REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Echahid Hamma Lakhdar – El Oued

Faculté des Sciences Exactes
Département d'informatique

N°d'ordre :.....
Série :.....

Mémoire

Présenté en vue de l'obtention du diplôme de master en Informatique
Option : **Intelligence Artificielle et Systèmes Distribués**

Proposition d'une méthode de segmentation adaptative de la parole arabe

---

Par :

BEKKOUCHE Dhiya Elhak
SAYAH Zaoui

Soutenu le :

Devant le jury :

XXXXXXXX	Maître de conférences A	Université d'El Oued	Président
XXXXXXXX	Maître de conférences A	Université d'El Oued	Examinateur
ZAIZ Faouzi	Maître de conférences A	Université d'El Oued	Rapporteur

Proposition d'une méthode de segmentation adaptative de la parole arabe

BEKKOUCHE Dhiya Elhak
SAYAH Zaoui

Dédicace

Tout d'abord, je veux rendre grâce à Dieu, le Clément et le Très Miséricordieux pour son amour éternel. C'est ainsi que je dédie ce mémoire à :

ma mère pour sa tendresse et mon père pour sa patience et encouragement
ma femme pour leur soutien et mes chers fils qui m'ont beaucoup perturbé,
mes frères et sœurs,
tous ceux que j'aime,
tous mes amies.

Dédicace

Je dédie ce modeste travail à mes très chers parents qui n'ont pas cessé de m'encourager durant toutes mes études, je souhaite que dieu me les garde ;

A ma petite famille, mon épouse de m'avoir supporté pendants toute ces années d'étude, à mon petit garçon ***Naïl El Ridhouane*** *et à mon petite ange* ***Ritel*** *. En reconnaissances de tous les sacrifices consentis par tous et par chacun pour me permettre d'atteindre cette étape de ma vie ;*

À mes frères et leurs petites familles, ma petite sœur ***Manar****, à qui je souhaite de réussite dans ses études ;*

À mon collègue de ce travail et à sa petite famille, à qui je souhaite une grande réussite dans sa vie ;

À mes amis d'étude, ainsi qu'à toutes les personnes qui m'ont aidé à la réalisation de ce travail.

Remerciements

A Dieu, le tout puissant, nous rendons grâce pour nous avoir donné santé, patience, volonté et surtout raison.

En premier lieu, je tiens à remercier mon encadreur Mr. ***ZAIZ Faouzi*** *qui m'a aidé et conseillé durant ce travail.*

Je remercie également tous les enseignants du département de l'informatique de l'université d'EL Oued pour leurs aides et encouragements.

Enfin, je remercie tous ceux qui en soutenu, encouragé et donné l'envie de mener à terme ce travail.

Résumé

Il y a longtemps, l'utilisation de la parole comme Interface Homme-Machine (IHM) s'est imposée dans de nombreux domaines car c'est un moyen naturel de communiquer pour les humains. Durant ces dernières années, deux applications dans le domaine du traitement de la parole ont connu des progrès considérables, la reconnaissance vocale et la synthèse de la parole.

Nous allons nous intéresser dans cette thèse au domaine de la reconnaissance vocale, qui est une technique informatique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. Ce processus utilise un ensemble de phases (près-traitement, segmentation, extraction des caractéristiques, classification et poste-traitement) afin d'extraire le résultat finale. Parmi ces phases, la phase de segmentation est une phase très cruciale du fait qu'elle définisse les unités de base à reconnaître par le système. En plus, la représentation d'un signal produite est affectée directement par le matériel d'acquisition utilisé (type de microphone et carte son).

Malheureusement, les méthodes actuelles de segmentation automatiques de la parole ne permettent pas de produire un signale de parole segmenté identique sur des ordinateurs différents. Il est donc nécessaire de disposer une méthodes de segmentation adaptative qui va résoudra cette hétérogénéité. Pour cela, on s'intéresse dans ce travail à proposer une méthode de segmentation qui s'adapte au matériel et donne un résultat similaire sur des différente machines.

Mots Clés: reconnaissance vocale, parole arabe, segmentation, adaptative.

Abstract

Over time, the use of the speech as Human Machine Interface (HMI) has emerged in many areas because it is a natural way to communicate to people. In recent years, two applications in the field of speech processing have made considerable progress, speech recognition and speech synthesis.

We will focus in this thesis to the field of speech recognition, which is a computer technique allows to analyze the human voice captured by a microphone to transcribe it in the form of a text readable by a machine. This process uses a set of phases (close-treatment, segmentation, feature extraction, classification and post-treatment) to retrieve the final result. Among these phases, the segmentation phase is a crucial stage because it defines the basic units to be recognized by the system. In addition, the representation of a signal produced is directly affected by the acquisition of equipment used (type of microphone and sound card).

Unfortunately, current methods of automatic segmentation of speech does not allow to produce a segmented same speech signals on different computers. It is therefore necessary to have an adaptive segmentation methods which will solve this heterogeneity. For this, we are interested in this work to propose a segmentation method that fits the material and gives a similar result on different machines.

Keywords: speech recognition, Arabic word, Segmentation, Adaptive.

الملخص

منذ وقت بعيد ، كان استخدام الكلام كواجهة بين الكمبيوتر والإنسان امر ضروري في العديد من المجالات، لكونه الوسيلة الطبيعية للتواصل مع الناس. في السنوات الأخيرة عرف نوعان من التطبيقات في مجال معالجة الكلام تقدما كبيرا، التعرف على الكلام وتركيب الكلام.

سنركز في هذه الأطروحة على مجال التعرف على الكلام ، الذي يعتبرمن تقنيات الكمبيوتر التي تسمح بتحليل صوت الانسان الناتج من خلال التقاطه عن طريق الميكروفون والذي يتم تحويله فيما بعد لكتابة في شكل نص مقروء من قبل الجهاز. تتكون هذه العملية من مجموعة مراحل (قبل المعالجة ، التقسيم ، استخراج الخصائص ، التصنيف والمعالجة البعدية)وذلك من اجل الحصول على النتيجة النهائية. من بين هذه المراحل ، مرحلة التقسيم والتي تعتبر مرحلة حاسمة لأنها تحدد الوحدات ألأساسية التي من خلالها يمكن للنظام ان يتعرف عليها. بالإضافة إلى ذلك فان تمثيل إلاشارة الناتجة يتأثر مباشرة بالمعدات المستخدمة في الادخال (نوع ميكروفون وبطاقة الصوت).

لسوء الحظ ، الأساليب المتبعة حاليا في التجزئة الآلية للكلام لا تسمح بإنتاج إشارات مجزأة للكلام متطابقة بين أجهزة كمبيوتر مختلفة. لذلك فمن الضروري أن يكون هناك طرق تجزئة تتكيف بطريقة آلية مع مختلف الاجهزة ، لتحل مشكلة عدم التجانس. وعليه فإننا ركزنا اهتمامنا في هذا العمل لاقتراح طريقة تقسيم تتأقلم مع جميع الأجهزة وتعطي نتيجة متطابقة على أجهزة مختلفة.

الكلمات المفتاحية : التعرف على الصوت، الكلام العربي، التقسيم، التأقلم.

Table des matières

Liste des figures

Liste des tableaux

Liste des algorithmes

Acronymes

AMR	Analyse Multi-Résolution
AMREC	Analyse Multi-Résolution de l'Enveloppe Complexe
BBN	Bolt Beranek and Newman
CAN	Convertisseur Analogique Numérique
DFT	Discret Fourier Transform
DTW	Dynamic Time Wraping
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
HNR	Harmonic-to-Noise Ratio
LPC	Linear Predictif Coding
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel-Scale Frequency Ceptral Coefficients
PCI	Peripheral Component Interconnect
PCMCIA	Personal Computer Memory Card International Association
PLP	Perceptual Linear Predective
PPZ	Passage Par Zéro
RAP	Reconnaissance Automatique de la Parole
SP	Signal de Parole
SRAP	Système de Reconnaissance Automatique de la Parole
TF	Transformée de Fourier
TFCT	Transformée de Fourier à Court Terme
TPPZ	Taux de Passage Par Zéro

Contexte de l'étude

Au fil des temps, l'utilisation de la parole comme Interface Homme-Machine (IHM) s'est imposée dans de nombreux domaines car c'est un moyen naturel de communication pour les humains. Durant ces dernières années, deux applications dans le domaine du traitement de la parole ont connu des progrès considérables, la reconnaissance vocale et la synthèse de la parole. Nous allons nous intéresser dans cette thèse au domaine de la reconnaissance vocale, qui est une technique informatique permettant d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. Ce processus utilise un ensemble de phases (près-traitement, segmentation, extraction des caractéristiques, classification et post-traitement) afin d'extraire le résultat final. Parmi ces phases, la phase de segmentation est une phase très cruciale du fait qu'elle définisse les unités de base à reconnaître par le système. En plus, la représentation d'un signal produit est affectée directement par le matériel d'acquisition utilisé (type de microphone et carte son).

Malheureusement, les méthodes actuelles de segmentation automatiques de la parole ne permettent pas de produire un signal de parole segmenté identique sur des ordinateurs différents. Il est donc nécessaire de disposer une méthode de segmentation adaptative qui va résoudra cette hétérogénéité. Pour cela, on s'intéresse dans ce travail à proposer une méthode de segmentation qui s'adapte au matériel et donne un résultat similaire sur des différente machines.

Motivations et objectifs

Aujourd'hui, la reconnaissance automatique de la parole est un domaine de recherche bien établi, à partir duquel émergent des technologies permettant le développement d'applications réelles. Le problème de la reconnaissance automatique de la parole consiste à extraire à l'aide d'un ordinateur l'information contenue dans le signal de parole. La technologie la plus utilisée depuis ces 20 dernières années est basée sur les modèles statistiques (modèle de Markov cachés – HMM) capables de modéliser simultanément les caractéristiques fréquentielles et temporelles du signal étudié. Dernièrement, une extension de ces modèles a été mise au point donnant naissance aux modèles hybrides. Ces derniers combinent la technologie des modèles de Markov cachés (HMM) et des réseaux de neurones artificiels (ANN). De nombreuses ont déjà montré l'efficacité de ces modèles en reconnaissance de la parole (continue ou isolée) indépendamment du locuteur pour de petits et grands vocabulaires.

Notre étude s'intègre dans le cadre du développement d'une méthode de segmentation adaptative de la parole arabe, néanmoins que le traitement de la parole arabe est encore à ses débuts auquel nous espérons apporter une contribution à travers ce travail. La reconnaissance par les méthodes les plus performantes de l'état de l'art reste insuffisante ; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer les performances des systèmes de RAP en mettant comme objectif l'augmentation du taux de reconnaissance, appliquant ainsi une nouvelle technique qui conduit à une segmentation correcte.

Le traitement de la parole arabe est encore à ses débuts, la raison pour laquelle, nous avons pensé à développer une méthode de segmentation adaptative qui permet d'avoir un taux de segmentation correcte pour garantir une bonne qualité de la parole segmenté pour la parole arabe toute en dépassant les contraintes matérielles ou environnementales de l'enregistrement.

Introduction

La parole est considéré comme un moyen primordial pour assurer la communication entre les êtres humains, sa simplicité en fait lui permet d'être le moyen de communication le plus utilisé dans la société humaine (facilité de parler à une personne que de lui écrire ou de lui schématiser). Néanmoins, le processus de communication chez l'être humain est effectué par des traitements très complexe réalisé par le cerveau, ce processus commence avec la production de la parole par l'appareil phonatoire et sa perception par l'appareil auditif jusqu'à sa compréhension, cette tâche est difficilement automatisable pour une machine.

Le développement continu de la technologie, notamment en informatique implique la nécessite de trouver des nouveaux moyens de dialogue homme machine, l'objectif visée par cette nécessité est comment diminuer la dépendance au clavier et les autres périphériques durant l'acquisition des données dans la machine, cette tâche est généralement très difficile et très lente. En général, le traitement automatique de la parole utilise des sources de connaissance qui sont nécessaires pour la compréhension de la parole, l'étude de l'aspect reconnaissance de la parole qui a pour objectif de décoder le signal de la parole en unités de bases (phonèmes, mots ...) sans en donner une signification (sans comprendre le sens des phrases construites). Actuellement, les systèmes de reconnaissance de la parole ont évolué et utilisent non seulement des connaissances en linguistique mais aussi des connaissances dans les domaines : Traitement du signal, Reconnaissance des formes...

La segmentation est l'une des phases les plus cruciales dans le processus de la reconnaissance de la parole parce qu'elle permet de préparer les segments de base nécessaires aux autres traitements. La problématique majeure de la segmentation réside dans la détermination des limites de différents segments de la parole. Dans ce travail nous allons proposer une méthode de segmentation adaptative permettant de bien délimiter les segments de la parole.

Ce document se compose de quatre chapitres. Le premier chapitre présente des notions de bases sur la parole en général et la RAP avec une vue sur le traitement du signal et les outils d'acquisition. Le second chapitre démontre les différentes méthodes de la segmentation. Celui-ci fait une synthèse des techniques existantes de segmentation automatique de la parole. Le troisième chapitre est consacré pour la modélisation et la mise en œuvre du système qui décrit en détail la méthode de segmentation adaptative. Le dernier chapitre montre les résultats obtenus par la méthode élaborée ainsi qu'une discussion et évaluation de la méthode. Enfin, nous dressons une conclusion de ce travail et proposons quelques perspectives qui nous croyons utiles et nécessaires pour améliorer et rendre le processus de segmentation efficace pour la parole arabe.

Chapitre 1

Reconnaissance Automatique de la Parole

1.1 Introduction

La parole est un moyen de communication très efficace et naturel utilisé par l'être humain. Depuis longtemps, l'objectif majeur des chercheurs est comment pouvoir s'adresser par ce même moyen à des machines pour les rendre plus intelligentes et interactives. La reconnaissance automatique de la parole (RAP) est un domaine multidisciplinaire d'étude actif depuis plusieurs années, le traitement automatique de la parole comprend deux branches principales : la (RAP) et la synthèse vocale. il est utilisé dans des nombreux domaines comme « Perception, Acoustique, Médecine, Électronique, Physique, Informatique et Traitement du signal» Il est évident qu'un outil de reconnaissance de la parole efficace facilitera l'interaction entre les hommes et les machines. Plusieurs applications possibles associées à un tel outil et sont amenées à connaître un grand essor. La plupart des ces dernières en RAP peuvent être classées par les catégories suivantes : commande et contrôle, accès à des bases de données ou recherche d'informations, dictée vocale et transcription automatique de la parole. En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis. Dans ce chapitre, on s'intéresse d'une part à introduire et présenter un état d'art sur la reconnaissance des parole et le traitement de signal de la parole. D'autre part à exposer le processus et les différentes approches de la RAP, en fin les domaines d'application des SRAP.

1.2 Généralités sur la parole

D'un point de vue physique, un son est une énergie qui se propage sous forme de vibrations dans un milieu compressible (l'eau, l'air, les matériaux solides), produit a partir d'une source sonore et capté par un récepteur sensible, il se propage à une certaine vitesse dans un milieu élastique (340 m/s dans l'air à 15 °C), appelée aussi «célérité», On parle alors de pression acoustique. Plus cette pression acoustique est forte et plus l'on entend le son fortement, la propagation du son diminue avec la distance. Ceci est dû à l'amortissement du système [6].On distingue deux types de son : le son analogique et le son numérique.

1.2.1 Son analogique

Lorsque le son est capté à partir d'un microphone, ce dernier transforme l'énergie mécanique (la pression de l'air exercée sur sa membrane), en une variation de tension électrique continue. Ce signal électrique dit « analogique » pourra ensuite être amplifié,

et envoyé vers un hautparleur dont la fonction est inverse, voir la figure (1.1). Le son analogique est généralement fixé sur des supports comme les bandes magnétiques, K7 audio..etc. Le problème rencontré au son analogique de faite il n'est pas traitable par l'ordinateur [19, 16].

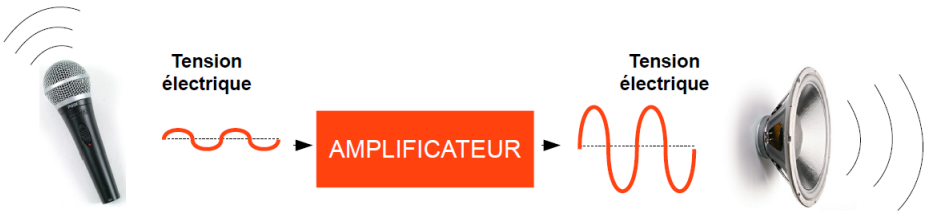

Figure 1.1: Exemple d'une chaîne analogique [16].

1.2.2 Son numérique

Avec l'informatique, lorsque ce même signal électrique est capturé à partir du micro, il est converti en une suite binaire (0,1), on parle alors de numérisation du signal. C'est la carte son qui s'en charge,ce processus est appelé numérisation, qui consiste donc à passer d'un signal continu (une variation de tension électrique) en une suite de valeurs mesurées à intervalles réguliers (discontinu) comme l'indique la figure (1.2), ce dernier est enregistré sur un support numérique tel que le disque dure [19].

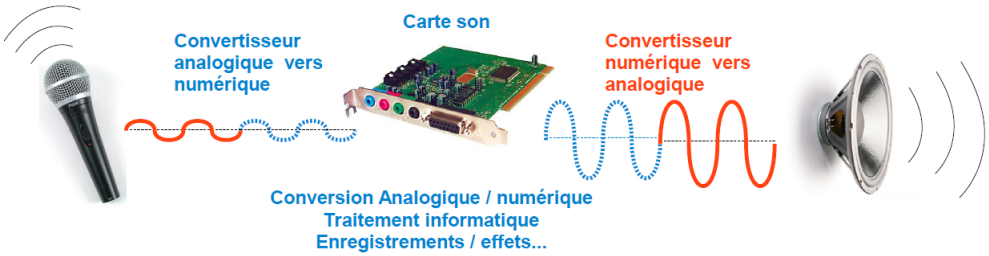

Figure 1.2: Exemple d'une chaîne numérique [19].

1.2.3 Convertisseur analogique / numérique

Un convertisseur analogique / numérique (CAN) est un appareil électronique qui se charge de convertir les tensions (un signal analogique continu) en chaînes de nombres binaires (un signal numérique discret) à chaque période de l'horloge d'échantillonnage d'une façon périodique. Les nombres binaires sont stockés sur un support d'enregistrement numérique sorte de mémoire, il s'agit de données multimédia (Figure 1.3) [7].

1.2.4 Mécanisme de la parole

Le processus de production de son est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique. Il y a une grande quantité d'organes et de muscles qui coopèrent afin de produire des sons des langues naturelles. Un son est produit lorsque l'air contenu dans les poumons est contraint à traverser un ou plusieurs résonateurs de l'appareil phonatoire (le pharynx ,la cavité buccale ,la cavité labiale ,fosses nasales le larynx) et donc a travers des cordes vocales [17], voir figure

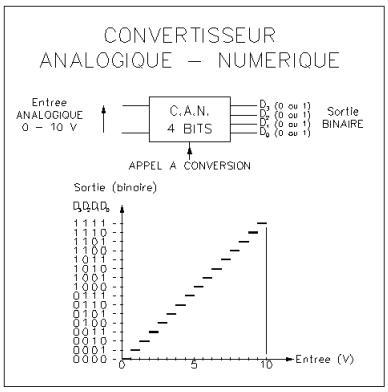

Figure 1.3: Conversion analogique numérique [7].

(1.4). La vibration quasi-périodique des cordes vocales permet la production de toutes les voyelles et aussi de certaines consonnes dites sonores ou voisées comme les sons /b/ et /d/. L'appareil phonatoire, émetteur d'informations, qui vont être captée et analysée par un récepteur appelé l'appareil auditif [17, 6]. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur [7].

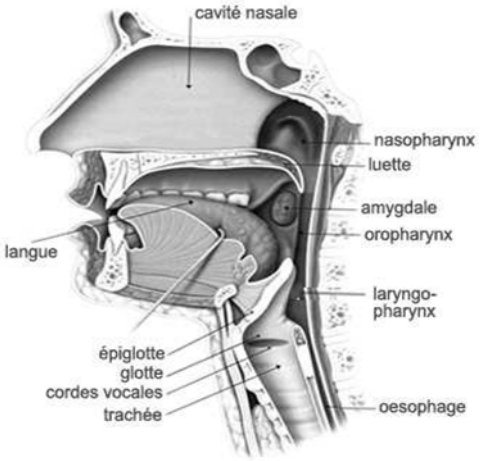

Figure 1.4: L'appareil phonatoire humain [17].

A. L'information vocale

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le timbre et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application, les phonèmes pour les machines de dictée par exemple. La parole est surtout contenue dans les deux premiers formants, mais l'information proprement dite provient des transitions for mantiques. En général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100Hz-5KHz (300Hz-3.4KHz pour la téléphonie) [19, 8].

B. Phonétique et caractéristiques des sons de parole

Cette partie décrit certaines caractéristiques des voyelles et des consonnes. Notons que ces caractéristiques sont très utiles pour la réalisation de la segmentation manuelle et la compréhension des erreurs d'étiquetage et de segmentation automatiques. Pour

différencier les différents types de consonnes, trois indices peuvent être pris en compte : la présence et la durée du silence, la présence du bruit ainsi que la position des formants. L'identification perceptive des sons (voyelles ou consonnes) est basée sur des indices acoustiques pertinents qui permettent de distinguer visuellement les différentes classes de sons. Pour les voyelles, ces indices correspondent à la position, la largeur de bande et l'intensité des formants. Nous avons tenu à présenter cette discipline car elle est fondamentale dans la détermination des caractéristiques des sons. Il existe trois types de la phonétique [15] :

- **Phonétique articulatoire :** La phonétique articulatoire est l'étude des sons du langage humain envisagé sous l'angle de la production. Cette discipline nécessite une connaissance de la physiologie des organes de la phonation et du rôle des différents organes dans la production des sons du langage [15, 17, 10].

