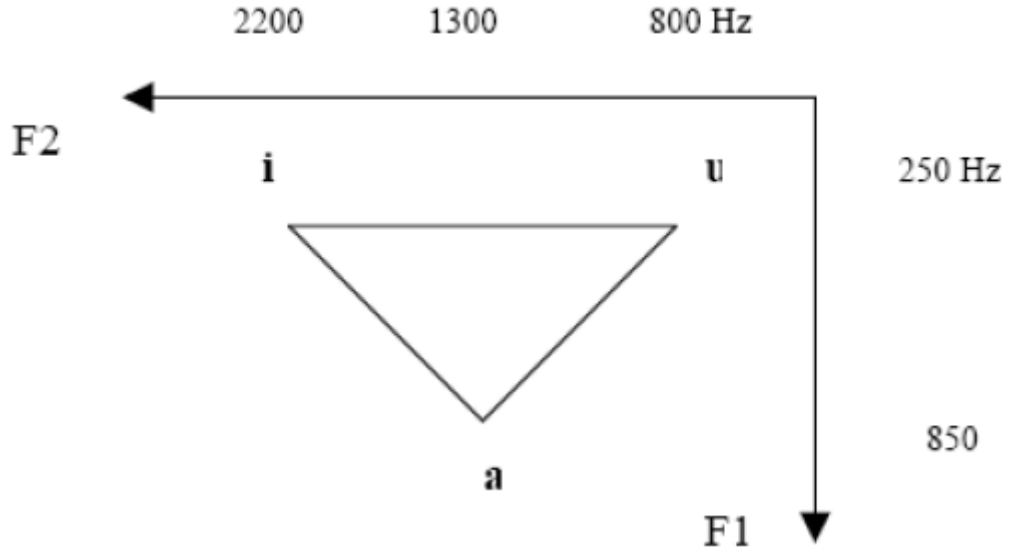

Figure 10 : Triangle articulatoire

On peut donc interpréter une augmentation de F1 comme le résultat d'une ouverture articulatoire et une augmentation de F2 comme une antérioration de l'articulation. Pour les voyelles antérieures, un accroissement de la labialisation (arrondissement des lèvres) se traduit par une baisse de F2 et de F3 à articulation linguale constante. La chute de F1 est un indice moins ambigu, car elle n'est pas liée à une éventuelle postérioration de l'articulation. La description que l'on donne à ces voyelles est la suivante [18]:

/a/ : est une voyelle centrale ouverte, elle se prononce en ouvrant largement la bouche et en conservant la langue dans une position horizontale.

/i/ : est une voyelle antérieure fermée qui se prononce en portant le devant de la langue en avant et en l'étalant largement tandis que l'arrière frôle presque le palais.

/u/ : est une voyelle postérieure fermée qui se prononce en contractant la langue au fond de la bouche et en avançant les lèvres qui s'arrondissent jusqu'à presque ce qu'elles se joignent.

Les voyelles longues sont : alif al-MADD ou /a:/, ya al-MADD ou /i:/ et waw el –MADD ou /u:/.

2.3.2 Classification articulatoire des consonnes

Traditionnellement, on dit que l'arabe se caractérise par un consonantisme riche et un vocalisme pauvre. Les 28 consonnes de l'arabe occupent pratiquement tout l'espace phonatoire. On les produit en n'importe quel point de cet espace depuis les lèvres jusqu'au larynx.

Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.	Gr.	Ph.
أ	E	خ	x	ش	ʃ	غ	ɣ	ن	n	ُ	u
ب	b	د	d	ص	ʂ	ف	f	ه	h	ِ	i
ت	t	ذ	ð	ض	ɖ	ق	q	و	w	ً	an
ث	θ	ر	r	ط	ʈ	ك	k	ي	j	ٌ	un
ج	ʒ	ز	z	ظ	ðˤ	ل	l	ء	a	ٍ	in
ح	ħ	س	s	ع	ʕ	م	m	ـ	a	ْ	silence

Figure 11 : Correspondance graphème phonème de la langue arabe suivant l'IPA

Le système phonétique de la langue arabe comprend cinq types de syllabes qui sont CV, CVV, CVC, CVCC et CVVC classées en fonction de traits ouverts et fermés [19]. Une syllabe est ouverte (respectivement fermée), si elle se termine par une voyelle V (respectivement par une consonne C). Les cinq types de syllabes sont comme suit :

1. la syllabe CV ex. /ka/ : "comme"
2. la syllabe CVV ex. /laa/ : "non"
3. la syllabe CVC ex. /hum/ : "ils"
4. la syllabe CVVC ex /riih/ : "vent"
5. la syllabe CVCC ex /bahr/ : "mer"

Les quatre premiers types de syllabes se produisent en début, au milieu et en fin du mot. Le plus fréquent étant le type CV. Le cinquième ne se produit qu'en fin de mot ou en isolé. Cette structure est totalement différente de celle des dialectes, ou un grand nombre d'autres combinaisons est possible [20][21].

Les consonnes sont caractérisées par leurs modes et leurs lieux d'articulation [18]. Nous décrivons ci-dessous les consonnes par grandes classes phonétiques et nous reviendrons plus tard sur la description des consonnes particulières à la langue arabe.

2.3.2.1 Les plosives

Physiologiquement, une plosive est caractérisée par:

1. la formation d'une fermeture à l'intérieur de la cavité vocale par un ou plusieurs articulateurs à l'endroit où le conduit de pression est bloqué et qui apparaît comme un vide sur le spectrogramme.
2. la brusque libération de cette pression qui apparaît comme une barre d'explosion ou burst sur le spectrogramme.

- / ت / : est une plosive aspirante alvéodentale non voisée. Le / ت / apparaît sur le spectrogramme comme un burst de durée relative de 20 à 40 msec, plus long avec les voyelles longues. Le burst du / ت / est sous la forme d'une barre verticale suivie par un intervalle de faible bruit.
- / ك / : est une plosive postpalatale non voisée et aspirante. Le / ك / apparaît sur le spectrogramme comme un burst sous forme d'une barre verticale suivie d'un petit bruit de friction, le tout d'une durée comprise entre 60 et 80 msec. Parfois le burst est double.
- / ب / : est décrit comme une occlusive bilabiale voisée non aspirante mais en réalité, il est en variation libre c'est-à-dire qu'il peut être voisé ou non. Le voisement du / ب / apparaît sur le spectrogramme, en basse fréquence vers 50 Hz, avec une durée de 60-110 msec.
- / د / : est une plosive alvéodentale voisée non aspirante de durée de 80-100 msec. Elle apparaît sur le spectrogramme comme un /t/ excepté pour le voisement.

2.3.2.2 Les fricatives

Les fricatives sont produites dans la cavité vocale par une constriction étroite qui rend la circulation d'air turbulente. Acoustiquement, les fricatives non voisées possèdent en général

un haut bruit aléatoire et les fricatives voisées possèdent des structures de résonance faibles qui apparaissent comme des ombres de formants faibles avec un léger bruit.

- / ج / : est un phonème voisé alvéopalatal et affriqué qui peut être non voisé en position finale. Phonétiquement, le / ج / est une combinaison de deux phones /d/ et /z/. Cette combinaison plosive-fricative apparaît sur le spectrogramme comme une plosive suivie d'une fricative voisée
- / ف / : est une fricative labiodentale non voisée qui apparaît comme un faible bruit aléatoire, sa durée varie entre 80 et 120 msec.
- / ث / : est une fricative interdentale non voisée, qui apparaît comme un bruit aléatoire plus fort que le / ف / et de durée de 80 à 120 msec.

	Occlusives	Emphatiques	Fricatives	Nasales	Liquides	Glides
Labiales	ب b		ف f	م m		و w
Interdentales		ظ V	ذ ث f v			
Dentales	ت د t d	ط ض T D		ن n	ر ل r l	
Sifflantes		ص S	س ز s z			
Palatales	ج J		ش X			ي y
Vélaires	ك k		خ غ h g			
Uvulaire	ق q					
Pharyngales			ح ع H e			
Glottales	ء '		ه h			

Figure 12 : L'alphabet Arabe

- / س / : est une fricative sifflante dentale non voisée qui apparaît aussi comme un bruit aléatoire de durée 100-170 msec, en hautes fréquences.
- / ش /: est une fricative chuintante palatale non voisée qui apparaît comme un bruit aléatoire de durée 100-150 msec, en hautes fréquences.

- / خ / : est une fricative vélaire non voisée qui apparaît comme un bruit aléatoire regroupé sous forme de structure de formant dans la bande 1000-5000. La durée relative du / خ / est de 100-160 msec.
- / ذ / : est décrit comme une fricative interdentale voisée, de durée relative 100-160 msec.
- / ز / : est une fricative sifflante dorsoalvéolaire voisée qui semble avoir trois faibles structures de formant FI vers 250 Hz, F2 à 1600 Hz et F3 à 2400 Hz. En hautes fréquences, / ز / contient un bruit aléatoire à partir de 3000 Hz avec une durée relative de 100-160 msec.
- / غ /: est décrit comme une fricative grasseyée voisée qui possède deux allophones. Sur le spectrogramme, / غ / apparaît comme un bruit à structure de formants étalée en basses fréquences et dont la limite supérieure se situe vers 6400 Hz.

2.3.2.3 Les nasales

La nasalité est définie en terme physiologique comme étant la formation d'une ou plusieurs fermetures orales et le passage de l'air à travers le nez. Au cours de la production des nasales les 2 cavités orale et nasale sont donc normalement utilisées. En Arabe, il y 2 consonnes nasales le / م / et le / ن / décrites comme suit:

- / م / : est une nasale bilabiale voisée de durée de 70-90 msec. Elle possède des résonances faibles qui apparaissent comme des formants F1 à 250 Hz, F2 à 1000 Hz et F3 à 2700 Hz et parfois, d'autres résonances plus faibles juste au-dessus de F1, en gros, le / م / semble être similaire au / ب /, excepté pour la qualité nasale caractérisée par un modèle de formants nasaux.
- / ن / : est une nasale alvéodentale voisée de durée 80 à 100 msec. Comme le / م /, le / ن / possède des résonances faibles apparaissant comme des formants F1 à 250 Hz, F2 à 1500-1600 Hz et F3 vers 2800-3000 Hz.

