RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole **N**ationale **P**olytechnique

Département d'Electronique

Projet de Fin d'Etudes

Pour l'obtention du diplôme
D'Ingénieur d'Etat en Electronique

Thème

Analyse sonagraphique des consonnes fricatives [s] et [ŝ] et leurs opposées [z] et [ž] en vue de la RAP en Arabe Standard

Proposé et dirigé par :

Dr M. GUERTI

Etudié par :

Mr LOUNIS Hocine

Devant le jury composé de :

M	B. BOUSSEKSOU	CC ENP	**Président**
Mme	M. GUERTI	MC ENP	**Promotrice**
M	R. ZERGUI	CC ENP	**Examinateur**

Soutenu le 24 Juin 2007

E.N.P 10 Avenue Hassen Badi EL HARRACH - ALGER

Dédicaces

Je dédie ce travail à :

- mes parents qui m'ont soutenu, orienté, aidé et encouragé le long de mes études
- mes frères et sœurs avec qui j'ai passé les plus beaux moments de ma vie
- tous mes amis : Sofiane , Cherif , Walid , Moh , Hakim , Lyes , Abd razak , Rachid

et une dédicace spéciale a Nihad pour leur soutien constant et surtout leur encouragement pour que ce travail aboutisse.

- A mes proches

HOCINE

Remerciements

Tout d'abord je remercie Dieu de m'avoir donne la force et le courage d'accomplir ce travail.

J'exprime ma profonde gratitude à ma promotrice ***Mme M.Guerti*** *Maître de conférence à l'ENP qui m'a encadre et dirige tout au long de ce projet, qui m'a octroyé de précieux conseils, et qui a été disponible et patiente.*

Je remercie aussi Mr le président ainsi que les nombres du jury qui ont bien voulu m'a faire l'honneur d'examiner ce travail.

Mes remerciements vont également à tous mes enseignants à l'Ecole Nationale Polytechnique qui ont contribué à ma formation.

Je remercie tous ceux, qui de prés ou de loin, m'ont apporté leur contribution pour la réalisation de ce travail.

ملخص:

يندرج هذا العمل في إنجاز طريقة للتشخيص الأوتوماتيكي لطبقة صوتية غير مفخمة من نوع (fricatives- إحتكاكية) للغة العربية و هي: ش ، س ، ج ، ز. ولقد قدمنا هذا العمل في شكل نموذج بياني مطور بواسطة MATLAB-7 .هذا الإنجاز يبدأ بإنشاء تقارب كبير بين المستعمل و المعالجة الكيفية المنجزة .الهدف الأساسي من هذا الإنجاز يكمن في التحليل الطيفي للإشارة الصوتية، لاستخراج تواتر التجاوب لمجرى الفم (التواتر، الشريط الناقل).نتائج تحليل التواتر و الشريط الناقل لمجرى الفم استعملت في نظام التشخيص عن طريق المقارنة التقليدية مع مصدر مسجل.

كلمات مفاتيح:
عربية فصحة - حروف ساكنة غير مفخمة - التشخيص الأوتوماتيكي للكلام- الأصوات الإحتكاكية.

Résume :

Ce travail consiste à réaliser un outil de la reconnaissance automatique d'une classe des sons fricatives non emphatiques de l'Arabe Standard ce que concerne les consonnes suivant : س[s],ش[š],ز[z],ج[ž]. Cet outil est présenté sous forme d'une interface graphique développée avec MATLAB 7, dans un environnement Windows XP. La conception du ce logiciel vise d'abord à créer le plus d'interactions possibles entre l'utilisateur et le déroulement d'un traitement spécifique. La fonction principale de ce logiciel est l'analyse fréquentielle du signal de parole dans le but d'une extraction des résonances formantiques du conduit vocal (leurs fréquences, leurs bandes passantes). Les résultats des paramètres formantiques (temps, bandes passantes, fréquences) ont été introduits dans un système de reconnaissance par une méthode de comparaison traditionnelle avec les références enregistrées.

Mots clés : Arabe Standard – Consonnes Non Emphatique de l'Arabe Standard – étude acoustique – formants – Reconnaissance Automatique de la parole – les sons fricatives.

Abstract :

This work consists in producing a automatic recignition outil for fricatives no emphasis sound in the Standard Arabic (س[s],ش[š],ز[z],ج[ž]). This tool is presented in the form of a graphic interface developed with MATLAB 7, in an environment Windows XP. The design of the this software initially aims at creating the most possible interactions between the user and unfolding of a specific treatment. The principal function of this software is the frequential analysis of the signal of word in the goal of an extraction of formantic resonances of the vocal tract (their frequencies, their band-widths). The results obtained for the formantic parameters (time, wide bande, frequency) were introduced in the recognition systems by comparative with enregistrated references.

Key words : standard Arabic – Consonnes no emphasis of Standard Arabic – acoustic studing – formants – automatic speech recognition – fricatives sound.

SOMMAIRE

Chapitre 2 : *Techniques d'analyse du signal de parole*

Chapitre 3 : *les méthodes d'extraction de formants*

Chapitre 4 : *Mise en œuvre du système SAFAS*

Liste des figures

Liste des tableaux

Liste des abréviations

TAP : Traitement **A**utomatique de la **P**arole

RAP : Reconnaissance **A**utomatique de la **P**arole

AS : Arabe **S**tandard

TOP : Transcription **O**rthographique de la **P**arole

DAP : Décodage **A**coustico-**P**honétique

HMM: Hidden Markov Model

F_0 : Fréquence **F**ondamontale

$F_{1...5}$: les **F**ormants

SAFAS : Système **- A**nalyse **– F**ricatives **- A**rabe **S**tandard

LPC : Codage **P**rédictif **L**inéaire , « **L**inear **P**redictive **C**oding »

AR : Auto**R**égressif

DFW : Dynamic **F**requency **W**raping

TFCT : Transformée de **F**ourier à **C**ourt **T**erme

MFCC : Mel **F**requency **C**epstral **C**oefficient

INTRODUCTION GÉNÉRALE

Introduction générale

Les outils de traitement automatique de la parole sont de plus en plus nombreux et plus performants. Cependant, il n'existe pas d'outils répondant exactement aux besoins de chacun car suivant les travaux que nous effectuons nous avons besoin de traitements spécifiques répondant à nos attentes.

Le but de ce travail est la réalisation d'un outil automatique de traitement répondant aux nous besoins dans le domaine de la RAP. La fonction principale de ce logiciel est l'analyse fréquentielle du signal de parole dans le but d'une extraction des résonances formantiques du conduit vocal (leurs fréquences, leurs bandes passantes) et la reconnaissance automatique d'une classe des sons fricatives non emphatiques de l'AS [s], [š], [z], et [ž].

Il convient de rappeler que la définition même des formants est discutée. Les formants ont été définis au départ comme des maxima du spectre vocalique : « "The spectral peaks of the sound spectrum are called formants. (Fant 1960) ».

Ainsi définis, l'estimation des formants exacte est impossible car dépendante de facteurs comme la position de la fenêtre d'analyse vis à vis des périodes de la fréquence fondamentale, le degré de lissage du spectre, etc. Une définition plus rigoureuse des formants est en tant que fréquences de résonance de la fonction de transfert acoustique du conduit vocal.

Il existe plusieurs méthodes pour la mesure directe de la fonction du transfert du conduit vocal et, en conséquence, de ses résonances. Cependant, ces méthodes sont complexes et demandent une mise en oeuvre spéciale (excitation extérieure). Mêmes les plus rapides comme temps de mesure imposent des contraintes trop lourdes pour l'étude de la parole.

Les méthodes les plus courantes pour le traitement du signal de la parole sont les analyses spectrales réalisées soit par Transformée de Fourier à Court Terme, soit par prédiction linéaire ou soit par évaluation des coefficients cepstraux. . Pour notre part, nous nous sommes limités au calcul des résonances de manière classique (à partir du signal vocal pré-enregistré), en utilisant la fonction de transfert du conduit vocal calculée par le modèle de prédiction linéaire.

Notons que l'analyse en trajectoires formantiques des signaux de parole est indispensable pour la recherche. Il n'existe pas une méthode totalement efficace pour permettre de bonnes estimations de ces trajectoires.

Notre but dans cette étude n'est pas de réaliser un outil très performant mais répondant à nos besoins. Il a été conçu avec le maximum d'interactions entre l'utilisateur et le traitement recherché.

Il est basé sur une méthode de détection qui donne de bons résultats à condition d'ajuster correctement les paramètres d'entrée de ce dernier, Evidemment notre logiciel de détection des formants n'est pas toujours sûr et dépend souvent de la qualité du signal à analyser. Cet outil permet une représentation graphique de l'estimation des formants du signal de parole ainsi qu'un affichage de toutes les étapes du traitement et une représentation du spectrogramme pour vérifier la qualité de l'estimation et a la fin de reconnaissance automatique des consonnes étudiées.

Comme ce logiciel est destiné au traitement automatique du signal de parole, nous avons jugé nécessaire d'introduire dans le chapitre 1 quelques rappels sur la production des sons de la parole suivie d'une généralité sur Le système phonétique de l'Arabe Standard (AS).

Dans le chapitre 2, nous avons décrit brièvement toutes les techniques qui ont permis de mettre en œuvre cet outil.

Dans le chapitre 3, nous avons présenté un état de l'art relatif aux méthodes de calcul des formants et définir les différentes unités acoustiques.

Dans le chapitre 4, nous avons commençons par la description d'un corpus de références et d'un procédure de segmentation par suit nous avons présenté notre logiciel suivi de la validation des résultats obtenu avec cet outil.

Chapitre 1 :

Généralités sur la parole Et L'Arabe Standard

1.1. Introduction

Le Traitement Automatique de la Parole (TAP) présente un fort potentiel d'amélioration de l'interaction entre les humains et les machines, et entre les humains qui utilisent les machines. L'industrie du traitement de la parole se compose de la reconnaissance de la parole, du texte, de la biométrie de la voix, ainsi que des applications, des plates-formes et des services connexes.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

Pour mieux comprendre cette particularité, nous avons commencé dans ce chapitre par définir les deux mécanismes phonatoire et auditif de l'être humain, nous présenterons ensuite les propriétés spécifiques du signal vocal, ensuit les notions fondamentales et les classes des sons de l'Arabe Standard (AS), et nous terminons par la Transcription Orthographique et Phonétique (TOP) des consonnes de l'AS

1.2. Production de la parole

La parole, très souvent considérée comme activité propre de l'homme, est rarement étudiée comme fonction biologique (fig. 1.1).

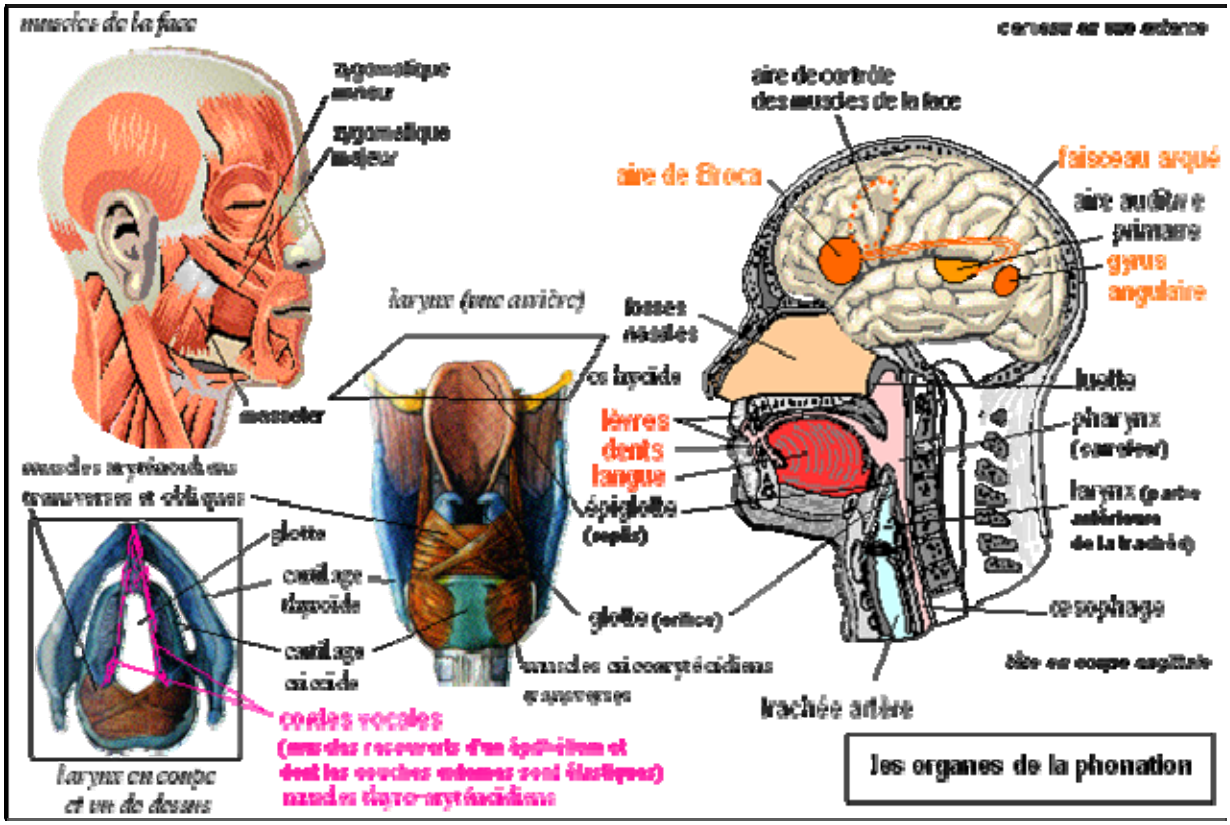


Figure 1.1 : Les organes de la phonation [1]

Elle fait sans aucun doute partie du travail de relation. Elle est considérée comme un moyen de communication avec les autres. Elle met en jeu des organes de phonation et est une véritable gymnastique des muscles du larynx, du pharynx, de la langue et des parois de la cavité buccale d'une façon générale. L'organe essentiel de la phonation est le larynx, extrémité différenciée de la trachée artère comprenant des pièces squelettiques cartilagineuses et des muscles. Les "cordes vocales" sont des muscles intrinsèques au larynx (c'est-à-dire reliant les différentes pièces squelettiques entre elles) qui comportent un revêtement souple et élastique formant une muqueuse. Ces cordes vibrent au passage de l'air provoquent des sons audibles. la hauteur des sons dépend d'abord de la tension des cordes vocales (fig.1.1), elle-même liée à l'action de certains muscles intrinsèques (par exemple les muscles cricoaryténoïdiens qui déterminent la hauteur des sons en augmentant la tension longitudinale des cordes vocales) et extrinsèques (reliant le larynx aux structures anatomiques voisines) qui modifient aussi la forme du larynx et modulent la voix (le larynx se déplaçant naturellement vers le haut quand la voix monte et vers la bas lorsqu'elle descend, il est nécessaire, pour contrôler sa voix de maintenir son larynx dans une position la plus fixe possible...). Mais interviennent aussi les muscles de la cage thoracique ou de l'abdomen. On parle ainsi d'appareil vocal subglottique pour désigner les poumons et les muscles de la cage thoracique, de l'abdomen, du dos et de la poitrine. Cet appareil, que les chanteurs appellent leur "appui" ou "soutien", confère sa puissance à la voix. Le pharynx (partie de la gorge située entre la bouche et l'œsophage) ainsi que les cavités buccales et nasales agissent comme des résonateurs qui atténuent certaines fréquences. Un chanteur expérimenté possède 4 ou 5 bandes de fréquences dominantes ou formants [1].

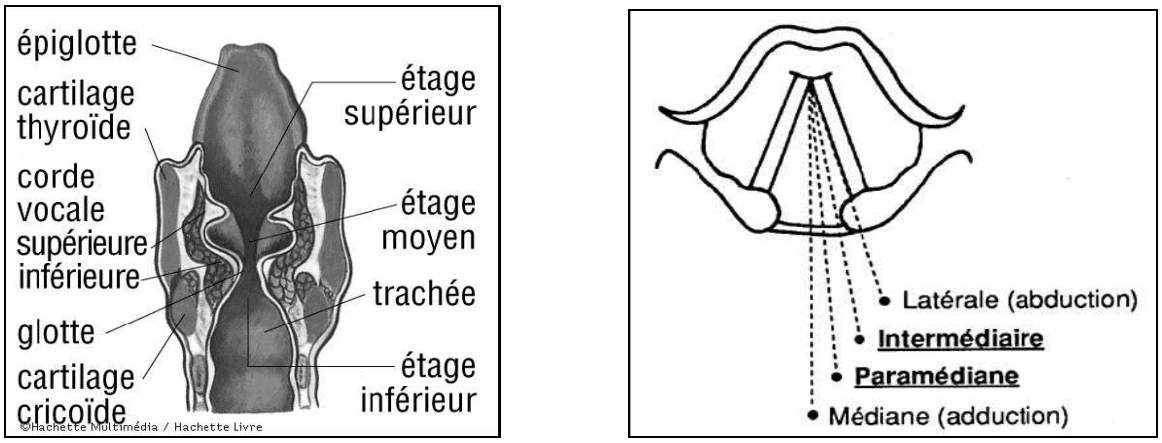


Fig. 1.2 : les cordes vocales [2]

A l'attaque d'un son les deux cordes vocales sont en contact, l'une de l'autre et la glotte est fermée. Lorsque les poumons expulsent l'air qu'ils contiennent, la pression sous la glotte augmente, écarte progressivement les cordes vocales en les poussant de bas en haut. La glotte finit par s'ouvrir et l'air s'y engouffre. La fermeture de la glotte se fait par retour élastique des cordes lorsque la pression de l'air dans la glotte diminue. La vibration acoustique résulte du hachage du courant d'air passant dans la glotte modulée par l'élasticité et la vibration de l'épaisseur de la lame des cordes vocales (la fermeture se fait d'abord dans leur partie inférieure puis par la partie supérieure). La voix est donc plus comparable à une série d'applaudissements qu'à la vibration d'une corde d'un instrument de musique. Une voix enrouée peut résulter d'un manque d'élasticité des cordes qui referment la glotte de façon moins symétrique. La fréquence fondamentale de la voix (hauteur du son) est associée à des fréquences secondaires plus élevées (harmoniques) et dépend directement du nombre de cycles d'ouverture-fermeture de la glotte par seconde.

1.2.1. La production des sons du point de vue articulatoire

La majorité des sons du langage sont le fait du passage d'une colonne d'air venant des poumons, qui traverse un ou plusieurs résonateurs de l'appareil phonatoire (figure 1.3). Les résonateurs principaux sont :

- le pharynx ;
- la cavité buccale ;
- la cavité labiale ;
- les fosses nasales ;

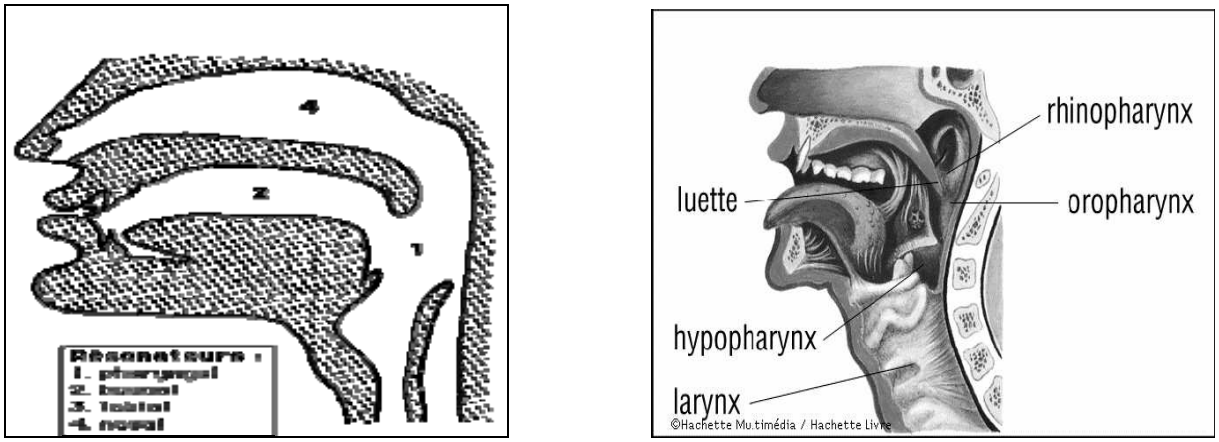


Fig. 1.3 : Les résonateurs principaux du conduit vocal [2]

La présence ou l'absence d'obstacles sur le parcours de la colonne d'air modifie la nature du son produit. C'est, entre autres, en classant ces obstacles éventuels que la phonétique articulatoire dégage les différentes classes de sons décrites ci-dessous.