- **Phonétique acoustique :** La phonétique acoustique s'intéresse à la transmission des sons en tant que signaux acoustiques. Dans ce processus, plusieurs éléments peuvent expliquer l'origine des indices acoustiques [15, 10]. Les paramètres acoustiques des sons les plus utilisés sont le pitch et les formants. Ces paramètres acoustiques permettent de différencier des classes acoustiques de sons. Par exemple, le spectre d'un son voisé contient plus de composantes (formants) en basse fréquence qu'en haute fréquence alors que le spectre d'un son non voisé présente une accentuation vers les hautes fréquences [6].

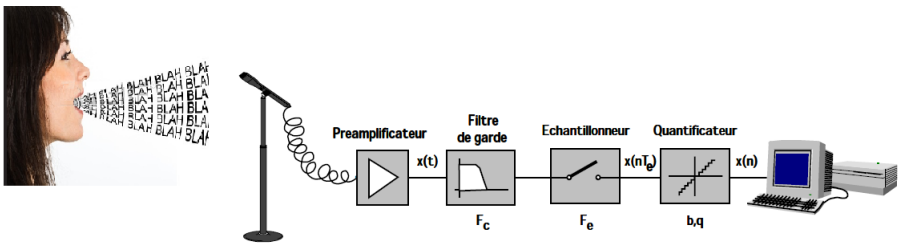

Figure 1.5: Enregistrement numérique d'un signal acoustique [15].

- **Phonétique perceptive :** La phonétique perceptive tente de comprendre et de décrire la perception de la parole humaine. Pour ce faire, elle a recours à la physiologie du système auditif et à la psycho-acoustique. Cette dernière est l'étude de la perception auditive en fonction du stimulus. Elle s'intéresse à la façon dont les ondes sonores sont captées par le système auditif et la manière dont elles sont interprétées par le cerveau [15].

1.3 Bruit

On appelle bruit tout phénomène perturbateur ou un signal nuisible qui se superpose au signal utile en un point quelconque d'une chaine de mesure ou d'un système de transmission. Il constitue donc une gêne dans la compréhension du signal utile ou a l'interprétation d'un signal, qui est dans notre cas, la parole. En physique, en acoustique et en traitement du signal, bien que le bruit soit, par nature, aléatoire, il possède certaines caractéristiques statistiques, spectrales ou spatiales.on distingue trois types de bruit [6, 8] :

1.3.1 Bruits additifs

Les bruits additifs sont dus à la multiplicité des systèmes de communication dans un même environnement. Plusieurs émetteurs et plusieurs receveurs pouvant être confinés dans un même espace, les messages de tous les émetteurs peuvent donc se trouver en concurrence sur une même voie sans que les récepteurs possèdent un mécanisme infaillible pour isoler le message qui leur est destiné. L'émetteur et le récepteur peuvent aussi se trouver en présence d'un ou de plusieurs équipements générant un bruit de fond de force variable [6].

1.3.2 Bruits convolutionnels

Les bruits convolutionnels (ou multiplicatifs) sont dus à la distorsion induite par la voie de communication. Ils résultent de la mauvaise qualité d'un ou de plusieurs éléments de support du signal ou, tout simplement, de son étroitesse en bande passante. La qualité de la transmission varie cependant très peu au cours d'une même communication. De manière plus générale, le bruit convolutionnel est présent dans toute application de RAP par l'intermédiaire du microphone utilisé pour la saisie de la voix ou lorsque le microphone utilisé pour l'enregistrement est placé assez loin du locuteur. Ou bien dépend aux milieux d'enregistrement qui sont de mauvaise qualité et peuvent provoquer des phénomènes de réverbération [6].

1.3.3 Bruits physiologiques

D'autres bruits peuvent également être considérés dans le domaine de la RAP car ils sont spécifiques à l'être humain lors de sa phase de production de parole. Dans un environnement bruité la personne essaie, lui, de s'adapter aux conditions sonores rencontrées en modifiant sa méthode de production de parole. Ce qu'on appel l'effet Lombard. Cette accentuation de la voix pose cependant un problème majeur aux systèmes de RAP car les spectres de tous les phonèmes peuvent être modifiés ce qui a pour effet de nettement amoindrir les taux de reconnaissance [6].

Tableau 1.1: Exemples de différentes classes des bruits.

Propriétés	**Types**
Structure	Continu/ intermittent /impulsif.
Type d'interaction	Additif/convolutif.
Comportement temporel	Stationaire/Non-stationnaire.
Bande de fréquence	étroit/large.
Dépendance	Corrélé / Décorrélé/large.
Propriété spatiales	Cohérent/Incohérent/large.

Le signal de la parole est un phénomène de nature acoustique porteur d'un message. L'information d'un message parlé réside dans les fluctuations de l'air, engendrées, puis émises par l'appareil phonatoire. Ces fluctuations constituent le signal vocal. Elles sont détectées par l'oreille qui procède à une certaine analyse. Les résultats sont transmis au cerveau qui les interprète [7]. D'autre part, le signal vocal représente la combinaison d'éléments simples et brefs du signal sonore appelés phonèmes, qui permettent de distinguer les différents mots. La parole est un signal réel, continu, d'énergie finie et non stationnaire

avec une structure complexe et variable avec le temps.Généralement le traitement de signal de la parole se présente dans les deux fameux domaines : la synthèse de la parole et la Reconnaissance de la parole [7].

1.3.4 Caractéristique du signal de la parole

En plus de ses caractéristiques, en tant qu'onde longitudinale, qui sont la fréquence, la longueur d'onde, et la vitesse de propagation qui dépend du milieu matériel de propagation.le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre. L'opération de numérisation, requiert successivement : un filtrage de garde, un échantillonnage, et une quantification. La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés fc, fe, b, et q [7, 8]. Le son a par conséquent d'autres caractéristiques, qui sont :

A. L'amplitude

L'amplitude d'un son correspond à la variation de pression maximale de l'air engendrée par les oscillations, et donc au volume sonore. L'amplitude d'une vibration peut être exprimée objectivement par le calcul des variations de pression d'air (exprimée en Micron Bar et convertie en watt/cm2). On utilise toutefois plus fréquemment une unité de mesure relative, le décibel (dB), pour rendre compte de l'intensité d'un son [7, 8].

B. La fréquence

La vitesse des mouvements d'aller et de retour des vibrations est responsable de la sensation de hauteur. Plus les mouvements vibratoires sont rapides, plus le son sera aigu. À l'inverse, un mouvement plus lent engendre un son plus grave, voir la figure (1.6). Lorsque la longueur du cycle appelée longueur d'onde ou période augmente, la fréquence en cycles par seconde diminue et vice versa.De façon objective, la hauteur d'un son correspond à sa fréquence et est exprimée en cycles par secondes ou Hertz. Un son comportant 100 cycles par seconde, soit 100 mouvements complets d'aller et de retour par rapport au point de repos, aura une fréquence de 100 Hertz [7, 8].

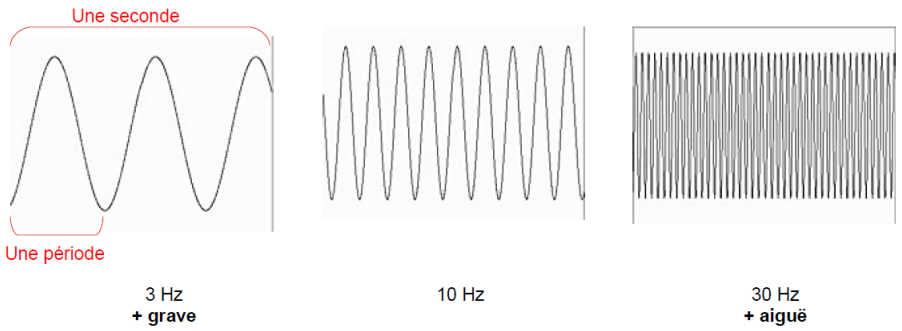

Figure 1.6: Représentation d'une fréquence [7].

C. La hauteur

La hauteur d'un son est la qualité liée à la fréquence de l'onde sonore(la fréquence des vibrations de la source sonore). Plus la fréquence est élevée, plus le son perçu est aigu et, inversement, plus la fréquence est faible, plus le son perçu est grave [7].

D. L'énergie d'un son (intensité)

C'est la qualité qui fait distinguer un son fort d'un faible. L'intensité est liée à la pression de l'air en amont du larynx, qui fait varier l'amplitude des vibrations sonores. Souvent l'énergie observée dans un segment voisé est plus importante que celle observée dans un segment non voisé [7].

E. Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix. Il provient en particulier de la résonance dans la poitrine, la gorge, la cavité buccale et le nez sont les amplitudes relatives a des harmoniques du fondamental qui déterminent le timbre du son [7, 8]. Les éléments physiques du timbre comprennent :

1. Les relations entre les parties du spectre, harmoniques ou non ;
2. les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont L'énergie est limitée à une ou plusieurs bandes de fréquence) ;
3. L'évolution dynamique globale du son ;
4. L'évolution dynamique de chacun des éléments les uns par rapport aux autres.

F. Fréquence d'échantillonnage

Lorsqu'un son est numérisé, le signal analogique (continu) qui entre dans l'ordinateur est mesuré, un certain nombre de fois par seconde (d'ou la discontinuité). Le son est donc découpé en tranches", ou échantillons (en anglais « samples ») (figure 1.7). Le nombre d'échantillons disponibles dans une seconde d'audio s'appel la fréquence d'échantillonnage exprimée en hertz. Pour traduire le plus fidèlement possible le signal analogique de notre micro, il faudra prendre le plus grand nombre de mesures possible par seconde. Autrement dit, plus la fréquence d'échantillonnage sera élevée, plus la traduction numérique du signal sera proche de l'original analogique [7].

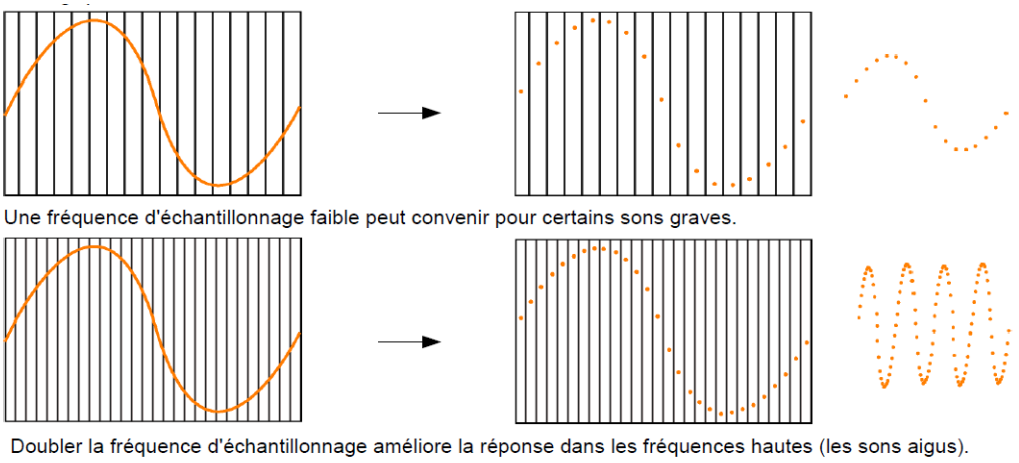

Figure 1.7: Exemple d'une fréquence d'échantillonnage [7].

G. Fréquence fondamentale (ou pitch)

La fréquence fondamentale *F0* (ou pitch) joue un rôle important dans la parole. C'est elle qui véhicule une grande partie de l'information prosodique. L'intensité de la voix et les durées successives des syllabes complètent ces informations, elle peut faire ressortir bien des caractéristiques du locuteur, mais participe aussi à la caractérisation de la langue elle-même, par la manière dont elle est utilisée pour différencier les divers éléments syntaxiques comme les énoncés (interrogatifs, exclamatifs ou déclaratifs), l'importance de certains mots, ou bien même pour caractériser les différences lexicales entre les mots. Elle s'étend approximativement de 70 à 600 Hz pour l'homme, la femme et l'enfant (voir tableau 4.1), cette différence est due à la différence de la taille des cordes vocales ; les hommes adultes ont généralement une voix plus grave et des cordes vocales plus longues, soit entre 17 et 25 mm. Celles des femmes se situent entre 12,5 et 17,5 mm. [3, 7].

Tableau 1.2: Les fréquences fondamentales pour l'homme,femme et enfant.

Sexe/Age	**Plage de la fréquence fondamentale**
Homme	de 70 à 250 Hz
Femme	de 150 à 400 Hz.
Enfant	de 200 à 600 Hz

La figure(1.8) représente l'évolution de la fréquence de vibration des cordes vocales dans la phrase "les techniques de traitement numérique de la parole". La fréquence est donnée sur une échelle logarithmique, les sons non-voisés sont associés à une fréquence nulle [15].

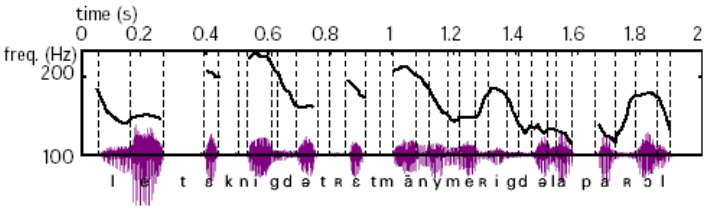

Figure 1.8: Évolution de la fréquence de vibration des cordes vocales [15].

H. Résolution et quantification (bit)

Une autre caractéristique importante est la résolution numérique du son, soit le nombre de « niveaux » ou de « paliers » qu'il est possible d'enregistrer pour reproduire l'amplitude du signal. Avec une résolution de 16 bit, on dispose de 216, soit 65535 valeurs possibles pour traduire l'amplitude du son. Ainsi, plus la résolution est élevée, meilleur sera la dynamique (l'écart entre le son le plus faible et le plus fort qu'il est possible de reproduire). Dans la figure (1.9) la zone bleue montre qu'en doublant la résolution, on est plus proche de la courbe « analogique », soit le signal parfait que l'on souhaite reproduire.

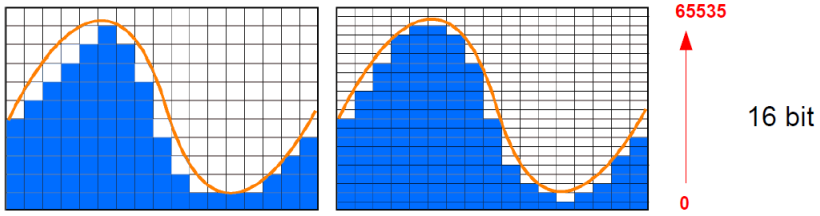

Figure 1.9: La résolution numérique du son [10].

La figure (1.10) explique que la quantification consiste en une deuxième phase où le chiffre de l'amplitude prélevé sera arrondi à l'entier le plus proche.

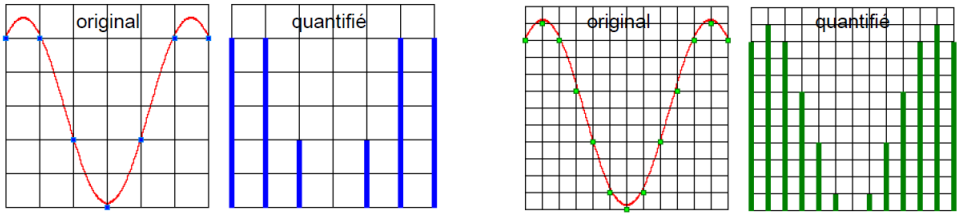

Figure 1.10: La quantification numérique du son [10].

Le tableau 4.2 donne des exemples pour les qualités de son largement utilisés avec la fréquence correspondante et le nombre des bits nécessaire pour les codifier.

Tableau 1.3: Quelques exemples de résolutions fréquemment utilisées.

Qualité du son	**Fréquence correspondante**	**Nombre de bits**
Son qualité téléphone	8000 Hz	8 bit.
Son qualité radio FM	22050 Hz	16 bit.
Son qualité CD	44100 Hz	16 bit.
Son qualité DVD	48000 Hz	24 bit.

1.3.5 Représentation du signal

Le signal de la parole est un vecteur acoustique porteur d'informations d'une grande complexité, il est représentable sous plusieurs formes, tout dépend l'utilité de différentes composantes fréquentielles du signal et la signification sur le plan perceptuel, on cite les représentations suivants :

A. Audiogramme

L'échantillonnage transforme le signal à temps continu $x(t)$ en signal à temps discret $X(nTe)$ défini aux instants d'échantillonnage, multiples, entiers de la période d'échantillonnage Te, celle-ci est elle-même l'inverse de la fréquence d'échantillonnage fe. Pour ce qui concerne le signal vocal, le choix de fe résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz [15, 10]. La figure 1.11 représente l'évolution temporelle, ou audiogramme du signal vocal pour les mots : **(بسم الله الرحمان الرحيم)**.

Figure 1.11: Audiogramme de signaux de parole [15].

B. Transformée de Fourier à court terme

La transformée de Fourier à court terme (TFCT) est obtenue en extrayant de l'audiogramme a une trentaine de ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant une transformée de Fourier sur ces échantillons. La figure (1.12) illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière [15, 10]. La forme générale de ces spectres, appelée enveloppe spectrale, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés formants et anti-formants [11, 6].

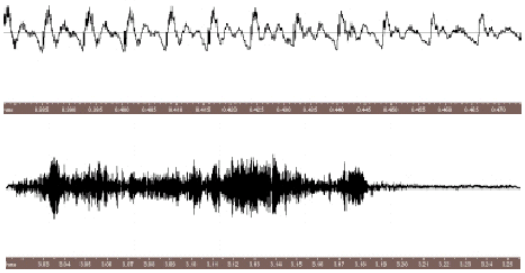

Figure 1.12: Exemples de son voisé (haut) et non voisé (bas) [15].

C. Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme. Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné, une transformée de Fourier rapide étant régulièrement calculée à des intervalles de temps rapprochés [6, 8]. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence [10]. On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms), ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants, voir figure (1.13). Les périodes voisées y

apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales [7, 15].

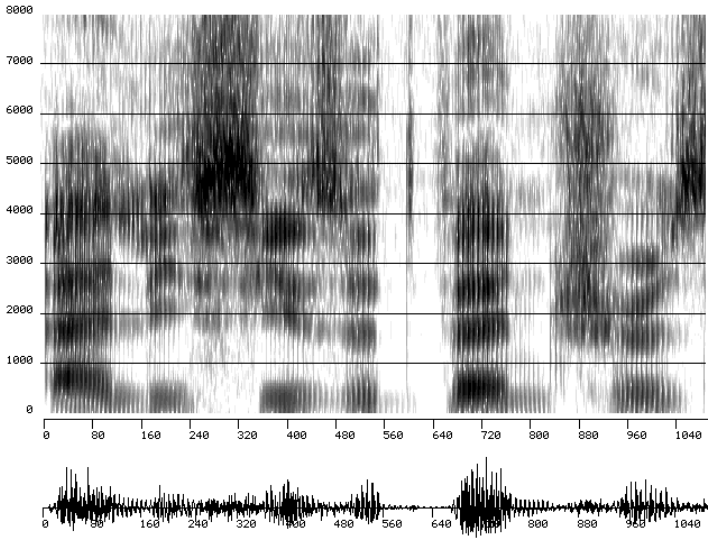

Figure 1.13: Spectrogrammes et évolution temporelle [15].

1.3.6 Méthodes d'analyse du signal vocal

L'arsenal des méthodes d'analyse et de traitement du signal est considérable. Nous présentons les méthodes générales couramment utilisées pour l'analyse du signal de la parole, puis les méthodes utilisées pour l'extraction des paramètres. Les méthodes d'analyse du signal de la parole peuvent être divisées en deux grandes classes, les méthodes paramétriques et les méthodes non paramétriques [8].

A. Méthodes temporelles

Le signal de la parole jouissant de quelques propriétés, sont exploitables à partir de la représentation temporelle. Les méthodes d'analyses temporelles se basent essentiellement sur la mesure du max, du min, du nombre des passages par zéro, de la fonction d'auto corrélation, du calcul de l'énergie et autres. Du fait que le signal vocal est considéré comme étant un signal quasi-stationnaire, le traitement doit se faire sur des tranches de 5 à 30 ms [8].

B. Méthodes d'analyse spectrale

Les propriétés spectrales du signal vocal présentent un intérêt majeur pour la perception auditive. L'analyse spectrale est une technique qui met en évidence les caractéristiques fréquentielles des signaux et permet en première approximation de séparer la contribution de la source de celle du conduit vocal. Elle est souvent utilisée dans les techniques d'analyse synthèse du signal vocal, notamment dans l'analyse par banc de filtre. C'est une méthode bien adaptée pour le calcul du pitch et des formants [13, 17].

C. Méthodes non paramétriques

Ces méthodes sont basées principalement sur le calcul de la transformée de Fourrier, soit sur le signal direct, soit sur sa fonction d'auto corrélation. Le calcul de la TF permet l'obtention de la densité spectrale de la puissance, qui nous mène ainsi l'extraction des paramètres nécessaires à l'analyse et la synthèse du signal vocal [13, 8].