2.3.2.4 Le vibrant

- / ر / : est un vibrant apicoalvéolaire lingual voisé qui comporte une vibration accentuée de la pointe de la langue. La durée du / ر / est de l'ordre de 80-100 msec, il possède des structures de formants qui sont interrompus par des intervalles verticaux très courts de silence de l'ordre de 10 msec qui peuvent être interprétés

physiologiquement comme le résultat de la frappe du bout de la langue contre le palais.

2.3.2.5 Le latéral

- / ل / : est un phonème lingual qui possède deux allophones : le plus commun de ces allophones est une latérale (/munharif/) dentale voisée de durée 80-120 msec, qui possède des structures de formants similaires à celles des voyelles F1 à 300 Hz, F2 à 1500-1600 Hz et F3 à 2400-2500 Hz. L'autre allophone est une consonne emphatique latérale postdentale qui se produit dans un environnement extrêmement limité.

2.3.2.6. Les semi-voyelles

- / و / : est une semi-voyelle bilabiale de durée 80-100 msec. Il possède des structures de formants similaires à celles du /u/ et /u : / avec F1 à 350 Hz, F2 à 950 Hz et F3 à 2100 Hz. Les débuts et les fins de F2 de / و / glissent vers les phones précédant et suivant.
- / ي / : est une semi-voyelle palatale de durée 80-100 msec. Le / ي / possède des formants similaires à ceux du /i/ et /i:/ avec FI à 275 Hz, F2 à 1900 Hz et F3 à 2650 Hz. Généralement, les débuts et les fins de / ي / glissent vers les phones précédentes et suivantes.

2.3.2.7 Les consonnes glottales et pharyngales

Les consonnes glottales et pharyngales se distinguent des autres consonnes par le fait qu'elles ont des lieux d'articulations verticaux. Un lieu d'articulation vertical est défini comme un ensemble de localisations anatomiques qui vont du palais jusqu'à la glotte inclusivement par opposition au lieu horizontal où les emplacements sont entre les lèvres et la luette.

Ces consonnes sont plus difficiles à étudier parce que leurs points et leurs manières d'articulation sont dans la région laryngale et pharyngale qui ne sont pas facilement accessibles.

L'Arabe comporte deux consonnes glottales / ه /et / ء /et deux consonnes pharyngales / ح / et

/ ع / que nous allons décrire dans ce qui suit :

- **Les consonnes glottales**

- / ه / : est décrit comme une fricative glottale non voisée de durée relative 100-160 msec qui apparaît le plus souvent comme un bruit à structure de formants et qui devient voisée en milieu intervocalique.
- / ء / : est décrit comme une plosive glottale non voisée dont la structure acoustique est très dépendante du contexte de production et de sa position à l'intérieur du mot. Initialement, le / ء / apparaît sur le spectrogramme sous forme variée. Dans certains cas, il est sous forme d'un burst suivi d'un petit intervalle de silence de durée 15-20 msec ou bien suivi d'un faible bruit. En milieu non intervocalique, le / ء / apparaît comme un intervalle de silence de durée 65-85 msec suivi d'un burst de durée environ 15 msec. En position finale, le / ء / est en variation libre et apparaît comme un burst qui peut être suivi ou non d'un bruit faible, ce burst est précédé par un intervalle de silence de durée 80-120 msec.

➢ **Les consonnes pharyngales**

- / ح / : est une fricative pharyngale non voisée de durée 100-150 msec qui devient voisée en milieu intervocalique. Lors de la production du / ح / une constriction est formée par le dorsum de la langue contre la paroi postérieure du pharynx et c'est cette constriction qui différencie principalement le / ح / du / ه /. Acoustiquement, le / ح / apparaît comme un bruit plus fort que celui du / ه /.
- / ع /: est décrit comme une fricative pharyngale voisée dont la structure est très dépendante du contexte de production: en position initiale, le / ع / apparaît sur le spectrogramme comme une sorte de "burst" - durée 40-50 msec - dont l'intensité est quelque part entre 1450 et 1550 Hz.

2.3.2.8 Les consonnes emphatiques

Le système phonétique de l'Arabe tire son originalité de la présence des consonnes emphatiques: la langue arabe est souvent appelée la langue du / ض /. Un phonème qui n'existe qu'en Arabe et qui est d'ailleurs difficile à prononcer. Les quatre articulations définies traditionnellement comme emphatiques sont / ض /, / س /, / ظ / et / ت /. Leur nombre varie d'un auteur à un autre. Le sentiment des non linguistes est généralement que les sons emphatiques sont prononcés avec fermeté et possède donc une autre tonalité.

2.4 L’évolution de la reconnaissance vocale Arabe

La Reconnaissance de la Parole Arabe remonte aux années 70. Depuis cette date, plusieurs solutions ont été proposées. Nous nous contenterons ici d'énumérer quelques exemples à titre non exhaustif.

Pour appréhender cette thématique plusieurs approches ont été tentées entre autres l’approche par déformation temporelle (DTW), les modèles de Markov cachés, les réseaux de neurones ainsi que certaines techniques hybrides.

Pour la conception d’un système de reconnaissance des chiffres arabe, DAHMANI a utilisé l’algorithme de Bridle et Nqkagawa : la programmation dynamique en une seule passe (the one-pass dynamic programming) qu’est l’approche la plus utilisée actuellement dans les systèmes de reconnaissance par rapport à celle de Sakoe ‟the two level dynamic programming‟ et celle de Myers ‟ the level building dynamic programming‟.

Les résultats obtenus ont donné un taux de l’ordre de 42 % de chaînes, ce faible score peut s’expliquer par le fait que les locuteurs participant à l’élaboration de la base n’étaient pas entraînés.

Dans le cas général, l’élasticité temporelle, formalisée mathématiquement par l’algorithme Dynamic Time Warping (DTW) s’est rapidement confrontée aux grands problèmes de la reconnaissance automatique de la parole. Comment faire face à la variabilité due aux locuteurs et au contexte d’enregistrement, comment élaborer une construction sémantique et non simplement lexicale et donc comment gérer de très grands dictionnaires ?

Des recherches ont été menées sur la reconnaissance automatique des chiffres et de l’alphabet arabe. En 1985, Hagos [23] et Abdullah [24], ont mené séparément des recherches sur la reconnaissance des chiffres arabe.

Hagos a conçu un système de reconnaissance indépendant du locuteur indépendant arabe chiffres arabes. Son système est basé sur les paramètres LPC pour l'extraction de caractéristiques, sa méthode utilise la technique du log de vraisemblance pour les mesures de similarité.

Abdullah a développé une autre technique de reconnaissance de chiffres arabe, celle-ci utilise comme algorithme d’extraction de paramètres : la pente-positive (positive-slope) la durée de

passage par zéro (zero crossing duration). Il a obtenu un taux de reconnaissance de l'ordre de 97%.

Al-Otaibi [25] a mis au point un système automatique de reconnaissance des voyelles en arabe. Il a étudié la nature syllabique de la langue arabe en termes : de types de syllabe, de structure syllabique et des règles de contrainte primaire.

Amiar [26] présente un système de reconnaissance basé sur l'utilisation des Modèles de Markov Cachés (MMC) discrets. Dans ce système de reconnaissance, le signal vocal est analysé et représenté par un ensemble de vecteurs caractéristiques, obtenus par analyse prédictive linéaire LPC. Les expériences menées ont permis d'atteindre un taux de reconnaissance égal à 87,43%

Dans une tentative d'amélioration du taux de reconnaissance, Bourouba [27] propose d'étendre la méthode HMM en la combinant avec l'algorithme DTW afin de combiner les avantages de ces deux puissantes techniques de reconnaissance. Les caractéristiques du signal sont extraites en utilisant les MFCC et l'analyse LPC. Les expériences montrent que le Système DTW/GHMM augmente la reconnaissance moyenne de 2 % par rapport à la reconnaissance traditionnelle à base des HMMs.

Un système de reconnaissance automatique basé sur les réseaux de neurones artificiels a été conçu et testé dans le cas des chiffres arabes par Y.A. Alotaibi [28][29]. Ce système avait pour tâche la reconnaissance de la parole isolée et a été mis en œuvre pour d'une part le cas multi locuteurs (le même ensemble de locuteurs a été utilisé dans les phases de test et d'apprentissage) et d'autre part dans le cas du mode indépendant du locuteur (c'est-à-dire les locuteurs utilisés pour la phase de reconnaissance sont différents de ceux utilisés pour l'apprentissage). Ce système de reconnaissance a atteint 99,5% de réponses correctes dans le mode multi locuteurs et 94,5% en mode indépendant du locuteur pour la parole non bruitée.

Un autre algorithme conçu par Djemili [30] et al. pour la reconnaissance des chiffres en arabe. Le système ainsi conçu, utilise les paramètres acoustiques du signal de parole comme éléments d'entrée pour les réseaux de neurones multicouches.

Moaz [31] et al, proposent une approche à reconnaître les lettres de l'alphabet arabe en utilisant des réseaux neuronaux artificiels en utilisant la perception multicouches, qui est l'un des algorithmes de classification les plus couramment utilisés pour les réseaux de neurones.

Les principales caractéristiques du signal de parole sont extraites à l'aide de la technique de l'analyse en composantes principales. Ce système a réalisé un taux de détection de 96 %.

Le chapitre suivant sera dédié aux différentes expérimentations que nous avons menées dans nos recherches pour l'implémentation d'un système de reconnaissance automatique de la parole arabe, cas des mots isolés aussi bien par l'approche globale que par l'approche analytique.

2.5 Conclusion

La description de l'appareil phonatoire nous permet d'expliquer comment se produit la parole. Il est donc important d'identifier une classification articulatoire des consonnes et des voyelles et de l'art qui en découle.