Pour un petit nombre de réalisations, l'air ne provient pas des poumons, mais de l'extérieur, par inspiration. Une articulation peut aussi être engendrée par une variation de pression entre l'air interne et l'air externe à la cavité buccale, voir même par une variation de pression purement interne. (Figure 1.4)

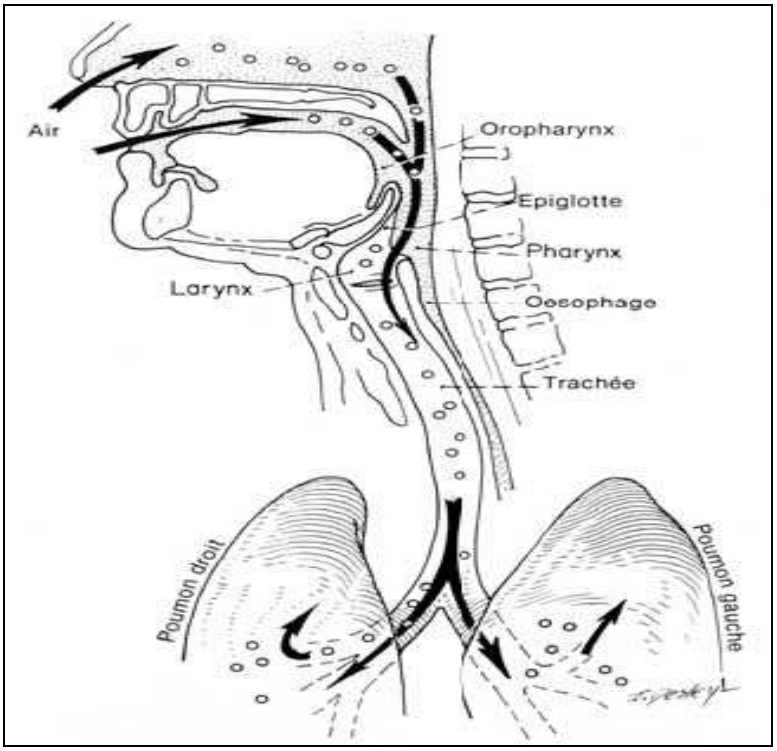


Fig.1.4 : l'ensemble des organes de l'appareil phonatoire humain [3].

1.2.2 Consonnes et voyelles

La distinction entre voyelles et consonnes s'effectue de la manière suivante :

- si le passage de l'air se fait librement à partir de la glotte, on a affaire à une voyelle.
- si le passage de l'air à partir de la glotte est obstrué, complètement ou partiellement, en un ou plusieurs endroits, on a affaire à une consonne.

Avant d'aller plus loin, on signalera que le passage des consonnes aux voyelles ne se fait pas de manière abrupte, mais sur un continuum. On distinguera ainsi des articulations intermédiaires, comme les vocoïdes (par exemple les semi-voyelles) ou les spirantes.

1.2.3. Point d'articulation et mode d'articulation

La distinction entre mode d'articulation et point d'articulation est particulièrement importante pour le classement des consonnes.

1.2.3.1. Le mode d'articulation

Est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- libre passage, ou mise en vibration, de l'air au niveau de la glotte (sourde ou sonore).
- libre passage, ou non, en un point quelconque (le point d'articulation) des cavités supra-glotiques (voyelle ou consonne).
- passage par une voie unique ou deux voies différentes (orale ou nasale).
- passage, dans le conduit buccal, par une voie médiane ou latérale (la plupart des articulations opposées aux latérales).

1.2.3.2 Le point d'articulation

Est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. De manière générale, on peut dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air.

Le point d'articulation peut se situer aux endroits suivants :

- les lèvres (articulations *labiales* ou *bilabiales*).
- les dents (articulations *dentales*).
- les lèvres et les dents (articulations *labio-dentales*).
- les alvéoles (c'est-à-dire les gencives internes des incisives supérieures, articulations *alvéolaires*).
- le palais (vu sa grande surface, on peut distinguer des articulations *pré-palatales*, *médio-palatales* et *post-palatales*).
- le voile du palais (palais mou, articulations *vélaires*).
- la luette (articulations dites *uvulaires*).

- le pharynx (articulations *pharyngales*).
- la glotte (articulations *glottales*).

1.2.3.3. Sourdes et sonores

Une réalisation est dite *sourde* lorsque les cordes vocales ne vibrent pas; si celles-ci entrent en vibration, la réalisation sera dite *sonore*. Les cordes vocales sont des replis musculaires situés au niveau de la glotte.

La vibration des cordes vocales est le résultat d'une obstruction de la glotte : celles-ci vibrent sous la pression de l'air interne qui force un passage entre elles.

1.2.3.4. Orales et nasales

Au carrefour du pharynx, le passage de l'air peut s'effectuer dans une ou deux directions, selon la position du voile du palais :

- si le voile du palais est relevé, l'accès aux fosses nasales est bloqué, et l'air ne peut traverser que la cavité buccale.
- si le voile du palais est abaissé, une partie de l'air traversera les fosses nasales (l'autre partie poursuivant son chemin à travers la cavité buccale).

Les réalisations du premier type sont dites *orales*, celles du second type *nasales*. Pour plus de détails. (Figure 1.5).

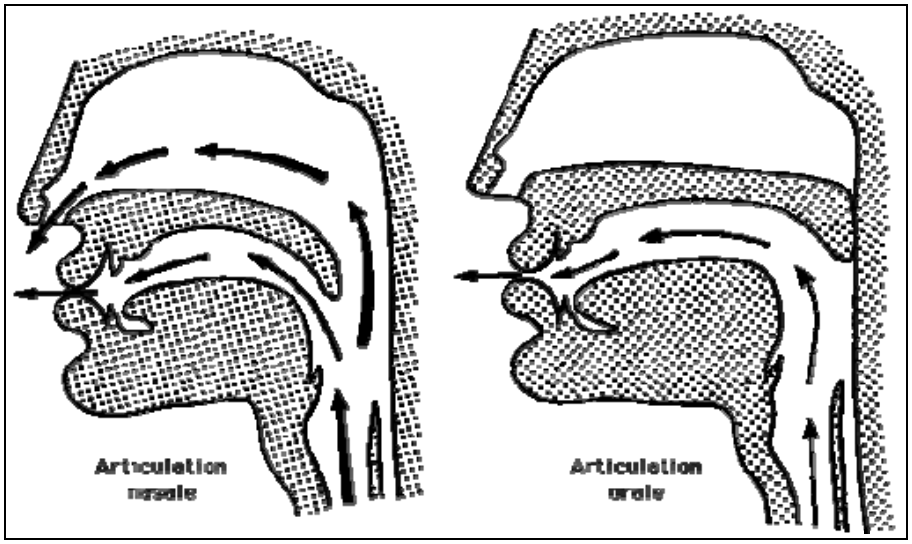


Fig. 1.5 : Articulations nasales et orales

1.2.4. Fonctionnement acoustique de l'appareil vocal humain

Le signal de parole est le résultat de la propagation d'une onde acoustique dans un tuyau de forme variable. Ce tuyau est mis en forme par un ensemble de muscles agissant sur des structures rigides, semi-rigides ou mous comme la langue ou les lèvres. De même, les sources d'excitation sont produites par des compressions exercées sur les poumons, les effecteurs laryngés, les parois du conduit vocal dont la tension est en permanence régulée et contrôlée par des structures musculaires spécifiques.

Le contrôle complexe de cette activité musculaire permet au locuteur de mettre en forme le contenu spectral et temporel du signal de parole. L'auditeur récupère la trace audible de cette suite de gestes complétée occasionnellement par leur trace visible excitateur et résonateur. L'appareil phonatoire humain fonctionne donc comme un système acoustique.

Un système acoustique comporte généralement deux parties : un excitateur et un résonateur, qui est le volume dans lequel se propage l'excitation. L'excitateur délivre un signal source dont certaines composantes vont être affaiblies ou renforcées dans le résonateur, c'est sa fréquence de résonance. Elle varie selon le volume de la cavité et la surface de l'ouverture du résonateur.

L'appareil vocal humain est constitué d'un excitateur, le complexe glotte/cordes vocales, et d'un ensemble de résonateurs qui sont :

- le pharynx ;
- la cavité buccale ;
- la cavité labiale ;
- les fosses nasales ;

Un des problèmes spécifiques à la phonation est que, souvent, le résonateur réagit sur l'excitateur et le signal source s'en trouve modifié.

1.2.5. Les formants

Lorsqu'un excitateur entre en vibration, il fournit un signal, dont le résonateur va amplifier certaines composantes. On obtient alors des formants qui sont un facteur fondamental dans la caractérisation du timbre. Ils servent, justement, à « former » ce dernier.
Le nombre des formants, selon les caractéristiques du résonateur (volume, forme et ouverture), est variable: d'un seul à (théoriquement) une infinité. Néanmoins, du point de vue perceptif, seuls quelques-uns d'entre eux jouent un rôle central au niveau de la parole. Par exemple, on peut caractériser toute voyelle en ne prenant en compte que ses trois premiers formants. (Pour une réalisation de la voyelle [i] par exemple, les trois premiers formants pourraient se situer respectivement à 300, 2200 et 3000 Hz.)
En fait, un formant ne peut jamais être ramené à une fréquence fixe (sinon de manière conventionnelle, en effectuant une moyenne par exemple, comme pour la voyelle [i] ci-dessus). Il s'agit plutôt d'une bande de fréquences qui sera d'autant plus large que le système est amorti. Ces régions formantiques apparaissent très clairement sur les spectrogrammes.
Pour définir les caractéristiques d'un résonateur (ce qu'on appelle, par abus de langage, sa fréquence), on envoie, à travers celui-ci, un bruit blanc, formé du mélange de toutes les fréquences. On verra alors clairement sur le spectrogramme du bruit coloré ainsi obtenu quelles zones fréquentielles seront amplifiées par le résonateur.
Parmi toutes les représentations phonétiques, les formants ont une place privilégiée. Les cartes formantiques sont largement utilisées pour :

- caractériser la production. La représentation formantique est centrale pour prédire les espaces vocaliques [19], pour caractériser les réalisations vocaliques ou consonantiques en fonction de divers paramètres contextuels [42] ou prosodiques [43].

- synthétiser : On utilise les formants pour contrôler la production de sons artificiels et établir des règles de génération [21].

- inverser : Les techniques d'inversion articulatori-acoustique prennent de manière privilégiée les formants comme caractérisation du spectre [22, 23, 31].

1.3. Audition - perception

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu. Par ailleurs, le rôle fondamental que joue l'audition dans le processus même de production de la parole.

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'***oreille interne*** qui envoie au cerveau l'influx nerveux qui en résulte; le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.[4]

L'appareil auditif comprend l'***oreille externe***, l'***oreille moyenne***, et l'***oreille interne.*** Le conduit auditif relie le pavillon au tympan (c'est un tube acoustique de section uniforme fermé à une extrémité), son premier mode de résonance est situé vers 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences (Fig. 1.6).

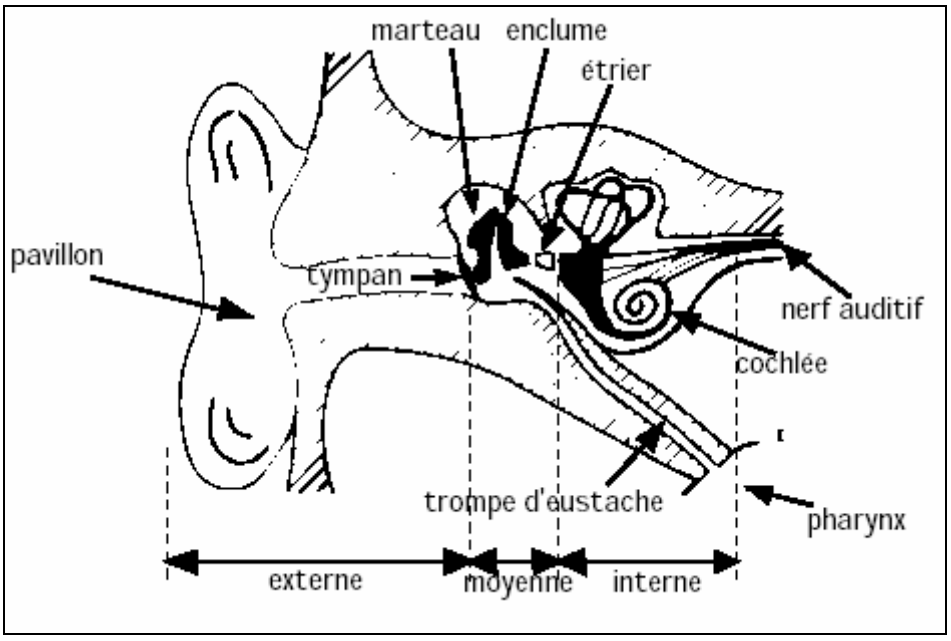


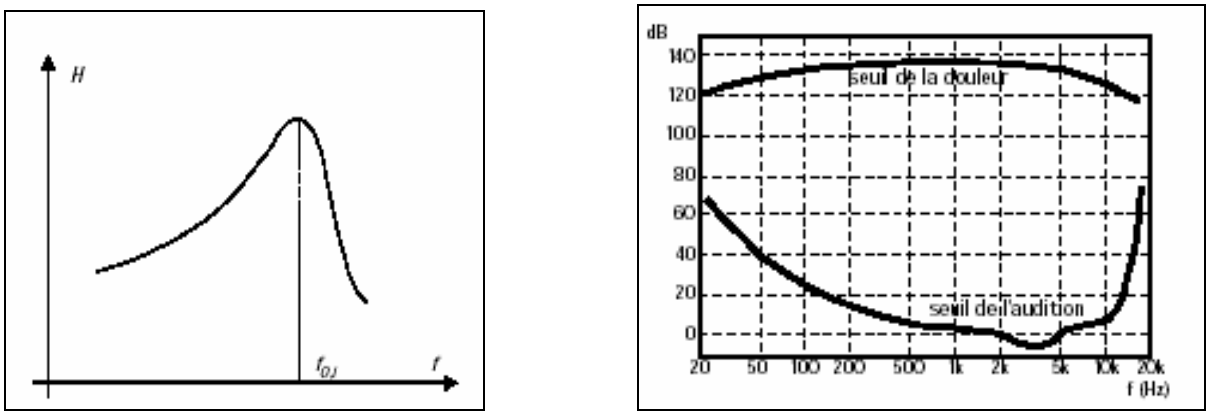
Fig. 1.6 : Le système auditif [3]

Le mécanisme de l'oreille interne (***marteau***, ***étrier***, ***enclume***) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la ***cochlée***. Celle-ci contient la ***membrane basilaire*** qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur

et à mesure que l'on se rapproche de l'apex de la cochlée; elle est le support de l'***organe de Corti*** qui est constitué par environ 25000 ***cellules ciliées*** raccordées au nerf auditif. La réponse en fréquence du conduit au droit de chaque cellule est esquissée à la figure (1.7-a). La fréquence de résonance dépend de la position occupée par la cellule sur la membrane; au-delà de cette fréquence, la fonction de réponse s'atténue très vite. Les fibres nerveuses aboutissent à une région de l'écorce cérébrale appelée ***aire de projection auditive*** et située dans le lobe temporal. En cas de lésion de cette aire, on peut observer des troubles auditifs. Les fibres nerveuses auditives afférentes (de l'oreille au cerveau) et efférentes (du cerveau vers l'oreille) sont partiellement croisées : chaque moitié du cerveau est mise en relation avec les deux oreilles internes.

Il reste très difficile de nos jours de dire comment l'information auditive est traitée par le cerveau. On a pu par contre étudier comment elle était finalement perçue, dans le cadre d'une science spécifique appelée ***psychoacoustique***. Sans vouloir entrer dans trop de détails sur la contribution majeure des psychoacousticiens dans l'étude de la parole, il est intéressant d'en connaître les résultats les plus marquants. [3]

Ainsi, l'oreille ne répond pas également à toutes les fréquences. La figure (1.7-b) présente le champ auditif humain, délimité par la courbe de ***seuil de l'audition*** et celle du ***seuil de la douleur***. Sa limite supérieure en fréquence (16000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (32000 Hz).



(a) **(b)**

Fig. 1.7 (a) : Réponse en fréquence d'une cellule ciliée. (B) : Le champ auditif humain [3]

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La figure (1.8 .a) fait apparaître les courbes d'égale impression de puissance auditive (aussi appelée ***sonie,*** exprimée en ***sones***) en fonction de la fréquence. Elles révèlent un maximum de sensibilité dans la plage [500 Hz, 10 kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

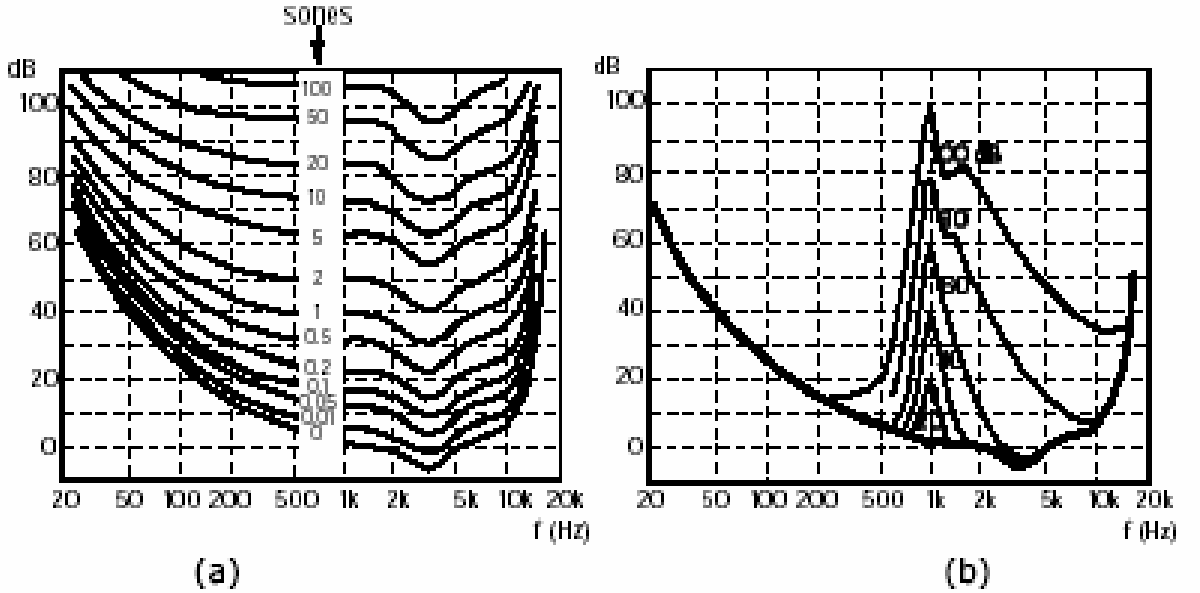


Fig. 1.8 (a) : Courbes isosoniques en champ ouvert. (b) : Masquage Auditif par un bruit à bande étroite.[3]

Enfin, un son peut en cacher un autre. Cette propriété psychoacoustique, appelée ***phénomène de masquage***, peut être visualisée sous la forme de courbes de masquage (Fig. 1.8.b), qui mettent en évidence la modification locale du seuil d'audition en fonction de la présence d'un signal déterminé (un bruit à bande étroite centré sur 1 kHz dans le cas de la figure (1.8.b).

Une modélisation efficiente des propriétés de masquage de l'oreille permet de réduire le débit binaire nécessaire au stockage ou à la transmission d'un signal acoustique, en éliminant les composantes inaudibles.

Remarquons pour terminer que ce qui est perçu n'est pas nécessairement *compris.* Une connaissance de la langue interfère naturellement avec les propriétés psychoacoustiques de l'oreille. En effet, les sons ne sont jamais prononcés isolément, et le contexte phonétique dans lequel ils apparaissent est lui aussi mis à contribution par le cerveau pour la compréhension du message.

Ainsi, certains sons portent plus d'information que d'autres, dans la mesure où leur probabilité d'apparition à un endroit donner de la chaîne parlée est plus faibles, de sorte qu'ils réduisent l'espace de recherche pour les sons voisins. Les sons sont organisés en unités plus larges, comme les mots, qui obéissent eux-mêmes à une syntaxe et constituent une phrase porteuse de sens. Par conséquent, c'est tout notre savoir linguistique qui est mis à contribution lors du décodage acoustico-phonétique. Les sections qui suivent ont précisément pour objet la description linguistique du signal de parole.

1.4. PROPRIETES SPECIFIQUES DU SIGNAL VOCAL

La grande difficulté de la reconnaissance automatique de la parole provient du caractère même du processus de la communication parlée et des propriétés intrinsèques du signal vocal, les messages vocaux subissent une série de transformations depuis 1' idée à émettre jusqu' au signal acoustique, ce qui correspond à un codage très complexe. Le décodage d'un tel message est à l'évidence particulièrement difficile [5] (Fig. 1.9).