D. Méthodes paramétriques

Les techniques basées sur l'analyse spectrale présentent quelques limitations, liées à l'hypothèse que le signal est nul au-delà de la fenêtre d'analyse. Pour remédier à ce problème, des méthodes paramétriques sont apparues. Parmi ces méthodes on trouve les méthodes dites autorégressive [13, 8].

E. Codage Prédictif Linéaire

C'est une méthode de type essentiellement temporel qui permet de calculer des coefficients appelés coefficients de la prédiction linéaire [8]. Le codage prédictif linéaire (LPC) est une méthode de codage et de représentation de la parole. Elle repose principalement sur l'hypothèse que la parole peut être modélisée par un processus linéaire. Il s'agit donc de prédire que le signal à un instant n à partir des p échantillons précédents. La parole n'étant cependant pas un processus parfaitement linéaire, la moyenne que constitue la somme pondérée du signal sur p pas de temps introduit une erreur qu'il est nécessaire de corriger par l'introduction du terme *e(n)*. Le codage par prédiction linéaire consiste donc à déterminer les coefficients ak qui minimisent l'erreur *e(n)*, ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage [6].

1.3.7 Méthodes d'extraction des paramètres

Ces méthodes consistent à extraire les paramètres essentiels qui caractérisent généralement le signal de la parole à savoir l'énergie, la fréquence fondamentale et les formants.

A. Extraction de la fréquence fondamentale (pitch)

L'extraction de la fréquence fondamentale (ou pitch) est comme son nom l'indique fondamentale. Les variations de la fondamentale pour un locuteur donné constituent ce qu'on appelle la prosodie. Celle-ci influe considérablement sur l'oreille humaine pour permettre la différentiation entre locuteurs et ainsi la reconnaissance du locuteur [10, 6].

B. Extraction des formants

Le début d'utilisation des méthodes d'extraction des formants remonte à 1934. Ces formants sont les résonances du conduit vocal considéré comme un filtre et correspondant aux pôles de la fonction de transfert de ce dernie (figure 1.14)r. On désigne par formant (acoustique) d'un son de la parole l'un des maxima d'énergie du spectre sonore de ce son de la parole. Ce sont des paramètres privilégiés dans l'étude et l'analyse de la parole, ils apparaissent plus clairement pour les sons voisés [6, 8].

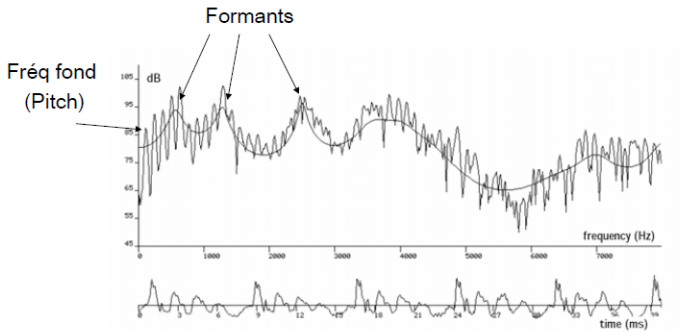

Figure 1.14: Exemple d'extraction des paramètres [6].

1.4 Outils d'acquisition de la parole

En informatique, le système d'acquisition de données représente l'interface entre le capteur et l'ordinateur. Ce système, composé de circuit imprimé et de logiciel, permet de recueillir automatiquement les informations analogiques provenant du capteur. En vu de faire rentré et stocké ces informations dans un support de stockage (CD, disque dur,...etc.) ce dernier peut être vu comme un système d'acquisition de données, figure(1.15).

Figure 1.15: Processus d'acquisition d'un signal.

1.4.1 Microphones

Un microphone est un capteur qui représente le premier élément de l'acquisition. Il est considéré comme un transducteur électroacoustique, qui est un dispositif transformant les ondes sonores (oscillation acoustique) en signal électrique (signal audio). Bien qu'un microphone soit un obstacle à la propagation des ondes sonores, pour l'acquisition du signal de la parole, ce microphone est un capteur comportant un organe sensible aux variations de pression dues à l'onde sonore (figure 1.16). Ces variations de pression sont utilisées pour exercer une force sur un système ne pouvant pratiquement pas se déplacer sans cette condition (existence de la force) [16].

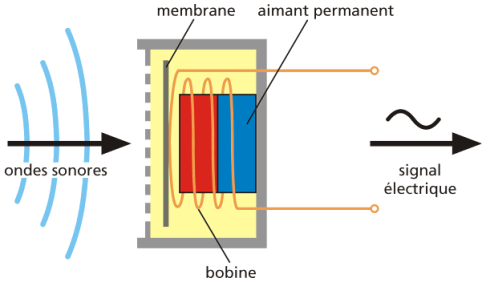

Figure 1.16: Fonctionnement d'un microphone dynamique.

A. Type des microphones

Il existe plusieurs types de microphone (Microphone : à charbon, à condensateur, à magnétostriction, électrodynamique, électronique, thermique, ionique). On prend le microphone à condensateur comme exemple. Ce dernier se trouve dans un circuit comprenant une résistance et un générateur. L'intensité du courant dans le circuit dépend de ces variations .Ce genre de microphone est le plus performant parmi les microphones disponibles, en plus son avantage majeur est sa petite taille ainsi que sa simple construction.

B. Problèmes liées aux microphones

le microphone est un organe de capture sensible à la moindre variation de pression, il peut aussi capter des ondes sonores latérales, qui constituent un bruit à la parole originale destiné à être exploiter, qui va constituer par la suite une défaillance au capture, engendrant une perturbation au traitement du signal projeté, en plus de cela, on peut constater l'influence du microphone lui-même sur la qualité de signale transformé, tout à fait normal, due à la nature de chaque microphone et son principe de fonctionnement (défaillance matérielles).

1.4.2 Cartes son

Une carte son est une carte d'extension d'ordinateur. La principale fonction de cette carte est de gérer tous les sons émis pour les envoyer vers les haut-parleurs ou reçus par l'ordinateur. Elle se présente sous la forme d'un périphérique que l'on peut connecter à l'ordinateur sur un bus PCI, PCI Express, PCMCIA (pour ordinateur portable), USB ou Firewire (bus informatique) [16].

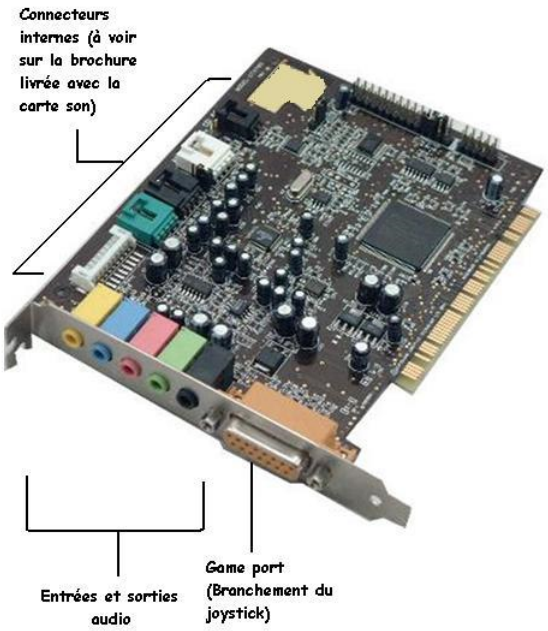

Figure 1.17: Exemple d'une carte son.

A. Rôle d'une carte son

Une fois le signal analogique, issu du microphone arrive à l'entrée MIC de la carte son, il doit passer par un circuit de conditionnement, qui permet l'amplification et le fil-

trage de ce signal, après quoi la conversion Analogique-Numérique est effectuée, dans le but de rendre l'information récupérée, traitable par le système numérique (microordinateur). Cette conversion comprend l'échantillonnage, la quantification et le codage. Après la conversion Analogique-Numérique, la carte son passe à la mémorisation des données numérisées dans un espace mémoire ou tampon (buffer) sous forme de valeurs numérique. Ces données seront présentées par des vecteurs comportant une série de chiffres. On utilise ce genre de mémorisation plusieurs fois pour un même mot prononcé selon le choix de la taille du dictionnaire voulu, attribuée à l'apprentissage des données. La conversion Analogique-Numérique doit passer par les différentes étapes tels que l'échantillonnage, la quantification et le codage.

1. **L'échantillonnage :** Est une technique qui permet de mesurer l'amplitude de l'onde à des intervalles de temps réguliers. Les impulsions représentent les amplitudes instantanées du signal à chaque instant. Plus le nombre d'échantillons est grand et plus le signal sera représenté finement. Pour déterminer la fréquence d'échantillonnage idéale dans les applications d'échantillonnage et de déplacement des hauteurs, le manque de hauteur libre nécessite un filtrage passe bas des échantillons avant que ceux-ci ne soient déplacés vers le haut[6]. A cause des horloges des convertisseurs qui ne sont pas stables, leurs voltages ne sont pas linéaires, les filtres introduisent de la distorsion de phase et ainsi de suite [7].
2. **La Quantification :** Quantifier un signal consiste à placer les amplitudes des échantillons sur une échelle de valeurs à intervalles fixes. Chaque impulsion correspond donc à un nombre binaire unique. -Une quantification à n bits permet d'utiliser $2n$ valeurs différentes. -Pour 8 bits, on a 256 valeurs et pour 16 bits, on a 65536 valeurs. La transformation d'une valeur physique (en volts) en une valeur binaire introduit donc une distorsion. De même lorsque l'impulsion dépasse la valeur maximale prévue [7, 6].
3. **Le codage :** Dans la littérature technique, ce terme englobe indifféremment toutes les méthodes de compression, les paramétrages d'échantillonnage et de quantification. En principe, le codage désigne le type de correspondance que l'on souhaite établir entre chaque valeur du signal analogique et le nombre binaire qui représentera cette valeur [6]. Par exemple, dans le codage PCM chaque valeur binaire (impulsion) est codée telle quelle sans compression. Ceci explique la taille importante des fichiers WAV ou AIFF. En 16 bits / 44 kHz stéréo (codage des CDaudio), 1 minute de musique PCM correspond à 10 MO de données numériques [7].

B. Différents types des cartes son

Il existe une multitude de types de cartes son, on cite les cartes suivantes :

1. **Cartes son multimédia non professionnelle**
   elles sont des interfaces audio-numériques grand public, plus adaptées à sortir le son vers les hauts parleurs qu'à véritablement enregistrer quelque chose hormis une conversation par internet ou à faire fonctionner un logiciel d'apprentissage de langue [18].

   a. Carte son intégrée : C'est celle que l'on trouve par défaut sur la carte mère, que ce soit un ordinateur de forme tour, desktop ou portable. Souvent basée autour d'un chipset de type AC'97 ou Realtek [18].

   b. Carte son multimédia(interne) : Carte interne de type PCI ou ISA (pour les modèles les plus anciens) elle s'installe dans un boitier d'ordinateur, tour ou

desktop, jamais dans un portable. Le modèle phare, dont on emploie parfois le nom comme terme générique est la soundblaster [18].

c. Carte son multimédia(externe) : Certains constructeurs de cartes multimédia comme Hercules ou Soundblaster proposent des cartes multimédia avec un boitier externe pour faciliter les branchements. En général elles sont de qualité supérieures aux cartes multimédia internes et souffrent moins des rayonnement électroniques que l'on rencontre à l'intérieur des boitiers d'ordinateur. Orientées grand-public [18].

2. **Cartes son professionnelles**
   les véritables professionnels en studio n'ont généralement pas d'interface audio-numérique à proprement parler. Ils ont tout un tas d'appareils distincts qui remplissent bien mieux, et pour beaucoup plus cher, toutes les fonctions de nos cartes son. ce genre d'équipement a montrer une qualité supérieure [18].

   a. Carte son pro interne : Carte interne de format PCI ou ISA (pour les plus anciennes) elle ressemble furieusement à une bonne vieille Soundblaster [18].

   b. Carte son pro externe : Beaucoup plus pratiques à l'usage, compatibles avec les ordinateurs portables, ces interfaces externes ont le vent en poupe [18].

C. Différentes limitations des cartes son

Comme chaque dispositif électronique, la carte son présente une multitude de différences qui caractérise chaque génération matérielle et logicielle, en plus le domaine où dont elle est destiné à être utiliser, soit professionnel ou non, et chaque firme pratique sa propre philosophie industrielle qui reflète la qualité fournie pour chaque type de cartes son, par conséquence on remarque la différence entre la qualité des types du son produit par le processus de conversion A/N , en plus de l'impacte de logiciels utilisés.

1.5 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (RAP) est une technique permettant une machine de comprendre, d'analyser et de traiter des informations fournies oralement par un utilisateur humain [19]. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale). En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis, comme un humain le ferait. Cette technologie utilise des méthodes informatiques des domaines du traitement du signal et de l'intelligence artificielle [2]. L'objectif de la reconnaissance automatique de la parole consiste à extraire l'information contenue dans un signal de la parole.

1.5.1 Approches de la reconnaissance de la parole

Il existe trois approches utilisées dans la reconnaissance de la parole : l'approche globale, l'approche analytique et l'approche statistique (hybride) [9]. Elles se différencient

principalement par la nature et la taille des unités de base projetées au traitement et de mettre ces dernières en correspondance avec le signal de la parole [15].

A. Approche globale : Dans l'approche globale, l'unité de base est le mot, elle permet au système d'avoir une image acoustique de chacun des mots qu'il devra identifier par la suite. lors de la phase d'apprentissage. Néanmoins il est limité aux petits vocabulaires prononcés par un nombre restreint de locuteurs [8].

B. Approche analytique : C'est une voie de recherche fondamentale qui concerne la reconnaissance et la compréhension de la parole continue, multi-locuteurs, à grand vocabulaire et langage peu contraint[15]. L'objectif de cette approche est de détecter et d'identifier les composantes élémentaires (phonèmes, syllabes, ...). Cette méthode, basée sur l'identification d'éléments phonétiques, engendra ces années là un recours massif aux traitements du type d'intelligence artificielle pour pallier aux erreurs de décodage des phonèmes [7, 6].

C.Approche statistique (hybride) : L'approche statistique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Ces niveaux sont représentés par des modèles de Markov cachés (HMM). Les unités acoustiques modélisées peuvent être des mots comme dans l'approche globale ou des unités plus courtes telles que le phonème comme dans l'approche analytique, donc elle a combiné entre les deux approches [15].

1.5.2 Problèmes rencontrés aux Systèmes de reconnaissance de la parole

La reconnaissance automatique de la parole (RAP) a des nombreux problèmes d'un point de vue théorique. Leur complexité fait que seuls des sous-problèmes ont pu être à ce jour résolus. Ces solutions partielles correspondent à des contraintes plus ou moins fortes, et les systèmes existants supposent une coopération plus ou moins grande des utilisateurs [15]. Pour classer les systèmes de reconnaissance automatique, on a généralement recours aux critères suivants [7] :

A. Mode d'élocution (continuité)

La production d'un son est fortement influencée par les sons qui le précédent et le suivent en raison de l'anticipation du geste articulatoire. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. Évidement il est plus simple de reconnaître des mots isolés bien séparés par des périodes de silence que de reconnaître la séquence de mots constituant une phrase. En effet, dans ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés [9].

B. Une grande variabilité

Elle a un contenu phonétique égal, le signal vocal est très variable pour un même locuteur (variabilité intra locuteur) ou pour des locuteurs différents (variabilité interlocuteur) [9, 15].

1. Variabilité intra-locuteur : Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution en détermine la durée. Toute affection de l'appareil phonatoire peut altérer la qualité de la production. Un rhume teinte les voyelles nasales ; une simple fatigue et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps : l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix de leurs sceaux [6].

2. Variabilité interlocuteur : Elle est encore plus flagrante. Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonnantes, modifient la production acoustique. En plus, il y a la hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. Enfin toute parole s'inscrit dans un processus de communication où entrent en jeu de nombreux éléments comme le lieu, l'émotion, l'intention, la relation qui s'établit entre les interlocuteurs. Chacun de ces facteurs détermine la situation de communication, et influe à sa manière sur la forme et le contenu du message [6].

C. L'environnement et le niveau de bruit ambiant

Indépendamment de ce qui précède, l'environnement acoustique et les conditions de prise du son, constituent un facteur important : la présence du bruit, même stationnaire, dégrade en général fortement les performances des systèmes de reconnaissance.De plus si ce bruit est intense, il induit une augmentation de la variabilité chez le locuteur. Donc il constitue une gêne dans la compréhension de la parole [6, 8].

D. Le langage

La prise en compte de la syntaxe du langage produit par l'utilisateur sera plus facile pour un langage (rigide), très contraint, que si toute la souplesse de la langue naturelle parlée peut être rencontrée [8].

E. Taille du vocabulaire

Un grand vocabulaire pose une difficulté majeure pour la grammaire (la complexité du langage autorisé) [8].

1.5.3 Phases de système de reconnaissance de la parole

Généralement un système de reconnaissance de la parole se constitue de six modules principaux dont chacun assure une tâche primordiale, et éventuellement ces tâches sont composés par des sous tâches, l'ensemble du système coopère dans son intégralité pour accomplir la mission essentiel, qui se caractérise à la reconnaissance de la parole, la schématisation ci dessous montre les différentes étapes mentionnés (figure 1.18) [19].

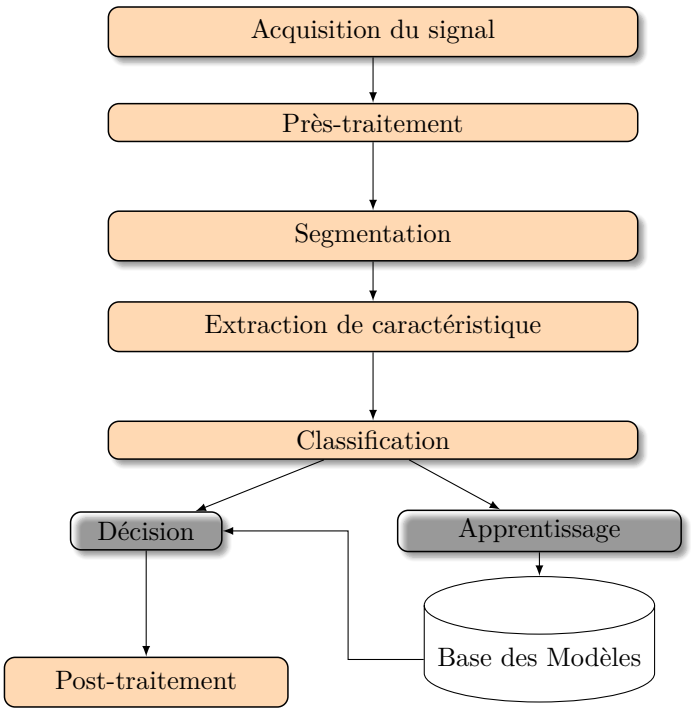

Figure 1.18: Schéma général d'un SRAP [19].

A. Acquisition du signal

La première phase consiste à transformer le signal de parole déjà capté via un microphone en une séquence de vecteurs acoustiques, pour qu'il soit traitable par un ordinateur, après un filtrage, le signal doit tout d'abord être numérisé à travers une carte spécialisée. La numérisation sonore repose sur deux paramètres qui sont la quantification et la fréquence d'échantillonnage. Le choix de la fréquence d'échantillonnage est un facteur déterminant pour la définition de la bande passante représentée dans le signal numérisé [19].

B. Pré-traitement

Après l'acquisition du signal, une phase de pré-traitement ou de filtrage est très cruciale, a pour but de retirer les distorsions ou les bruits provenant du matériel ou de l'environnement du locuteur. Ce module est aussi appelé «traitement du canal de transmission». Du fait de sa complexité et du peu d'amélioration qu'il apporte, cette phase n'est pas toujours intégrée aux systèmes. Cependant, la recherche de meilleurs traitements du canal de transmission sera impérative pour l'amélioration des systèmes de reconnaissance vocale [19].

C. Segmentation

La segmentation est donc une étape majeure dans la reconnaissance des mots parlés. Selon un schéma classique des modèles de reconnaissance des mots parlés, la segmentation du signal de parole intervient entre la phase de pré-traitement et l'extraction de caractéristiques. La segmentation correspond au découpage du signal acoustique en unités discrètes linguistiques. Des difficultés apparaissent notamment dans plusieurs contextes comme l'acquisition de la langue maternelle, la reconnaissance de mots parlés et l'acquisition d'une seconde langue, du fait de la double articulation. Cependant, des stratégies ont été mises en évidence pour pallier à ce problème de segmentation. On peut définir

plusieurs types de segmentation (organisation du segment le plus court au segment le plus long)[19, 13] :

1. en phonèmes ;
2. en syllabes ;
3. en mots ;
4. en groupes inter-pausaux (segments délimités par deux pauses silencieuses) ;
5. en locuteurs et tours de parole.

D. Extraction des caractéristiques

Une fois que la parole est segmentée l'extraction des paramètres devient nécessaire pour interpréter l'information cachée derrière ce signal, qui est appelé aussi un vecteur de caractéristique ou descripteur qui pourront être utilisées pour le traitement de signal vocal pour la reconnaissance. Il existe des approches, et pour chaque approche il y a plusieurs techniques (qui vont être décrites dans le prochain chapitre) [19].

E. Classification

La classification ou la reconnaissance dans un Système RAP regroupe les deux tâches d'apprentissage et de décision. Elles tentent toutes les deux, à partir de la description en paramètres extraits dans l'étape précédente, d'attribuer une forme acoustique à un modèle (ou à une classe) de référence. la classification doit assurer la propriété de compacité et la séparabilité [19].