Notre intérêt se porte, bien évidemment, sur la phonétique de la langue Arabe, branche de la linguistique, qui étudie les sons utilisés dans la communication verbale et l'évolution de la reconnaissance vocale arabe.

Chapitre 3 : Les outils mathématiques et logiciels pour le traitement de la parole

3.1 Introduction

Dans ce chapitre nous décrirons les différents outils nécessaires au traitement de la parole. Nous commencerons par un bref aperçu sur le traitement à court-terme, ensuite nous exposerons les concepts de base de la paramétrisation du signal vocal pour calculer les moments spectraux des consonnes et voyelles de ce dernier. Nous nous baserons sur deux algorithme, celui de ZCR et celui de niveau d'énergie.

3.2 Traitement du signal à court terme

3.2.1 Définition

Le signal parole est un processus aléatoire non stationnaire, or les outils de traitements du signal conventionnels sous-entendent la stationnarité du signal, alors on va exploiter le fait que le signal parole soit quasi stationnaire sur des courts segments de parole appelés "frames" en anglais. Ces derniers sont des tranches temporelles de 10 à 45 ms, d'où l'appellation de l'analyse à cour-terme.

3.2.2 Fenêtrage

Généralement, le découpage du signal dans le domaine temporel équivaut à multiplier le signal par une fonction rectangulaire, ce qui équivaut à une convolution dans le domaine fréquentiel entre le spectre du signal analysé et celui de la fenêtre. Dans la majorité des cas, la fenêtre rectangulaire s'avère trop brutale. En effet, il a été démontré que toute variation rapide dans le domaine temporel correspond à des hautes fréquences dans le domaine fréquentiel qui se traduit par des ondulations sur le spectre. Alors, on lui préfère d'autres fenêtres plus douces. Parmi les fenêtres les plus utilisées on trouve [32][33] :

Bartlett :

$$w(n)=\begin{cases}2n/(N-1) \text{ pour } 0\le n\le (N-1)/2\\ 2-2n/(N-1) \text{ pour } (N-1)/2<n\le N-1\\ 0 \text{ ailleurs}\end{cases}$$

Hanning :

$$w(n)=\begin{cases}0.5-0.5\cos(2\pi n/(N-1)) & \text{pour } 0\le n\le N-1\\ 0 & \text{ailleurs.}\end{cases}$$

Hamming :

$$w(n)=\begin{cases}0.54-0.46\cos(2\pi n/(N-1)) & \text{pour } 0\le n\le N-1\\ 0 & \text{ailleurs.}\end{cases}$$

Blackman :

$$w(n)=\begin{cases}0.42-0.5\cos(2\pi n/(N-1))+0.08\cos(4\pi n/(N-1)) & \text{pour } 0\le n\le N-1\\ 0 & \text{ailleurs.}\end{cases}$$

Où N représente la langueur de la fenêtre, et n un échantillon du signal.

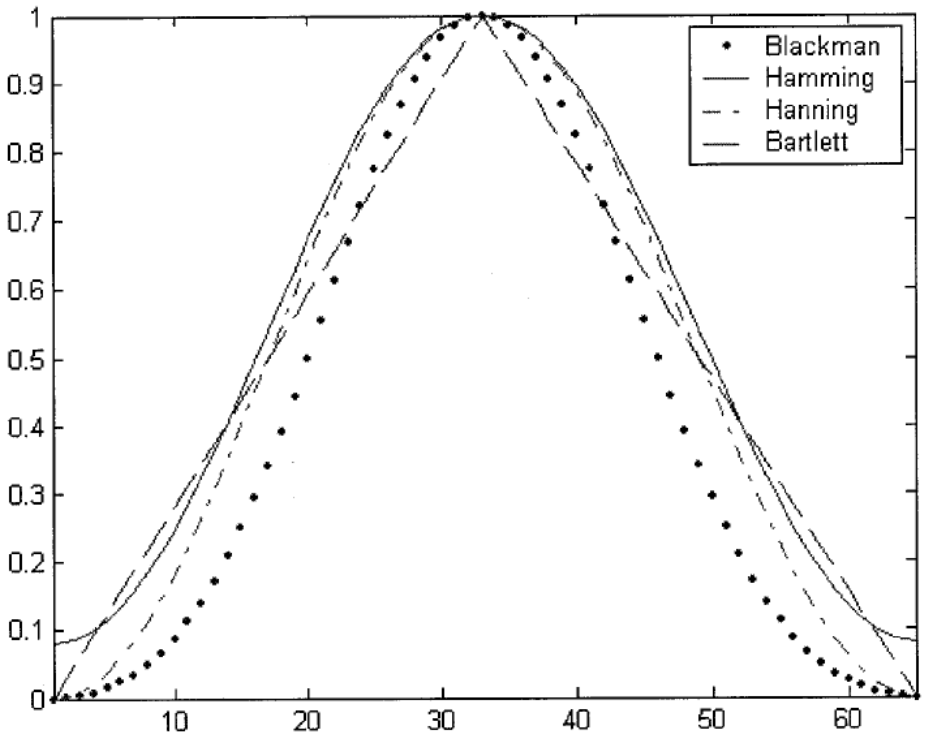

Figure 13 : Exemples de fenêtres de pondération

En pratique, la fenêtre de Hamming est souvent la plus utilisée. La Figure 14 est une illustration de son spectre et celui de la fenêtre rectangulaire. Dans cette figure, on voit clairement que la fenêtre de Hamming permet une grande atténuation en dehors de la bande passante comparativement à la fenêtre rectangulaire d'où son avantage.

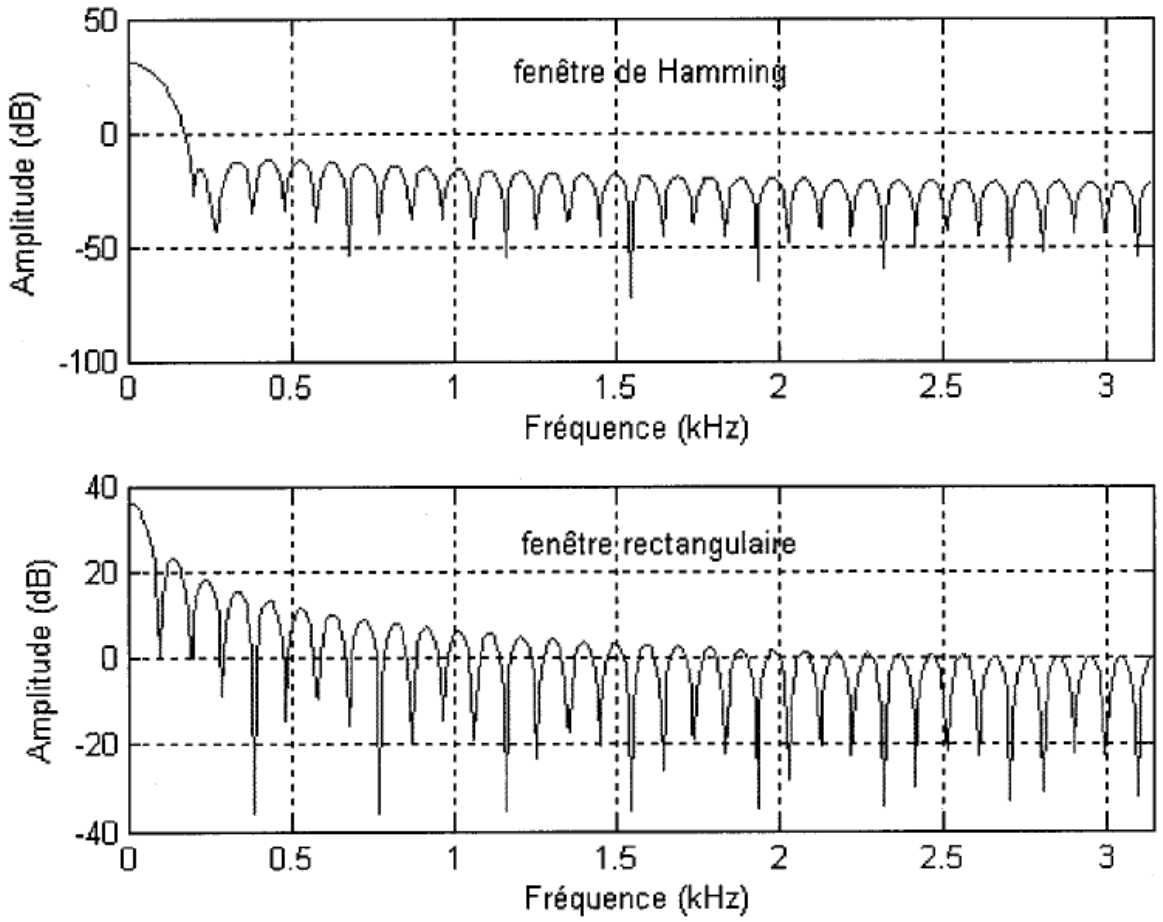

Figure 14 : Spectre des fenêtres de Hamming et Rectangulaire

Lors du traitement du signal, on peut prendre des fenêtres avec recouvrement (overlapped) ou non. La région de recouvrement peut varier de 0 à 75% de la taille de la fenêtre N.