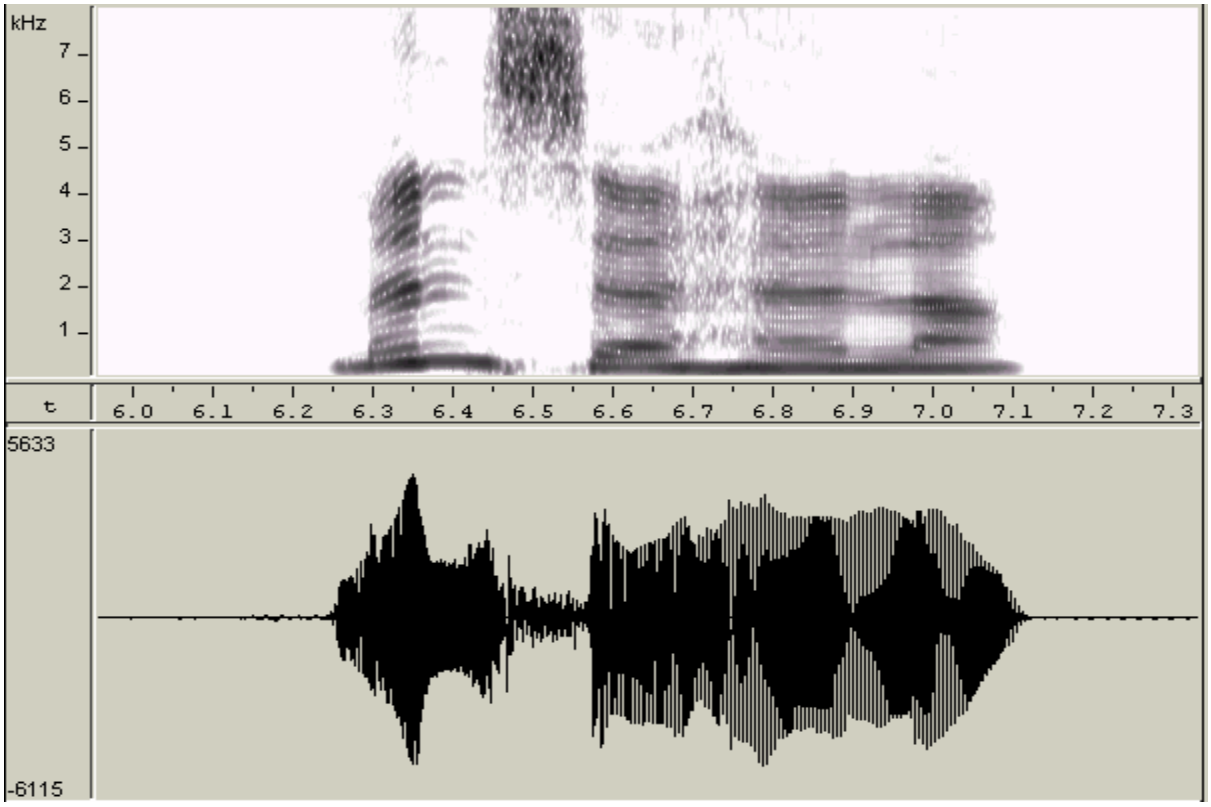


Fig. 1.9 : Spectrogramme et signal temporel de la phrase /men sahala/

Le signal vocal, tel qu'il apparaît sur la figure 1.9 possède des propriétés très spécifiques et qui se résument par :

- La continuité ;
- La variabilité ;
- La redondance ;
- La grande liberté du langage parlé ;

1.4.1. La continuité

Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont 1' occurrence est aléatoire, il peut très bien y avoir des intervalles de silence au milieu d'un mot et aucun intervalle entre deux mots successifs. Il est donc très difficile de déterminer le début et la fin des mots composant la phrase [4].

1.4.2. La variabilité

La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que se soit pour un même locuteur ou plusieurs. Pour un même locuteur, des différences importantes de prononciations peuvent apparaître suivant 1' état émotionnel du sujet et l'intensité de sa voix ; celui - ci peut crier ou murmurer, être enroué ou enrhumé. De même, des contrastes considérables peuvent se manifester entre plusieurs locuteurs suivant l'âge, le sexe, 1' origine géographique et le milieu social. On peut ajouter aussi les perturbations apportées par le microphone (selon le type, la distance, 1' orientation) et l'environnement (bruit, réverbération), etc. [5].

1.4.3. La redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, en effet, de réduire au maximum cette redondance afin de diminuer 1' encombrement en mémoire et de limiter les durées du traitement, lequel doit se faire en temps réel. A 1' inverse, le débit ne doit pas être trop faible pour conserver un bon rapport signal / bruit. Une valeur de 100 ou 50 bits/s paraît convenir à la reconnaissance.

1.4.4. La grande liberté du langage parlé

La syntaxe du langage parlé est généralement moins stricte que celle du langage écrit. Les programmes de reconnaissances évolués doivent obligatoirement en tenir compte si 1' on veut qu'ils soient utilisables en pratique [5].

1.5. Décodage Acoustico - Phonétique

Il sert à décoder le signal acoustique en unités linguistiques (phonèmes, syllabes, les mots…).

Phonème: élément sonore d'un langage donné, déterminé par les rapports qu'il entretient avec les autres sons de ce langage.

Par exemple, le mot " cou " est formé des phonèmes " keu " et " ou ". Il en existe une trentaine en français. Cette notion est assez importante en reconnaissance vocale.

1ère partie : Faire apparaître les segments du signal

1ère étape : segmenter le signal en segments élémentaires et étiqueter ces segments. Le principal problème est de choisir les unités sur lesquelles portera le décodage.

- Si des unités longues telles que les syllabes ou les mots sont choisies, la reconnaissance en elle-même sera facilitée mais leur identification est difficile.
- Si des unités courtes sont choisies, comme les phones (sons élémentaires), la localisation sera plus facile mais leur exploitation nécessitera de les assembler en unités plus larges.

Les phonèmes constituent un bon compromis, leur nombre est limité : ils sont donc souvent utilisés. Mais le choix dépend également du type de reconnaissance effectuée : mots isolés ou parole continue. Cela sera abordé plus loin.

2ème étape : identifier les différents segments en fonction de contraintes phonétiques, linguistiques… Il faut que le système ait intégré un certain nombre de connaissances : données articulatoires, sons du français, données phonétiques, prosodiques, syntaxiques, sémantiques …

Deux sortes d'outils sont utilisées :

- Les outils de reconnaissance de formes structurelle (ex : grammaires déterministes)
- Les outils provenant de systèmes experts (souvent associés pour de meilleures performances).Un système expert effectue les interprétations et déductions nécessaires grâce à la modélisation préalable du raisonnement de l'expert (domaine de l'intelligence artificielle).

2ème partie : Reconnaissance des mots isolés

Retrouver les phonèmes et les mots dans un signal vocal est une réelle difficulté pour la reconnaissance vocale. De ce fait, séparer tous les mots prononcés par des silences permet de simplifier le problème.

1.5.1. Les techniques

Deux approches :

Dans **l'approche globale**, l'unité de base est le mot (donc non décomposable). Cette méthode fournit une image acoustique de chaque mot (Figure 1.10) 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 restreint de locuteurs (les mots peuvent être prononcés de manière différente suivant le locuteur).

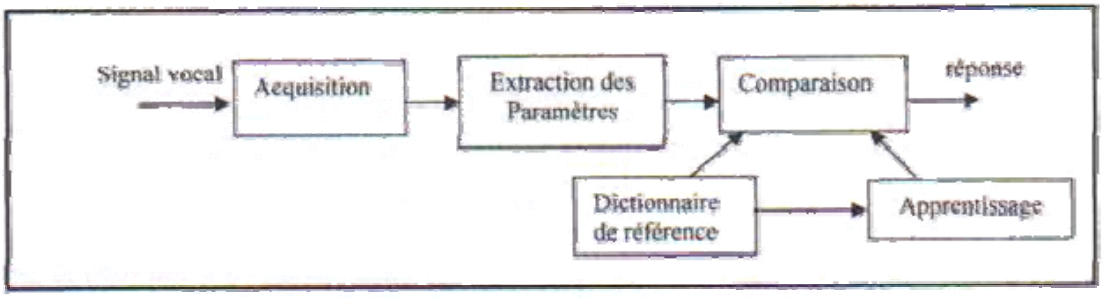


Figure 1.10 : Schéma synoptique d'un système de reconnaissance de la parole
Selon une approche globale

L'approche analytique, qui tire parti de la structure des mots, identifie les composantes élémentaires (phonèmes, syllabes, …). Celles-ci sont les unités de base à reconnaître.
Cette approche est plus générale que la précédente pour reconnaître de grands vocabulaires, il suffit d'enregistrer dans la mémoire de la machine les principales caractéristiques des unités de base (Figure 1.11).

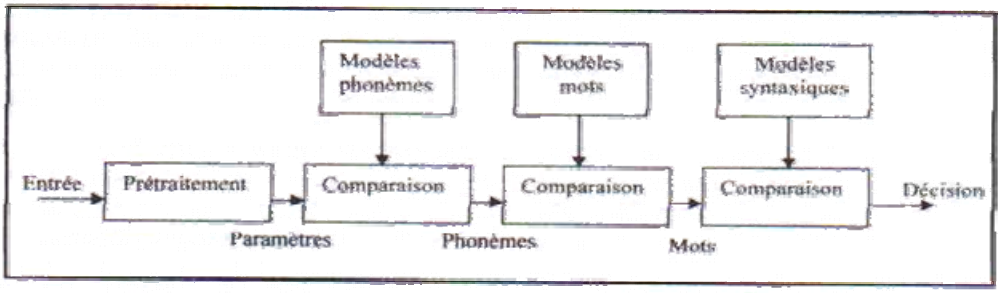


Figure 1.11 : schéma synoptique d'un système de reconnaissance de la parole
Selon une approche analytique

Pour la reconnaissance de mots isolés à grand vocabulaire, la méthode globale ne convient plus car la machine nécessiterait une mémoire et une puissance considérable pour respectivement stocker les images acoustiques de tous les mots du vocabulaire et comparer un

mot inconnu à l'ensemble des mots du dictionnaire. Il est de plus impensable de faire dicter à l'utilisateur l'ensemble des mots que l'ordinateur a en mémoire.

C'est donc la méthode analytique qui est utilisée : les mots ne sont pas mémorisés dans leur intégralité, mais traités en tant que suite de phonèmes.

1.5.2. Principe général de la méthode globale et analytique

Le principe est le même que ce soit pour l'approche analytique ou l'approche global, ce qui différencie ces deux méthodes est l'entité à reconnaître : pour la première il s'agit du phonème, pour l'autre du mot, le tableau suivant présent les Avantages et les inconvénients des méthodes globales et analytiques (Tableau 3.1).

On distingue dans ce cas deux phases:

- **La phase d'apprentissage :** un locuteur prononce l'ensemble du vocabulaire, souvent plusieurs fois, pour créer en machine le dictionnaire de références acoustiques. Pour l'approche analytique, l'ordinateur demande à l'utilisateur d'énoncer des phrases souvent dépourvues de toute signification, mais qui présentent l'intérêt de comporter des successions de phonèmes bien particuliers.
- **La phase de reconnaissance :** un locuteur prononce un mot du vocabulaire. Ensuite la reconnaissance du mot est un problème typique de reconnaissance de formes. Tout système de reconnaissance des formes comporte toujours les trois parties suivantes:
  - Un capteur permettant d'appréhender le phénomène physique considéré (dans notre cas un microphone),
  - Un étage de paramétrisation des formes (par exemple un analyseur spectral),
  - Un étage de décision chargé de classer une forme inconnue dans l'une des catégories possibles.

CRACTERES	Méthode globale	Méthode analytique
Taille du vocabulaire	limitée	indépendante
Taux de reconnaissance actuel	très élevé (> **95** %)	faible
Indépendance vis-à-vis de la langue	oui	non
Traitement de la parole continue	impossible	possible
Exploitation et mise en oeuvre	facile	difficile
Problèmes de segmentation	simple	très difficile
Adaptation au locuteur	difficile	relativement facile
Domaine d'application	spécialisé	vaste

Tableau 1.1 : Avantages et inconvénients des méthodes globales et analytiques [5].

1.6. Notions fondamentales sur les sons de l'Arabe Standard

La recherche en traitement automatique de la parole et notamment en reconnaissance, dans une langue donnée doit nécessairement passer par l'étude de sa composante phonétique. Cette étude nous permet de dégager les principales caractéristiques relatives aux différents phonèmes et ainsi de cerner l'ensemble des paramètres acoustiques, en vue de les exploiter dans l'élaboration d'un système de reconnaissance de la parole [9].

1.6.1. Le système phonétique de l'Arabe Standard

L'Arabe Moderne ou l'Arabe Standard est, la langue de communication commune à l'ensemble du monde arabe. II s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel. La langue arabe appartient à la famille des langues sémitiques. L'étude de la grammaire arabe a commencé très tôt au milieu du 11 éme siècle de l'hégire et a donné lieu à d'énormes productions, avant de connaître une période de stagnation qui a duré plusieurs siècles. Ces dernières années, elle connaît un regain d'intérêt, entre autres dans le domaine du traitement automatique**.**

1.6.1.1. Phonétique et Phonologie de la langue arabe

Nous présentons ci-dessous certaines caractéristiques phonétiques de l'AS.

- **Le système vocalique**
  - Le système vocalique comprend trois voyelles brèves / [a]/ [u]/ [i]/, et trois voyelles longues /[A]/[U]/[I]/ qui s'opposent aux précédentes par une durée plus importante sur le plan temporel. L'ensemble des voyelles brèves et longues est dit oral car elles sont émises sans l'intervention de la cavité nasale. Elles sont généralement classées selon le degré d'ouverture du conduit vocal (ouvert /[a]/, fermé /[i]/) et sa position de constriction (/[i]/ antérieure, /[u]/ postérieure) [10].
  - Ces voyelles peuvent avoir des timbres différents selon leur contexte d'apparition :
  - dans un contexte emphatique (au contact des consonnes ص/[S]/, ض/[D]/ , ط//[T]/, ظ/[Z]/), le point d'articulation des voyelles est reporté à l'arrière.
  - après les consonnes labiales م/[m]/ et ب/[b]/ les voyelles sont plus arrondies et se rapprochent du phonème /[u]/.

- **Le système consonantique**

  - L'arabe standard contient 28 consonnes qui correspondent chacune à un phonème, les consonnes de l'arabe sont classées selon leur mode d'articulation (occlusif, fricatif, nasal, glissant ou liquide), leur lieu d'articulation (labial, dental ou vélo-palatal) et leur voisement (sonore ou sourd). Nous proposons de les grouper en fonction de leurs équivalences dans les autres langues :
  - Les phonèmes spécifiques à l'arabe qui n'ont pas d'équivalent dans les langues européennes /[S]/, ض/[D]/ , ط//[T]/, ظ/[Z]/, ح/[H]/,ق /[q]/, ع/[ɛ]/.
  - Les phonèmes qui ont des équivalents dans la langue française :ت /[t]/ , س/[S]/,ش/[O]/,غ/[G]/,ك/[K]/,ج/[j]/,ف/[f]/,ب/[b]/,ز/[z]/,د/[d]/, ل/[l]/,م/[m]/,ن/[n]/,و/[w]/,ي/[y]/ .
  - Les phonèmes qui ont des équivalents dans plusieurs langues telles que l'espagnol, l'allemand ou l'anglais : ر/ [r]/,د / [v]/, ح / [h]/.

1.6.1.2. Particularités phonologiques

Les caractéristiques phonologiques de l'arabe sont l'emphase, la gémination et le madd :

- **l'emphase** est habituellement utilisée pour rendre compte des manifestations prosodiques liées à l'accentuation volontaire d'une syllabe, les consonnes ظ /[Z]/, ط/[T]/, ض/[D]/,.ص/[S]/) sont dites emphatiques, Certaines des études affirment que le phénomène de l'emphase dépasse le cadre de la voyelle (ou des voyelles) adjacente(s) et se propage aux phonèmes voisins comme dans le mot $[C_1V_1C_2V_2...]$ ([C]=consonne,[V]=voyelle), si $[C_I]$ est emphatique, alors la synthèse est plus naturelle quand la propagation de l'emphase arrive jusqu'à$[C_2]$. En revanche, il existe des divergences sur la portée de cette propagation, en d'autres termes, sur la taille du segment sonore affecté par la consonne emphatique.
- **la gémination** est symbolisée par le signe de la chadda qui signifie le dédoublement de la consonne. Une consonne géminée est *un son* unique pour lequel les organes de phonation ne changent pas de position (les lèvres ne se referment pas après le premier /b/ dans /kabbara/), d'où la transcription /kab:ara/ qui est plus appropriée Dans beaucoup de langues, ce phénomène permet de mettre en relief un mot dans son contexte, alors qu'il s'avère être un élément distinctif sur les plans morpho-sémantiques en langue arabe

  حضر[haDara] (il a assisté) est différente de حضّر[haDDara] (il a préparé) la deuxième consonne est géminée.

- **le madd** concerne l'allongement des voyelles. H est provoqué par la présence d'une voyelle longue (و /[U]/, أ/[A]/ ou ي/[I]/) La lecture de textes arabes est régie par des règles phonologiques qui ont trait à la contraction des sons, leur élision et à l'assimilation homo-organique des nasales. Certaines de ces règles sont obligatoires, d'autres facultatives ou réservées à certains types de textes, comme le Coran. Exemple le mot نام/ nAma, la voyelle [A] représente le madd dans le mot. [10]

1.6.2. Classification des sons

La taxonomie des sons est définie de deux manières, grâce à la phonétique et à la phonologie. Alors que la phonétique peut être considérée comme véritablement descriptive, associant chaque son de la langue à un symbole et à une classe, la phonologie s'intéresse, elle, à la description des interdépendances entre sons et au codage effectif des mots du langage lors du processus d'oralisation. La phonologie essaie donc plus particulièrement d'expliquer les différences qui peuvent exister entre la transcription phonétique d'un mot du langage et la transcription phonétique exacte du mot qui est effectivement prononcer.

1.6.2.1. Description des voyelles

Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle [4].

Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais se sont principalement les formants en basses fréquences qui caractérisent les voyelles.

1.6.2.2. Description des consonnes

Les consonnes sont caractérisées par la présence de bruits sans définition périodique précise.
On distingue deux types de consonnes :

• Les occlusives

Les phonèmes de cette la classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont donc constituées de deux parties successives, une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion, au moment du relâchement.

Les occlusives peuvent être voisées, à la manière des voyelles, ou sourdes, c'est-à-dire non voisées. Les occlusives voisées peuvent également être appelées occlusives sonores. Il existe deux types d'occlusives :

- Les occlusives nasales sont produites avec la participation de la cavité nasale.
- Les orales sont produites avec le velum en position relevée, c'est-à-dire que l'air ne passe pas dans les fosses nasales.

• Les fricatives

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

Les semi consonnes ou semi voyelles combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte. Mais le relâchement soudain de cette position produit une friction qui est typique des consonnes [4].

1.6.2.1. Modes et lieux d'articulation

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- libre passage, avec mise en vibration, de l'air au niveau de la glotte (sonore ou sourde).
- libre passage, ou non, en un point quelconque (le lieu d'articulation) des cavités supra-glottique (voyelles ou consonnes) ;
- passage par une voie unique ou deux voies différentes (orale ou nasale) ;
- passage, dans le conduit buccal, par une voie médiane ou latérale (la plupart des articulations opposées aux latérales).

Le lieu d'articulation est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. Il peut se situé aux endroits suivants :

- les lèvres (articulation labiale ou bilabiale).
- les dents (articulation dentales).
- les lèvres et les dents (articulation labio-dentale).
- les alvéoles (les gencives internes des incisives supérieures, articulations alvéolaires).
- le palais (vue sa grande surface, on peut distinguer des articulations pré-palatales, médio-palatales et post-palatales).
- le voile du palais (palais mou, articulation vélaires).
- la luette (articulations dites uvulaires).
- le pharynx (articulations pharyngales).
- la glotte (articulations glottales).

1.6.2.2. Transcription Orthographique Phonétique (TOP)

La TOP permet de représenter le texte tel qu'il sera prononcé par le système. La complexité de cette tâche varie selon la langue traitée. Ainsi, la transcription de l'arabe ou de l'espagnol est relativement directe par rapport à celle de la langue française qui présente de nombreuses ambiguïtés de prononciation que seul le contexte syntaxique permet de lever (Tableau 1.2).

Mode	Type de phonème		Phonèmes Arabes	Transcription Arabisante	Lieux d'articulation
Occlusives	Voisées		ب د	b d	bilabiale alvéodentale
	Non-Voisées		ق ت ك ء	q t k ,	uvulaire alvéodentale postpalatale glottale
	Voisée	Emphatiques	ظ	ḏ	alvéolaire
	Non-Voisée		ط	ṭ	alvéodentale
Fricatives	Voisées		ز ذ غ ع ج	z ḍ ġ ʿ ǧ	sifflante dorsoalvéolaire interdentale uvulaire pharyngale
	Non-Voisées		س ث ف ش خ ه ح	s ṯ f š h h h	sifflante dentale interdentale labiodentale chuinchante palatale vélaire glottale pharyngale
	Voisée	Emphatiques	ص	ṣ	dorsealvéodentale sifflante
	Non-Voisée		ض	ḍ	interdentale
Nasales	Voisées		م ن	m n	bilabiale alvéodentale
Liquide	Voisée		ل	l	dentale
Affriquée	Voisée		ج	ǧ	alvéopalatale
Vibrante	Voisée		ر	r	apicoalvéolaire
Semi-voyelles	Non-Voisées		و ي	w y	bilabiale palatale

Tab 1.2 : Transcription orthographique et phonétique Des consonnes de l'Arabe Standard [11]

1.7. Conclusion

Ce bref aperçu de la façon avec laquelle est produite la parole et l'audition de ce parole et nous avons passé en revue les principales caractéristiques acoustiques et phonétiques du signale de la parole. Nous avons aussi présenté quelque notion de base sur le problème de la Reconnaissance Automatique de la Parole (RAP) qui peut être résolu par deux approches : globale ou analytique, et on a fini par présenté le système phonétique de l'Arabe Standard (AS) et sont classification de sons.