1. Apprentissage : L'étape d'apprentissage est l'une des étapes les plus importantes dans le processus de reconnaissance c'est l'étape de la construction du dictionnaire de référence ,En outre, l'apprentissage est dit supervisé, si la tâche d'apprentissage est guidée par un superviseur (concepteur) ou apprentissage non supervisé, si les classes sont créés automatiquement, sans l'intervention d'un opérateur, à partir d'échantillons de référence et de règles de regroupement [19, 11].

2. Décision : La décision est l'ultime étape de la reconnaissance. A partir de la description en paramètres, elle recherche, parmi les modèles d'apprentissage en présence, ceux qui sont les plus "proches", et cela en un temps aussi court que possible. La décision peut conduire à un succès si la réponse est unique (un seul modèle répond à la description de l'image acoustique). Elle peut aussi conduire à une confusion (substitution) si la réponse est multiple (plusieurs modèles correspondent à la description). Comme elle peut conduire à un rejet de la forme si aucun des modèles ne correspond à sa description [19].

F. Poste-traitement

Cette phase consiste à faire une sélection de la solution en utilisant des niveaux d'information plus élevés (syntaxique, lexicale, sémantiques...etc.) [2, 12]. Le poste-traitement se charge également de vérifier si la réponse est correcte (même si elle est unique) en se basant sur d'autres informations non disponibles au classificateur [19].

1.5.4 Domaine d'applications

La technologie de la reconnaissance de parole est utilisée dans des nombreux domaines, cette dernière permet de faciliter plusieurs tâches, on cite à titre d'exemple quelques domaines d'applications et éventuellement les logiciels utilisés :

1. **Services vocaux :** utilisation " mains libres", par exemple un téléphone de voiture ou bien le commande et le contrôle de services d'accès aux bases de données (exemple : Dragon Naturally Speaking, JARVIS, MacSpeech).
2. **Télécommunication :** De nombreuses applications pourraient bénéficier de la reconnaissance de la parole pour faciliter l'accès à des données ou des services sur lignes téléphoniques (exemple : Voxygen TTS Server), notamment l'automatisation des services de renseignements [9, 11].
3. **Saisie des données :** grâce à une interface vocale qui donne une liberté et une rapidité de mouvement. Pendant qu'il observe un processus complexe (exemple : Dragon Naturally Speaking, IBM ViaVoice, IIVoxForge) [9, 11].
4. **Avionique :** A bord des avions, les tâches étant de plus en plus complexes et le tableau de bord de plus en plus réduit, la parole permet au pilote d'avoir à sa disposition un moyen supplémentaire d'interaction avec la machine, sans cependant gêner l'accomplissement des tâches courantes qui requièrent de sa part toute son attention visuelle [9, 8].
5. **Application bureau/PC :** Donner la possibilité de RAP par applications, et aux environnements des stations de travail, le contrôle et l'interaction homme-PC, comme exemple la dictée vocale (exemple : Dragon Naturally Speaking, MacSpeech, Talking desktop) [9, 11].
6. **Applications médicales et légales :** Utilisant la RAP pour la rédaction de rapports ou remplissage de formulaires, la reconnaissance de caractères et de séquençage de l'ADN (exemple : Hidden Markov Model Toolkit, Dialoca Santé) [11].
7. **Production et fabrication :** Où la RAP peut être utilisée pour la commande de contrôle vocal de processus de fabrication (par exemple, pour l'accès aux systèmes de contrôle de qualité) et apporter une aide au tri et envoi de paquets [9].
8. **Autres applications :** Telles que l'aide aux handicapés (exemple : Finger Voix) (contrôle par voix, machines à parler vocale) et l'utilisation de la RAP dans les jeux électroniques [10, 8].

1.6 Conclusion

La parole est le moyen utilisé pour communiquer les pensées entre les humains qui sont les seuls être vivants à utiliser un tel système structuré, Il est évident de commencer par des définitions de notions de bases qui seront utiles pour bien comprendre l'essentiel de ce chapitre et la suite de cette thèse. Il s'agit de faire un rappel théorique concis mais suffisamment complet, à savoir des généralités sur la parole et les principales caractéristiques acoustiques du signal de la parole. Nous avons aussi présenté quelques techniques de base d'analyse et de modélisation du signal de la parole. Ces techniques peuvent être utilisées pour mettre en évidence les caractéristiques fréquentielles de ce signal et pour mettre en œuvre, en exploitant ces caractéristiques, nous avons ensuite présenter les outils d'acquisition de la parole avec leurs variétés et leurs limitations. On a également expliqué la RAP avec les différentes approches et les problèmes rencontrés dans ce processus, à la fin de ce chapitre nous avons cité les principaux domaines de SRAP avec des exemples de logiciels employés dans la RAP. Le chapitre suivant aborde les différentes méthodes de segmentation.

Chapitre 2

Traitement de la parole

2.1 Introduction

Nous consacrons ce chapitre à la phase de la segmentation de la parole pour laquelle nous allons présenter un aperçu sur les recherches déjà faites dans ce domaine. On commençant par une vue sur les différentes approches de la segmentation. Ensuite, nous allons définir les classes de segmentations selon la notation de la connaissance à priori du contenu du signal et nous citons les méthodes les plus connues de chaque classe. Finalement, nous expliquons la paramétrisation et l'extraction des caractéristiques du signal d'une segmentation automatique de la parole.

2.2 Généralité de la segmentation

Pour bien comprendre la segmentation, il est évident de la définir et de connaitre ses différentes paramétrés :

2.2.1 Définition de la segmentation

La segmentation de la parole est une opération nécessaire dans le traitement de la parole, qui consiste à découper le signal en segments assez extrêmement homogènes pouvant être transcrits en unités de base (phonème, syllabe...). Ou on trouve ces unités variées selon la nature du segment considéré (figure 2.1). Il existe plusieurs types de segmentation selon la taille du segment traité, on cite quelques types de segment du plus court au segment le plus long [14] :

• **Segment en voisé/non-voisé :** Les sons voisés se résulte par la vibration des cordes vocales. Les voyelles sont généralement voisées, cependant les consonnes peuvent être voisées ou non ;

• **Segment en phonèmes :** Cette technique consiste à délimiter la continuité acoustique d'un signal à une séquence de segments d'un ensemble discret et fini d'éléments, qui est l'alphabet phonétique de la langue (exemple : le mot "arab" en le divise en " a,r,a et b") ;

• **Segment en syllabes :** La syllabe est l'unité structurante de la langue, elle est décomposée en 3 parties : l'attaque,le noyau et la coda. On trouve quelquefois une difficulté de segmentation d'une phrase en syllabes à cause de la caractéristique facultative des consonnes ;

• **Segment en mots :** il est difficile de segmenter un message en ses constituants élémentaires. Pour résoudre cette complexité, on se base à la reconnaissance de mots prononcés

isolés qui sont séparés par des silences de durée supérieure à quelques dixièmes de secondes, elle applicable par l'approche globale ;
• **Segment en locuteurs et tours de parole :** La segmentation selon le locuteur apparait pour résoudre l'ambiguïté entre plusieurs locuteurs, il s'agit de segmenter en tours de parole pour chaque locuteur) ;
• **Segment en groupes inter-pausaux :** segments délimités par deux pauses silencieuses.

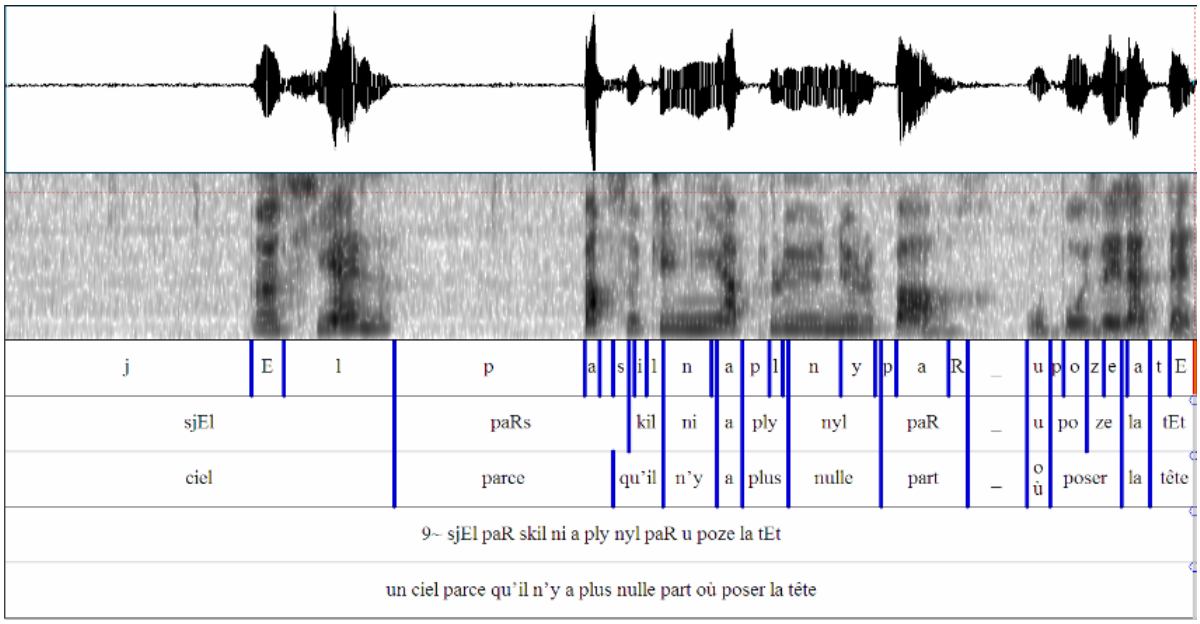

Figure 2.1: Exemples de segmentation de parole continue.

2.2.2 Paramétrisation du segment

Pour chaque segment un vecteur de paramètres (traits acoustiques) est extrait, ces paramètres peuvent être[6] :
- **pertinents :** Extraits de mesures suffisamment fines, ils doivent être précis mais leur nombre doit rester raisonnable (éliminer la redondance des données) afin de ne pas avoir de coût de calcul trop important dans le module du décodage.
- **discriminants :** Ils doivent donner une représentation caractéristique des sons de base et les rendre facilement séparables.
- **robustes :** Ils ne doivent pas être trop sensibles à des variations de niveau sonore ou à un bruit de fond.

2.3 Présentation de langue arabe

L'arabe est une langue parlée par plus de 250 millions de personnes classé sixième mondialement. Elle est une langue officielle dans moins 22 pays. C'est aussi la langue de référence pour plus d'un milliard de musulmans. Comme son nom l'indique, la langue arabe est la langue parlée à l'origine par le peuple arabe. C'est une langue sémitique (comme l'hébreu, l'araméen et le syriaque). Au sein de cet ensemble, elle appartient au sous groupe du sémitique méridional. Le développement de la langue arabe a été associé à la naissance et la diffusion de l'islam. L'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. Ce développement s'est accompagné d'une rapide et profonde évolution (en particulier dans la syntaxe et l'enrichissement lexical). Elle a un alphabet de vingt-huit lettres, dont vingt-cinq représentent des consonnes et trois représentent les voyelles longues (/ ا/ ي/ و /). Chaque lettre apparaît souvent en quatre formes selon qu'elle soit en début, en milieu ou en fin de mot, ou isolée[19, 15]. Dans la phonologie l'alphabet arabe est classé selon des consonnes et des voyelles.
Les consonnes : Une consonne est un phonème dont la prononciation se caractérise par une obstruction totale ou partielle en un ou plusieurs points du conduit vocal. Elle est généralement précédée ou suivie d'une voyelle [19].

Les voyelles : Lors de la prononciation des voyelles, l'air émis par les vibrations des cordes vocales passe librement à travers le conduit. On distingue trois types de voyelles [19] :

1. les voyelles courtes «" ’", " ِ", " ً"»,
2. les voyelles longues «" ا", " ِي", " و" »
3. les semi-voyelles « sekune et tanwin ».

le tableau (2.1) illustre les différentes classes des sons arabes classés selon le mode d'articulation.

Tableau 2.1: Classification des sons arabes selon le mode d'articulation [15].

Mode d'articulation							
Voyelle orale	Semi voyelle	Fricative	Son combiné	Occlusive	Nasale	Littérale liquide	Vibrant
ـَ	و	ف	ج	ب	م	ل	ذ
’	ي	د		ض	ن		
ـِ		ث		ر			
أ		ظ		ل			
أو		س		ت			
إي		ز		ط			
		خ		ك			
		ح		ف			
		ص					
		ش					
		ع					
		غ					
		ه					

2.4 Approches de la segmentation

Il existe trois approches permettant d'aborder la reconnaissance de la parole. L'approche globale, l'approche analytique et l'approche hybride (statistique). La première approche basée sur la décomposition de la parole continue la plupart du temps en utilisant un décodage acoustico-phonétique utilisé par des modules de niveau linguistique. La deuxième basé sur l'identification globale d'un mot ou une phrase en utilisant la notion de comparaison de l'identifient avec des références enregistrées. L'introduction des méthodes statistiques basées sur des modèles de Markov à donner une intégration pour la reconnaissance de la parole continue et le traitement des grands vocabulaires[15].

2.4.1 Approche globale

Dans l'approche globale le mot ou la phrase sont utilisés comme des entités élémentaires (donc non décomposable) et en les comparant avec des références enregistrées. Cette opération est faite lors de la phase d'apprentissage, où chacun des mots est prononcé une ou plusieurs fois, cette méthode fournit une image acoustique de chaque mot à identifier et permet donc d'éviter l'influence mutuelle des sons à l'intérieur des mots. Elle se limite aux petits vocabulaires prononcés par un nombre limité de locuteurs (les mots peuvent être prononcés de manière différente tout dépend de locuteur). Dans le cas de mots isolés à grand vocabulaire, cette dernière n'est plus valable puisque, elle demande une mémoire et une puissance considérable pour stocker les images acoustiques de tous les mots du vocabulaire et comparer un nouveau mot avec l'ensemble des mots du dictionnaire [15, 17].

2.4.2 Approche analytique

L'approche analytique décompose le problème de la parole continue en des composants élémentaires et des unités acoustiques courtes (phonèmes, syllabes...) qui seront les unités de base à reconnaître. Cette méthode consiste à enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base. Une petite phrase de très courte durée, peut aussi être considérée comme un mot [15, 8]. Pour reconnaître de grands vocabulaires. Un exemple classique de cette approche est l'analyse par traits : des indices acoustiques sont calculés à partir du signal de la parole, ils permettent de faire des hypothèses locales sur certains traits phonétiques, comme le voisement, la nasalisation, le lieu d'articulation ou le degré d'ouverture du conduit vocal. En fonction de ces traits, le signal acoustique est segmenté et une identification phonétique des segments est réalisée. Le décodage acoustico-phonétique ainsi obtenu est exploité par des modules d'ordre linguistique. Les niveaux lexicaux, syntaxique ou sémantique utilisent des sources de connaissances spécialisées et sont organisés avec le module acoustique dans des architectures montantes ou descendantes. C'est donc la méthode analytique qui est utilisée car les mots ne sont pas mémorisés dans leur intégralité mais traités en tant que suite de phonèmes, tandis qu'elle est restée au stade expérimental a cause de leur faiblesse qui provient d'un processus de décision trop précoce, à savoir une segmentation préalable à l'identification ou une identification phonétique sans prise en compte des niveaux linguistiques [15, 17].

2.4.3 Approche hybride (statistique)

Aujourd'hui, la décomposition du problème de la reconnaissance de la parole continue est devenue classique d'après la formalisation statistique simple proposée par F. Jelinek. Cette approche est basé sur des modèles de Markov cachés (HMM) le principe est de

rechercher la suite des mots prononcés la plus probable parmi toutes les suites de mots acoustiques. L'approche statistique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Dans cette approche, les unités acoustiques utilisées peuvent être des mots comme dans l'approche globale ou des unités plus courtes telles que le phonème comme dans l'approche analytique. D'après la modélisation markovienne qui est plus générale que l'alignement temporel dynamique et tient compte en plus de non linéarité temporelle du processus la variabilité acoustique de la production de la parole. Elle devient applicable à la reconnaissance de la parole continue qu'après l'augmentation de la puissance des ordinateurs et de la taille des bases de données disponibles [15, 17].

2.5 Classes des méthodes de la segmentation de la parole

Les méthodes de la segmentation acoustique de la parole se divisent en deux grandes classes de méthodes selon leur démarche au connaissance à priori du contenu linguistique de ce signal ou non [13]. Tel que :

• La première classe est dite sans contrainte linguistique dans la mesure où elle n'est pas nécessaire de connaître l'étiquetage du corpus de la parole.

• La deuxième classe est dite avec contrainte linguistique qui nécessite une connaissance a priori de cette transcription.

Par la suite nous allons voir plus en détails chacune de ces classes de méthodes avec leurs sous classes.

2.5.1 Segmentation sans contrainte linguistique

Cette classe englobe toutes les méthodes qui permettent de segmenter un signal de la parole sans connaissance à priori du contenu linguistique de ce signal. Cette classe peut être également divisée en deux sous-classes. Une contenant les méthodes locales de segmentation partielle qui sont dédiées à certaines classes de sons spécifiques. La deuxième regroupe les méthodes locales et les méthodes globales qui cherchent à segmenter tout le signal [15, 17]. Nous citons quelques méthodes qui appartiennent à cette classe.

A. Détection de ruptures de stationnarité dans le temps

Ces méthodes posent des hypothèses sur le signal de parole pour chercher à détecter des ruptures correspondant à des discontinuités de stationnarité. Parmi les hypothèses de ces méthodes c'est que la parole est une suite d'unités stationnaires, chaque unité étant caractérisée par un modèle autorégressif. Ces modèles utilisent un critère statistique pour détecter de changements de paramètres tel que le changement de paramètres c'est pour trouver une rupture de stationnarité, par exemple l'algorithme de Brandt l'un de ces méthodes [15, 17].

B. Détection d'activité vocale

Les méthodes de détection d'activité vocale a comme rôle de définir précisément les zones contenant de la parole à partir des échantillons du signal de la parole. Qui donne une segmentation silence/parole. Ces méthodes sont très appliquées dans les domaines de la compression, du codage et de la reconnaissance de la parole. Pour traiter la parole bruitée, il y a des méthodes simples mais aussi

des méthodes plus complexes. Parmi les méthodes simples d'identification des segments silence/parole la méthode qui se base sur la comparaison des amplitudes du signal de parole avec le niveau du bruit, où le niveau de bruit est calculée avec les valeurs absolues des amplitudes sur une portion du silence. Une autre méthode simple est basée sur la fonction d'énergie à court-terme qui est calculée par la somme du signal multiplié par une fonction de fenêtrage sur N trames. La détection des trames contenant de la parole et les trames du silence se fait par l'application d'un seuil à cette fonction [13].

C. Détection de voisement/non voisement

Il existe plusieurs méthodes pour effectuer la segmentation par détection de voisement/non voisement. Les méthodes effectuant cette détection peuvent être classées en deux catégories [13] :

- les méthodes temporelles comme le HNM.
- les méthodes fréquentielles basées sur les ondelettes.

Le classement des segments du signal en segment voisement ou non voisement se fait selon les valeurs d'une mesure HNR (Harmonic-to-Noise Ratio) locale de l'énergie, le taux de passage par zéro (ou ZCR), l'analyse des coefficients d'autocorrélation et l'analyse du spectre. Par exemple la détection du voisement selon le nombre de passages par zéro est de faire la courbe qui passe par les milieux des segments, puis de détecter les passages par zéro de cette courbe. Cette dernière est considérée comme une estimation grossière du contenu des basses fréquences du signal de la parole. Telle que la courbe se caractérise par peu de passages à zéro dans le cas de signaux voisés,.

D. Segmentation fricatif/non-fricatif

Cette segmentation cherche à identifier un bruit de friction. La statistique du nombre de passages par zéro de la dérivée du signal est la techniques utilisé pour déterminer le bruit. La raison de l'orientation vers ce type de techniques est que les fricatives ne présentent que peu ou pas d'énergie en dessous de 2 kHz. Les voyelles et les consonnes liquides et nasales se distinguent par une énergie concentrée dans le bas du spectre pour cela on utilise cette méthode [15, 17].

E. Segmentation par ondelettes

La segmentation par ondelettes sont des méthodes d'analyse du signal sur de variante en temps et en fréquence. Les méthodes les plus connues de ces méthodes sont la segmentation par paquets d'ondelettes, par ondelettes de Malvar et l'AMR (Analyse Multi-Résolution). Les méthodes de segmentation par paquets d'ondelettes favorisent la résolution fréquentielle par rapport à la résolution temporelle, Par contre les ondelettes de Malvar favorisent la résolution temporelle par rapport à la résolution fréquentielle et l'AMR est une analyse fréquentielle dyadique. Une amélioration de l'AMR qui est l'AMREC (Analyse Multi-Résolution de l'Enveloppe Complexe d'un signal) basée sur l'analyse locale de la présence d'énergie du signal autour d'une fréquence choisie par l'utilisateur. l'AMREC peut être utilisée pour analyser le signal de parole dans trois canaux : canaux basses fréquences, fréquences intermédiaires et hautes fréquences. À partir des coefficients en sortie de l'AMREC dans ces trois canaux, on peut identifier certains traits acoustiques, segmenter le signal de la parole, ou combiner l'identification et la segmentation de manière à reconnaître certains phonèmes [15, 17].