3.2.3 Énergie à court- terme

Un des outils qui permettent de fournir une représentation fidèle des variations de l'amplitude du signal vocal x(n) dans le temps est l'énergie court terme (voir Figure l5).
En général, elle est définie par [34] :

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)w(n-m)\right]^2$$

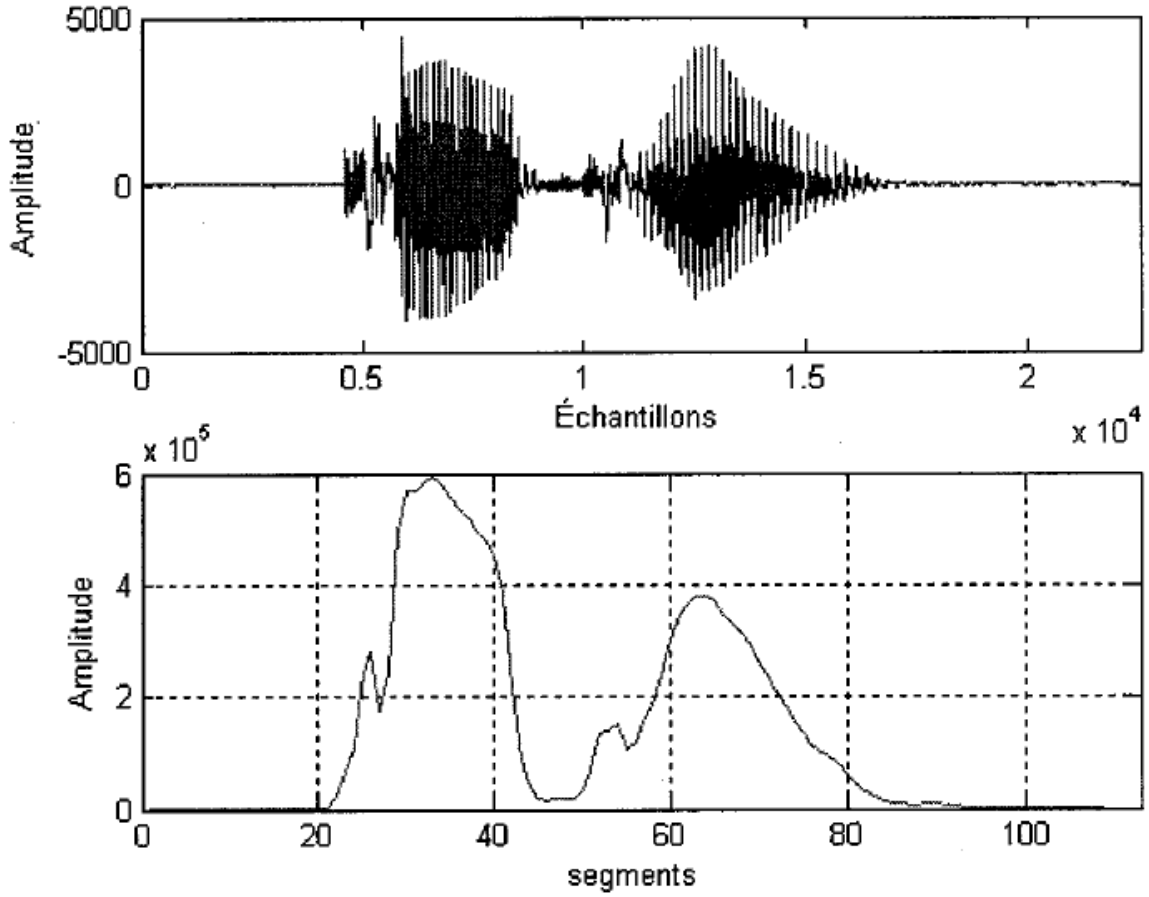

Figure 15 : L'énergie à court terme d'un signal vocal

En absence du bruit de mesure, l'énergie s'avère un outil efficace pour séparer la parole du silence [35].

3.2.4 Amplitude moyenne

L'énergie à court terme avec une élévation au carré pour chaque échantillon, est très coûteuse en termes de temps de calcul. En pratique, on préfère utiliser une autre forme qu'on appelle l'amplitude moyenne, elle est définie par :

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)| w(n-m)$$

3.2.5 Puissance à court terme

La puissance à court terme d'un segment de parole de longueur N est définie par:

$$P_n = \frac{1}{N} \sum_{m=-\infty}^{+\infty} \left[x(m) w(n-m) \right]^2$$

Il faut noter que l'énergie court terme et la puissance court terme fournissent à un facteur près (1/N) la même information.

3.2.6 Le taux de passage par zéro à court terme

Un autre outil très utile de traitement de la parole est le taux de passage par zéro (zéro crossing rate en anglais) Pour un signal échantillonné, il y'a passage par zéro lorsque deux échantillons successifs sont de signes opposés. Le taux de passage par zéro court terme est estimé par la formule [36] :

$$Z_n = \frac{1}{2}\sum_m \left|\text{sgn}\left[x(m)\right] - \text{sgn}\left[x(m-1)\right]\right| w(n-m)$$

avec

$$\text{sgn}\left[x(m)\right] = \begin{cases} 1\,, si & x(m) \geq 0. \\ -1\,, si & x(m) < 0. \end{cases}$$

et

$$w(n) = \begin{cases} \dfrac{1}{N} & , 0 \leq n \leq N \\ 0 & , \text{ailleurs.} \end{cases}$$

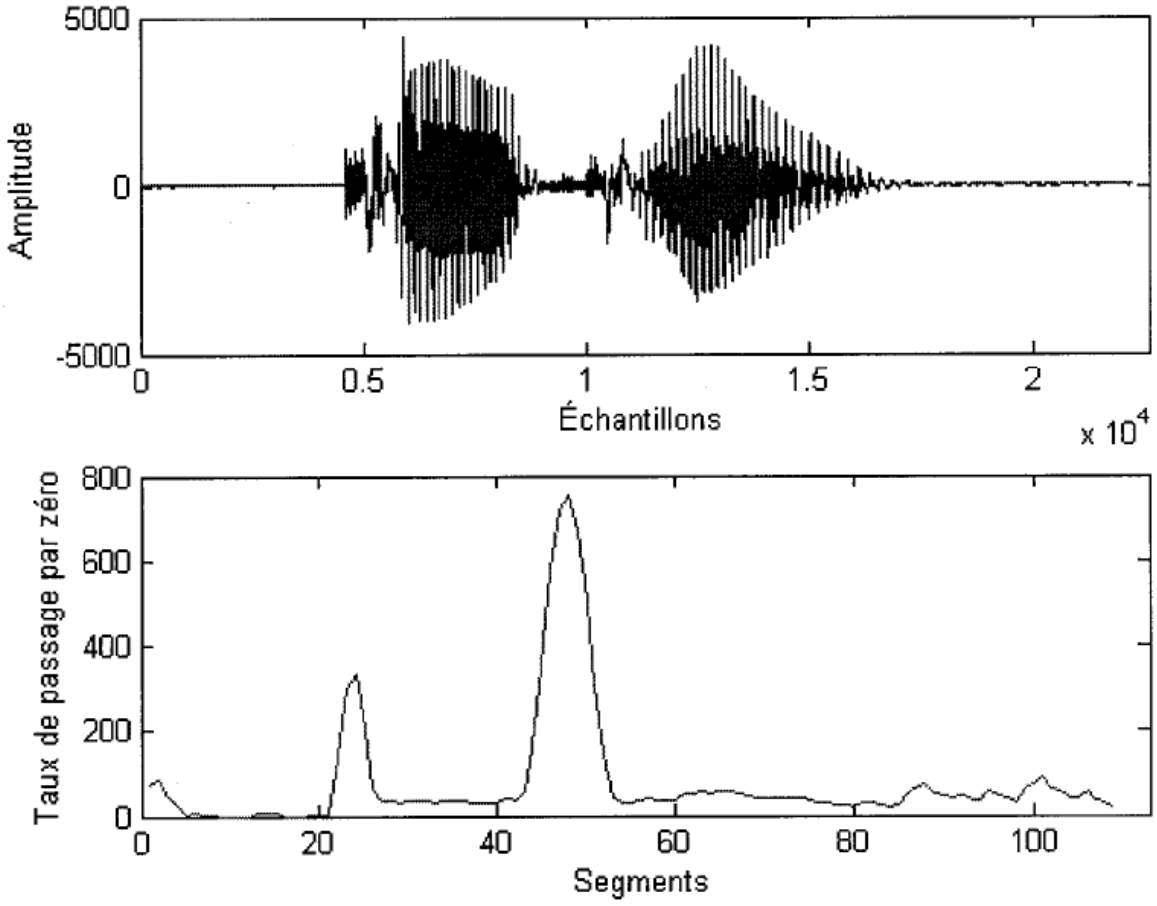

Figure 16 : Le taux de passage par zéro d'un signal vocal

Une caractéristique pour le taux de passage par zéro, est qu'il est élevé pour le son non voisé et faible pour le son voisé. Le taux de passage par zéro constitue un outil important pour la classification voisé/non voisé, et pour la détection du début et la fin de la parole dans un signal vocal.

3.2.7 La FFT : Transformée de Fourier Rapide

La Transformée de Fourier Rapide (notée par la suite FFT) est simplement une Transformée de Fourier Discrète (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.

Pour calculer une TFD, on doit calculer N valeurs X(k) :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{nk}{N}}$$

et ceci pour k∈[0,N−1].

Si on effectue le calcul directement sans algorithme efficace, on doit effectuer:

- N^2 multiplications complexes
- N(N−1) additions complexes

Il existe différents algorithmes de FFT. Le plus connu est sûrement celui de Cooley-Tukey (appelé aussi à entrelacement temporel ou à «decimation in time») qui réduit à :

(N/2)*log2(N) le nombre de multiplications.

Il existe deux versions de l'algorithme:

- FFT avec entrelacement temporel,
- FFT avec entrelacement fréquentiel.

L'algorithme nécessite que N soit une puissance de 2. Le principe de l'algorithme consiste à décomposer le calcul de la TFD d'ordre $N=2^l$ en l étapes successives.

En pratique la transformée en Z est remplacée par une transformée de Fourier Rapide (ou FFT) qui possède les mêmes propriétés de linéarité que la transformée en Z.

La phase de la FFT du signal de parole ne contient pas d'informations suffisamment pertinentes pour la reconnaissance de la parole, alors il est judicieux de garder juste la partie réelle, c'est ce qu'on va faire en prenant seulement le module au carré de la FFT.