Il est nécessaire de situer le formant dans le phénomène de la production de la parole et l'importance de cette grandeur dans la caractérisation d'un son. Donc il est intéressant à la fin de voir comment procéder pour extraire cette grandeur à partir du signal acoustique de la parole. Ce qui fera l'objet du chapitre 2.

Chapitre 2 :

Techniques d'analyse Du Signal vocal

2.1. Introduction

Nous avons vue dans le chapitre précédent que La parole est un signal réel, continu, d'énergie finie. Sa structure est complexe et variable dans le temps tantôt périodique pour les sons voisés, tantôt aléatoire pour les sons fricatifs, tantôt impulsionnel dans les phases explosives des sons occlusifs.

Dans ce second chapitre nous essayons d'illustrer les déférents technique d'analyse du signal de la parole et pour mieux comprendre le fonctionnement de la production de la parole ainsi que la complexité de ce signal, d'ou la multitude de méthodes et techniques existantes dans ce domaine.

2.2. Les paramètres pertinents du signal de parole

Le traitement du signal vocal a pour but de fournir une représentation moins redondante de la parole que celle obtenue par codage de l'onde temporelle tout en permettant une extraction précise des paramètres pertinents du signal de parole tels que :

- Pour la source :
  - Période du fondamental
  - Amplitude Ao.
- Pour le conduit vocal :
  - Période des formants Ti =1/Fi , i=1,2,… (Fi : fréquences des formants
  - Amplitude Ai
  - bandes passantes Bi.
- L'extrême variabilité du signal vocal est due à la :
  - Complexité du couplage (source/conduit) ;
  - Grande dynamique et variété des voix ;
  - Variation rapide de la parole.
- Cette variabilité est liée directement au locuteur :
  - à son âge, son sexe et à son accent géographique
  - et son état physique (fatigue, maladie) et émotionnel (content, triste, nerveux)

Toutes ces propriétés complexes du signal vocal rendent ce dernier difficiles à traiter d'où la multiplicité des méthodes de traitement. Ces méthodes de traitements sont très nombreuses, nous n'en citeront que celles utilisées dans cette étude.

2.3. Les techniques de traitement du signal de parole

0n classe habituellement, les différentes méthodes de traitement du signal en trois catégories :

- Les transformées usuelles : transformée discrète de Fourier et transformée en Z.
- Les méthodes fondées sur la dé convolution ‘‘source/ conduit’’ cepstre et codage prédictif linéaire (LPC) qui s'appuient sur un modèle même simplifié de production de la parole.
- Les méthodes basées sur un modèle de perception (filtre). [12]

L'objectif poursuivi dans le domaine de traitement du signal est la transmission ou l'enregistrement de ce signal vocal, ou encore sa synthèse ou sa reconnaissance. Vu les caractéristiques de ce signal, son interprétation se complique et les données à traiter augmentent.

Avec le développement des calculateurs et des circuits numériques spécialisés le traitement analogique du signal a subi un déclin important vis à vis du traitement numérique. La stabilité et la précision des systèmes numériques n'est plus a démontrer. On fait appel à certaines techniques ou certains outils mathématiques**.** [13], [14], [15]. Ces techniques et outils mathématiques ne sont malheureusement applicables qu'à des signaux stationnaires et discrets, ce qui n'est pas le cas du signal de parole. D'ou la nécessité d'un pré traitement de ce dernier.

2.4. Aspect fréquentiels de la parole

La parole est constituée de plusieurs éléments appelés phonèmes, dépendants les uns des autres. On peut caractériser ces phonèmes ainsi que leurs liens grâce à leurs aspects fréquentiels : présence ou non de fondamental laryngé, formants, transitions phonétiques ainsi que bruits d'explosion et de friction. [12] [13]

2.4.1- le fondamental laryngé ou pitch (F_0)

La fréquence fondamentale constitue une caractéristique très importante de nombreux signaux environnementaux comme les sons de la parole. Elle correspond à la fréquence de vibration des cordes vocales lors de la production des voyelles ou des consonnes voisées. Elle génère des variations prosodiques, c'est à dire de mélodie et d'intonation, qui contribue à l'identification du sexe, de l'âge et de l'identité du locuteur, ainsi qu'à la signification du message prononcé. Le fondamental se trouve dans un registre grave, et différent selon la voix.

2.4.2. Les formants

Les formants sont des zones fréquentielles dont l'intensité est renforcée.

Chaque voyelle est reconnaissable par l'amplification d'harmoniques déterminés du son laryngé, appelés formants. La composition formantique de chaque voyelle est indépendante de la hauteur de son fondamental. Ainsi, que l'on soit un homme, une femme ou un enfant, on prononce les mêmes voyelles.

Les formants sont caractérisés par leurs fréquences de résonance et une bande passante. (Le tableau (2.1) représente les formants des voyelles de la langue française)

Voyelles françaises	1er formant [Hz]	2ème formant [Hz]	3ème formant [Hz]
[i]	280	2300	2950
[e]	350	1950	2550
[ɛ]	450	1800	2470
[a]	660	1350	2380
[ɑ]	620	1150	2250
[ɔ]	480	1050	2250
[o]	360	780	2230
[u]	290	850	2270
[y]	290	1800	2140
[ø]	360	1450	2290
[œ]	490	1380	2270
[ə]	480	1400	2200

Tableau 2.1 : valeurs des formants F1, F2et F3 des voyelles françaises [1]

Les caractaires de chaques formants sont :

- **Le 1er formant (F_1): La** zone formantique de F_1 est située entre 250 et 750Hz. Le premier formant F1 correspond à l'aperture de la voyelle (ouverture de la mandibule).
- **Le 2ème formant (F_2)** : La zone formantique de F_2 est située entre 750 et 2500Hz. C'est surtout ce deuxième formant qui est nécessaire pour l'intelligibilité du langage, et en particulier dans la zone située autour de 2KHz. Il exprime la position plus ou moins avancée de la langue.

- **Le 3ème formants (F_3) :** Le troisième formant est beaucoup moins caractéristique de la voyelle que le premier et le deuxième, car sa hauteur fréquentielle varie peu pour la majorité des voyelles.

Il est aussi à noter que le troisième formant donne de l'information sur l'arrondissement des lèvres.

Les valeurs des premiers et deuxièmes formants permettraient aux auditeurs d'identifier les voyelles orales. Leurs valeurs respectives rendent compte des propriétés du résonateur buccal et du résonateur pharyngal. Ce sont les formants les plus graves et il arrive que le premier formant se confonde avec le fondamental, particulièrement lorsqu'il s'agit de voix de femmes ou d'enfants dont la fréquence naturelle de la voix est plus élevée.

Les voyelles sont souvent représentées positionnées sur un plan, dont les axes sont les formants F1 et F2. Elles tracent alors un triangle dont les extrémités sont occupées par les voyelles "extrêmes", c'est-à-dire [**a**], [**u**], [**i**]. Ce triangle représente également, de manière assez grossière, les positions de la langue dans la bouche selon deux axes (figure 2.1) :

- Antérieur à postérieur ;
- Fermé à ouvert ;

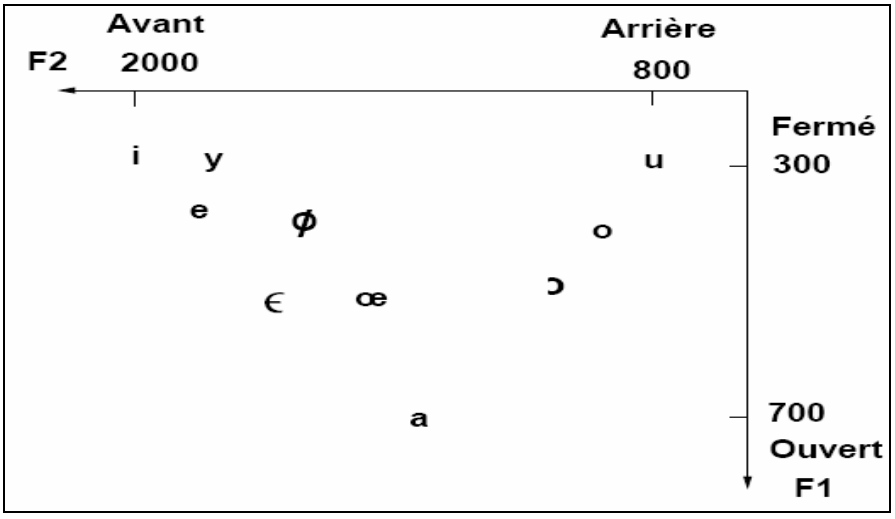


Fig. 2.1 : Représentation des voyelles dans le plan F1 - F2 [1]

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un **spectrogramme**. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. Ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants.

2.5. Représentations spectrales du signal de parole

Il existe plusieurs représentations spectrales du signal de parole, parmi elles nous citerons

2.5.1. Spectre obtenu par FFT

Tout son est la superposition de plusieurs ondes sinusoïdales. Grâce à la *FFT* , on peut isoler les différentes fréquences qui le composent. On obtient ainsi une répartition spectrale du signal (figure.2.2).

Les valeurs des formants sont calculées automatiquement dans le signal de parole au moyen d'un lissage spectral.

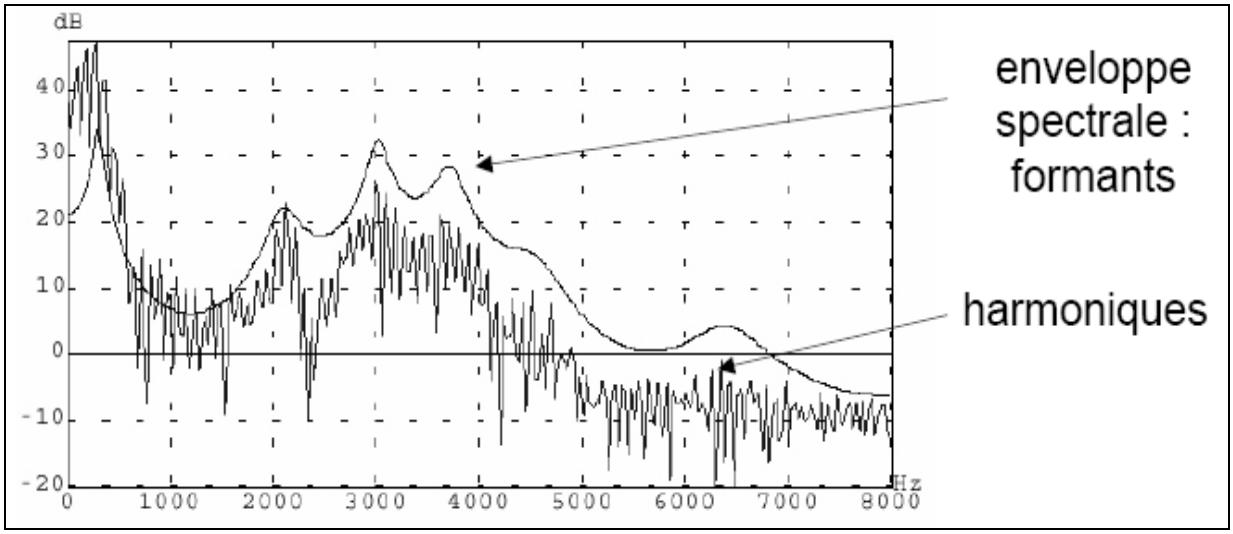


Fig. 2.2 : Spectre obtenu par transformée rapide de Fourier (FFT) [1]

2.5.2. Spectre obtenu par prédiction linéaire (LPC)

Le spectre obtenu par LPC est plus lisse et permet ainsi de repérer plus facilement les formants (figure 2.3).

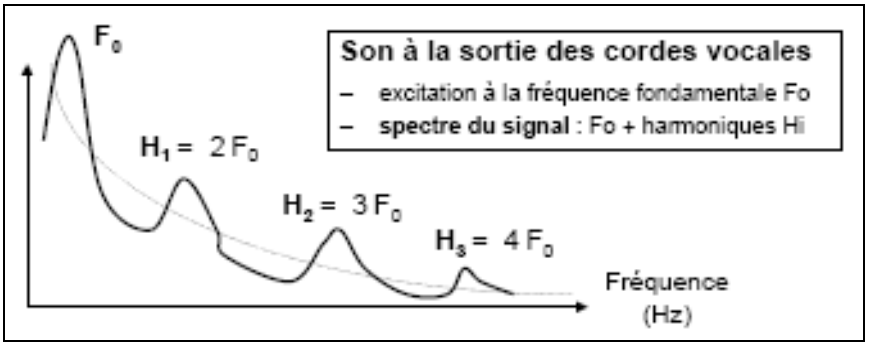


Fig. 2.3 : Spectre lissé obtenu par prédiction linéaire (LPC) [2]

Pour estimer les fréquences des formants, on calcule le spectre d'amplitude correspondant au modèle LPC et on cherche les fréquences correspondant aux pics spectraux.

2.5.3. Le Spectrogramme

Le spectrogramme est un outil de visualisation utilisant la technique de la transformée de Fourier et donc du calcul de spectres. Il a commencé à être largement utilisé en 1947, à l'apparition du sonagraphe, et est devenu l'outil incontournable des études en phonétique pendant de nombreuses années.

L'apparition de l'informatique puis d'écrans graphiques de bonne qualité a permis d'abandonner tout matériel comme le sonagraphe mais la technique du spectrogramme est encore aujourd'hui largement utilisée dans de nombreux domaines, du fait de sa simplicité de mise en oeuvre et des résultats intéressants qu'elle procure.

On parle de spectrogramme *à larges bandes* ou *à bandes étroites* selon la durée de la fenêtre de pondération. Les spectrogrammes à bandes larges 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. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bandes étroites 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.

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à tout instant (figure 2.4).

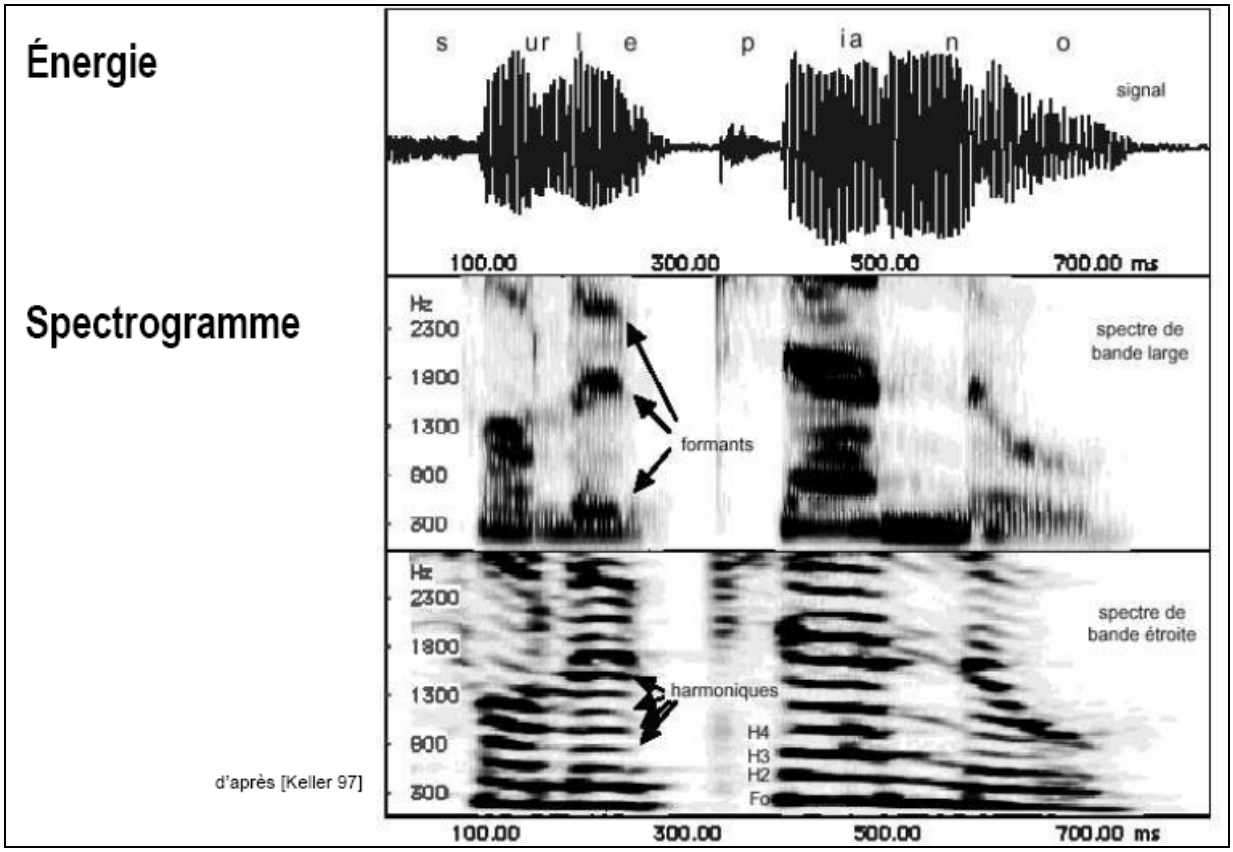


Fig. 2.4 : Spectrogramme et signal temporel [2]

2.5.4. Intérêts de la représentation fréquentielle du signal de parole

La représentation fréquentielle de la parole est d'une très grande importance dans le domaine de la communication parlée. Elle a permis l'extraction des paramètres pertinents du signal de parole comme la fréquence fondamentale et les formants. Ces paramètres sont d'une importance capitale dans de nombreux domaines comme :

- les différentes méthodes de synthèse ;
- la reconnaissance automatique de la parole et du locuteur ;
- l'identification automatique des langues ;
- Et bien d'autres domaines ;

2.6. Présentation de quelques méthodes de prétraitement et traitement du signal de parole

2.6.1. Les méthodes de prétraitement du signal vocal

Avant d'être analysé, le signal analogique de parole doit subir un pré traitement qui se présente suivant trois étapes : échantillonnage, pré accentuation et fenêtrage.

2.6.1.1. Echantillonnage :

C'est un processus qui permet de transformer un signal continu en un ensemble de valeurs discrètes affectées aux instants Ti. On dit qu'on a numérisé ou discrétiser le signal analogique, c'est une représentation discrète.

D'après le théorème de Shannon, la perte d'information entre le signal continu et le signal discret correspondant est quasiment nulle si et seulement si la fréquence d'échantillonnage :

$$Fe \geq 2 \cdot F_{\max}$$

F_{max} est la fréquence maximale du spectre du signal

Pour le signal de parole, le choix de la fréquence d'échantillonnage résulte d'un compromis. Son spectre peut s'étendre jusque 12khz ; il faut donc en principe choisir une fréquence *Fe* égale à 24khz au moins. Cependant le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtre préalable.

Cependant, *Fe* peut être choisie suivant la technique utilisée. *Fe* peut varier de 6khz à 16khz pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole. Par contre pour le signal audio (parole et musique) ,on exige une bonne représentation du signal jusque 20khz.**[4]**

L'opération inverse d'échantillonnage est dite interpolation (Fig. 2.5).

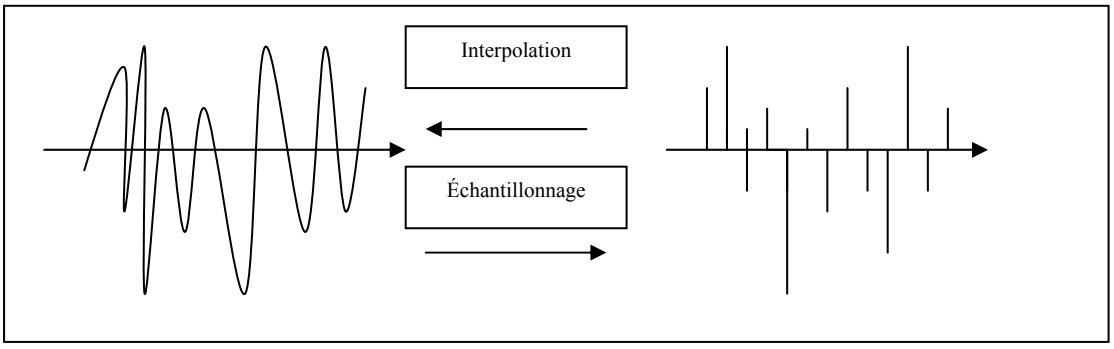


Fig. 2.5 : l'échantillonnage et l'interpolation d'un signal [1]

2.6.1.2. Pré accentuation

Le spectre du signal en sortie présente une atténuation de –6db/oct due aux influences de la source d'excitation et des lèvres. Pour compenser cette atténuation, on introduit le signal vocal dans un filtre de pré accentuation afin d'égaliser les aigus toujours plus faibles que les graves.