F. Détection des variations spectrales

Le principe de ces méthodes est basé sur la décomposition fréquentielle du signal sans connaissance a priori de sa structure fine. Il s'agit donc de transformer le signal original de la représentation temporelle à une représentation fréquentielle par la transformation de Fourier selon la formule [19] :

$$F(w) = \int_{-\infty}^{+\infty} f(t)e^{-jwt}dt$$

Où $j^2 = -1$ et $f(t)$ est la fonction temporelle

2.5.2 Segmentation avec contrainte linguistique

Les méthodes de segmentation avec contrainte linguistique sont des méthodes basées sur l'utilisation de l'information linguistique ou l'étiquetage pour poser les frontières d'unités acoustiques. Le but de ces méthodes est d'étendre des frontières de phones pour obtenir un nombre de segments égal au nombre d'étiquettes de la séquence phonétique. Chaque segment acoustique est en effet caractérisé par son étiquette linguistique. D'ou les marques de frontière entre ces segments acoustiques qui représentent les marques de transition entre les phones constituant le signal. La segmentation d'un signal de la parole par les méthodes avec contraintes linguistique utilisant une mesure de distorsion permet d'évaluer la partition de la séquence de trames acoustiques du signal en un nombre de classes d'équivalence représentées par les symboles linguistiques, nous citons ci-dessous quelques méthodes de segmentation du signal de la parole qui tiennent compte de la description phonétique :

A. Segmentation par Dynamic Time Wraping

La méthode de segmentation d'alignement temporel dynamique DTW (Dynamic Time Wraping) utilise un algorithme de synthèse de parole pour générer une forme de référence du signal de la parole à segmenter. Les instants de transition entre les segments phonétiques de ce signal sont connus et prédéterminés grâce au dictionnaire qui a permis d'effectuer la synthèse. Le principe de cette algorithme est de mesurer la similarité de manière optimale entre deux vecteurs, n'est pas obligatoirement de même longueur, un vecteur cepstral source avec un vecteur cepstral cible selon un critère de ressemblance acoustique. D'autre terme l'algorithme DTW permet d'aligner les séquences de trames acoustiques de deux signaux de la parole d'une manière à minimiser la distorsion spectrale entre les trames acoustiques alignées (figure.2.2). Enfin en déduire la segmentation du signal de parole grâce à cet alignement [5, 4].

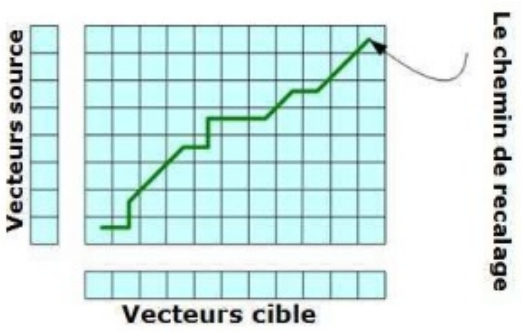

Figure 2.2: Alignement temporel par la DTW décrivant le chemin de similarité entre les vecteurs source et cible [5].

L'avantage de cette méthode de segmentation est qu'elle ne requiert pas (comme nous le verrons dans les méthodes de segmentation à bases de modèles neuronaux ou markoviens) un apprentissage des modèles des classes d'équivalence phonétiques. Son inconvénient majeur réside dans le fait qu'elle ne peut être facilement adaptée à la segmentation des énoncés d'un locuteur particulier.mais son inconvénient est de disposer d'un synthétiseur de parole de langue similaire à celle de corpus de parole à segmenter. De plus, elle reste moins performante que la segmentation par HMM.

B. Segmentation par Markov cachés

Le préambule de la méthode de segmentations par modèle de Markov cachés HMM ($HiddenMarkovModelsouHMMs$) était dans les années 1960 − 1970 par Baum et ses collaboratrices, est définie par un processus stochastique déterminé par le quintuplé [5].

$$\lambda = (S, \delta, T, G, \pi) \text{ ou}$$

- S : est un ensemble de N états,
- δ : est un alphabet de M symboles,
- $T = S \times S \rightarrow [0, 1]$: est la matrice de transition, indiquant les probabilités de transition d'un état à l'autre on note $P(s \rightarrow s_0)$ la probabilité de transition de l'état s vers l'état s_0
- $G = S \times \delta \rightarrow [0, 1]$ est la matrice de génération, indiquant les probabilités de génération associées aux états on note $P(o|s)$ la probabilité de générer le symbole o appartenant à δ à partir de l'état $s \in S$.
- $\pi : S \rightarrow [0, 1]$ est un vecteur de probabilités initiales de visite

Il n'y a pas de règle stricte pour choisir l'architecture du HMM, par conséquent nous trouvons des travaux sur l'apprentissage dynamique du nombre d'états d'un Modèle de Markov Caché à des observations continues au traitement de signal et au traitement d'images. La procédure de génération d'une séquence $o_1...o_T$ de symboles à l'aide d'un HMM consiste à partir d'un état s en suivant la distribution π , de se déplacer d'état en état suivant les probabilités de transition, et générer un symbole sur chaque état rencontré en utilisant la distribution de probabilité de génération associée à l'état. Lorsqu'un symbole a été généré, on choisit une transition sortante suivant la distribution de probabilité de transition associée à l'état courant, et la procédure est réitérée jusqu'à la $T^{ième}$ génération de symbole (figure 2.3) [7, 5].

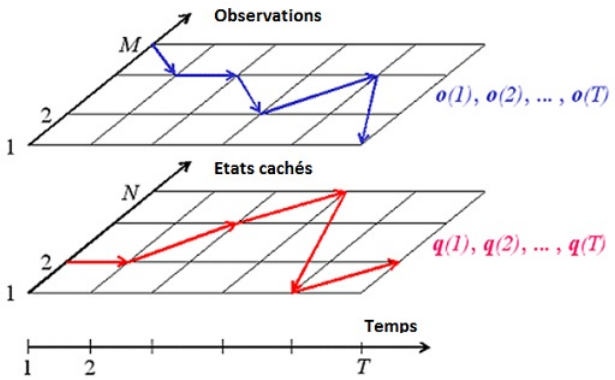

Figure 2.3: Séquence caché et observé de la HMM [5].

Ces méthodes HMM ont présenté une grande amélioration à la qualité et les performances des systèmes de reconnaissance et de synthèse vocales et ils ont montré une capacité à traiter de grands corpus de parole continue. Cette approche statistique est constituée de deux étapes. La première étape est d'apprendre des modèles (HMM) des unités acoustiques. La deuxième étape est le décodage ou alignement. Il est constitué de deux types : un décodage sans contrainte linguistique ou un décodage avec contrainte linguistique. Le premier cherche à trouver les unités acoustiques contenues dans le signal de la parole et les frontières de chaque unité, qui produisent une segmentation du signal pouvant contenir des erreurs de substitutions, d'omissions et d'insertions. Le deuxième type de décodage utilise la connaissance à priori de l'étiquetage et cherche à trouver les frontières de chaque étiquette. On utilise ce genre de décodage dans les systèmes de synthèse vocale afin de créer les dictionnaires d'unités acoustiques.

C. Segmentation par réseaux de neurones

Ce modèle de segmentations proposé par (Vorstermans et al. 1996) utilise des réseaux de neurones artificiels pour estimer les probabilités à posteriori des marques de frontières phonétiques et des classes phonétiques larges de la langue. Cette méthode de segmentation basée sur le même principe qui utilise les modèles de Markov cachés HMM, pour déterminer les frontières de phones des signaux de parole après l'estimation des paramètres des modèles sur un corpus d'apprentissage, un alignement est effectué entre la séquence des trames du signal à segmenter et la séquence des modèles associés au contenu linguistique de l'énoncé [13].

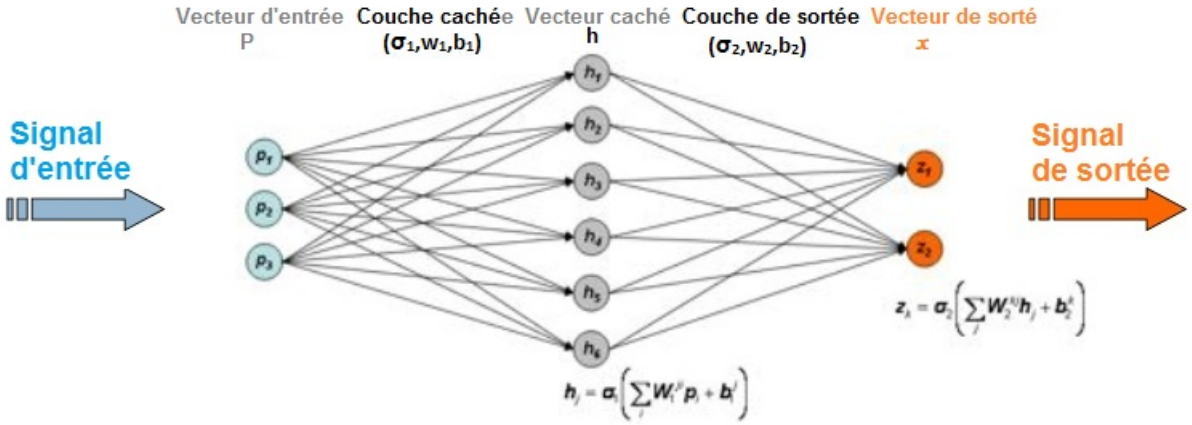

Figure 2.4: Exemple de traitement de signal en réseaux neurones.

2.6 Paramétrisation des caractéristiques de signal

L'extraction du maximum d'informations utiles à partir du signal de la parole obtenue avec l'exactitude le plus possible d'information nécessite cette phase pour faire la tâche de fenêtrage sur les trames de signal et la tâche d'extraire des paramètres et des coefficients représentatifs du signal de la parole. Selon chaque méthode de segmentation il y a des techniques d'extraction de caractéristique à employer (voir la figure 2.5).

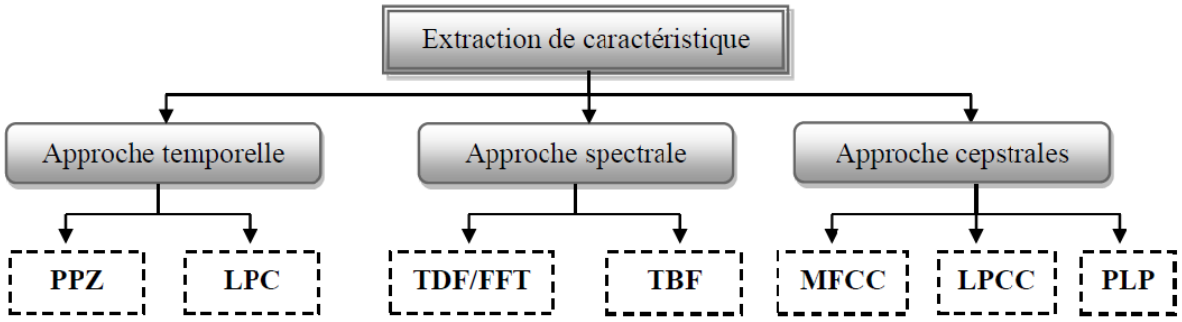

Figure 2.5: Schéma présentant les différentes méthodes d'extraction de caractéristique [19].

2.6.1 Groupement en trames (Frame Blocking)

Le signal acoustique continu est segmenté en trames de N échantillons, avec un pas d'avancement de M trames (M inf N) , où le chevauchement de deux trames consécutives est N – M échantillons. En pré-traitement les valeurs utilisées pour M et N sont respectivement 10 et 20. Dans la plupart du temps à la pré-accentuation du signal en appliquant l'équation de différence du premier ordre aux échantillons $x(n)$ avec l'equation suivante :

$$x'(n) = x(n) - kx(n-1)$$

Où, k représente un coefficient de pré-accentuation qui peut prendre une valeur entre le zéro et un : $0 < \mathrm{k} < 1$.

2.6.2 Fenêtrages

L'extraction du maximum d'informations à partir des trames résultantes de la phase précédente et avec le problème de risque de discontinuités d'information aux frontières de ces trames nécessite une phase de fenêtrage qui consiste à réduire ce problème. Le fenêtrage se fait selon des fonctions ou formules est d'appliquer sur l'ensemble des échantillons prélevés dans la fenêtre du signal original de façon à diminuer les effets de bord. Il existe plusieurs fenêtre, on va citer les plus utilisées : Dans tout ce qui suit, on définit $w(n)$ comme fenêtre où $0 < n <$ N-1 et N représente le nombre d'échantillons dans chacune des trames, alors le résultat du fenêtrage est le signal x_a, donné par la formule[5] :

$$x_a = x(n)w(n),\ 0 < n < \text{N-1}$$

- Rectangulaire : $w(n) = \begin{cases} 1 & si \quad 0 < \mathrm{n} < N-1 \\ 0 & sinon \end{cases}$
- Triangulaire : $w(n) = \begin{cases} \frac{2n}{N-1} & si \quad 0 \leq n \leq \frac{N-1}{2} \\ \frac{2(N-n-1)}{N-1} & si \quad \frac{N-1}{2} < n \leq N-1 \\ 0 & sinon \end{cases}$
- Hamming : $w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & si \quad 0 < n < N-1 \\ 0 & sinon \end{cases}$
- Hanning : $w(n) = \begin{cases} 0.5 - 0.5\cos(\frac{2\pi n}{N-1}) & si \quad 0 < n < N-1 \\ 0 & sinon \end{cases}$

- Blackman : $w(n) = \begin{cases} 0.42 - 0.5\cos\frac{2\pi n}{N-1} + 0.08\cos\frac{4\pi n}{N-1} & si \quad 0 < n < N-1 \\ 0 \quad sinon \end{cases}$

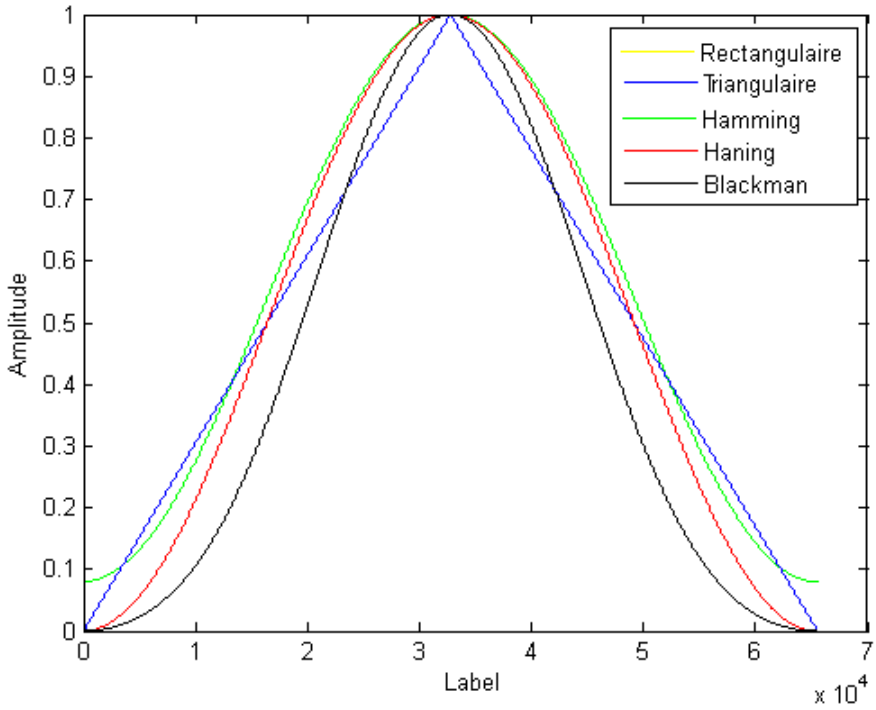

Figure 2.6: Exemples représentatifs des formules du fenêtrage.

2.6.3 Taux de passage par zéro

Cette technique se base sur le calcul du passage par zéro du signal, de construire des histogrammes d'intervalles de fréquence. On ne tient pas compte dans cette méthode à l'énergie du signal mais à son signe. Pour un signal échantillonné, il y a un passage par zéro lorsque deux échantillons successifs sont des signes opposés.(dont *(sgn)* est la fonction signe définie par la fonction ci-dessous). Le calcul du taux de passage par zéro (TPPZ) du signal de la parole permet de faire la distinction d'une part entre le signal de la parole (information utile) et le bruit, et d'autre part avec le silence. Une caractéristique pour le taux de passage par zéro, est qu'il est élevé pour la parole et très faible pour le silence. Le taux de passage par zéro constitue un outil important pour la classification parole/silence , et pour la détection du début et la fin de la parole dans un signal vocal [1].

$$sgn[x(n)] = \begin{cases} +1 \quad si \quad \text{x(n) >=0} \\ -1 \quad si \quad \text{x(n) <0} \end{cases}$$

la figure 2.7 montre le passage par zéro dans un signal de parole dont le signe se change après à chaque fois qu'il traverse l'axe des x.

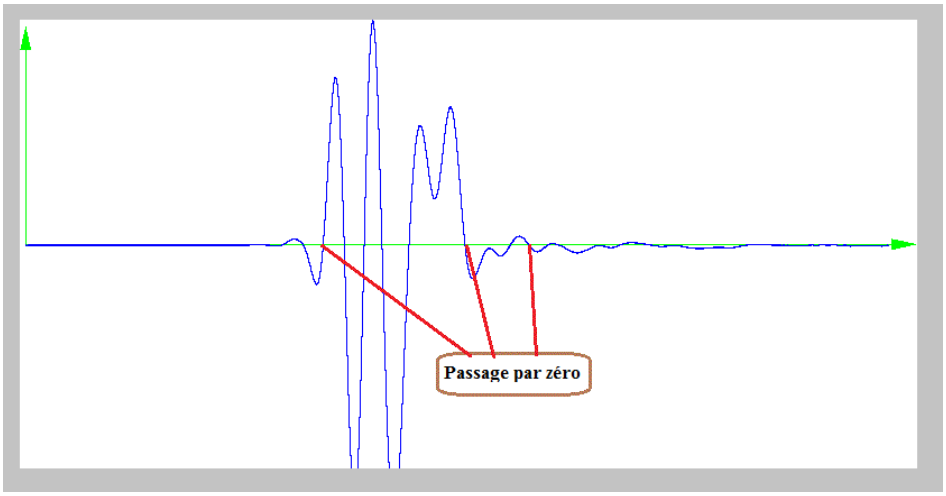

Figure 2.7: Figure présentant le passage par zéro dans un signal.

2.6.4 Transformée de Fourier rapide (Fast Fourier Transform)

L'objectif de la transformée de Fourier rapide FFT (Fast Fourier Transform) est de convertir du domaine temporel au domaine fréquentiel chacune des trames, de N valeur . La FFT est un algorithme rapide pour le calcul de la transformée de Fourier discret (DFT) où les valeurs obtenues sont appelées le spectre, cette algorithme est définie par la formule suivante[19, 5] :

$$x[k] = \sum_{n=0}^{N=1} x_a[n] \frac{-2j\pi}{N} kn,\ 0 \leq k \leq N-1$$

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

2.6.5 Filtrage sur l'échelle Mel

Dans un doublement de la fréquence, l'oreille perçoit moins d'une octave après 500Hz. L'échelle de Mel où le « Mel » est une unité représentative de la hauteur perçue d'un son. Le rôle de l'échelle Mel est qu'elle est assez proche des échelles issues d'études sur la perception sonore et sur les bandes passantes critiques de l'oreille . Tel que dans l'échelle de mesure Mel, la correspondance est approximativement linéaire sur les fréquences au-dessous de 1kHz et logarithmique sur les fréquences supérieures à celle-ci, définie par la formule suivante[19, 5] :

$$M = 2595 \log_{10}(1 + \frac{f}{700})$$

2.6.6 Extraction des coefficients

Les coefficients ceptraux sont les coefficients les plus utilisés qui représentent au mieux le signal de la parole, qui sont aussi appelés ceptres. Les deux méthodes les plus connus pour l'extraction du ceptres sont : la méthode d'analyse spectrale MFCC (Mel-Scale Frequency Ceptral Coefficients) et la méthode d'analyse spectrale LPCC (Linear Prediction Cepstral Coefficients). Comme le signal de la parole est transformé en une série de vecteurs calculés pour chaque trame. Il existe d'autres types de coefficients qui sont surtout utilisés dans des milieux bruités, nous citons par exemple les coefficients

PLP (Perceptual Linear Predective) ces coefficients permettent d'estimer les paramètres d'un filtre autorégressif en modélisant au mieux le spectre auditif [15, 5].

A. Coefficients cepstraux

Les coefficients cepstraux appelés Mel Frequency Cepstral Coefficients (MFCC) donnent de bons résultats dans leur utilisation en domaine de reconnaissance automatique de la parole. Les MFCC sont une représentation d'un signal de la parole pré-traité et regroupé en trames lissés par une opération fenêtrage (rectangulaire, triangulaire....) les résultats passent par une transformation FFT Mel, où l'échelle Mel redistribue les fréquences selon une échelle non linéaire qui simule la perception humaine des sons. les coefficients sont donnée par la formule suivante :

$$M_{els} = (\frac{1000}{\log 2})log(1 + \frac{f}{1000})$$

Dans ce qui suit, nous décrivons chacune des étapes nécessaires pour l'obtention d'un vecteur caractéristique tiré des coefficients MFCC, tel qu'il est illustré dans la Figure (2.8).

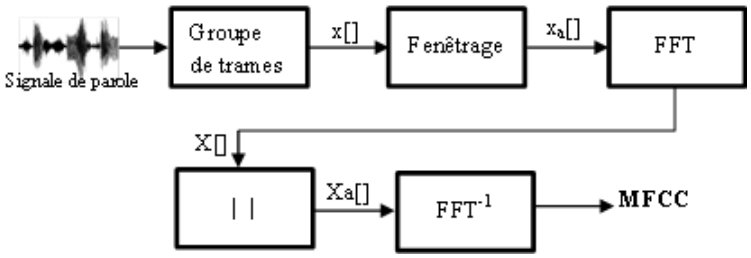

Figure 2.8: Exemples de calcul des coefficients cepstraux MFCC.