3.2.8 Moments spectraux

La forme générale du spectre a été paramétrée avec des moments spectraux qui sont dérivés de moments statistiques qui sont parfois appliqués à l'analyse de la forme d'un histogramme. Les moments statistiques peuvent être calculés comme suit :

$$moyenne = \overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$$Déviation\ Standard = S_x = \left(\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}\right)^{1/2}$$

$$skewness = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^3}{\frac{n}{S_x^3}}$$

$$kurtosis = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^4}{\frac{n}{S_x^4}}$$

où Xi et n représentent respectivement les échantillons et le nombre d'échantillons.

Pour calculer les moments spectraux, un spectre est traité comme un histogramme de sorte que Xi représentent les intervalles de fréquence et n est la valeur en dB à une fréquence donnée.

Le spectre de puissance ou la puissance spectrale est le principal outil de traitement du signal, et des algorithmes d'estimation du spectre de puissance ont trouvé des applications dans des domaines tels que le radar, sonar, sismique, biomédical, les communications et le traitement du signal de parole.

La puissance spectrale normalisée est :

$$p(f_k) = \frac{P(f_k)}{\sum_{k=0}^{N/2} P(f_k)}$$

où $P(f_k)$ est la puissance spectrale, $f_k = 2f_{Nq}\ k/N$, k=0,1,…,N/2, N=256. f_{Nq} indique la fréquence de Nyquist (11025 Hz).

Grâce à cette puissance, on peut trouver les moments spectraux: Moyenne (Mean), Déviation Standard (Standard Deviation), Coefficient d'asymétrie (skewness), Coefficient d'aplatissement (Kurtosis).

La moyenne ou le centroïde spectral est un paramètre essentiel dans la perception du timbre. C'est une mesure utilisée en traitement numérique du signal pour caractériser un spectre. Il indique où se situe le centre de gravité du spectre.

La moyenne :

$$\mu = \sum_{k=0}^{N/2} p(f_k) f_k$$

La déviation standard ou l'étendue spectrale ou l'écart-type est l'étendue du spectre autour de sa valeur moyenne. L'écart-type de la distribution est la principale mesure de la variabilité. Elle a plusieurs propriétés d'intérêt, dont on cite que lorsque la distribution de l'échantillon est normale, le pourcentage de partitions qui relèvent de toute valeur spécifiée d'un écart-type peut être calculé. Un skewness positif indique que la queue sur le côté droit est plus longue que le côté gauche et le nombre des valeurs se situent à gauche de la moyenne. Une valeur nulle indique que les valeurs sont relativement uniformément réparties des deux côtés de la moyenne, typiquement (mais pas nécessairement) impliquant une distribution symétrique.

La déviation standard :

$$\sigma = \sqrt{\sum_{k=0}^{N/2} (f_k - \mu)^2 p(f_k)}$$

Le skewness mesure la symétrie du spectre autour de sa moyenne. Si les partitions sont réparties symétriquement autour de la moyenne, l'asymétrie est nulle. Lorsque la distribution des scores s'étend de la moyenne vers les grandes valeurs de la distribution normale (Figure 17), elle est dite asymétrique positivement. Le complément de cette asymétrie est négatif, qui se produit lorsque les valeurs de la moyenne s'étendent vers les valeurs petites de la distribution normale.

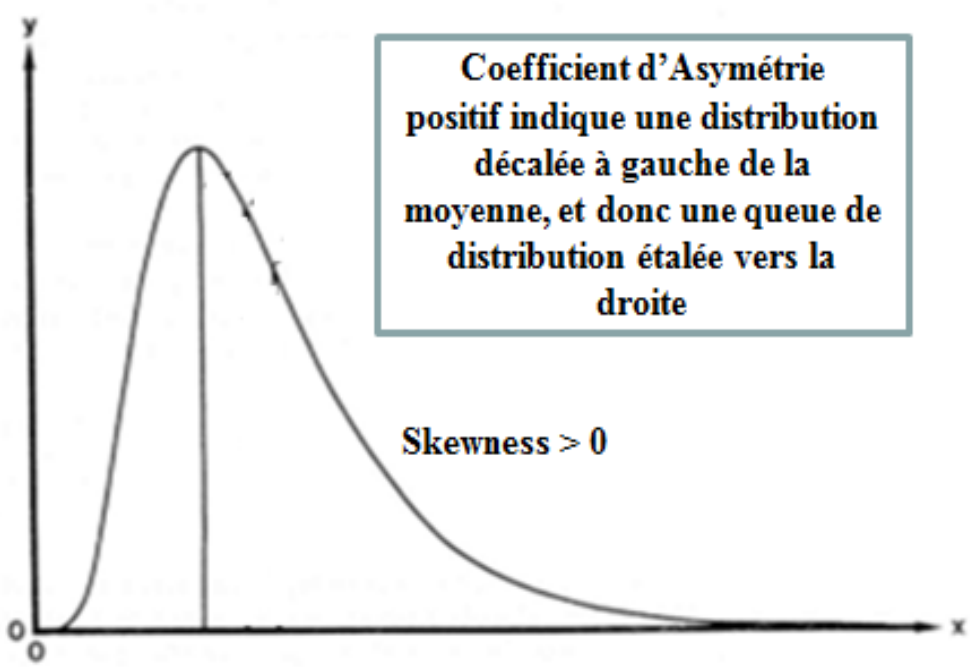

Figure 17 : Représentation de l'asymétrie

Le coefficient d'asymétrie :

$$s = \sum_{k=0}^{N/2} \left(\frac{f_k - \mu}{\sigma}\right)^3 p(f_k)$$

Le kurtosis a été introduit dans les années quatre-vingts pour détecter les transitoires dans les signaux sonar. Après cette application isolée, il a été récemment réintroduit dans la communauté du traitement du signal pour distinguer les différents types de signaux. Il mesure l'aplatissement du spectre autour de sa moyenne. Il reflète l'aplatissement de la distribution, avec une courbe normale ayant une valeur de kurtosis-3. Un haut aplatissement conduit à un nombre supérieur à 3 alors que la platitude conduit à une estimation d'aplatissement entre zéro et trois (Figure 18). Certains chercheurs soustraient la valeur 3 de l'estimation d'aplatissement dans l'ordre que zéro représente la valeur de l'aplatissement lorsque la distribution est normale.

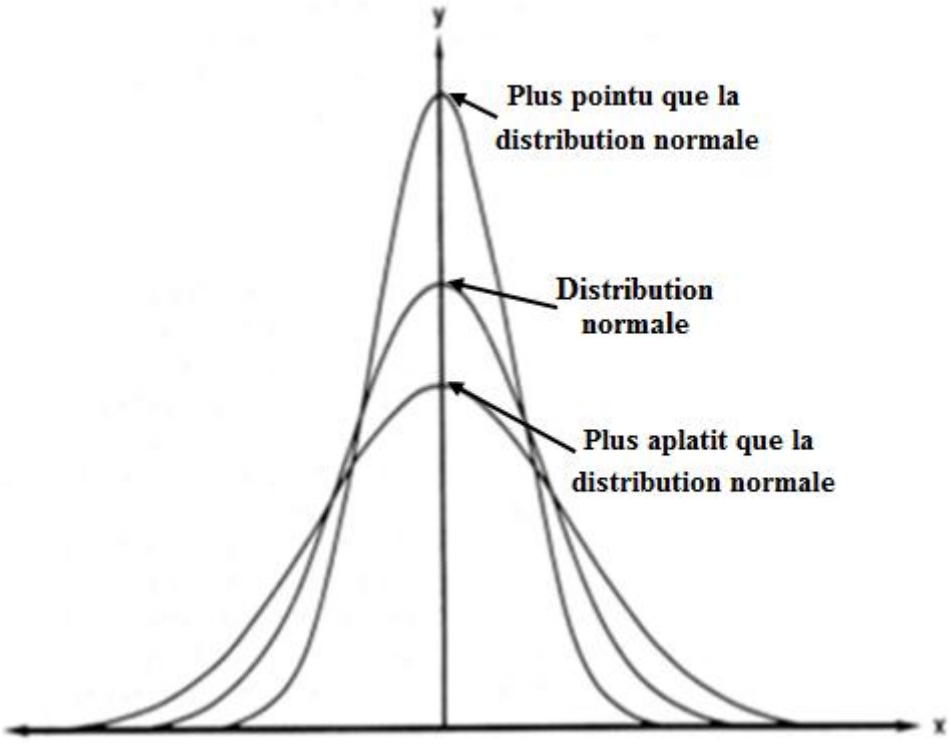

Figure 18 : Représentation de l'aplatissement

Le coefficient d'aplatissement :

$$k = -3 + \sum_{k=0}^{N/2} \left(\frac{f_k - \mu}{\sigma}\right)^4 p(f_k)$$

3.3 Outils logiciels

3.3.1 MATLAB

MATLAB (matrix laboratory) est un langage de programmation de quatrième génération et un environnement de développement, il est utilisé à des fins de calcul numérique. Développé par la société The MathWorks, MATLAB permet de manipuler des matrices, d'afficher des courbes et des données, de mettre en œuvre des algorithmes, de créer des interfaces utilisateurs, et peut s'interfacer avec d'autres langages comme le C, C++, Java, et Fortran. Les utilisateurs de MATLAB (environ un million en 2004) sont de milieux très différents comme l'ingénierie, les sciences et l'économie dans un contexte aussi bien industriel que pour la recherche.

Nous avons utilisé MATLAB pour appliquer les algorithmes suivants :

- Extraction du signal utile
- Affichage de la puissance lisse du signal
- Séparation automatique d'un signal en consonne et voyelle
- Visualisation des consonnes et des voyelles séparément dans le domaine temporel et fréquentiel
- Affichage des moments spectraux pour ces Consonnes et Voyelle
- Développement d'une interface qui englobe tout ce qui précède

3.3.2 PRAAT

Praat est un logiciel libre scientifique gratuit conçu pour la manipulation, le traitement et la synthèse de sons vocaux (phonétique). Il a été conçu à l'institut de sciences phonétiques de l'université d'Amsterdam par Paul Boersma et David Weenink.