Ce filtre est défini par une fonction de transfert dont la transformée en Z est **:**

$$X(z) = 1 - a \cdot z^{-1} \qquad (2.1)$$

Le paramètre d'accentuation '*a*' est tel que : $0.90 < a < 0.98$
en pratique en choisie $a = 0.95$

2.6.1.3. Fenêtrage

La parole est un phénomène non stationnaire, c'est à dire, ses propriétés statistiques changent continuellement dans le temps. Cependant, l'observation du signal de la parole indique qu'il n'évolue pas ou peu sur des durées de quelques millièmes de secondes. On peut donc considérer ce signal comme étant stationnaire durant ce temps qu'on appellera fenêtre (stationnarité locale) [1], [13], [15].

Le fenêtrage consiste à délimiter la durée de ce dernier en le multipliant par une fenêtre allant de 20 à 30ms. En effet, sur cette durée, on estime que le signal n'a pas le temps de varier et conserve au moins une durée de la période du fondamental.

Ces fenêtres doivent être glissantes de manière à conserver les échantillons importants à traiter et se recouvrir afin de diminuer la perte d'informations aux bords de ces dernières.

Il existe plusieurs sortes de fenêtres: Hamming, Hanning, rectangulaire et autres.

La fonction générale de la fenêtre, peut s'écrire de la forme suivante :

$$W_H = \begin{cases} \alpha + (1-\alpha).\cos(2\pi n / N) : pour : |n| \le N/2 \\ 0 \cdots ailleur \end{cases} \qquad (2.2)$$

n : le n^{éme} échantillon.
N : le nombre d'échantillons.

On obtient la fenêtre de :

Hanning ;	Pour : α=0.5
Hamming ;	Pour : α=0.54
Rectangulaire.	Pour : α=1
Papoulis.	Pour :α=0

Le fenêtrage de type Hamming est le plus utilisée en parole, car les lobes latéraux de son spectre sont plus faibles que ceux des autres types de fenêtres et le lobe principal est deux fois plus large (Fig. 2.7) [1].

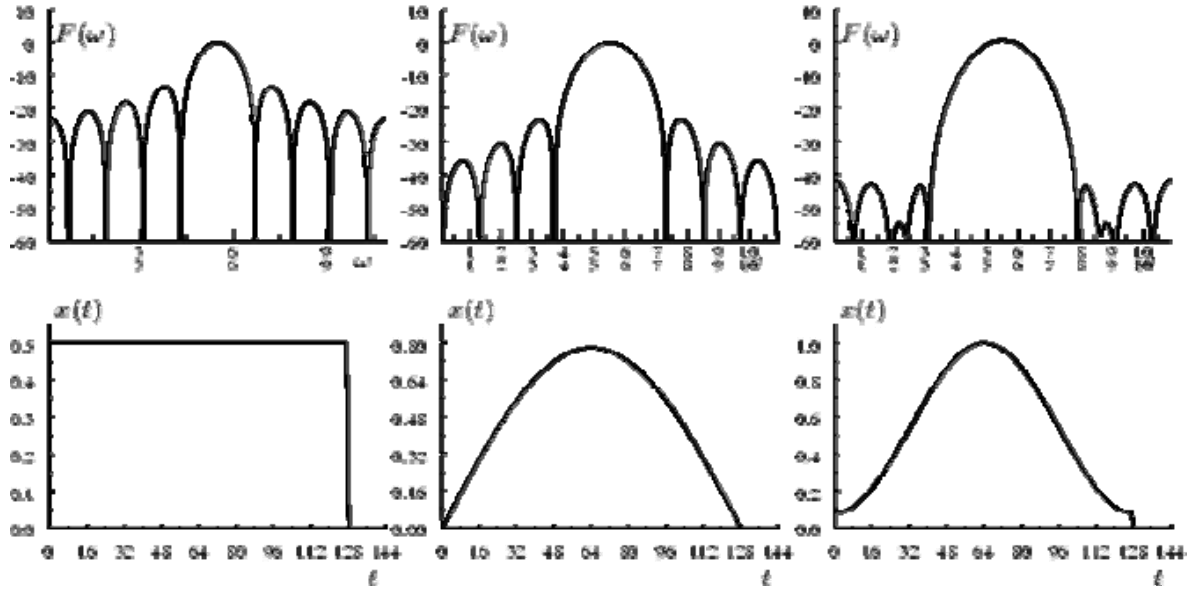


Fig. 2.6 : Lobe principal et lobes latéraux des fenêtres rectangulaire, de Papoulis et de Hamming (échelle logarithmique)[1]

La fenêtre de Hamming est définie par :

$$W_H = 0.54 + 0.46 \cos 2\pi nt \qquad (\text{avec } \alpha = 0.54) \qquad (2.3)$$

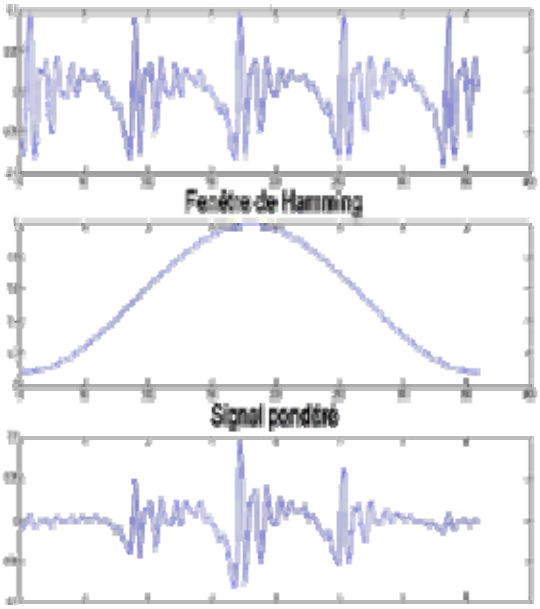


Fig.2.7: Signal, fenêtre de Hamming et signal pondéré par cette fenêtre

2.6.1.4. Recouvrement des fenêtres

Le signal issu du microphone est d'abord amplifié, filtré par un filtre anti-repliement et échantillonné par blocs de 16 à 32ms à des fréquences variant de 8KHz à 16KHz. [15]. Le signal est ensuite multiplié par une fenêtre d'analyse qui doit glisser sur tout le signal

Cette fenêtre doit glisser de manière à conserver les échantillons importants à traiter et se recouvrir afin de diminuer la perte d'informations aux bords de ces dernières.

Un recouvrement à moitié des fenêtres d'acquisition améliorées, de type Hamming (Fig.2.9), permet de raccorder les fenêtres en minimisant les « trous » et les « bosses » [16].

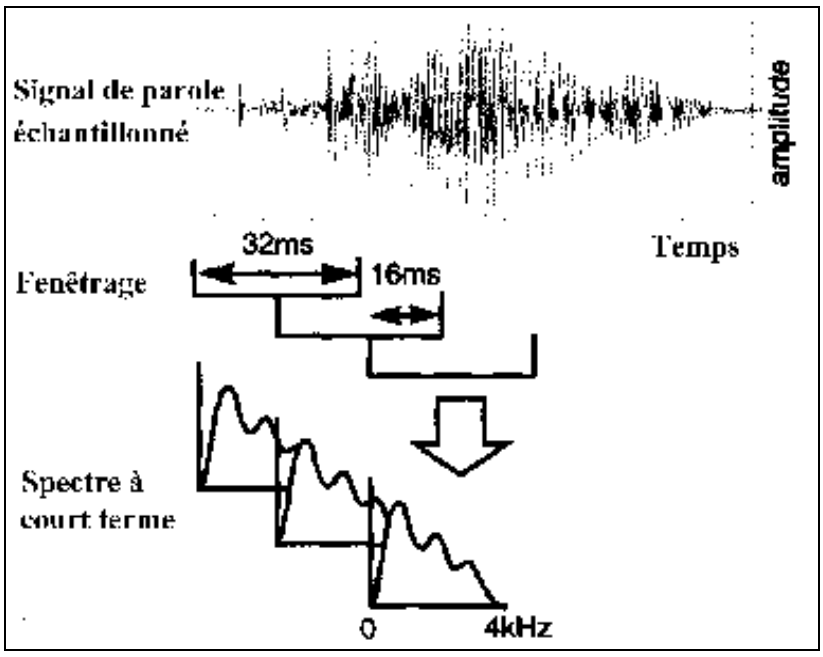


Fig. 2.8 : Effet du chevauchement des fenêtres de pondération

2.6.2 Technique de traitement du signal de parole

Dans cette présente étude nous avons utilisé la méthode par prédiction linéaire pour l'extraction des formants. Nous allons donc présenter cette méthode afin d'avoir une idée sur la façon avec laquelle nous avons calculé les formants.

Le « Linear Predictive Coding »ou LPC repose sur un modèle simple décrivant le comportement des organes vocaux lors de la synthèse d'un son. Ce modèle a été conçu à partir d'un modèle mathématique appelé « modèle autorégressif ». Nous allons donc commencer par présenter le modèle autorégressif afin de mieux comprendre le modèle LPC.

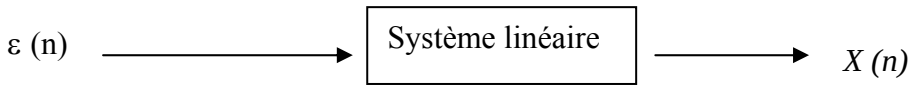
2.6.2.1. Le modèle autorégressif (AR)

Un processus AR peut être modélisé par la sortie d'un filtre linéaire et invariant dans le temps dont l'entrée est un bruit blanc ε (n) et tel que sa valeur à l'instant n ne dépend que de l'entrée ε (n) au même instant et des sorties aux instants précédents.

Son équation aux différences s'écrira :

$$X(n) + \sum_{i=1}^{p} a(i).X(n-i) = \varepsilon(n) \tag{2.4}$$

Sa fonction de transfère s'écrit comme suit :



$$H(z) = \frac{X(z)}{\varepsilon(z)} = \frac{1}{\sum_{P} a(i).z^{-i}} = \frac{1}{A(z)} \tag{2.5}$$

Notons que H(z) ne contient alors que des pôles et c'est pour cette raison que ce modèle est aussi appelé modèle tous pôles .

La condition de stationnarité du signal x(n) est équivalente à la condition de stabilité du filtre est que celle-ci n'est assurée que si les racines du polynômes H(z) (donc les coefficients a(i) du filtre), sont de modules inférieurs à 1.

2.6.2.2. Modèle AR et modèle de prédiction linéaire

Si l'entrée d'un modèle autorégressif est inconnue alors nous pouvons estimer ce dernier par un modèle ayant les mêmes propriétés appelées modèle de prédiction linéaire

Son équation aux différences $\hat{X}(n)$ où est l'estimé de *X(n)*, s'écrira alors :

$$\hat{X}(n) = -\sum^{p} \hat{a}(i).X(n-i) \tag{2.6}$$

$\hat{a}(i)$ les estimés des coefficients a(i) du filtre AR

P : ordre de prédiction.

On note *e(n)* l'erreur de prédiction définit comme suit :

$$e(n) = X(n) - \hat{X}(n). \tag{2.7}$$

De la relation (2.7) on trouve :

$$e(n) = X(n) + \sum_{i=1}^{p} \hat{a}(i).X(n-i)$$

$$\Rightarrow e(n) = \sum_{i=0}^{p} \hat{a}(i).X(n-i)..... \quad (2.8)$$

$$\hat{a}(0) = 1;$$

2.6.2.3. Pourquoi utilise-t-on le modèle autorégressif

On peut assimiler le mécanisme phonatoire à un système de transmittance :

Avec :

$$H(z) = G / A(z) \quad (2.9)$$

$$A(z) = \sum_{i=0}^{p} a(i).z^{-i} .. \quad (2.10)$$

a(0) = 1 et *A(z)* est un polynôme qui s'écrit comme suit:

$$X(n) + \sum_{i=1}^{p} a(i).X(n-i) = G.U(n) \quad (2.11)$$

G : le gain de ce système.

Dans le domaine temporelle :

$$X(z) = U(z).H(Z).. \quad (2.12)$$

Ce modèle de production d'un signal est appelé AR (autorégressif) avec :

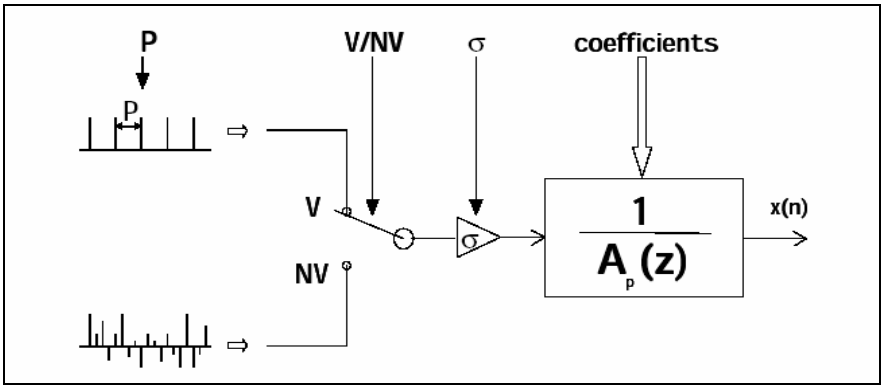


Fig. 2.9 : Modèle de production de la parole [1]

Si on suppose que notre système est excité par une excitation U (n) qui se présente comme :

- des **sons voisés** (ou sonores): l'excitation est un train périodique d'impulsions.
- des **sons non voisés**: l'excitation est un bruit blanc centré (de moyenne nulle et de variance nulle) (Fig.2.10**).**

La transmittance *H (z)* est celle d'un filtre polynomial, On définit le filtre inverse dont la transmittance est définie par :

$$A(z) = \sum_{i=0}^{p} a(i).\, z^{-i}.. \qquad (2.13)$$

$$a(0)=1$$

Ce filtre excité par le signal original, engendre en sortie l'erreur de prédiction.

2.6.2.4. Estimation des coefficients de prédiction linéaire

Le critère usuel pour l'estimation des coefficients du modèle de prédiction est la minimisation de l'erreur quadratique de ce dernier ou de la variance.

La variance est définit sous la forme suivante :

$$\sigma_e^2 = R_e(0) = \sum_{i,j=0}^{p} a(i).a(j).\overline{X(n-i).x(n-j)}. \qquad (2.14)$$

$$\sigma_e^2 = \sum_{i,j=0}^{p} a(i).a(j).R_x(i-j). \qquad (2.15)$$

La minimisation par rapport aux coefficients *a(i),* nous mène a calculé la dérivé partielle suivante par rapport à *a(i)* :

$$\frac{\partial \sigma_e^2}{\partial a(i)} = \sum_{j=0}^{p} R_x(i,j).a(j) = 0. \qquad (2.16)$$

D'ou :

$$\sum_{j=1}^{p} R_x(i-j).a(j) = -R_x(i).. \qquad (2.17)$$

L'autocorrélation est l'une des méthodes les plus utilisées des **LPC**, la variance de l'erreur de prédiction, sous forme quadratique :

$$\sigma_e^2 = [1, \underline{a}].R_{xx}^p \begin{bmatrix} 1 \\ \underline{a} \end{bmatrix}. \tag{2.18}$$

Avec :

$$a = [1, a(1), a(2),..., a(p)]$$
$$\underline{a} = [1, a(1), a(2),..., a(p)]$$

La méthode d'autocorrélation assure la stabilité du modèle AR et conduit à un système de matrice de Toeplitz (symétrique et égalité des éléments diagonaux de la matrice) qui s'écrit :

$$R_{XX}(P) = \begin{bmatrix} R_{XX}(0) & R_{XX}(1) & . & . & . & R_{XX}(p) \\ R_{XX}(1) & R_{XX}(0) & . & . & . & R_{XX}(p-1) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ R_{XX}(p) & R_{XX}(p-1) & . & . & . & R_{XX}(0) \end{bmatrix}. \tag{2.19}$$

On pose :

$$R_x = [R_{xx}(1), R_{xx}(2),..., R_{xx}(p)]$$

On écrit dans ce cas :

$$R_{xx}(p) = \begin{bmatrix} R_{xx}(0) & R_x \\ R_x & R_{xx}^{(p-1)} \end{bmatrix}. \tag{2.20}$$

On aura donc d'après la forme quadratique de la variance de l'erreur :

$$\sigma_e^2 = \sigma_x^2 + 2.\underline{a}'.R_x + \underline{a}'\, R_{xx}^{(p-1)}\, \underline{a} \tag{2.21}$$

$$R_{xx}^{(p-1)}\, \underline{a} = -R_x. \tag{2.22}$$

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(1) & R_{xx}(2) & . & R_{xx}(p-1) \\ R_{xx}(1) & R_{xx}(0) & R_{xx}(1) & . & R_{xx}(p-2) \\ . & . & . & . & . \\ R_{xx}(p-1) & . & . & . & R_{xx}(0) \end{bmatrix}. \begin{bmatrix} a(1) \\ a(2) \\ . \\ a(p) \end{bmatrix} = -\begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ . \\ R_{xx}(p) \end{bmatrix}.$$

2.7. Conclusion

Les généralités sur la parole citées dans le « chapitre I » nous ont permis de mieux comprendre le fonctionnement de la production de la parole ainsi que la complexité de ce signal, d'ou la multitude de méthodes et techniques existantes dans ce domaine. Notons que les différentes méthodes numériques de traitement citées dans cette partie ont toutes été utilisées dans cette étude.

Chapitre 3 :

Les méthodes d'extraction De Formants

3.1. Introduction

D'après l'analyse du signale de la parole que nous avons présenté dans le seconde chapitre Donc il est intéressant de voir comment procéder pour extraire les formants à partir du signal acoustique de la parole, et ces déférents méthodes Ce qui fera l'objet du ce chapitre.

Et nous avons commencée dans ce chapitre par présenté un état de l'art relatif aux méthodes de calcul des formants, et les méthodes d'extraction de formants, et nous avons citer à la fin quelques outils logiciel qui permettent de visualiser la forme d'onde et le spectrogramme d'un signal ou de paroles.

3.2. État de l'art

Il convient de rappeler que la définition même des formants est disputée. Les formants ont été définis au départ comme des maxima du spectre vocalique :

"The spectral peaks of the sound spectrum are called formants." [Fan60].

Ainsi définis, leur estimation exacte est impossible car dépendante de facteurs comme la position de la fenêtre d'analyse vis à vis des périodes de la fréquence fondamentale, le degré de lissage du spectre, etc. Une définition plus rigoureuse des formants est en tant que fréquences de résonance de la fonction de transfert acoustique du conduit vocal :

"Traditionnellement, dans le domaine de la parole, les termes de formant et de pôle sont employés de manière interchangeable ..." [13]

Il existe plusieurs méthodes pour la mesure directe de la fonction du transfert du conduit vocal et, en conséquence, de ses résonances. Cependant, ces méthodes sont complexes et demandent une mise en oeuvre spéciale (excitation extérieure). Même les plus rapides comme temps de mesure imposent des contraintes trop lourdes pour l'étude de la parole en situation. Pour notre part, nous nous limitons donc au calcul des résonances de manière classique (à partir du signal vocal pré-enregistré).

Les méthodes utilisées pour la détermination des trajectoires formantiques se séparent selon deux approches principales spectrales et directes

3.2.1. Méthodes spectrales.

La caractéristique commune de ces méthodes est l'analyse fréquentielle par trame. Elles comportent deux étapes souvent découplées :

1. *L'analyse locale* est l'étape qui aboutie au calcul des candidats pour les formants dans chaque trame.

2. *L'analyse dynamique* où le suivi de formants proprement dit est l'étape pendant laquelle les candidats sont reliés pour former une trajectoire.

Ce découplage est un avantage car l'analyse locale est généralement très précise L'inconvénient tient à la complexité de la deuxième étape.

3.2.2. Méthodes directes

Ces méthodes ne passent pas par une analyse spectrale typique. Elles évaluent les paramètres d'un modèle autorégressif directement à partir du signal. Étant des méthodes récurrentes, la recherche des candidats et la construction des trajectoires sont simultanés.

Leurs inconvénients sont liés au coût de calcul et au manque de précision. Elles peuvent travailler aussi par trame mais alors les candidats choisis dans une trame sont employés comme valeurs initiales pour la recherche des candidats dans la trame suivante.