B. Coefficients de codage prédictif linéaire

Le codage prédictif linéaire appelé (Linear Prediction Coding, LPC) est un outil utilisé la plupart du temps dedans et pour représenter l'enveloppe spectrale. C'est l'une des plus puissante méthode de codage qui fournit des évaluations extrêmement précises des paramètres de la parole. Cette méthode revient à déterminer les coefficients basés sur l'hypothèse que la parole peut être modélisée par un processus linéaire tel que prédire le signal à un instant n à partir des p échantillons précédents (voir le formule ci-dessous). Le passage du signal de la parole par une phase pré-traitement détermine les groupes trames tel que chaque trame est transformé en une série de vecteurs calculés qui construisent une matrice d'auto-corrélation ou le calcul coefficients LPC se fait par une développement de Taylor d'où d'ailleurs ils sont appelés coefficients de prédiction (Linear Prediction Cepstral Coefficients (LPCC)) [15, 13].

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n)$$

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

C. Coefficients linéaire prédictif perceptuel

Le linéaire prédictif perceptuel appelé Perceptual Linear Prediction (PLP) permet d'estimer les paramètres d'un filtre auto-régressif en modélisant au mieux le spectre auditif. C'est une méthode inspirée du principe de prédiction linéaire. Elle est basée sur la combinaison entre le principe de prédiction linéaire et la représentation du signal qui suit l'échelle humaine de l'audition. Elle se base sur le spectre à court terme du signal de la parole. On utilise une fenêtre de longueur 10 à 30 ms pour chaque trame [15, 5].

2.7 Evaluation des méthodes de la segmentation

L'évaluation des performances d'une méthode de segmentation est une étape importante car elle permet d'une part de connaître les limitations de la méthode et donc de prévoir les améliorations adéquates. D'autre part, elle est un moyen facile pour comparer plusieurs méthodes de segmentation. Si on prend l'approche globale, on constate qu'elle est basée sur la décomposition de la parole continue la plupart du temps en utilisant un décodage acoustico-phonétique utilisé par des modules de niveau linguistique, l'unité de base est le mot ou la phrase. L'approche analytique est basé sur l'identification globale d'un mot ou une phrase en utilisant la notion de comparaison de l'identifient avec des références enregistrées, les composantes élémentaires sont plus courts comme les phonèmes et le syllabes. L'approche statistique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Ces niveaux sont représentés par des modèles de Markov cachés (HMM). Les unités acoustiques modélisées peuvent être des mots comme dans l'approche globale ou des unités plus courtes telles que le phonème tel qu'il est utilisé dans l'approche analytique, donc elle a combiné entre les deux approches.

Pour les classes des méthodes de segmentation avec ou sans contrainte linguistique on trouve que le méthode de segmentation avec contrainte linguistique est une méthode qui utilise l'information linguistique ou l'étiquetage pour apposer les frontières d'unités acoustiques comme exemple(Segmentation par DTW, Segmentation par HMM, Segmentation par réseaux de neurones, par contre dans le cas sans contrainte linguistique a pour objectif de générer une segmentation acoustique sans aucune connaissance a priori sur l'étiquetage du signal de parole. Elles conduisent donc à des insertions et des omissions des marques de segmentation. Certaines de ces méthodes sont plus adaptées à séparer certaines classes acoustico-phonétiques que d'autres. Parmi ces méthodes nous pouvons citer les suivantes (Détection de ruptures de stationnarité dans le temps, Segmentation fricatif/non-fricatif, Segmentation par ondelettes...). On général chaque méthode a ses avantages et ses inconvénients donc le choix se fait selon l'objectif voulu par la segmentation.

2.8 Conclusion

Dans ce chapitre nous avons vu les méthodes de segmentation de la parole les plus connues et ces classifications. Ensuite nous avons cités les techniques de paramétrisation et d'extraction de caractéristique du signal de la parole. Tel que le traitement automatique de la parole repose sur des données analogiques en fonction du temps. L'extraction des meilleurs paramètres aide, sans aucun doute, à ce traitement. L'intelligence artificielle peut intervenir pour trouver les paramètres pertinents ou utiliser n'importe quels représentants de la parole pour faire la segmentation ou la classification qui sera notre base de démarche pour expliquer notre méthode dans le chapitre suivant.

Chapitre 3

Conception et mise en oeuvre

3.1 Introduction

Les chapitres qui précèdent ont détaillé d'une part, un état de l'art sur la reconnaissance de parole et le traitement de signal avec ses caractéristiques, d'autre part, nous avons vue dans le deuxième chapitre les différentes approches de segmentation avec leurs classes les plus connus, la paramétrisation et l'extraction des caractéristiques du signal d'une segmentation automatique de la parole. Dans ce chapitre nous allons proposer la conception de notre système en commençant par la description de différentes étapes de la reconnaissance automatique de la parole, ensuite nous précisons la partie concerné par notre étude qui est la phase de segmentation et notamment la méthode proposé pour résoudre la problématique de la détermination d'un seuil automatiquement tout en exposant les algorithmes qui expliquent le processus réalisé. A la fin une conclusion de chapitre.

3.2 Algorithme proposé

La détection de présence d'une parole dans est une étape importante de segmentation dans les systèmes de RAP. Une fois le signal est enregistré dans un buffer dynamique on le fait découpé en trame de taille fixe qui regroupe la totalité brute du signal en question, l'élimination des trames qui expriment le silence dans le flux d'entrée (les trames de signal parole) permet de réduire efficacement le taux d'erreur du système. La plupart des techniques effectuent la classification de la parole sur la base des caractéristiques extraites de la trame en cours. Dans notre cas, la détection s'effectue en tenant compte de tous les trames qui représentent le signal parole. nous avons choisi les rapports d'énergies et le taux de passage par zéro pour déterminer les trames parole/non-parole et de supprimer les intervalles de silence en utilisant un seuil adaptatif, le seuil S est calculé pour chaque trame introduite en comparaison avec des caractéristiques de la totalité des trames de signal. Comme résultat de cet étape on obtient a la fin que les trames de parole et on ignore les trames non significatifs de silence pour avoir un taux de reconnaissance acceptable. Notre algorithme est composé par les étapes suivantes :

1. Segmentation du signal de la parole en entier (le signal du locuteur) en trames de taille fixe avec fenêtre rectangulaire et sans chevauchement.
2. Calcule de l'énergie (E[m]) de chaque trame :
3. Calcule de taux de passage par zéro de chaque trame(tppz[m]).
4. Calcule de rapport de l'énergie de chaque trame (rap E[m]).

5. Calcule de moyen de rapport d'énergie mrap(E[m]) et l'écart d entre le max rap(E) et le min rap(E).
6. Calcule de seuil S.
7. Traitement des trames de parole selon le seuil S pour définir les blocs parole/non-parole

3.3 Mise en œuvre du système

suite aux précédents travaux réalises notamment l'application " de proposition d'un modèle de descripteur structurel pour la voix arabe application saisie des notes " [19], le taux de reconnaissance obtenu est limité pour test précis et conditionné par un environnement d'acquisition et matériel défini car le choix de seuil est fixé, l'objectif de notre travail se concrétise à réaliser une méthode permettant de déterminer un seuil adaptatif qui se change avec chaque signal de parole pris dans les conditions suivantes : le locuteur prés ou loin de la source d'enregistrement l'environnement peu bruité ou calme ; la voie d'une femme, homme ou(enfant) puisque chacun entre eux a sa propre fréquence fondamentale ; les différentes support d'enregistrement (hétérogénéité matérielle). Cette technique est par la suite valable pour développer une application de reconnaissance de la parole arabe. Elle permet de séparer la parole de le silence et le bruit. Pour ce faire, nous avons enregistré des différentes lettres, mots et phrases arabe par des différentes personnes homme ,femmes et enfants où chaque séquence passe par une succession des étapes : acquisition,prés-traitement et segmentation pour quelle soit prête a exploiter dans une application quelconque de la RAP.

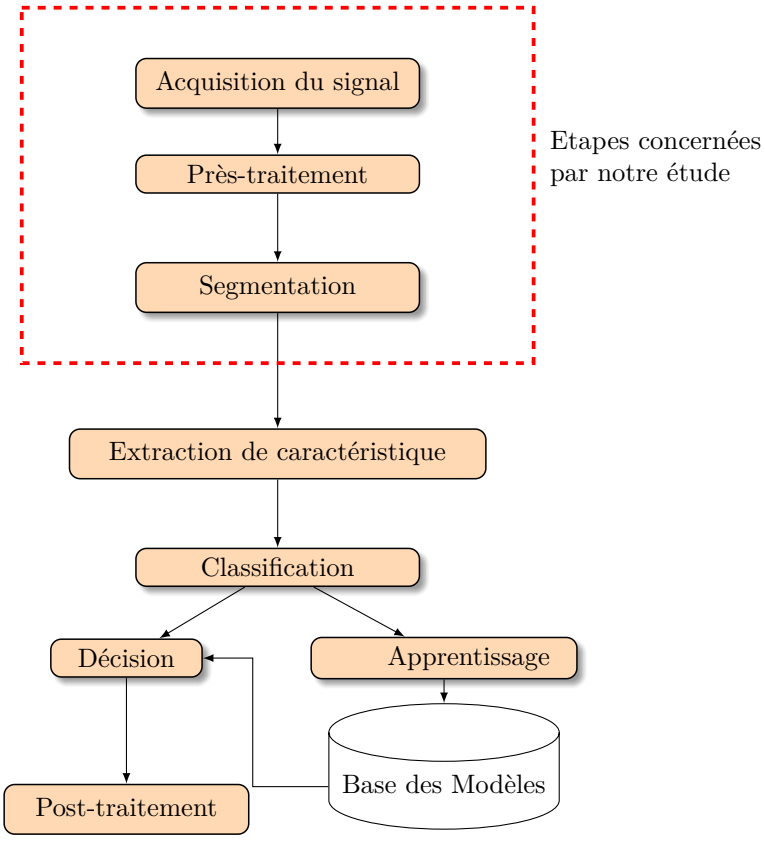

Figure 3.1: Schéma général d'un SRAP.

3.4 Description des étapes de système

Pour bien comprendre notre travail, il est évident de décrire l'ensemble des étapes réalises au cours de cette application en détaillant chaque phase.

3.4.1 Acquisition

La première phase de chaque processus de reconnaissance vocale est l'acquisition de parole. Ce module permet d'introduire le signal de parole à un micro-ordinateur via un microphone ainsi que la carte son comme périphérique interne afin d'enregistrer et exploiter ce signal pour qu'il soit destiné à une tâche précise. Dans notre application nous allons utiliser le signal acquit est échantillonné, quantifié et codé à l'aide de la carte son de l'ordinateur.

3.4.2 Pré-traitement

cette phase consiste à utiliser des filtres pour améliorer la qualités du signal acquis pour qu'il soit prêt à la phase suivante. Ces filtres permettant de diminuer les effets indésirables influant sur la qualité du signal de la parole prévenant des outils d'acquisition.

3.4.3 Segmentation

La phase de segmentation est la partie la plus importante dans le procédé de reconnaissance, une segmentation correcte et précise du signale donne une bonne description et classification qui résulte un taux acceptable de reconnaissance. la méthode de segmentation proposé permet d'extraire la parole de signal. Pour pouvoir réaliser cette tâche, le système passe par les étapes suivantes (voir la figure 3.3) : d'après le schéma ci-dessous, on explique chaque étape énoncé :

1. Dés qu'on obtient le signal déjà enregistré dans un tableau dynamique qui se détermine selon la longueur d'enregistrement et qui nous permet de manipuler les données facilement, on doit d'abord segmenter ce signal de la parole en entier (le signal du locuteur) autrement dit des trames de taille fixe avec une fenêtre rectangulaire et sans chevauchement ce qui permet de parcourir par la suite l'entière des trames .

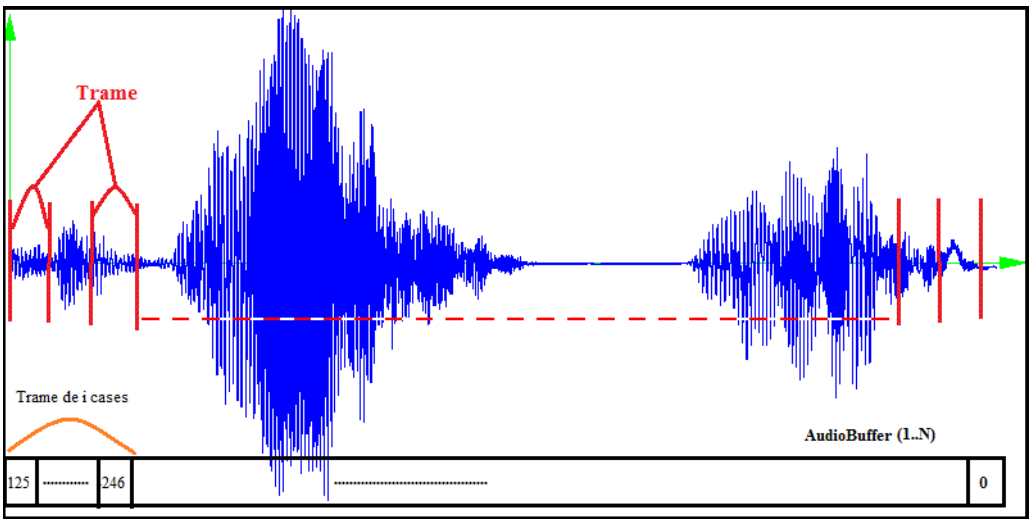

Figure 3.2: Division de l'audio-buffer en trames.

la taille de La décomposition du signal de parole en trames de taille définit par la formule suivante :

$$Fen = entier(FS * 15/1000);$$

Ou FS : La fréquence du signal utilisé (dans notre méthode est 44100 hz) et la durée d'échantillonnage est de (15 ms).

2. L'étape suivante est le calcul de l'énergie (E[m]) de chaque trame qui sert à la somme des amplitudes de cette trame, on déduit la moyenne de ladite trame par la formule suivante :

$$E[m] = \sum_{i=1}^{N_x} (x - MoyT)^2 \tag{3.1}$$

dont N : nombre d'amplitudes x de trame, i :[0..N]
pour calculer l'énergie de la trame, on commence d'abord par le calcul de la moyenne de l'amplitude de la trame par la fonction $MoyT$ définit par l'algorithme suivant :

Algorithme 3.1: Fonction de calcul de la moyenne de l'amplitude de la trame

Fonction *MoyT (Tab, D, F : entier)*: *Reel*
  $size, i, moy$: *Entier*;
  sum : *Reel*;
Début
  $size = F - D$;
  $sum = 0$;
  Pour $i = D$ **Jusqu'à** F **Faire**
    $sum = sum + Tab[i]$;
  FinPour
  Si $((size > 0)$ **Alors**
    $moy = sum/size$
  FinSi
  Sinon
  $moy = 0$;
  retourner $MoyT = moy$;
Fin

L'algorithme suivant permet de calculer l'énergie de la trame en cours de traitement :

Algorithme 3.2: Fonction de calcul de l'energie de la trame

Fonction *ENRG (Tab, D, F : entier)*: *Reel*
  E, moy : *Reel*;
Début
  $E = 0$;
  $moy = MoyT(tab, D, F)$;
  Pour $i = D$ **Jusqu'à** F **Faire**
    $E = E + (Tab[i] - moy)^2$;
  FinPour
  retourner $ENERG = E$;
Fin

3. Calcul du taux de passage par zéro de chaque trame(tppz[m]) qui détermine le rapport entre le nombre de passage du signal par zéro dans chaque trame sur la taille de la fenêtre déjà fixée au début, le taux est obtenu par la formule suivante :

$$TPPZ[m] = \frac{NPPZ}{FEN}.$$

dont ($NPPZ$:nombre de passage par zéro dans la trame et FEN : taille de la fenêtre)

Algorithme 3.3: Fonction de calcul de nombre de passage par zéro

```
Fonction NPPZ(Tab, D, F : entier): Entier
  | npz, i : Entier;
Début
  | npz = 0;
  | Pour i = D Jusqu'à F Faire
  |   | Si ((tab[i] * tab[i + 1]) < 0) Alors
  |   |   | npz = npz + 1
  |   | FinSi
  |   | Sinon
  |   | Si ((tab[i] * tab[i + 1]) = 0) Alors
  |   |   | npz = npz + 1;
  |   | FinSi
  |   |
  | FinPour
  | retourner NPPZ = npz;
Fin
```

4. Il est nécessaire de calculer le rapport d'énergie de chaque trame (rap E[m]) qui est un facteur déterminant pour le seuil, il s'agit d'un rapport de l'énergie de la trame sur le taux de passage par zéro de la même trame :

$$rapE[m] = E[m]/tppz$$

5. Une fois que le rapport d'énergie de chaque trame est calculé, on a besoin d'avoir le moyen des rapports d'énergie de toutes les trames $mrap(E[m])$, ce dernier est calculé par l'équation ci-dessous :

$$mrapE[m] = \frac{1}{N_x} \sum_{i=1}^{N_x} rapE[m]$$

6. Dans cette étape on calcule d'abord l'écart (δ) entre le maximum $maxrap(E)$ et le minimum $minrap(E)$ des rapports d'énergie pour qu'il soit un paramétré important dans le calcule de seuil :

$$\delta = maxrap(E) - minrap(E)$$

Ensuite on calcule le seuil S qui va être un seuil adaptatif pour chaque signal traité indépendamment, le seuil S est la somme de le moyen de rapport d'énergies et le coefficient α multiplié par l'écart δ :

$$S = mrap(E) + \alpha * (\delta) \quad dont(\alpha = 0.04)$$

L'algorithme suivant calcule le seuil S qui permet de déterminer les blocs parole et les blocs non parole :

Algorithme 3.4: Fonction de calcul du seuil S du signal de la parole

Fonction *SEUIL (Tab, Fen, NbFen : entier)*: *Reel*
  npz, pos : $Entier$;
  $S, E, rap, sumpar, rapmax, rapmin$: $Reel$;
Début
  $E = 0$;
  $rap = 0$;
  $sumrap = 0$;
  $rapmax = 0$;
  $rapmin = 500000$;
  Pour $i = 0$ **Jusqu'à** $NbFen$ **Faire**
    $pos = i * Fen$;
    $npz = NPPZ(tab, pos, pos + Fen)$;
    $E = ENERG(tab, pos, pos + Fen)$;
    $rap = \log 10(E/((npz/Fen) + 10^{-10})$;
    $sumrap = sumrap + rap$;
    Si $((rap > rapmax)$ **Alors**
      $rapmax = rap$
    FinSi
    Sinon
    Si $((rap < rapmin)$ **Alors**
      $rapmin = rap$
    FinSi

  FinPour
  $S = (sumrap/NbFen) + (0.04 * (rapmax - rapmin))$;
  retourner $SEUIL = S$;
Fin

7. On compare le rapport d'énergie de chaque trame avec le seuil S, si le rapport d'énergie de la trame traité $rap[E]$ est supérieur au Seuil S alors elle est considéré comme une trame de parole, si le nombre des trames de parole devient supérieur à 5 et suivi par une succession de 3 trames de silence, alors c'est un bloc de parole sinon est un bloc de silence.