Nous avons utilisé PRAAT pour :

- Faire des enregistrements : d'un corpus symétrique et d'un autre asymétrique
- Visualiser le spectrogramme des signaux enregistrés
- Faire la segmentation manuelle des consonnes et voyelles

3.4 Conclusion

Dans ce chapitre nous avons présenté quelques outils pour le traitement du signal vocal, qu'on appelle aussi analyse court-terme, en référence à l'utilisation des segments de parole de courte durée durant laquelle le signal est quasi-stationnaire pour pouvoir utiliser ces outils.

La parole est un signal redondant, ce qui lui confère une meilleure résistance au bruit. Cependant, les informations qu'il véhicule ne sont pas toutes pertinentes pour la reconnaissance de la parole. Ainsi, en pratique et dans le but de réduire le nombre de données à traiter, le signal est représenté par un ensemble limité de paramètres, c'est la paramétrisation du signal.

Pour l'extraction des paramètres et caractéristiques de signal, nous avons déterminé les moments spectraux qui nous donnent une idée sur le phonème.

Dans le chapitre suivant, nous développons les algorithmes que nous avons utilisés ainsi qu'une synthèse d'une interface graphique résumant notre projet.

Chapitre 4 : Expérimentations et Résultats

4.1 Introduction

Dans ce chapitre, nous présentons le déroulement de notre travail à partir de l'enregistrement du corpus dont la finalité est une interface graphique. Ceci est la mission de ce projet de fin d'étude.

4.2 Corpus

Pour caractériser le signal vocal en langue arabe, lors de l'introduction d'une voyelle, d'une consonne ou d'une syllabe sur le contexte CV, nous avons construit notre corpus en langue arabe standard moderne contenant les trois structures CV et CVCV symétriques et asymétriques.

Trois tranches d'âge sont considérées : moins de 15ans, les adultes et les âgés. Dix hommes et dix femmes locuteurs marocains ont été invités à prononcer, à cinq reprises, les consonnes /b/ /m/ /d/ /t/ /k/ /kh/ avec les trois voyelles /a/ /i/ /u/. L'enregistrement a été fait en utilisant le logiciel « Praat » (Figure 19) à une distance de 20 cm du microphone (NGS Microphone MS102; Sensibilité : 105 dB; Réponse de fréquence : 50-16000 Hz; Longueur du câble : 1,79 m) (Figure 20) dans une chambre isolée et calme. Le son est directement numérisé sur un PC avec une fréquence d'échantillonnage de 22050 Hz, une quantification linéaire sur 16 bits et une durée de 2 à 3 secondes pour chaque syllabe.

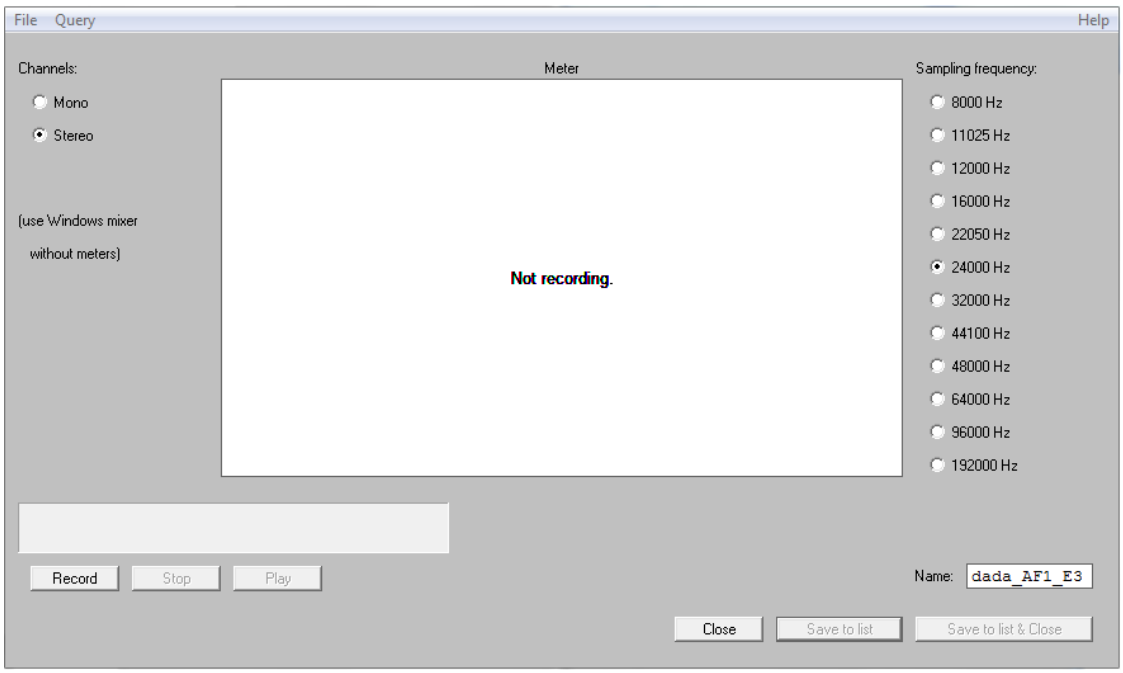

Figure 19 : Enregistrement de la syllabe /dada/ sous Praat

Figure 20 : NGS Microphone MS102

4.3 Extraction du signal utile

Nous devons éliminer le bruit au début et à la fin du signal qui présente le silence associé à un bruit de la salle d'enregistrement avant de segmenter ces derniers.

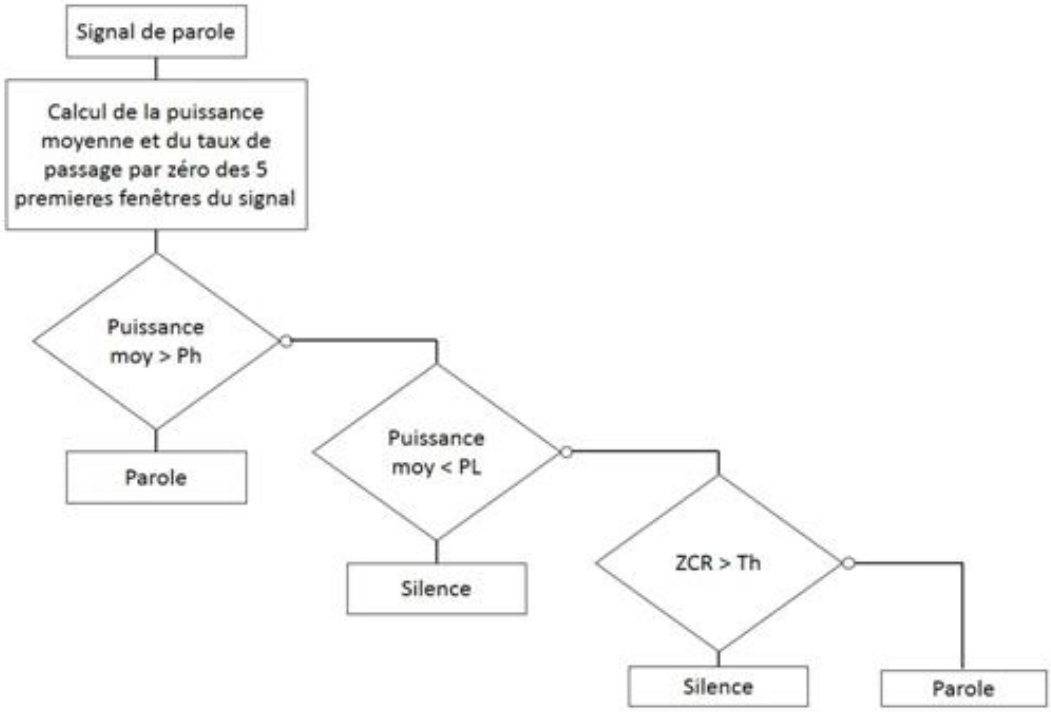

Figure 21 : Organigramme de détection du signal utile

notre algorithme de détection de la parole.

Avec :

- PL : Puissance moyenne basse du bruit
- Ph : Puissance moyenne haute du bruit
- Th : Taux de passage par zéro de bruit

La figure 22 présente cette extraction en se basant sur les deux algorithmes de ZCR et le niveau d'énergie.

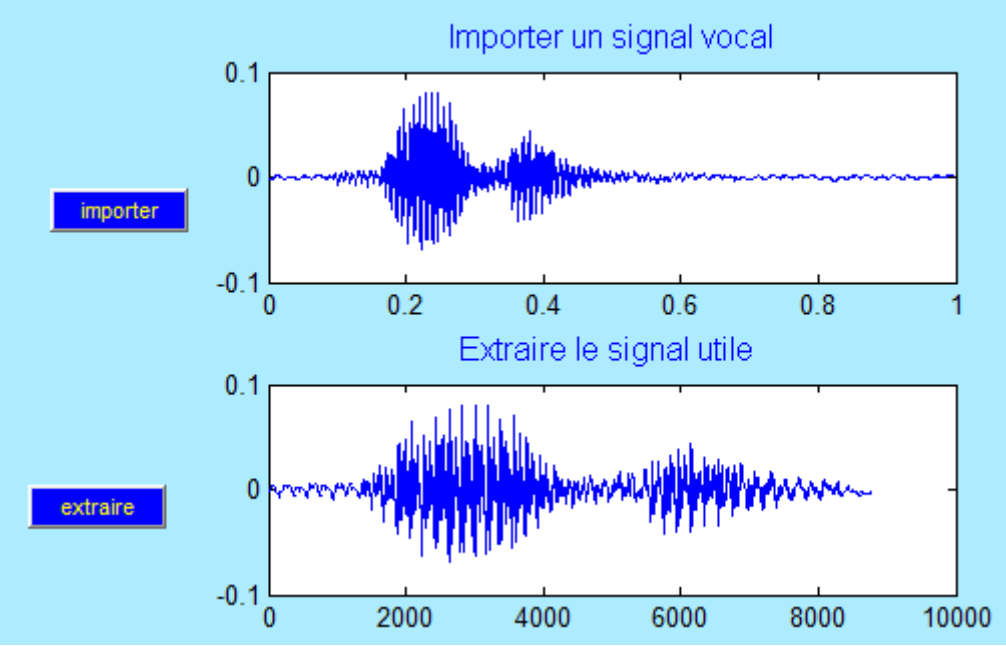

Figure 22 : Extraction du signal utile du mot enregistrer /dada/

4.4 Segmentation

Avant toute action d'analyse du signal, puisque nous ne pouvons pas travailler sur le signal tout entier, nous avons commencé par segmenter et étiqueter les enregistrements en consonnes-voyelles. Nous avons ensuite formé des dictionnaires pour chaque consonne et chaque voyelle.