La première étape des méthodes spectrales, l'extraction des formants, a généré une riche littérature. Nous nous limiterons ici à une énumération non exhaustive des principales solutions adoptées car nous considérons que même les plus basiques ont des performances suffisantes suivant le but recherché :

- racines du polynôme LPC [18, 19] ;
- maxima du spectre lissé par le calcul du cepstre ;
- zéros de la fonction de retard de groupe [20] ;
- centroïdes [21] ;
- "Multiband energy demodulation" [22]. Cette méthode récente s'affranchi du modèle autorégressif. Elle utilise un jeu de filtres de Gabor pour fournir les fréquences et les bandes passantes des formants ;
- méthodes coopératives (un vote combinant plusieurs solutions) [23, 24] ;

Par contraste, le problème du suivi proprement dit a connu moins de solutions originales :

- S. McCandless [18] a proposé en 1974 une première solution. Elle est basée sur un nombre de règles heuristiques pour choisir les candidats d'une trame en fonction des candidats retenus dans la trame précédente. Une description détaillée d'une version améliorée peut être trouvée dans [19]. Des versions de cette méthode ont été utilisées par la grande majorité des autres travaux.
- Une méthode significativement plus élaborée a été proposée au CRIN par Y. Laprie [25]. Les améliorations concernent la recherche des conflits (donc une segmentation

plus fine de la zone du suivi) et une optimisation globale (sur toute la longueur du suivi) dans la phase du lissage.

Cette méthode peut être décomposé en trois étapes :

- Proposition des trajectoires. Un nombre de trajectoires élémentaires continues sont construites par détection des contours dans le spectrogramme. Ces trajectoires élémentaires sont alors étiquetées en termes de formants et un ensemble de connections possibles sont évaluées. Pour gérer l'explosion combinatoire, une note est affectée à chaque connection et les n-meilleures solutions sont retenues.
- . Une régulation des trajectoires est effectuée par la méthode des contours actifs [35]. En donnant de la rigidité aux trajectoires, elle ajuste le compromis entre leur degré de lissage et leurs proximités par rapport aux maxima spectraux.
- La décision finale est prise en réévaluant les notes de chaque solution proposée auparavant.

- Une famille d'algorithmes de suivi a été proposée par G. Kopec en 1985 [26].

  Dans sa forme la plus simple, l'algorithme associe à un spectre une probabilité de présenter un certain formant (étape de détection) et une série de probabilités pour que ce formant se trouve dans une suite d'intervalles de fréquence (étape d'estimation). Les deux étapes sont implémentées, séparément ou conjointement, par des chaînes de Markov cachées (HMM) pour un formant ou pour un jeux de formants. À la suite de l'apprentissage des HMM, l'algorithme est capable de choisir pour une suite de spectres voisés donnée la séquence des fréquences des formants la plus probable. L'optimum ici est global, sur toute la séquence des spectres, non pas local à une paire de trames. La méthode présente les avantages et les inconvénients de l'apprentissage des HMM. L'avantage important est qu'il élimine tout critère heuristique de décision tout en fournissant des statistiques intéressantes sur les formants dans le corpus d'apprentissage. Mais la performance est fort dépendante de la qualité et de la dimension de ce corpus. D'autant plus que, une fois ce coûteux apprentissage fini, il n'y a plus aucune possibilité de modifier le seuil de détection ou la précision de l'estimation. Mais la présence du corpus d'apprentissage permet de chiffrer exactement l'erreur moyenne sur les trajectoires.

  G. Rigoll a proposé une série de méthodes directes. La première utilise un estimateur de Kalman pour évaluer les paramètres d'un modèle proche du modèle autorégressif appelé

FLPC [27]. La deuxième améliore le coût de calcul en utilisant un estimateur "quasilinéaire" [28]. À chaque instant d'analyse (même pour chaque échantillon de signal) une correction des valeurs des formants obtenus à l'instant précédent est calculée. Ainsi, pour un signal stationnaire, les valeurs des paramètres convergent, dans un processus itératif, vers les valeurs réelles des formants.

Enfin, M.J. Hunt propose en 1985 un algorithme pour la comparaison des spectres [29] basé sur une anamorphose fréquentielle (Dynamic Frequency Wraping). Cet algorithme peut être utilisé pour mettre en correspondance les formants d'une trame avec ceux de la trame suivante dans un suivi.

Outre l'utilisation à bon escient de connaissances a priori, autre problème fondamental du suivi des formants est l'évaluation objective des résultats. La comparaison simple avec des suivis de référence est rarement utilisée. La principale raison en est le coût du suivi effectué à la main. Si un corpus de test et un autre d'apprentissage de dimensions conséquentes doivent être passé en revue par l'expert, l'utilité même du suivi automatique est mise en doute. Le plus souvent, l'évaluation est subjective ou faite sur un petit corpus de test. Pendant la conception de notre algorithme, nous avons cherché des solutions pour que les traces de l'évolution fournies par l'algorithme soient les plus informatives pour l'évaluation.

3.3. Réalisation

Les problèmes rencontrés dans l'implémentation d'un suivi de formants se résument à deux compromis fondamentaux :

- Le choix des candidats pour la fréquence des formants. Plus la résolution spectrale est fine moins les omissions sont nombreuses mais plus grand est le risque des erreurs par insertion.
- La rigidité des trajectoires. Plus la réactivité du suivi aux candidats observés dans la trame courante est grande, plus grand est le risque de "dérailler" de la vraie trajectoire. Mais, si cette réactivité est trop faible, des variations utiles peuvent être gommées et les chances de récupérer après un "accident" de suivi sont réduites.

Les deux cas représentent donc des exemples typiques du problème plus général du réglage du compromis Pour l'extraction des candidats nous avons retenu une méthode. Elle consiste à calculer Les racines du polynôme approximant la fonction de transfert du conduit, car c'est la plus simple et nous a donné de bons résultats.

L'information portée par le signal de la parole est essentiellement contenue dans les formants. Il est donc nécessaire d'affranchir ces derniers de signaux indésirables tels que la source ou le « bruit »" de numérisation (enveloppe spectrale en dent de scie).

Les systèmes de prétraitement permettent donc d'améliorer la représentation des formants par « lissage »", et de fournir, à l'outil de comparaison, des vecteurs constitués de coefficients pertinents.

Les méthodes les plus courantes pour le traitement du signal de la parole sont les analyses spectrales réalisées soit par transformée de Fourier à court terme, soit par prédiction linéaire ou soit par évaluation des coefficients cepstraux. D'autres méthodes existent, telles que la représentation par formes d'ondes et la représentation en ondelettes de Morlet [30].

3.3.1. L'analyse spectrale par Transformée de Fourier à Court Terme (TFCT)

Le signal de la parole, échantillonné et préaccentué dans les hautes fréquences, est prélevé par une **fenêtre temporelle glissante** de type Hamming (Fig.3.1).

Puisque c'est la largeur de cette dernière qui détermine la résolution spectrale de l'analyse, il apparaît donc un **conflit** entre la résolution **temporelle** et la résolution **fréquentielle**, comme l'indique l'exemple suivant :

Pour une f_e = 12 kHz et N=256 $\Delta(f) = f_e/N = 47Hz \rightarrow$

et w(t)=N.Te=21ms,

N=40 $\rightarrow \Delta(f) = 300Hz$

Et w(t) =3.3ms.

On peut conclure qu'une analyse en bande étroite, d'une résolution fréquentielle de 50Hz environ, permet une bonne représentation de la structure harmonique du signal. Mais cette dernière se fait au détriment de la résolution temporelle qui se traduit par une intégration des évolutions temporelles rapides. Une transformée de Fourier est ensuite calculée pour chaque valeur de décalage de la fenêtre.

TFCT (temps continu) $$S(t,f) = \int s(\tau).w(\tau - t).e^{-i2\pi f\tau}.d\tau \qquad (3.1)$$

TFCT (temps discret) :

$$S(n,k)=\sum_{r} s(r)w(r-n).e^{-i.\frac{2\pi k r}{N}} \qquad (3.2)$$

Avec $k \in [0;N-1]$ et N le nombre de points prélevés.

En prenant le carré du module de la TFCT, on obtient alors un spectrogramme représentant la distribution énergétique dans le plan temps fréquence, puisque pour chaque instant « n », on dispose alors de l'énergie associée aux fréquences k=0, ..., N-1.

Il suffit alors d'appliquer un filtrage suivant une échelle de Mel ou de Bark, fréquemment utilisée en reconnaissance vocale, pour obtenir un superbe sonagraphe numérique . En conclusion, la TFCT présente l'avantage de vecteurs de paramétrisation constitués d'une vingtaine de composantes obtenus avec un faible volume de calcul donnant une image proche de celle du sonagraphe [30] (Fig.3.1).

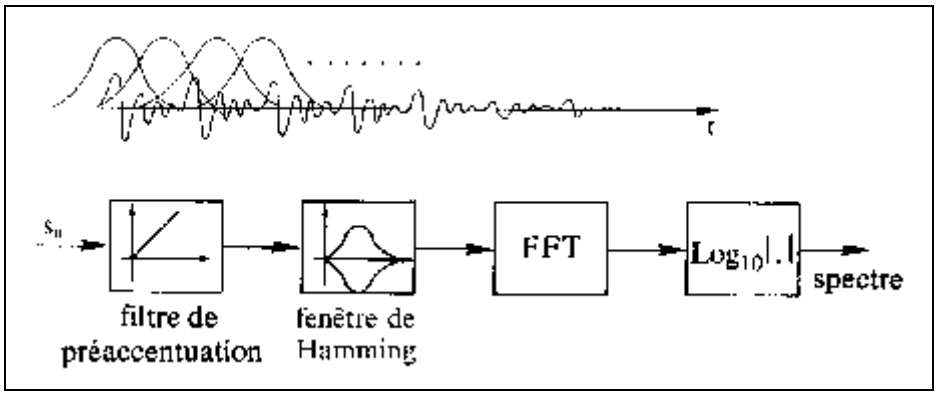


Figure 3.1 : Fenêtre temporelle glissante de type Hamming [30]

3.3.2. L'analyse par prédiction linéaire (LPC) et estimation des formant

La prédiction linéaire est une technique qui s'applique directement après l'échantillonnage et la quantification du signal de la parole. C'est une méthode permettant l'approximation du signal par un modèle. Pour cela, elle considère l'appareil phonatoire comme un modèle source-filtre linéaire.

Par conséquent, un échantillon de parole peut-être prédit par une combinaison linéaire d'un certain nombre d'échantillons précédents,

l'équation aux différences est donnée par :

$$x(m) = a_1 x(m-1) + a_2 x(m-2) + \ldots + a_p x(m-p) + e(m)$$

$$or$$

$$x(m) = \sum_{k=1}^{p} a_k x(m-k) + e(m) \qquad (3.3)$$

Où les $\mathbf{a_k}$ sont les coefficients de prédiction, **p** l'ordre du filtre de prédiction, **e(m)** représente l'erreur de prédiction du modèle LPC et **x(m)** le signal temporel à l'instant m.

La transformée en z de x(m) conduit vers la fonction de transfert $H_{lp}(z)$ du filtre tel que :

$$X(z)\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) = E(z) \rightarrow \frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = H_{LP}(z) \qquad (3.4)$$

Ce filtre est un filtre tout pôle. Le numérateur de sa fonction de transfert est une constante. L'entrée du filtre correspond à l'erreur du modèle e(m), Les fréquences correspondantes aux pôles sont des candidats aux formants qui peuvent être extraits à partir des racines z_i du polynôme $H_{lp}(z)$.

Si z_1 est une racine de $H_{lp}(z)$, alors le formant qui lui correspond (sa fréquence **F**, son amplitude **M** et sa bande passante BW) pourra être calculé par les expressions suivantes :

$$z_1 = \exp(a + j\omega_p) = \exp(a)\exp(j\omega_p) \qquad .(3.5)$$

$$F = f_s \times \omega_p / 2\pi$$

$$BW \approx f_s \times a / \pi$$

$$M = H_{LP}(e^{j\omega_p})$$

Où f_s correspond à la fréquence d'échantillonnage du signal.

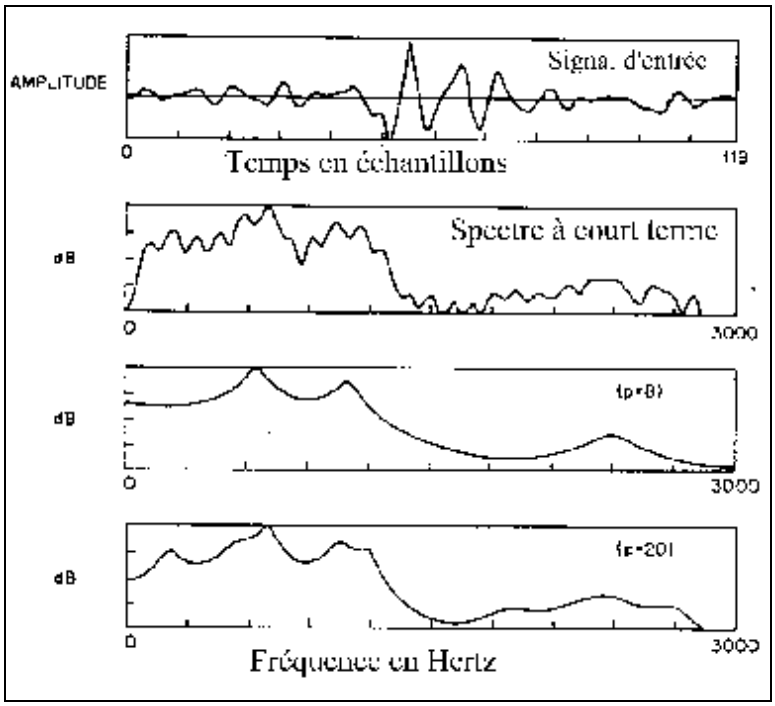


Figure 3.2 spéctre de la voyelle[a] [30]

3.3.3. L'analyse par évaluation des Coefficients Cepstraux

3.3.3.1. Le cepstre

Comme nous l'avons vu précédemment, le signal de la parole est porteur de différentes informations, dont le fondamental, qui ne sont pas toutes nécessaires à la compréhension du message. Ce signal est le résultat d'une **multiplication** du spectre d'entrée par la réponse fréquentielle du conduit vocal. Dans ce cas, **tout** le contenu spectral du signal source, modifié par le conduit vocal, ne peut pas être éliminé par filtrage, puisque précisément cette opération de filtrage est adaptée à l'extraction d'information localisée en fréquence. Une alternative consiste à utiliser la méthode « cepstrale ». En effet, celle-ci est basée sur l'opérateur logarithme qui transforme un produit en addition, ce qui permet alors de séparer la source du conduit vocal.

Le cepstre est défini comme la transformée de Fourier inverse du module d'un spectre exprimé en échelle logarithmique.

$$y_c(k) = TF^{-1}\left[Log_{10}\left|Y(f)\right|^2\right] \tag{3.6}$$

où : « c »" indique le domaine cepstral et « k » la variable cepstrale

Après la séparation du conduit vocal et de la source (Fig. 3.3-a), on procède à une opération de « filtrage » (Fig.3.3-b) pour éliminer toutes les fréquences dépassant un certain seuil (voir seuils de fréquences a, b et c sur la figure (3.4). Le retour au domaine fréquentiel s'effectue par une FFT dont on calcule ensuite l'exponentielle (Fig. 3.3-c). On obtient alors un spectre lissé constitué de l'information formantique [31] ,(Fig. 3.3).

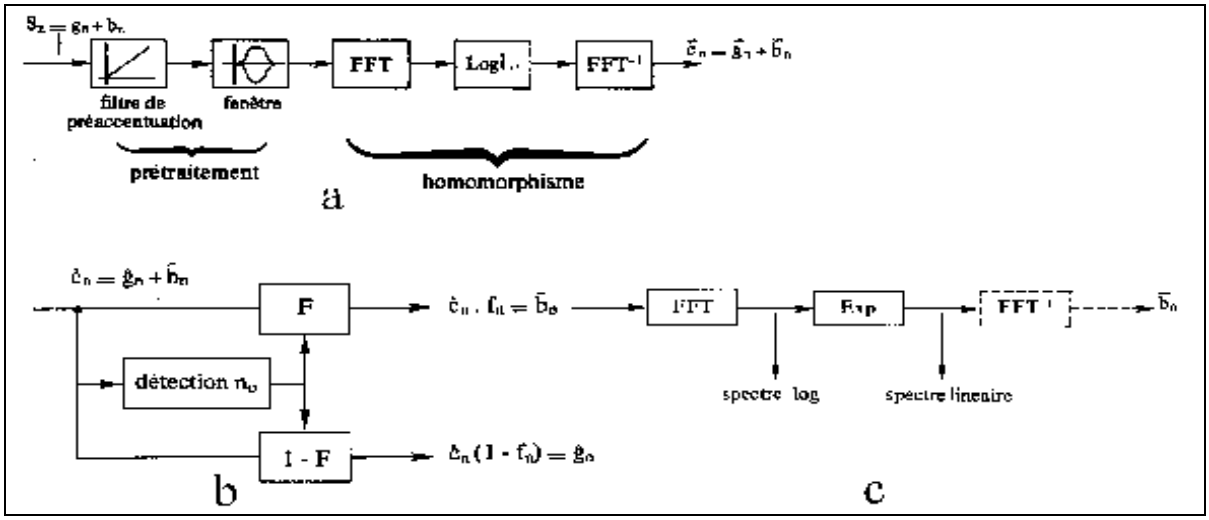


Figure 3.3 : a : Séparation du conduit vocal et de la source , b : Filtrage ,
C : Retour au domaine fréquentiel par FFT [13]

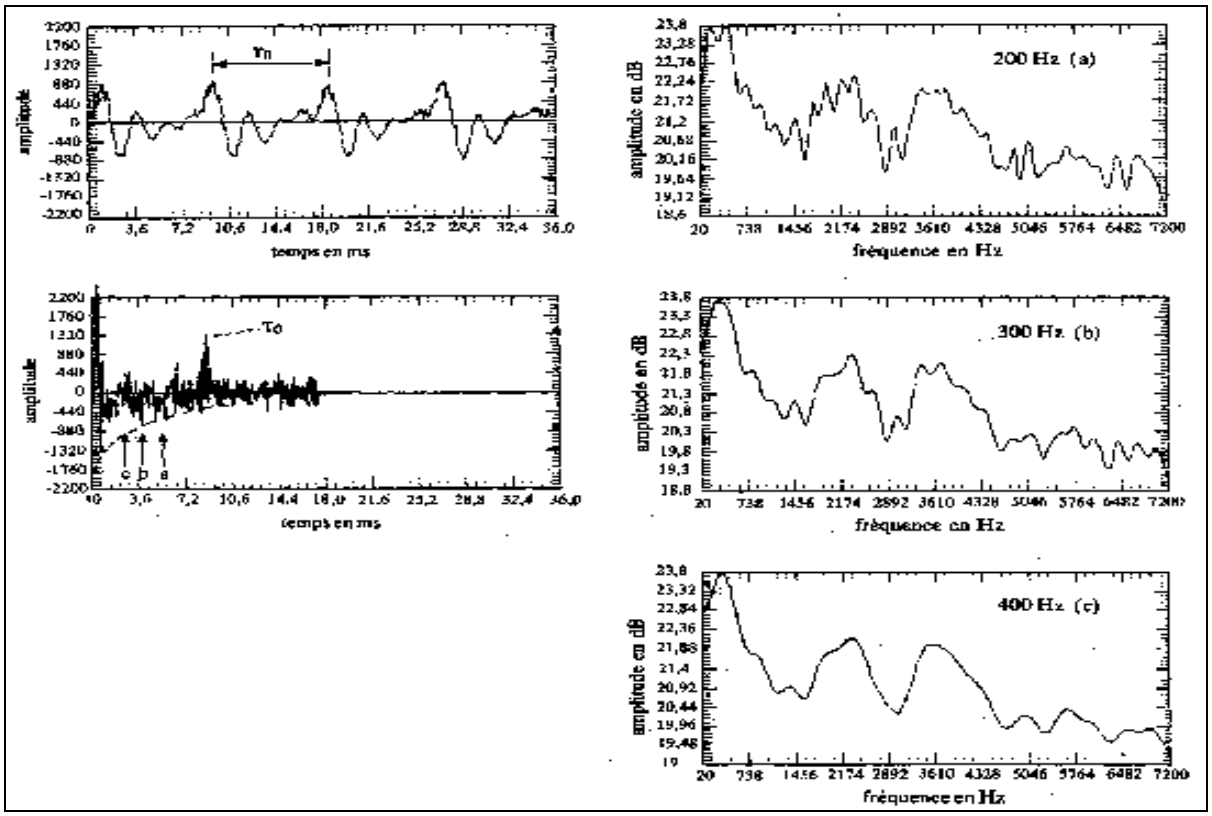


Figure 3.4 : Seuils de fréquences [13]

3.3.3.2. L'analyse MFCC

L'analyse MFCC consiste en l'évaluation de Coefficients « Cepstraux » à partir d'une répartition Fréquentielle selon l'échelle des Mels [13]. Mais il faut noter, dans l'analyse MFCC décrite ci-dessous, que les coefficients cepstraux obtenus ne correspondent pas exactement à la définition du cepstre défini précédemment. [13]

Cette technique est composée dans un premier temps d'une analyse spectrale ou d'une analyse LPC. L'étendue dynamique du spectre de puissance ainsi obtenu permet sa compression logarithmique afin de s'accorder avec la perception d'intensité de l'oreille humaine.