L'algorithme principal 3.5 illustre en détail cette étape :

Algorithme 3.5: Algorithme d'extraction des segments de la parole

Fonction *SEG (audioBuffer[], hs : reel)*: *Entier*
  $Seuil, seg, size, Fen, NbFen, npz, pos, dp, tp, ss : Entier$;
  $E, rap, sumrap, rapmax, rapmin : Reel$;
Début
  $size = audioBuffer.taille$;
  $Fen = entier(44100 * 15/1000)$; `/* taille du trame traité où FS:44100 c'est la Fréquence de Signal et 15ms la duré du trame */`
  $NbFen = entier(size - Fen/Fen)$;
  $npz = 0$; $pos = 0$;
  $seg = 0$; `/* initialisation de compteur de nombre des segments paroles */`
  $dp = 0$; `/* indicateur de debut de trame parole */`
  $tp = 0$; `/* compteur de nombre des trames parole dans le bloc traité */`
  $ss = 3$; `/* le nombre des trames succissives en silence détérminant la fin du bloc parole */`
  $Seuil = SEUIL(audioBuffer, Fen, NbFen)$;
  Pour $i = 0$ **Jusqu'à** $NbFen$ **Faire**
    $pos = i * Fen$;
    $npz = NPPZ(tab, pos, pos + Fen)$;
    $E = ENERG(tab, pos, pos + Fen)$;
    $rap = \log 10(E/((npz/Fen) + 10^{-10})$;
    Si $((rap > Seuil) et (ss = 3)$ **Alors**
      $ss = 0$;
      $dp = pos$;
      $tp = tp + 1$;
    FinSi
    Sinon
    Si $((rap > seuil) et (ss = 1)) ou ((rap > seuil) et (ss = 2))$ **Alors**
      $ss = 0$
    FinSi
    Sinon
    Si $((rap <= Seuil) et (ss < 3)$ **Alors**
      $ss = ss + 1$;
      Si $((ss = 3) et (tp > 5)) ou ((i = NbFen) et (tp > 5))$ **Alors**
        $tp = 0$;
        $Dessinerligne(((dp) * hs) + 5, 50, ((dp) * hs) + 5, 350)$; `/* debut de segment de parole */`
        $Dessinerligne(((i * Fen) * hs) + 5, 50, ((i * Fen) * hs) + 5, 350)$; `/* fin de segment de parole */`
        $seg = seg + 1$;
      FinSi
      Sinon
      Si $(ss = 3)$ **Alors**
        $tp = 0$
      FinSi
      Sinon
      Si $((rap > Seuil) et (npz >= 10)$ **Alors**
        $tp = tp + 1$
      FinSi

    FinSi

  FinPour
  retourner $SEG = seg$;
Fin

Nous résumons notre algorithme de la segmentation parole / non parole par la figure 3.3 :

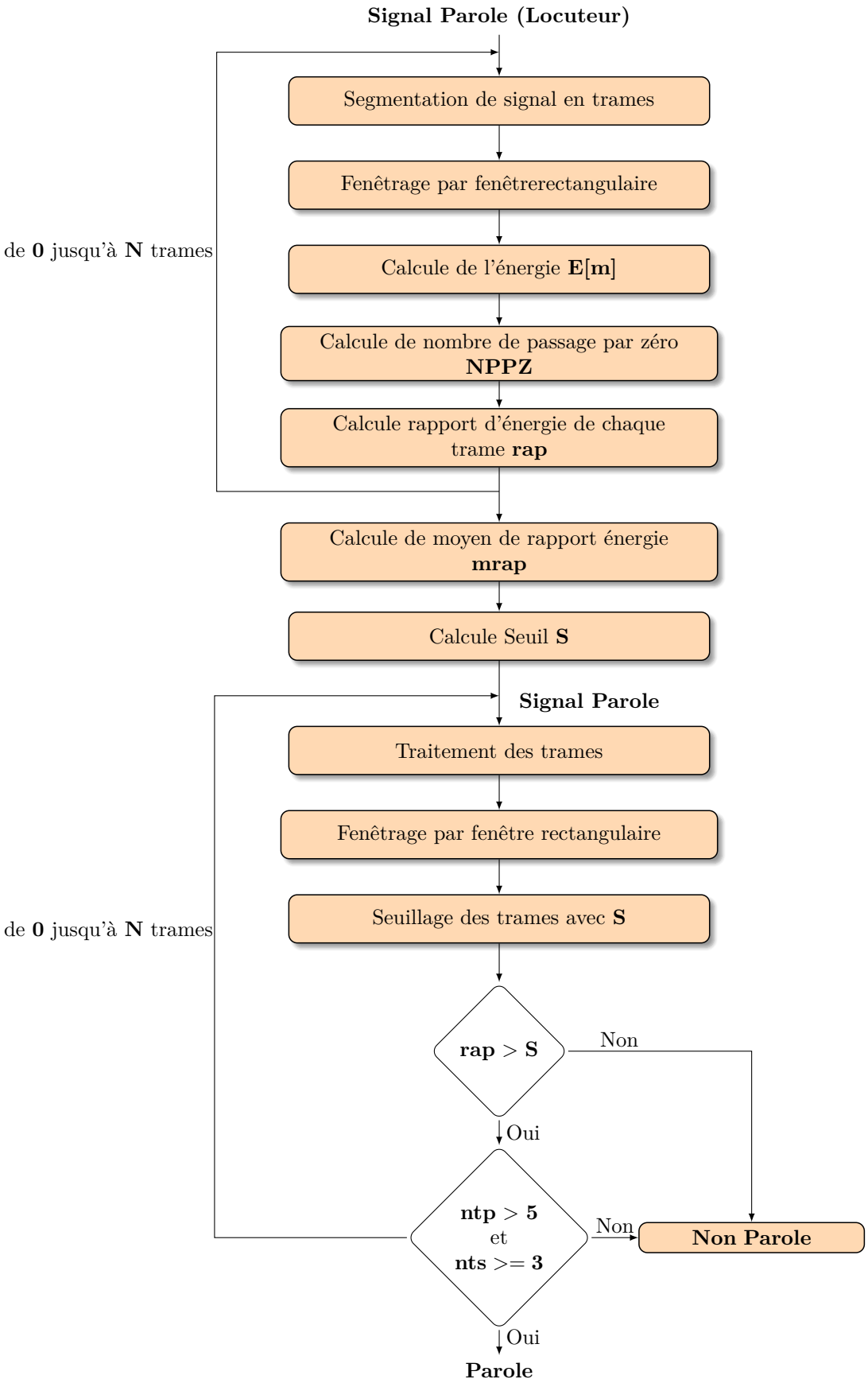

Figure 3.3: Algorithme général de détection parole/non-parole.

3.4.4 Extraction des caractéristiques

La phase d'extraction des caractéristiques a pour but de retenir les caractéristiques spécifiques des segments générés par la phase de segmentation, cette tâche est effectuée par le biais d'une des techniques d'analyse (statique, structurelle, hybride,...). Ces caractéristiques décrivent bien les segments de parole. Dans notre système, nous allons utiliser le modèle du descripteur structurel développé dans l'application " Proposition d'un modèle de descripteur structurel pour la voix arabe, Application saisie des notes" [19] pour extraire les caractéristiques structurelles d'un signal vocal.

A. Normalisation de signal vocal

La phase de segmentation génère des segments non homogènes qui se caractérisent par une variabilité dans la longueur et l'amplitude. Pour réduire cette hétérogénéité nous avons proposé une méthode de normalisation qui abaisse cette variabilité. Pour arriver à notre objectif, on a précisé une taille fixe (H=3000, L=30000) [19]. La figure 3.4 montre un exemple de normalisation.

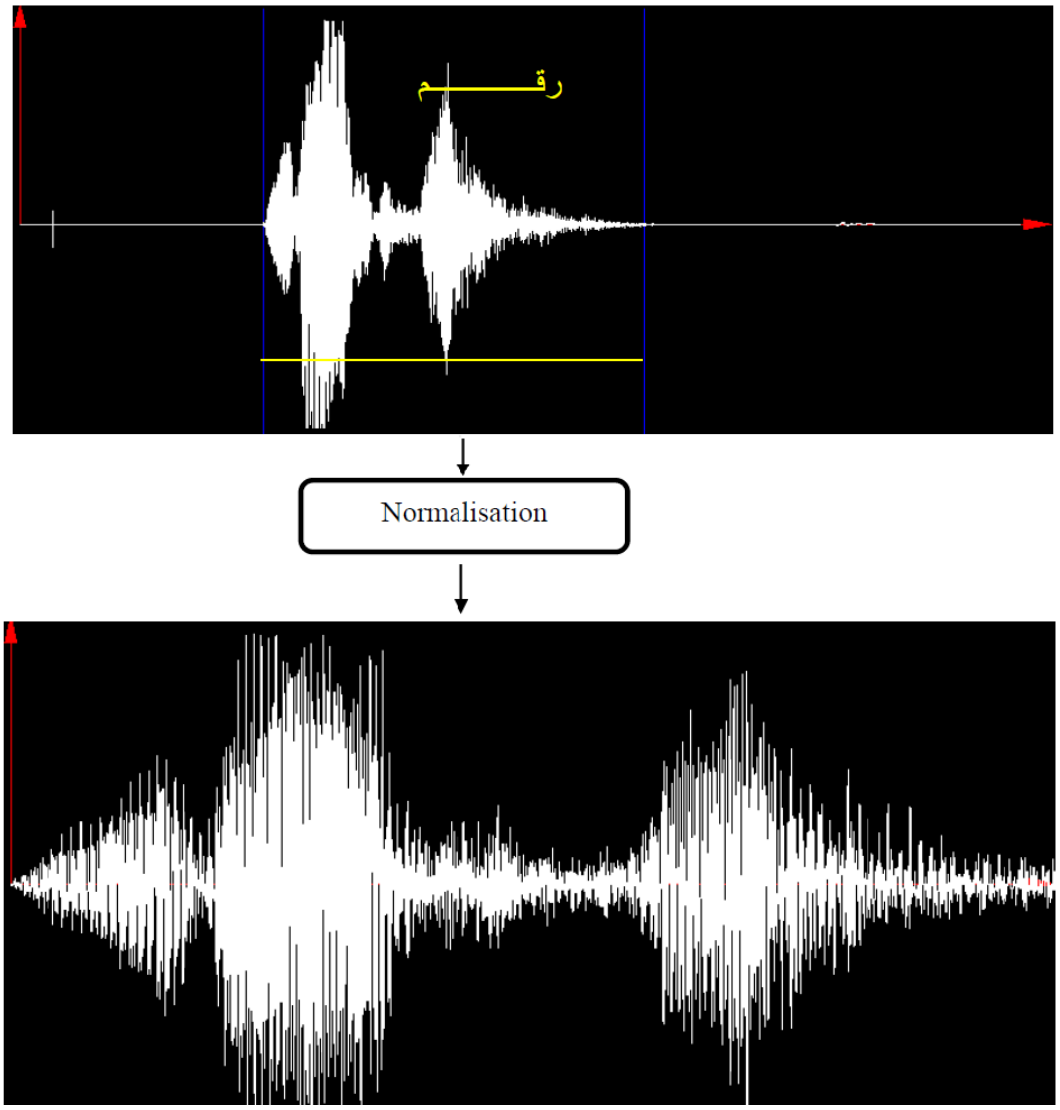

Figure 3.4: Exemple de normalisation du mot (رقم) [19].

B. Méthode d'extraction des caractéristiques

Pour retirer les vecteurs de caractéristiques des segments après leur normalisation par l'étape précédente nous avons employé une méthode structurelle. Le résultat de cette étape est des vecteurs de taille fixe égale à L/k. La description structurelle est obtenu par la division de la longueur d'une segment sur k (dans notre cas k=150), ensuite on calcule le moyen de chaque partie pour qui nous donne une taille L/k=200 [19].

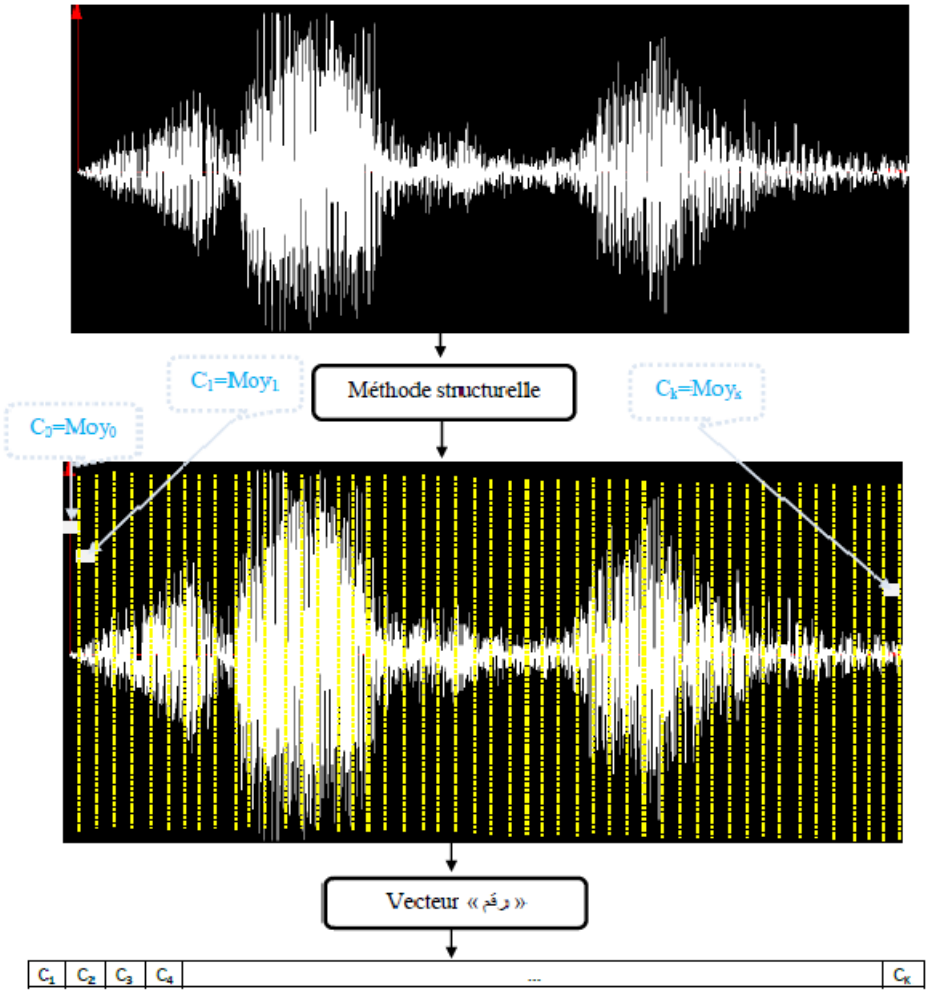

Figure 3.5: Exemple d'extraction des caractéristiques pour le mot (**رقم**) [19].

C. Classification

Pour vérifier la tâche de la phase de classification on utilise la méthode de classification, (FLC), cette dernière contient deux étapes : Apprentissage et Test (Décision). L'étape d'apprentissage sert à initialiser la base des modèles (dictionnaire), et le test consiste à affecter une classe pour chaque nouveau exemple donné (vecteur de modèle).le schéma suivant explique les étapes de cette phase [19].

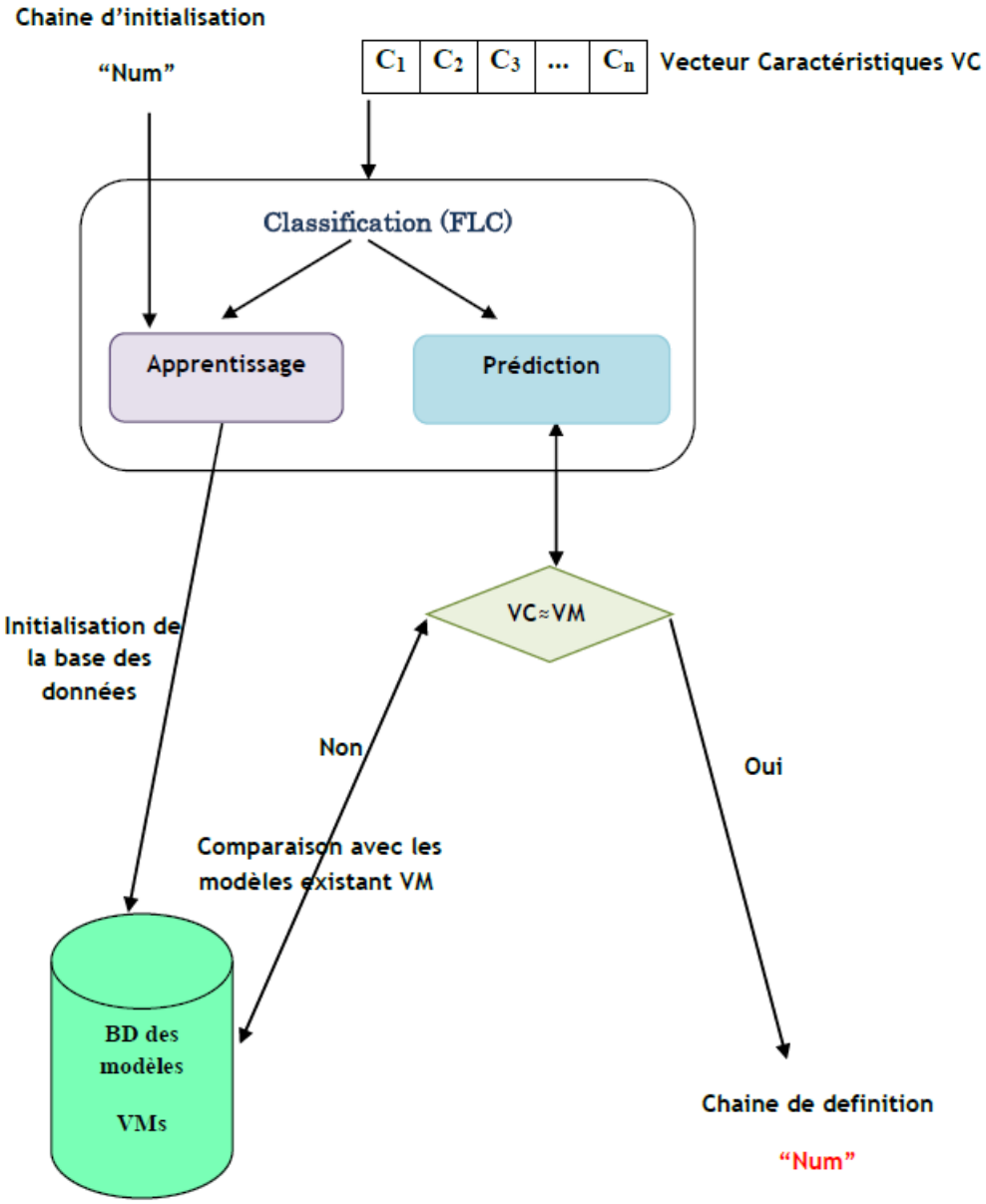

Figure 3.6: Les differentes étapes de la phase de classification [19].

3.5 Conclusion

Nous avons présenté dans ce chapitre l'architecture globale de la méthode de segmentation proposé qui peut conduire à une conception convenable d'un système de reconnaissance de la parole, ainsi que les différentes étapes en détails dont on a attribuer à chaque phase l'algorithme correspondant. Finalement, on obtient un signal segmenté d'une façon plus au moins correcte qui permet par la suite, d'améliorer les résultats des phases ultérieures (extraction de caractéristiques, classification, apprentissage et poste-traitement).Dans le prochain chapitre, nous allons mettre en évidence notre méthode pour montrer les résultats obtenus.

Chapitre 4

Résultats et discussion

4.1 Introduction

Après avoir achever la phase de conception du système proposé tout en détaillant les différentes étapes suivis et leurs algorithmes réalisés de chacune, il est évident de démonter la concrétisation du travail. Pour cela nous allons consacré ce chapitre à décrire l'environnement de la programmation, l'interface globale du système avec les fonctionnalités offertes. Ensuite nous allons exposer les résultats obtenus avec une comparaison des différents types d'enregistrements effectués.

4.2 Choix du langage de programmation

En matière de programmation, nous avons choisi le langage JAVA comme un environnement de mise en application de la méthode proposé, ce dernier possède une richesse de bibliothèques audio et offre une grande simplicité de manipulation de son et d'images. Il permet d'acquérir ou de traiter des fichiers audio avec aisance. Ce langage peut fournir les avantages suivants :

1. La portabilité des logiciels ;
2. L'exploitation des classes déjà développées ;
3. La possibilité d'ajouter à l'environnement de base des composants fournis par l'environnement lui-même ;
4. La quasi-totalité de contrôle de Windows (boutons, boites de saisies, listes déroulantes, menus...) qui sont représentés par classes.

4.3 Interface de système et fonctionnalité

Dans cette section, nous allons voir l'interface globale du système et les différentes étapes à suivre pour appliquer les traitements essentiels sur le signal de la parole.

4.3.1 Interface générale

Lorsqu'on lance cette application, nous aurons l'interface suivante (figure 4.1), dans cette étape l'utilisateur du système à la possibilité soit de charger un fichier déjà enregistré sur l'ordinateur ou sur un support de stockage, soit d'enregistrer directement une parole via le microphone pc :

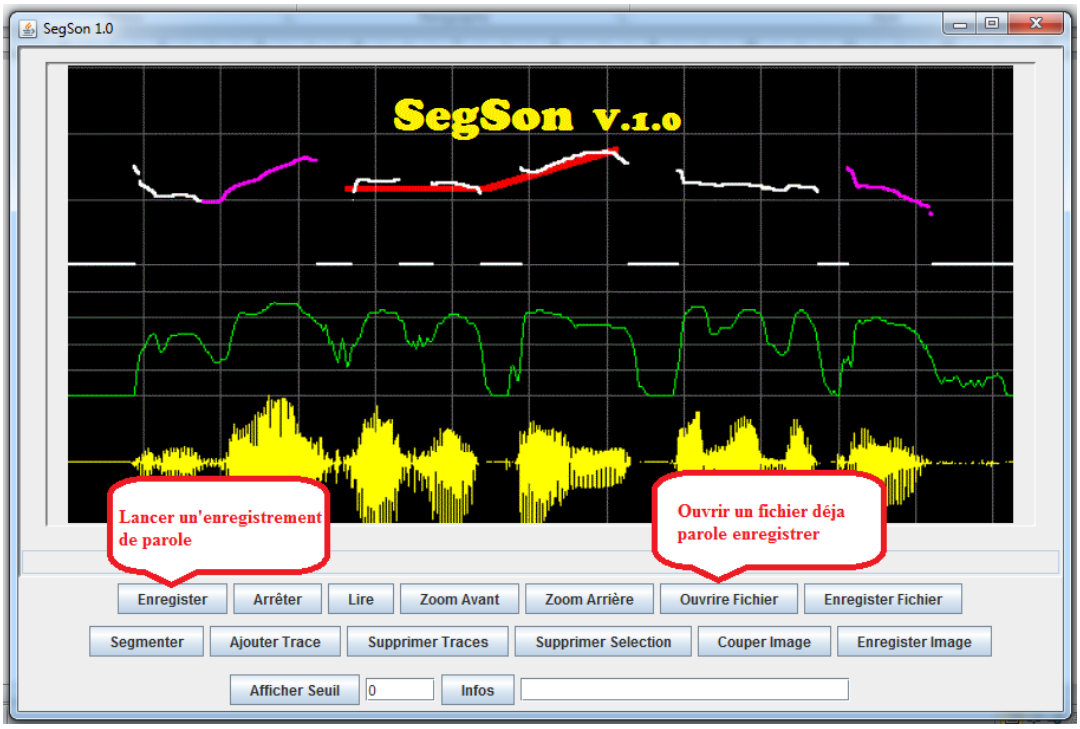

Figure 4.1: Lancement de l'application proposé.