Nous allons effectuer cette segmentation manuellement et automatiquement.

4.4.1 Segmentation manuelle

Nous avons réalisé cette tâche de segmentation manuellement en utilisant le logiciel «Praat» (Figure 22). Le début et la fin de chaque situation de production sont déterminés par l'apparition/disparition d'onde dans la première/dernière période sur le signal acoustique en fonction de la modification brusque de l'intensité (Al-Tamimi, 2004). Les voyelles sont déterminées par leurs puissances sonores qui sont plus grandes que celles des consonnes.

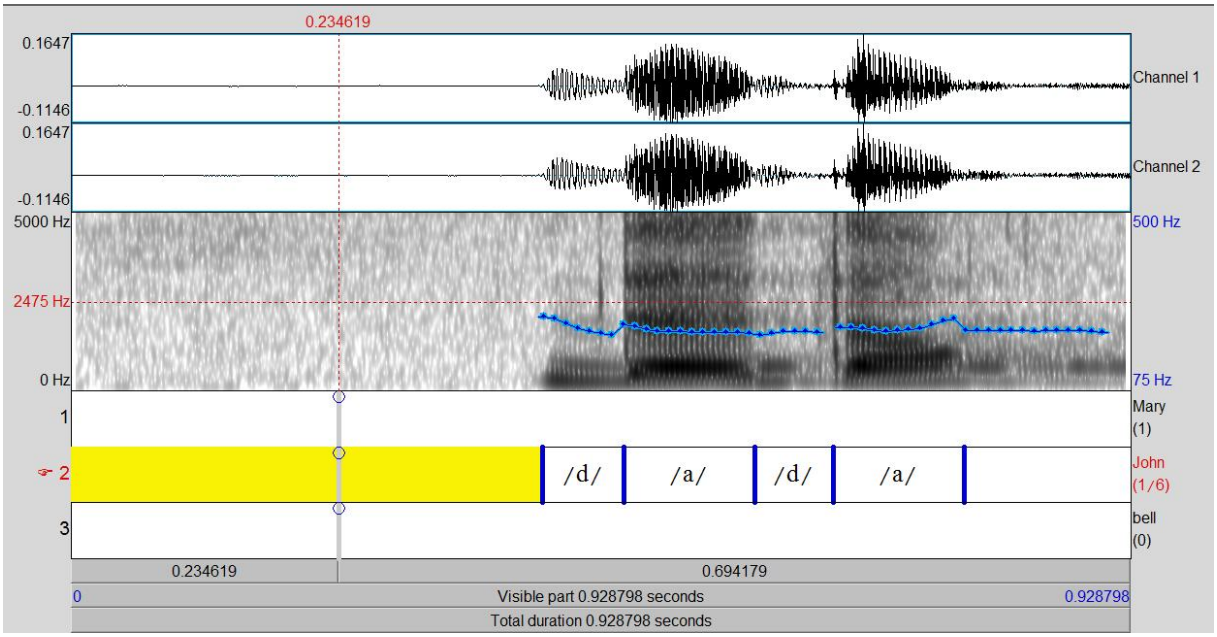

Figure 23 : Segmentation manuelle de la syllabe /dada/ pour le contexte CVCV

4.4.2 Segmentation automatique

Nous avons ensuite élaboré un programme sous MATLAB qui permet de séparer automatiquement les voyelles des consonnes. La figure 23 représente l'organigramme décrivant la détection de la transition consonne-voyelle et voyelle-consonne.

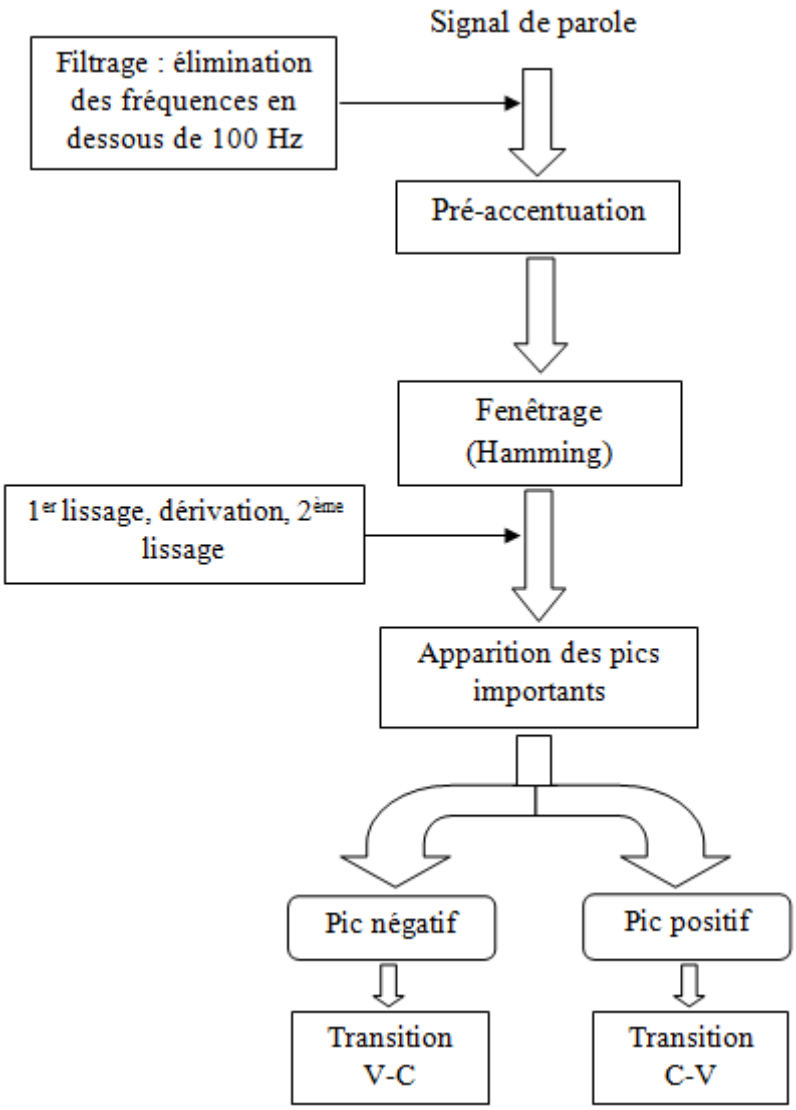

Figure 24 : Organigramme de détection des transitions C_V et V_C

Le signal de parole est bruité, donc il nous faut le filtrer pour éliminer le bruit de fond en calculant la FFT. Ensuite, nous coupons les harmoniques en-dessous de 100Hz pour un renvoi de la FFT inverse.

Le signal est lissé pour éviter les pics de fluctuation puis il est dérivé et lissé une deuxième fois pour le rendre facile à traiter. Ensuite, nous cherchons les pics les plus brutaux. Si un pic est négatif, la puissance a donc subi une forte baisse, c'est une transition Voyelle-Consonne. Si le pic est positif, c'est une transition Consonne-Voyelle (Figure 24).

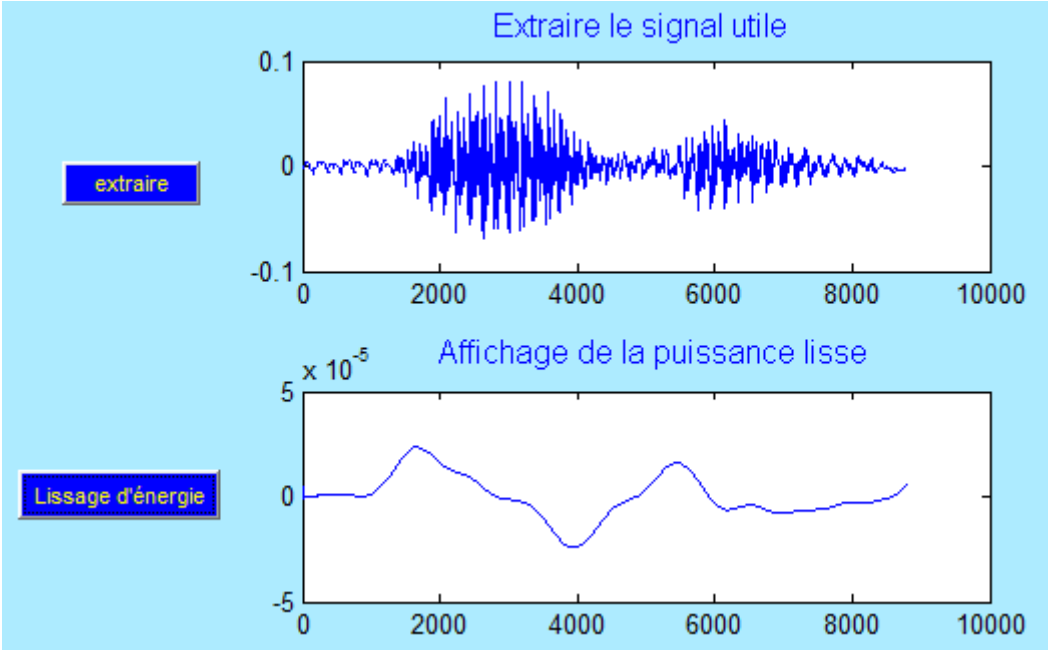

Figure 25 : Détection des transitions consonne-voyelle dans le mot /dada/

4.5 Taux de performance

Ce programme permet d'atteindre un taux de performance de 70%. Les transitions sont plus importantes pour les voyelles qui durent bien plus longtemps contrairement aux consonnes. Ceci explique la difficulté de détection des transitions.