Pour finir, une transformée discrète en cosinus (DCT) est appliquée afin d'obtenir les Nc coefficients cepstraux.

$$C(k) = \sum f(i).\cos\left(\frac{2\pi.i.k}{N} + 0.5\right) \tag{3.7}$$

Avec $k \in [0, Nc]$ et f(i) la ième des Nf sorties log du banc de filtres.

Ces derniers, ajoutés à des coefficients représentant l'information énergétique et de dérivées première et seconde, forment alors un vecteur acoustique. (Figure 3.5) [30].

Les avantages de cette méthode sont les suivants :

- Le nombre de données par vecteur est réduit. En pratique, pour 24 filtres (Nf=24), il a été montré que 12 coefficients cepstraux (Nc=12) suffisent pour représenter l'information, puisque l'enveloppe de la DSP varie lentement.
- Les valeurs des vecteurs sont relativement décorellées entre elles, ce qui est idéal pour la reconnaissance de forme. [30]

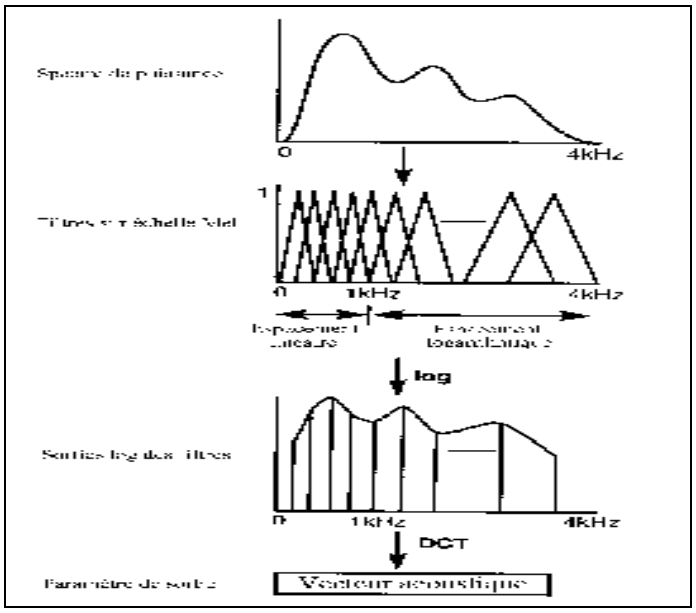


Figure 3.5 : Analyse MFCC [13]

3.4. Les outils d'analyse

Afin de réaliser la phase de segmentation semi-automatique nous avons besoin d'un outil d'analyse qui facilite cette phase.

Nous pouvons citer quelques outils logiciel qui permettent de visualiser la forme d'onde et le spectrogramme d'un signal ou de paroles, d'éditer et d'aligner des transcriptions orthographiques et phonétiques sur ce signal, tels que PRAAT, CLAN, speech analysis, Goldwave, Cool Edit , etc....

3.4.1. Le logiciel CLAN (Computertzed Language ANaalysis)

CLAN dont la traduction serait Analyse du langage par ordinateur. Chaque ligne/paragraphe correspond par exemple à une prise qui peut étre alignées avec le signal audio ou audiovisuel. CLAN facilite aussi l'alignement temporel entre transcription et signaux audio ou vidéo. II permet de communiquer des segments sonores au logicield'analyse. De plus, CLAN permet l'insertion et la représentation de catégories syntaxiques, morphologiques et phonétiques sous formes d'annotations interlinéaires. Le logiciel offre un nombre important de routines d'analyse linguistique, de recherche et de calcul statistique. CLAN offre également de nombreuses possibilités d'analyses automatiques sur les données transcrites telles que le calcul de fréquence, la recherche de mots, les analyses interactionnelle et **morphosyntaxique.** Il permet de communiquer des segments sonores au logiciel d'analyse et **d'annotation** phonétique.

IL est gratuit et accessible sur plusieurs environnements informatiques : Mac Classic, Mac Carbon (OSX), Windows, Unix [10].
Enfin, aucune possibilité d'import n'est envisageable alors qu'il est possible d'exporter un segment sonore vers PRAAT mais aucune transcription ou annotation n'est fournie dans la foulée : seul le son est exporté.

3.4.2. Le logiciel PRAAT

Le logiciel PRAAT a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam.
PRAAT est un logiciel d'analyse et de transcription phonétique (spectre, intonation, intensité etc.). Le logiciel comporte aussi des fonctionnalités importantes pour l'enregistrement, pour la manipulation et pour la synthèse de sons, pour la création d'algorithmes d'apprentissage, pour t'analyse statistique, ainsi que pour diverses expériences auditives. Praat est hautement portable, configurable et programmable [11]. En linguistique interactionnelle, le logiciel est utilisé pour divers types de transcription alignée de données sonores (éventuellement extraits d'une vidéo), pour aligner des transcriptions déjà réalisées en texte brut, mais aussi pour l'analyse et la transcription prosodiques. Avec ce logiciel, il est possible :

- d'enregistrer des fichiers audio qui pourront ensuite être analysés .
- de transcrire, d'étiqueter et de segmenter des données audio (que les enregistrements aient été effectués sous Praat ou qu'ils proviennent d'autres fichiers, au format WAV, par exemple).
- d'effectuer des analyses phonétiques et acoustiques au niveau segmentai (spectrogramme, analyse de formants, sonagrammes, etc.) et au niveau suprasegmental (pitch ou Fo, intensité et durée).
- de manipuler et modifier le signal de parole (utilisation de filtres , modification des contours intonatifs et de la durée, etc.) .
- de faire de la synthèse de la parole (créer des stimuli audio, synthèse articulatoire, analyse -synthèse de données modifiées, etc.).
- de faire des analyses statistiques à partir des études phonétiques (analyses de covariances, etc.). Nous pouvons résumer les fonctionnalités de ce logiciel dans la figure suivante (Figure 3.6).

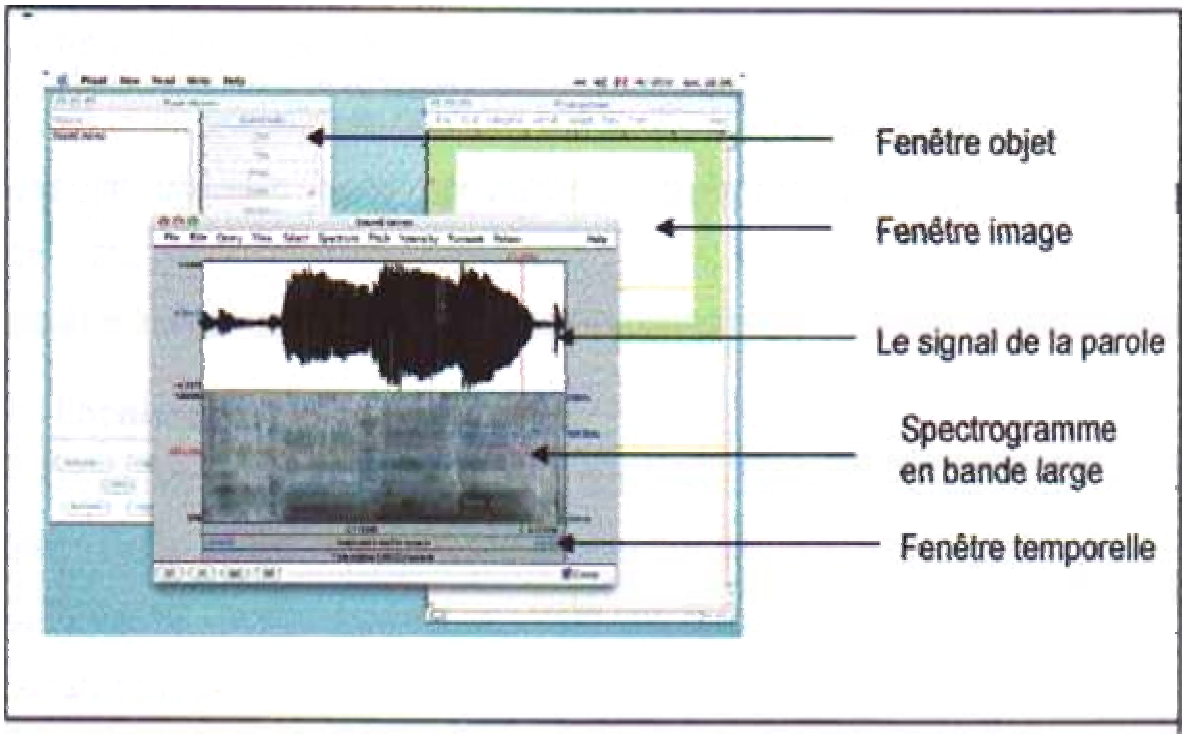


Figure 3.6 : Présentation du logiciel PRAAT

On peut aussi visualiser différentes courbes en surimpression sur le spectrogramme :

- La fréquence fondamentale : cochez « Show pitch » dans le menu « Pitch », et elle apparaît (c'est une courbe de couleur cyan). Sa valeur moyenne (Hz) s'affiche à droite.
- Les formants : cochez « Show formants» dans le menu « Formant », et ils apparaissent en pointillés rouges. Pour les afficher sur toute la longueur de la fenêtre, affichez la fenêtre « Formant Settings » du menu « Formant », et dans le champ « Maximum Duration », entrez la durée de la fenêtre, en secondes, à la place de la valeur initiale.
- Les périodes du signal sonore : cochez « Show pulses » dans le menu « Pulses ». Chaque période est représentée, sur l'enveloppe, par un trait bleu vertical

3.5. Conclusion

L'analyse par prédiction linéaire permet de passer d'un spectre échantillonné, donc « bruité »" à une représentation spectrale continue et « lissée ». La détection des formants en est alors plus aisée (voir figure 3.2).

Cette méthode présente l'inconvénient du choix du nombre de coefficients (8 à 14) à prendre en fonction de la fidélité par rapport au signal analysé. Notons que cette méthode de détection de formant a été adoptée dans cette étude car en plus de sa simplicité de mise en œuvre, elle procure des résultats satisfaisants.

Chapitre 4 :

MISE EN OEUVRE DU SYSTEME SAFAS

4.1. Introduction

La parole présente des caractéristiques ayant leurs origines dans les mécanismes de production.

L'objectif de notre travail consiste à réaliser un système d'analyse et de la reconnaissance automatique des sons fricatives non emphatique de l'Arabe Standard, ce qui concerne les sons suivant : س /[s]/ش/[š]/ز/[z]/ج/[ž]/. Il s'agit de trouver les propriétés nécessaires et suffisantes pour caractériser la classe des notre phonèmes par rapport à d'autres classes de sons.

Dans ce chapitre, la première partie nous intéresser aux l'analyse acoustique des sons étudies dans les différents contextes vocaliques nous permet d'extraire les paramètres pertinents et leurs effets sur les voyelles. En suit, nous expliquons la démarche expérimentale que nous avons appliquée, en exposant les grandes lignes introduites dans les étapes de l'élaboration de notre outil d'analyse. Nous examinons les algorithmes des différentes techniques d'analyse introduites dans notre logiciel.

4.2. Description des bases de données utilisées

4.2.1 Elaboration du corpus

Nous avons choisi, pour faciliter la segmentation, d'enregistrer un corpus de mots significatifs appartenants au vocabulaire arabe d'un locuteur masculin. Pour l'extraction de la totalité des phonèmes et diphones, et syllabe pris dans les trois positions (initiale, médiane et finale), et avec leurs trois voyelles (fatha, damma, et kasra), nous avons utilisé les enregistrements de prés de 72 phrases et expressions utilisant le vocabulaire arabe usuel. Nous justifions le choix de ce type de corpus (parole continue au lieu de l'utilisation de logatomes) par le fait qu'il est préférable d'étudier les segments dans un continuum vocal pour pouvoir prendre en considération les effets de coarticulation existants entres phonèmes.

4.2.2. Enregistrement de corpus

Les enregistrement du corpus on été effectué dans les conditions suivant :

- Les données sont échantillonnées sur 16kHz échantillons codés sur 16 bits.
- la chambre d'enregistrement est un peu bruyante.
- le type de parole : phrases continu en arabe.
- Le microphone utilisé : un microphone à large bande bidirectionnelle (il apport des bruit de l'autre direction).

- et les signaux acoustiques sont enregistrés en format (WAV).

- سمع صوتا غريبا	- شعر بالخوف
- سأل المعلم التلميذ	- كان شابا قويا
- سوف أذهب	- المشاهد
- مناسبة – مؤسسة	- عاش لمدة طويلة
- الحسنة تذهب السيئة	- الشعب الجزائري
- خمسة أسر	- الشعوب الحرة
- جلس يستمع إلى الراديو	- الشروق
- مارس الرياضة	- شوهد أمام النزل
- لبس ثوباً جديدا	- فراش الموت
- سبحان الله	- يناقش الأستاذ الموضوع
- من سور القرآن	- نوعية القماش عالية
- سعاد سلوكها غريب	- شفاء
- تقع سوريا في الشام	- شجار
- موسوعة	- تفشي الظاهرة
- السلطة القضائية مستقلة	- ظواهر المشينة
- استناد إلى نص	- تنتشر بسرعة
- يمارس السباحة	- يستحق الفوز
- الأساس الوحيد للقومية	- الاعتزاز بالهوية الوطنية
- تنعكس عليه	- الرموز الوطنية
- ستة و عشرون	- زيارة الأقارب
- سيرة ذاتية	- تنزل الثلوج
-حسب السن و الجنس	- بدون منازع
- و العرق يسيل	- جلس يستمع الراديو
- قالت له آسفة	- في ما جاء في البرنامج
- زعيم القبيلة	- بصوت جميل
- زار المريض	- شجرة الأرز
- مزارع	- المجاهد
- لابد لي من الزواج بهذه المرأة	- ما أعجب الحياة
- فاز في السباق	- ذهب ليعالج مرضه
- زبير بن الحارث	- و تعوذ جذوره إلى
- يزور جاره	- الجوع يقتلني
- السفارة	- وإذا النجوم انكدرت
- على تحسين وسائل النقل	- البرهان بالتراجع
- على الأساس الصحيح	- الدجاج ينتج البيض
- برب الناس	- جيجل مدينة سياحية
- المحيط الأطلسي	- ألم يجدك يتيما فآوى

4.2.3. Procédure de segmentation

Cette étape indispensable pour l'analyse a nécessité un intérêt particulier de notre part pour réduire le plus possible la marge d'erreur durant le calcul de la durée phonémique.
Les fichiers son segmentés en utilisant notre logiciel, l'image du spectrogrammes, ainsi que le fichier correspondant aux formant seront stockés dans des répertoires. La segmentation du corpus ont été effectués manuellement ce qui justifie le temps, relativement long, alloué a cette opération. La procédure adoptée pour isoler les phonèmes à dégager l'unité à étudier de l'onde temporelle, et enfin effectuer des tests de perception pour s'assurer de la qualité de la segmentation. Une fois cette étape achevée, nous passons à l'étiquetage phonétique et on va déterminées a la fin les valeurs moyenne de la durées et les formants de chaque phonèmes étudiés

4.3. Phonèmes étudiés

Bien que le système phonétique de l'Arabe Standard soit constitué de 28 consonnes et de 6 voyelles (3 brèves et 3 longues), nous nous somme limités dans notre étude aux voyelles brèves et à quelques consonnes de la langue. Les consonnes concernées sont : [s],[š],[z],[ž] . Le prélèvement des durées et formants de des phonèmes, à partir des signaux de parole enregistrés.

4.3.1. Etude acoustique

L'objectif de cette étude acoustique est de présenter l'intérêt de l'analyse acoustique pour l'étude phonétique de cette classe de son de l'AS. Il s'agit de trouver les propriétés nécessaires et suffisantes pour caractériser la classe des notre phonèmes par rapport à d'autres classes de sons.

Les fricatives en générale sont des bruits, c'est-à-dire événements apériodiques. Ce bruit résulte d'une turbulence aérodynamique qui prend naissance en un ou plusieurs points du conduit vocal en raison de la présence d'un fort resserrement ou d'un obstacle placé dans le flot d'air expiratoire.

Sonagraphiquement, avec un filtre de largeur 300 Hz, ce bruit de turbulence apparaître comme un ensemble de petites stries verticales plus ou mois longues , d'intensité variable, disposées aléatoirement

Pour notre étude d'une classe des sons fricatives non emphatiques, nous avons élaboré un tableau qui permet de distanguer chaque consonne étudiée dans les différents contextes (tableau 4.1). Et a partir des données mesurées, nous avons calcul les durées moyennes (D_{moy}) et les valeurs formantiques prélevées au niveau des formants F_1 ,F_2 ,F_3 , F_4 et F_5, et le tableau (tableau 4.1) donne ses valeurs moyenne calculés.

phonème	D_{moy}	F_1	F_2	F_3	F_4	F_5
[s]	155	1378	2954	4684	5780	6708
[š]	135	2329	3070	4149	5297	6342
[ž]	55.66	1071	2857	3737	4973	6007
[z]	127.5	0860	3035	4511	5778	6772

Tableau. 4.2 : les durées moyennes et les valeurs moyennes des formants de consonnes étudiées

4.4. Présentation générale du logiciel SAFAS

Cette interface a été baptisée **SAFAS** ce que signifié :

- S : pour Systéme ;
- A : pour Analyse ;
- F : pour Fricative ;
- A : pour Arabe ;
- S : pour Standard ;

La fonction principale de ce logiciel est l'analyse fréquentielle du signal de parole dans le but d'une extraction des résonances formantiques du conduit vocal (leurs fréquences, leurs bandes passantes).

- Les formants sont le résultat des estimations des fréquences des formants sur le spectre d'une fenêtre glissant le long du signal. Cette fenêtre a une durée choisie d'une part assez courte pour que les caractéristiques spectrales puissent être considérées comme étant stables et d'autre part assez longue pour contenir le maximum d'harmoniques possibles pour augmenter le contraste des formants sur le spectre.

- La durée, avec laquelle se déplace le signal, est choisie généralement de 5 à 10 ms et elle peut changer suivant la précision recherchée.

- Le nombre de points correspondant à la fenêtre du signal est déterminé avec la fréquence d'échantillonnage du signal.

- La méthode utilisée pour l'extraction des formants est basée sur le calcul des pôles de la fonction de transfert du modèle LPC. En effet, La prédiction linéaire est une méthode de dé convolution directe qui sépare le signal périodique d'excitation de la fonction de transfert qui module ce signal. Cette fonction est un modèle tous pôles autorégressif (AR) où chaque pôle correspond aux caractéristiques d'un formant c'est à dire à sa fréquence et à sa bande passante. Le nombre de pôles de cette fonction donne l'ordre du modèle.

Pour calculer les coefficients de ce modèle, nous avons utilisé des fonctions prédéfinies de la bibliothèque « Matlab7 » L'algorithme utilisé est très simple. Il consiste à déterminer d'abord par la commande "LPC" le polynôme du dénominateur de la fonction de transfert représentant le conduit vocal ; ensuite, en utilisant la commande "ROOTS" on détermine les pôles dans le plan "Z" et n'en prendre que ceux qui appartiennent au demi plan supérieur pour ne pas avoir les formants en double. Et à partir des positions de ces pôles on détermine les formants et les bandes passantes.

Une représentation des trajectoires formantiques sur le spectrogramme du signal analysé permet de vérifier la qualité de l'estimation.

La modélisation des formants a été réalisée à l'aide d'un moyen nage de la trajectoire par un polynôme mathématique d'ordre « n », présent dans la bibliothèque Matlab. Ce polynôme est actionné par la fonction « polyfit » de la bibliothèque Matlab. Les trajectoires formantiques lissées sont tracées sur les trajectoires obtenues au préalable.

Les paramètres du polynôme sont représentes sur la trajectoire du formant concerné afin de pouvoir les utiliser pour écrire éventuellement l'équation Fi(t) de la trajectoire formantique sous forme mathématique.

Ce logiciel Il offre la possibilité:

- de segmenter, de zoomer, de découper, d'écouter et de sauvegarder une portion de signal à partir d'un fichier préalablement enregistré ou disponible sur le PC ;
- d'effectuer différentes analyses temporelles et fréquentielles du signal de parole ;

- par le biais d'un spectrogramme : pour une représentation temps fréquence du signal
- par le biais de LPC, pour la détection des formants et calculée les moyennes formants ;
- De la reconnaissance automatique des consonnes étudiées ;

4.4.1 L'interface graphique de logiciel SAFAS

Elle est présentée sous forme d'un menu déroulant et de plusieurs fenêtres. Elle permet la, visualisation du signal temporel et du spectrogramme (Fig. 1.4) .