Pour effectuer un enregistrement de la parole et de la segmenter, on passe par les étapes suivante (figure 4.2) :

1. L'utilisateur clique sur le bouton " enregistrer" et commence sa parole ;
2. Lorsqu'il veut compléter sa parole, il clique sur le bouton "terminer", l'audiogramme s'affiche sur la console du système ;
3. Pour afficher les segments de parole , l'utilisateur clique sur le bouton "Segmenter".

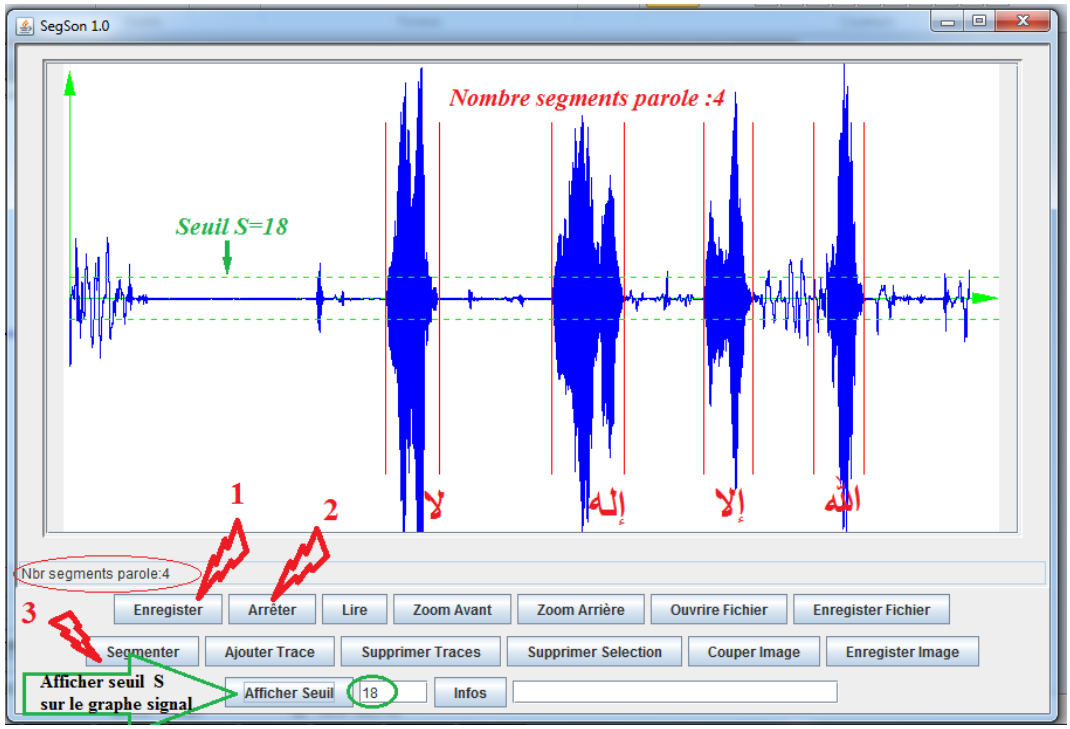

Figure 4.2: Étapes de segmentation de la parole, exemple de la phrase (لا اله الا الله).

On a appliqué une segmentation de la phrase (العلم نور والجهل ظلام) qui était prononcé par différentes personnes, les résultats obtenus sont présentés par la figure ci-dessous :

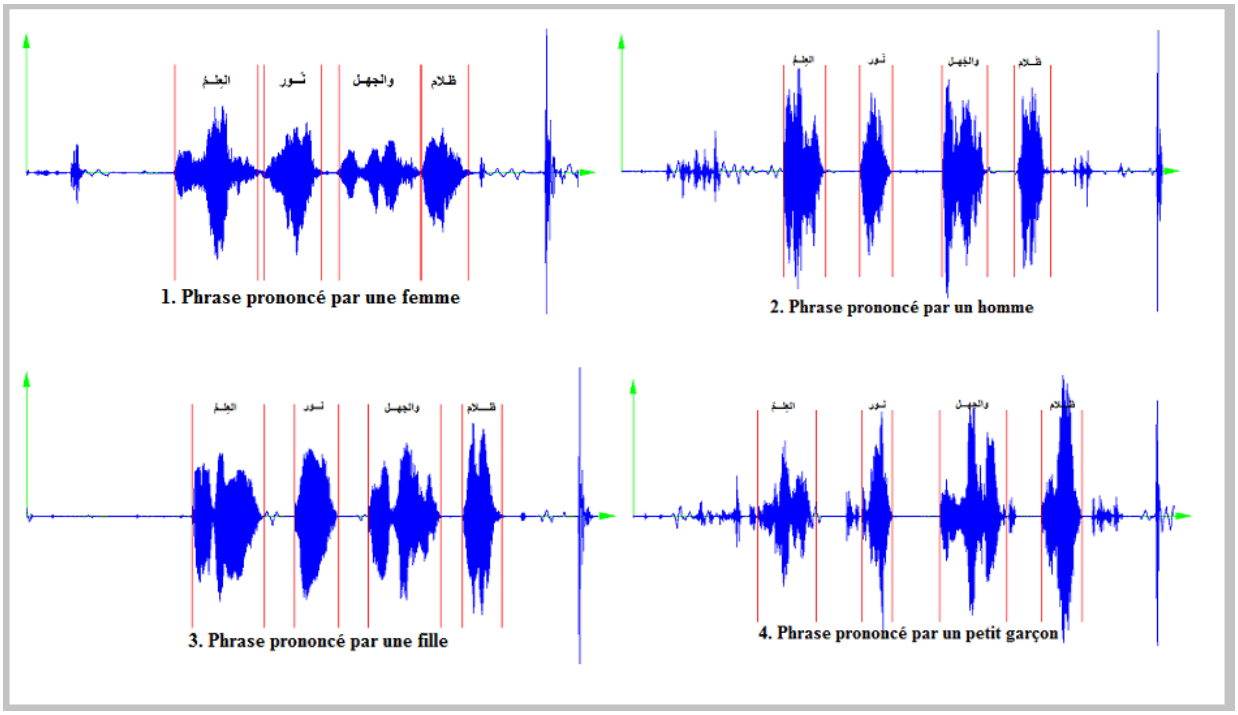

Figure 4.3: Exemple de segmentation de la phrase (العلم نور والجهل ظلام).

4.3.2 Test et bilan

Afin d'estimer les performances de notre méthode, nous l'avons appliquée sur un ensemble de test dont on a pris premièrement un seul mot (تقسيم) prononcé plusieurs fois par quatre personnes de différente âge et sexe. On a évalué le degré de segmentation correcte de chaque énoncé. Enfin, on a calculé le taux de segmentation (TS) correcte de chaque personne. La figure 4.4 montre un exemple de cette opération.

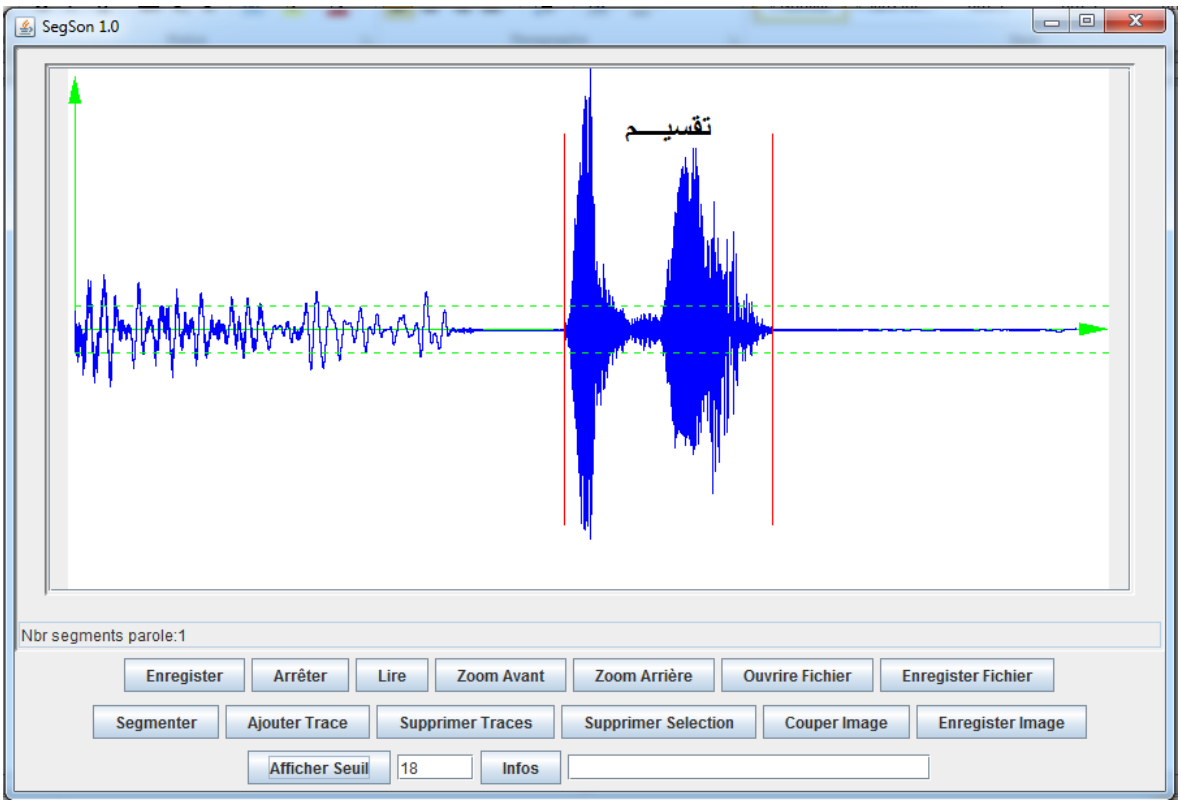

Figure 4.4: Segmentation du mot (تقسيم).

Le tableau 4.1 présente les résultats d'évaluation de chaque locuteur du test :

Tableau 4.1: Tableau comparatif du segmentation d'un mot (تقسيم).

Nbr test / personne	1	2	3	4	5	6	7	8	9	10	TS
Homme âgé	8	9	8.5	10	10	9	10	9.5	7	9.5	90.5%
Homme adulte	9	10	10	9.5	10	9	9.5	10	8	10	95%
Femme âgée	7	8	9.5	10	10	9	9	9.5	10	9	91 %
Fille	8	10	9.5	9	10	10	9.5	9	9.5	10	94.5%
Petit garçon	8	9	7.5	10	9.5	10	10	9.5	9	10	92.5%

Après à voir obtenu les résultats de la segmentation d'un mot, nous avons appliqué la même opération sur la phrase (تقسيم الكلام العربي) (voir la figure 4.5).

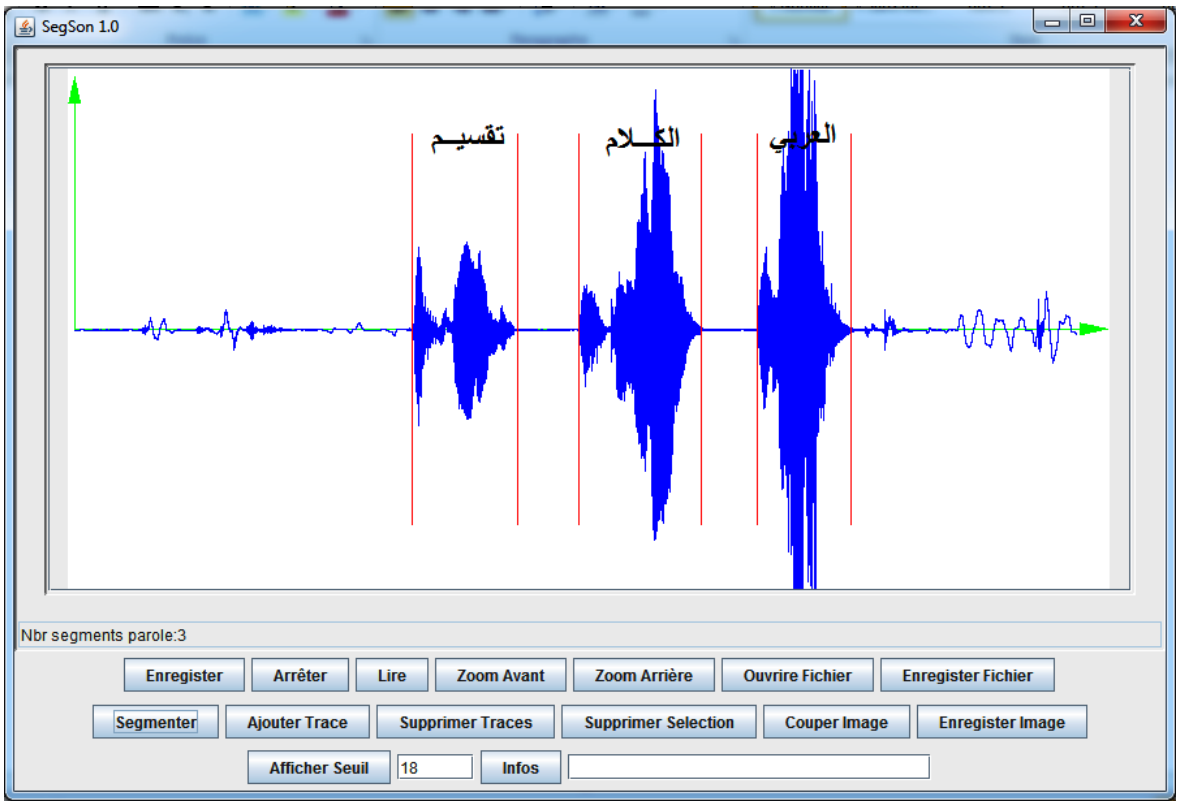

Figure 4.5: Segmentation de la phrase (تقسيم الكلام العربي).

les résultats d'évaluation de la segmentation de la phrase sont représentés par le tableau 4.2 suivant :

Tableau 4.2: Tableau comparatif du segmentation d'une phrase (تقسيم الكلام العربي).

Nbr test / personne	1	2	3	4	5	6	7	8	9	10	TS
Homme âgé	9	9.5	10	9	10	9	9	8	10	9.5	93%
Homme adulte	9.5	10	9	10	10	9	9.5	9.5	9	10	95.5%
Femme âgée	9	8	9	10	9	10	9	9.5	10	9	92.5%
Fille	8	9.5	10	9	10	9.5	10	9	10	10	95%
Petit garçon	9	9	9.5	9.5	10	9	10	9.5	9	9	93.5%

Nous avons remarqué que les résultats des tests présentés ci-dessus ont un taux de la segmentation dans le cas ou le locuteur est un adulte est plus élevé par rapport aux autres catégories, parce-qu'il prononce les mots plus clairement. De plus, d'après la figure 4.3 nous remarquons que les formes des mots sont très semblables pour toutes les catégories mais avec une différence remarquable dans les niveaux des amplitudes des mots entre les sexes (masculin et féminin). La même chose est constaté pour l'existence de bruit dans le signal.

4.4 Comparaison des résultats

Pour évaluer la performance de notre méthode nous avons fait une comparaison avec la méthode structurelle ([19]). Nous avons pris des phrases comme des échantillons pour la comparaison . Dans cette évaluation nous avons pris les critères de comparaisons suivantes :

1. Âge de locuteur ;
2. Milieu d'enregistrement (calme ou peu bruité) ;
3. Matériel d'exécution (PC1 :Laptop ASUS X54C , PC2 :Laptop MSI).

Le tableau suivant représente les résultats obtenus dans cette comparaison :

Tableau 4.3: Résultats de comparaison entre les deux méthodes

	Critères de comparaison	Taux de la segmentation	
		Méthode proposé	Méthode structurelle
Âge de locuteur	Homme adulte	95 %	91 %
	Femme âgée	91 %	82.5 %
	Petit garçon	93 %	68 %
Milieu	Calme	95.5 %	90 %
	Peu bruité	91 %	41 %
Matériel	PC1	95 %	90 %
	PC2	93.5 %	58.5 %

D'après les résultats représentés dans le tableau ci-dessus on constate la robustesse de notre méthode par rapport la méthode structurelle dont la performance de la segmentation est toujours élevé même avec le changement des critères d'évaluation par contre dans la méthode structurelle on remarque la baisse de taux de la segmentation avec chaque chargement de critères.

4.5 Conclusion

Dans ce chapitre, nous avons vu les critères sur lesquels nous avons choisit le langage de programmation, l'interface générale et les différentes fonctionnalités du système, ainsi que les résultats obtenus avec une discussion pour évaluer la performance de la segmentation dans différentes conditions d'enregistrements.

Conclusion

Dans cette étude, nous avons abordé le problème de séparation parole/non parole dans un signal de la parole arabe. Plusieurs méthodes ont été citées dans la littérature pour la segmentation du signal, mais aucune d'entre elles ne se démarque des autres, surtout pour les signaux des différentes sources (locuteur/matériel). Nous avons présenté par la suite une vue générale sur la langue arabe et les différents méthodes de la segmentation ainsi l'extraction des caractéristiques d'un signal de la parole.

Suite à la difficulté rencontrée dans la phase de la segmentation du signal de la parole engendrée par la variabilité des voies des locuteurs et les différents types des outils d'acquisition du signal, il est difficile de déterminer les limites des segments de la parole. De ce fait nous avons proposé une méthode de segmentation adaptative de la parole arabe. Il s'agit de décomposer le signal en des trames de taille fixe pour calculer l'énergie de chaque trame et les rapports d'énergies moyen, minimum et maximum de tous les trames de signal. De là, nous avons déterminé un seuil global pour extraire les segments de parole dans le signal.

Pour valider notre approche, nous avons effectué des tests avec différents locuteurs selon le sexe et la tranche d'âge sur des mots et des phases. Les résultats obtenues sont encourageants et montre la robustesse de notre méthode.

Comme perspective, on suggère l'amélioration de notre méthode pour qu'elle soit capable de segmenter le signal en phonèmes et syllabes.

Bibliographie

[1] R. AJGOU – « Reconnaissance automatique du locuteur à travers les canaux digitaux », Thèse, Université Mohamed Khider Biskra, 2016.

[2] J. Allegre – « Approche de la reconnaissance automatique de la parole », Tech. report, Conservatoire National Des Arts Et Métiers Centre Régional Languedoc-Roussillon, avril 2003.

[3] Y. Aziza – *Modélisation ar et arma de la parole pour une vérification robuste du locuteur dans un milieu bruité en mode dépendant du texte*, Mémoire, Université Ferhat Abbas de Setif 1, 2013.

[4] F. Bahja – « Détection du fondamental de la parole en temps réel : application aux voix pathologiques », Thèse, Université Mohammed V-Agdal Rabat.

[5] R. Benammar – « Traitement automatique de la parole arabe par les hmms : Calculatrice vocale », Thèse, Université Abou Bekr Belkaid Telemcen, 2012.

[6] A. Bendahmane – *Cours de traitement automatique de la parole*, vol. 1, Université Des Sciences Et De La Technologie D'Oran Mohamed BOUDIAF 1ière Année Master RFIA, 2014.

[7] K. Berbache – *Modèles de markov cachés : Application a la reconnaissance automatique de la parole*, Mémoire, Université Mouloud Mammeri de Tizi Ouzzou, 2014.

[8] S. E. Berchaoua – *Reconnaissance de la parole arabe par les supports vecteurs machines (svm)*, Mémoire, Université Hama Lakhdar D'el-Oued, 2013.

[9] O. Douib – *Reconnaissance automatique de la parole arabe par cmu sphinx 4*, Mémoire, Université Ferhat Abbas de Sétif 1 – UFAS (Algérie), 2013.

[10] T. Dutoit – *Introduction au traitement automatique de la parole*, vol. 1, Université de Nantes, 2000.

[11] J.-P. Haton – « La reconnaissance automatique de la parole », Tech. report, Université Henri Poincaré, Nancy 1, Janvier 2005.

[12] B. Jacoub – *Reconnaissance automatique de la parole présentation du module de décodage acoustique*, vol. 1, Université Paul Sabatier de Toulouse III, n° d'ordre 2127, 2010.

[13] S. Jarifi – « Segmentation automatique de corpus de parole continu dédie à la synthèse vocale », Thèse, Ecole nationale supérieure des télécommunications de Bretagne, 2007.

[14] F. Lamare – *Segmentation non supervisée d'un flux de parole en syllabes*, Mémoire, Institut de Recherche et Coordination Acoustique/Musique IRCAM, Paris, 2012.

[15] L. Lazli-Boukhalfa – « Système neuro-markovien basé sur la fusion de données floues et génétiques : Application pour la reconnaissance automatique de la parole », Thèse, Université Badji Mokhtar de Annaba, 2006.

[16] N. eddine Mesbahi – *Conception et réalisation d'un système de pilotage d'un véhicule par commande vocale*, Mémoire, Université Mohammed Khider de Biskra, 2011.

[17] S. Nefti – « Segmentation automatique de parole en phones. correction d'étiquetage par l'introduction de mesures de confiance », Thèse, Université de Rennes 1, 2004.

[18] Netophonix – « Forum wiki découverte », http ://wiki.netophonix.com/Cartes_son, juillet 2010, consulté le 10 avril 2016.

[19] T. Settou et M. OumElhana – *Proposition d'un modèle de descripteur structurel pour la voix arabe, application saisie des notes*, Mémoire, Université Hama Lakhdar D'el-Oued, 2015.