Les signaux des consonnes et voyelles qui ont été mal définis lors du découpage automatique ont été remplacés par ceux obtenus par le découpage manuel.

4.6 Interface graphique

Cette interface permet :

- D'extraire le signal utile du corpus enregistré (CVCV par exemple).
- De séparer les consonnes et voyelles par la segmentation automatique présentée dans la figure 23.
- D'afficher le signal, la FFT et les moments spectraux de chaque Consonne (ou voyelle).
- D'enregistrer ces consonnes et voyelles sous format *wav.

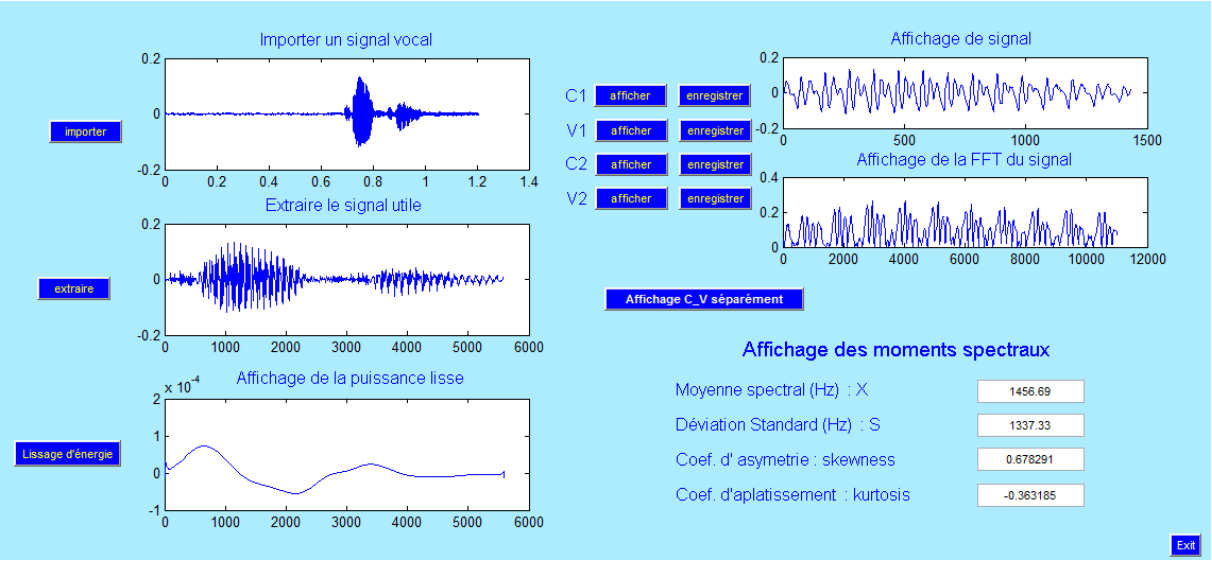

Figure 26 : Interface graphique globale

4.7 Conclusion

Après avoir enregistré notre corpus bruité, nous allons extraire notre signal utile CV ou CVCV afin de le segmenter par la suite sous la forme de consonnes et voyelles.

Notre interface peut afficher ces consonnes et voyelles séparément avec leurs FFT et leurs moments spectraux. Et à la fin, nous pouvons les enregistrer sous format *wav.

Conclusion générale

Ce projet de fin d'étude s'inscrit dans le domaine de la reconnaissance automatique de l'Arabe standard, domaine riche en applications potentielles. Nous avons résumé notre travail à l'aide d'une interface graphique qui intègre les algorithmes suivants :

- Algorithme qui extrait le signal utile.
- Deux algorithmes de séparation : un sur ZCR et l'autre sur la différence entre l'énergie de la consonne et de la voyelle.
- Algorithme pour éliminer le bruit du signal vocal.

Le signal acoustique est éminemment variable d'un locuteur à l'autre, ce qui rend très ardu le problème de reconnaissance de la parole. Le taux de performance est de 70%, il dépend de plusieurs phénomènes comme le lieu, le matériel d'enregistrement et les problèmes articulatoires des locuteurs.

Ce travail qui ne représente qu'un point de départ, nous lui prévoyons des possibilités d'évolution :

- Les machines à dicter et bureautique.
- Composition des numéros dans un téléphone portable.
- Mot de passe audio dans le domaine de sécurité : par la voix du locuteur.

Bibliographie

[1] Robert Paul, le petit Robert, Paris, 1976

[2] T. Dutoit, Introduction au traitement automatique de la parole, Faculté polytechnique de Mons, Belgique, 2000.

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

[4] Rachedi.J 2005. Reconnaissance et classification de phonèmes. Mastère, Laboratoire IRCAM, Paris.

[5] Marchal. A 2007. La production de la parole. (Lavoisier, Éd.) Hermès.

[6] Dutoit.T 2000. Introduction au traitement automatique de la parole. (Première Edition).

[7] Gargouri.D 2010. Contribution à l'estimation et à la poursuite des trajectoires de formants

[8] Dugand.P 1999. Phonétique et phonologie du Français. CEFISEM Nancy - Metz.

[9] Knoerr.H 2011. Récupéré sur aix1.uottawa.ca/~hknoerr/MunotNeve115119.pdf

[10] Linguistique UNIL 2011. Caractéristiques acoustiques des différentes réalisations. Récupéré sur http://www.unil.ch/ling/page13432.html

[11] Nguyen.N 2001. Rôle de la coarticulation dans la reconnaissance des mots. Année Psychologie, , Vol 101, pp. pp.125-154.

[12] O'Shaughnessy.D 2008. Formant estimation and tracking. Bouk Chapter in Springer Handbook of Speech Processing , pp.213-228.

[13] Anusuya.M.A 2011. Front end analysis of speech recognition: a review. Journal Int J Speech Technol , , Vol 14, pp.99-145.

[14] Calliope 1989. La parole et son traitement automatique (éd. CENT-ENST). Masson.

[15] S. Boulaknadel, ” Traitement Automatique des Langues et Recherche d‟‟Information en langue arabe dans un domaine de spécialité : Apport des connaissances morphologiques et 103 syntaxiques pour l'indexation”, THÈSE DE DOCTORAT ÉCOLE DOCTORALE : SCIENCES ET TECHNOLOGIES DE L‟‟INFORMATION ET DES MATÉRIAUX, Nantes, 2008.

[16] P. Bonaventura, ”INVARIANT PATTERNS IN ARTICULATORY MOVEMENTS”. The Ohio State University, 2003.

[17] D. E. KOULOUGHLI, ” Grammaire de l'arabe d‟‟aujourd‟‟hui‟‟, collection Langues pour tous”, Pocket 1994.

[18] M. Djoudi, ” Contribution à l'étude et à la reconnaissance automatique de la parole en Arabe standard”, Thèse du Centre de Recherche en Informatique de Nancy, INRIA Lorraine, 1991.
[19] A. Braham,” An Acoustic study of temporal organization in Arabic specific to Tunisian speakers”, PhD dissertation, (written in Arabic), university of Manouba, Tunis.
[20] D. Korichane et F. Wioland. ” De quelques aspects rythmiques de l'Arabe dialectal tunisien”. Actes des 16ème Journées d'Etudes sur la Parole, pages294-295, Hammamet, Tunisie, Octobre 1987.
[21] T. Benkirane et C. Cavé. ” Hiérarchie de sonorité et segmentation syllabique dans le parler arabe marocain”. Actes des 16ème Journées d'Etudes sur la Parole, pages 274-277, Hammamet, Tunisie, Octobre 1987.
[22] IPA,” Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet”, Cambridge University Press.
[23] Elias Hagos,” Implementation of an Isolated Word Recognition System”, UMI Dissertation Service, 1985.
[24] W. Abdulah and M. Abdul-Karim, ” Real-Time Spoken Arabic Recognizer”, Int. J. Electronics, 59(5),pp. 645–648, 1984.
[25] A. Al-Otaibi,” Speech Processing”, the British Library in Association with UMI, 1988.
[26] L. Amiar, M. Sellami, ” Un système basé sur une modélisation Markovienne pour la reconnaissance de la parole Arabe”, Laboratoire LRI- Département d‟Informatique-Université, SETIT Technologies of Information and Telecommunications, March 27-31, 2005 TUNISIA.
[27] E. H. Bourouba, M. Bedda , R. Djemili, ” Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM”, Department of electronic,Faculty of Engineering, University of Annaba, Algeria,Informatica 30 (2006) 373–384 373.
[28] Y. A. Alotaibi,” Investigating Spoken Arabic Digits in Speech Recognition Setting”, Journal of Information Sciences, 173(1–3), Elsevier, pp. 115–139, 2005.
[29] Y. A. Alotaibi,” High Performance Arabic Digits Recognizer Using Neural Networks”, Proceedings of the International Joint Conference on Neural Networks, pp. 670–6742003
[30] Rafik Djemili, Mouldi Bedda, and Hocine Bourouba, ” Recognition of Spoken Arabic Digits Using Neural Predictive Hidden Markov Models” , The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004.

[31] A.A. Moaz, R.H. El awady, ” Phonetic recognition of Arabic Alphabet letters using neural Networks” , International Journal of electric & Computer Sciences IJECS- IJENS vol: 11 No: 01, Faculty of Engineering , Mansoura Univesity , Egypte,2011
[32] T.Parsons, Voice and Speech Processing, McGraw-Hill, 1986.
[33] C.-S. Gardour, Traitement numérique des signaux, Ecole de technologie supérieur, 2001.
[34] T.-F. Quatieri, Discrete-Time Speech Signal Processing Principles ans Practice,Prentice Hall PTR, 2001.
[35] L.Rabiner and R. W. Scharfer, Digital Processing of Speech Signals, Prentice Hall, 1978.
[36] R. Boite et M.Kunt, Traitement de la parole,Presses Poluthechniques Romandes, Lausanne, 1987.