Il permet le tracé du spectrogramme du fichier visualisé en choisissant le nombre de points de la fenêtre d'analyse LPC

4.4.1.1. Le menu déroulant

Il est constitué des fonctions suivantes :

- ***ouvrir :*** permet l'ouverture et de visualiser d'un fichier de signal temporel à analyser disponible dans un certain emplacement dans le PC .Ce fichier est en général un fichier audio avec l'extension « .wav ».
- ***Zoom on :*** il permet d'agrandir la partie du signal sélectionné.
- ***Reset zoom :*** permet d'annuler le zoom préalablement effectué sur le signal
- ***sound*** : permet d'écouter une portion ou le fichier entier du signal temporel ouvert
- ***ANALYSE (LPC) :*** permet de démarrer l'analyse

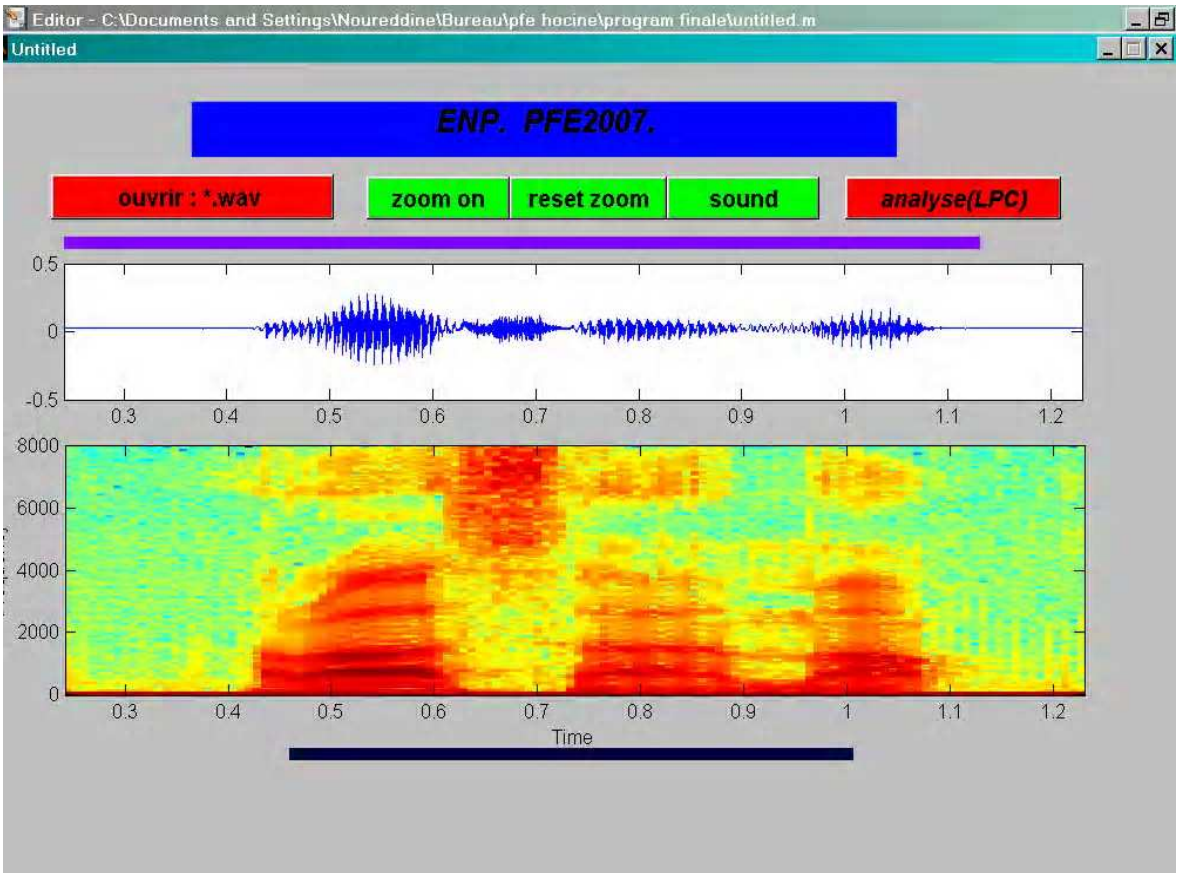


Fig 1.4 : L'outil d'analyse sonagraphique

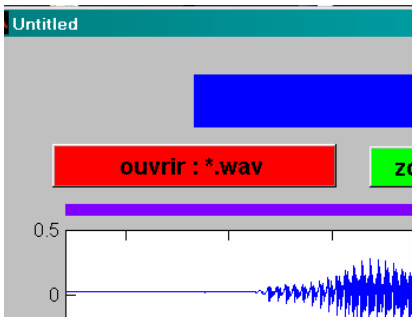
les paramètres nécessaires pour une analyse formantiques et une détection de trajectoires formantiques à l'aide de la LPC sont :

- **Nombre de formants** : permet de sélectionner le nombre de formants à détecter a partir de programme.
- **Ordre du filtre :** permet d'introduire l'ordre du filtre prédicteur de la LPC a partir de programme et nous avons utilisée dans ce cas l'ordre du filtre 13.
- . **Durée de la fenêtre :** permet d'introduire dans le programme la durée de la fenêtre d'analyse.
- **Garder le temps de la sélection** : permet de visualiser la trajectoire formantique en respectant le temps de la portion du signal sélectionné pour la détection de formants.

4.4.2 Exécution du logiciel SAFAS

4.4.2.1 Ouverture d'un fichier audio

La première chose à faire avant de procéder à une analyse est d'obtenir un son. Pour cela Pour ouvrir le fichier audio (.wav) dont vous voulez faire la transcription, Cliquez sur [Ouvrir : *.wav] dans l'interface.

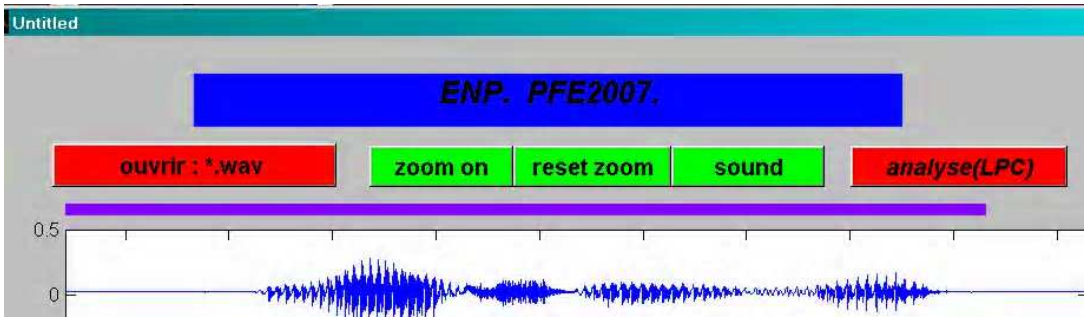


Une boîte de dialogue s'ouvre afin que vous puissiez choisir ouvrir .
Vous pouvez naviguer via le menu « regarder dans : » jusqu'au dossier contenant le fichier « .wav » que vous avez préalablement enregistré. Quand son nom apparaît dans le champ « Nom : » (après l'avoir sélectionné), cliquer sur le bouton « Ouvrir ».

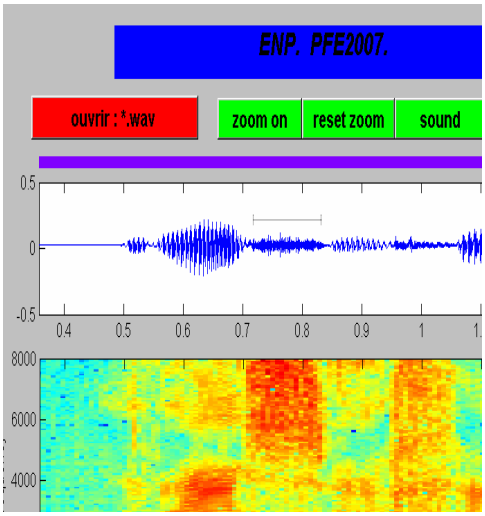
Une boîte de dialogue s'ouvre afin que vous puissiez choisir ouvrir :

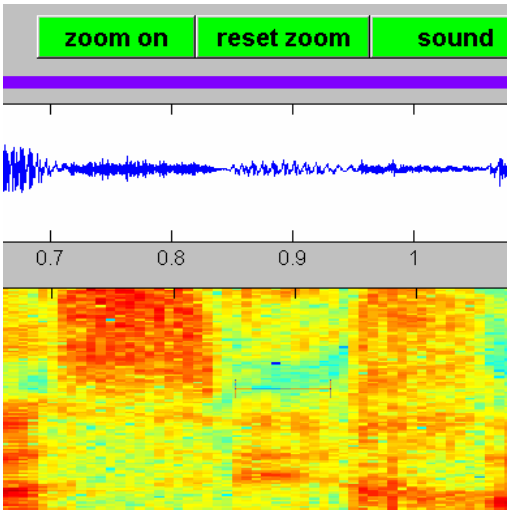
Vous pouvez naviguer via le menu « regarder dans : » jusqu'au dossier contenant le fichier « .wav » que vous avez préalablement enregistré. Quand son nom apparaît dans le champ « Nom : » (après l'avoir sélectionné), cliquer sur le bouton « Ouvrir ».

Une fois le fichier est chargé, le logiciel affiche le signal acoustique ainsi que son spectrogramme sur la figure et à partir de bouton [**Zoom on]** on peut faire la segmentation



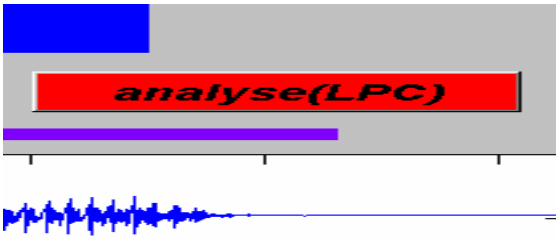
On peut faire la segmentation que sa soit dans le signal acoustique ou dans le spectrogramme



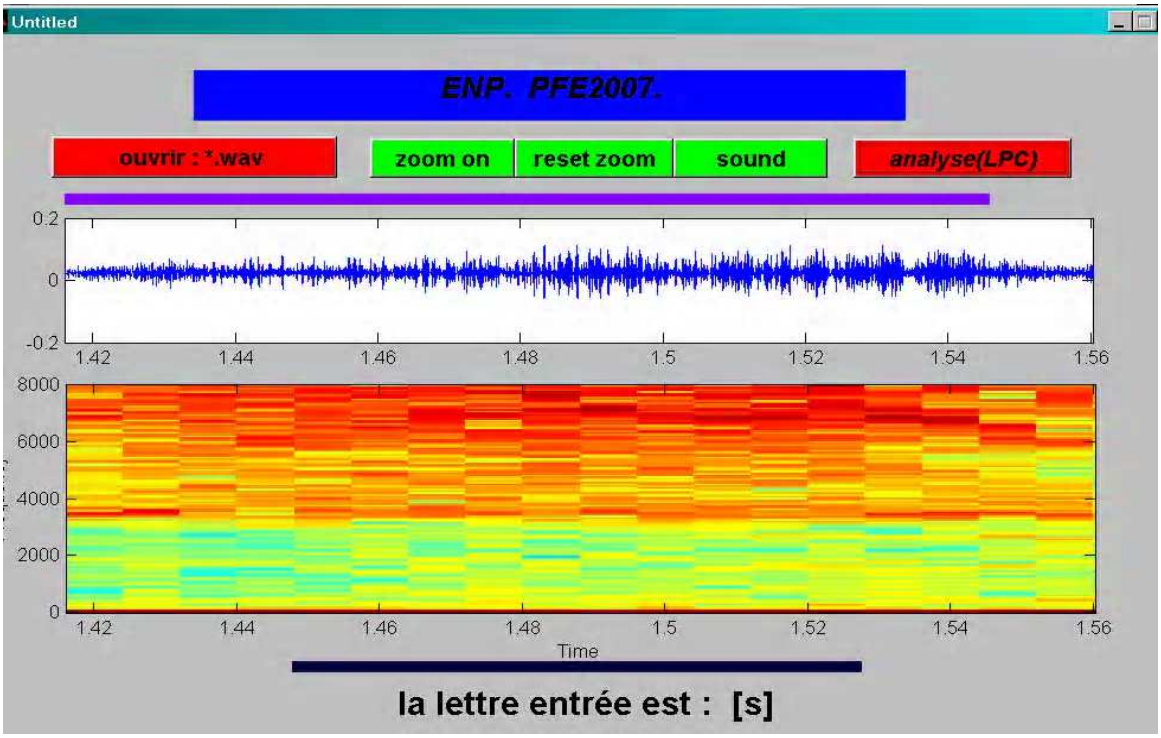


Et à l'aide de Bouton **[sound]** on peut écouter la partie sélectionnée, ou la totalité du signal pour simplifier la segmentation.

Puis Cliquez sur le bouton [**ANALYSE (LPC)]** pour démarrer l'analyse et la Reconnaissance Automatique de phonème entrée.



Et nous avons ajouté dans l'interface une espace pour afficher la résultant de reconnaissance.



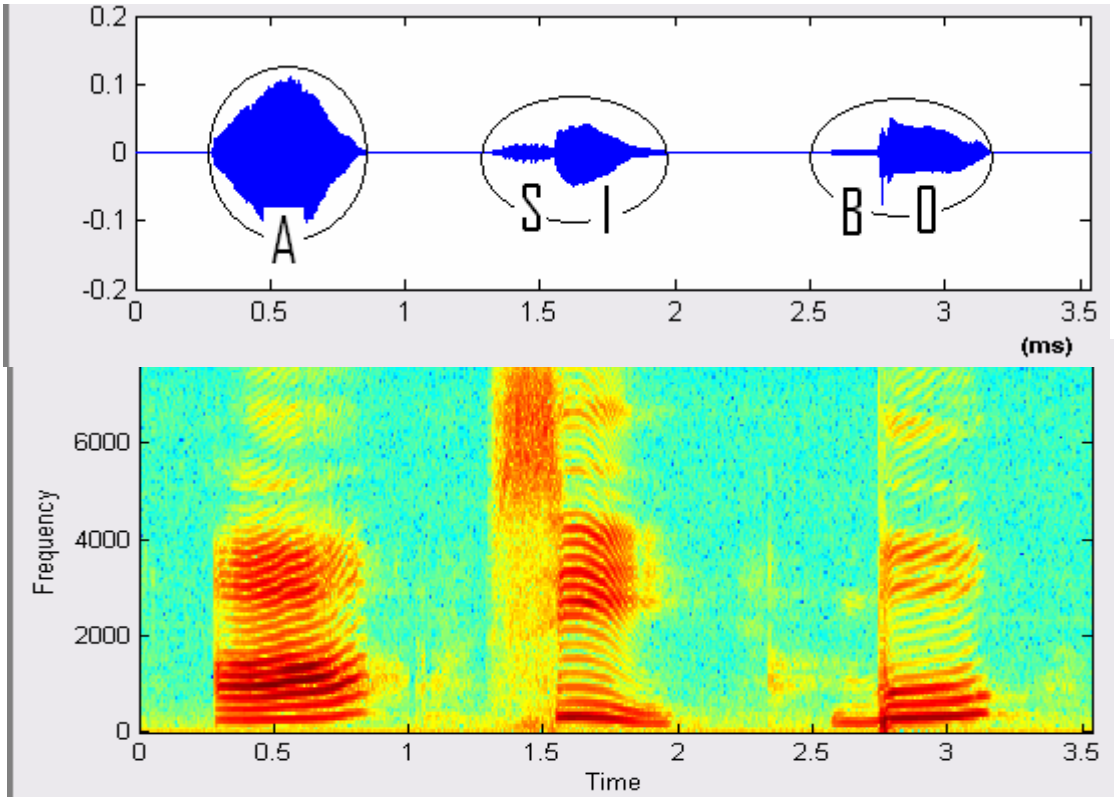
4.5. Validation des résultats:

Afin de vérifier l'efficacité de notre logiciel SAFAS, nous proposons de voir les résultats obtenus pour différents types de sons, a savoir : des voyelle, des fricatives, des plosives, etc. Nous comparerons ensuite les résultats obtenus avec SAFAS à ceux obtenus avec d'autres outils d'analyse comme winsnoori et praat qui sont deux logiciels d'analyse et de traitement très connus.

Les sons analysés sont :

- La consonne fricative /S/,), dans la syllabe /SI/ ;
- La voyelle orale /A/ ;
- La consonne plosive /B/, dans la syllabe /BA/ ;

Les conditions d'analyse sont similaires aux trois outils, à savoir : un fenêtrage de Hamming (car c'est la fenêtre que donne les meilleurs résultats), un ordre du filtre aux alentours de 18 et une fréquence d'échantillonnage de 16kHz



4.5.1. Résultats obtenus pour la consonne fricative /S/

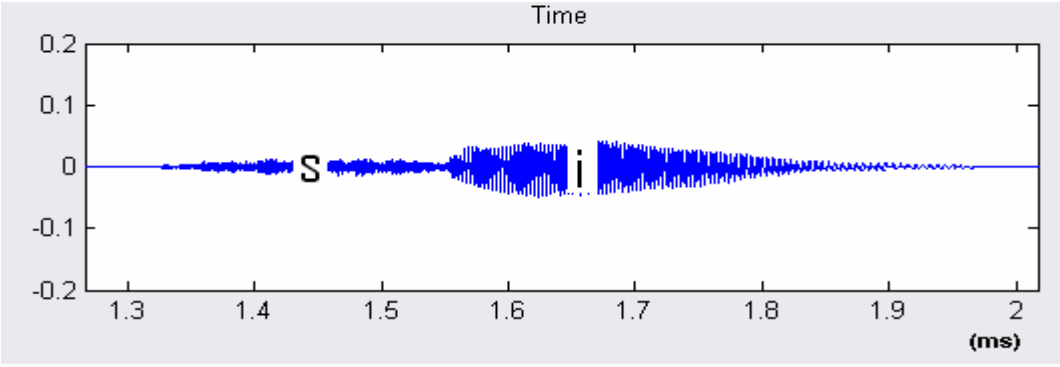
Cette consonne fricative est étudiée suivie par la voyelle orale /i/, dans la syllabe /si/.

Notons que dans la littérature pour le /i/ les cinq premiers formants sont aux alentours des valeurs suivantes :

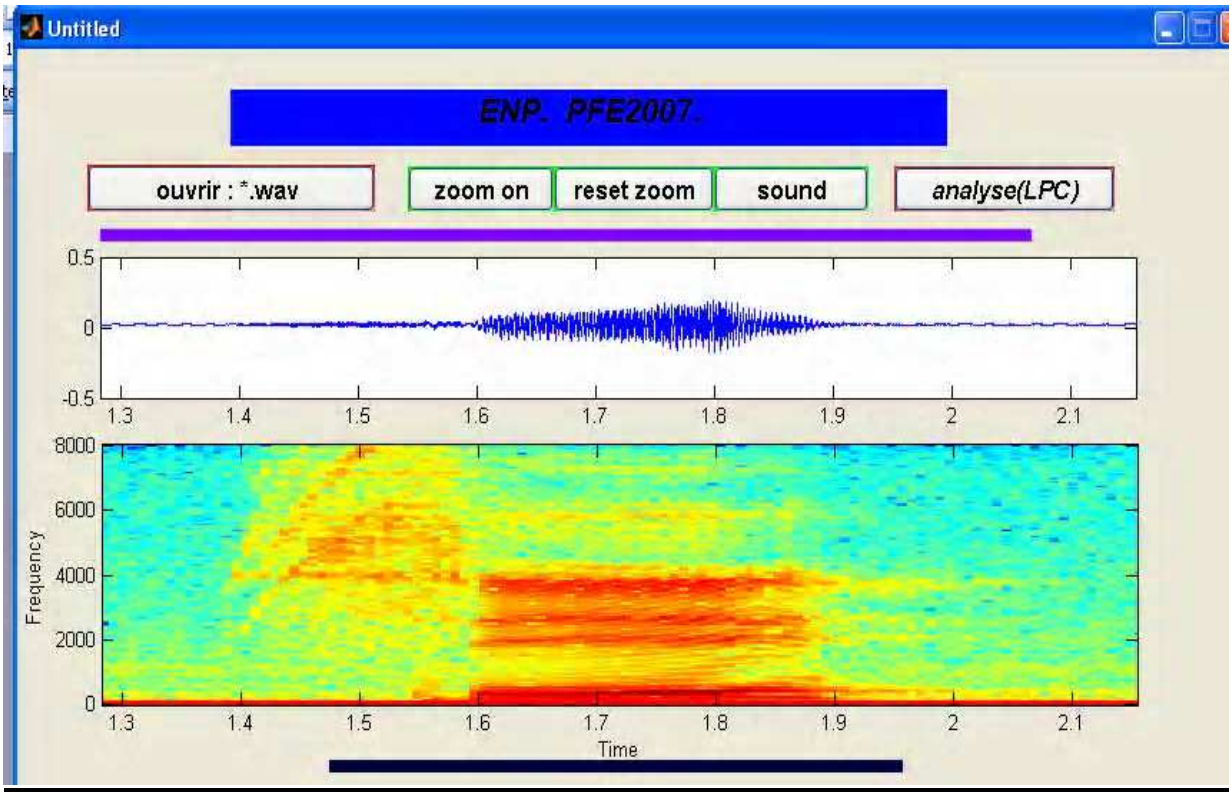
F1≈280Hz ; F2≈ 2200Hz ; F3≈2800Hz F4≈3400Hz F5≈3800Hz

Et pour le /s/ :

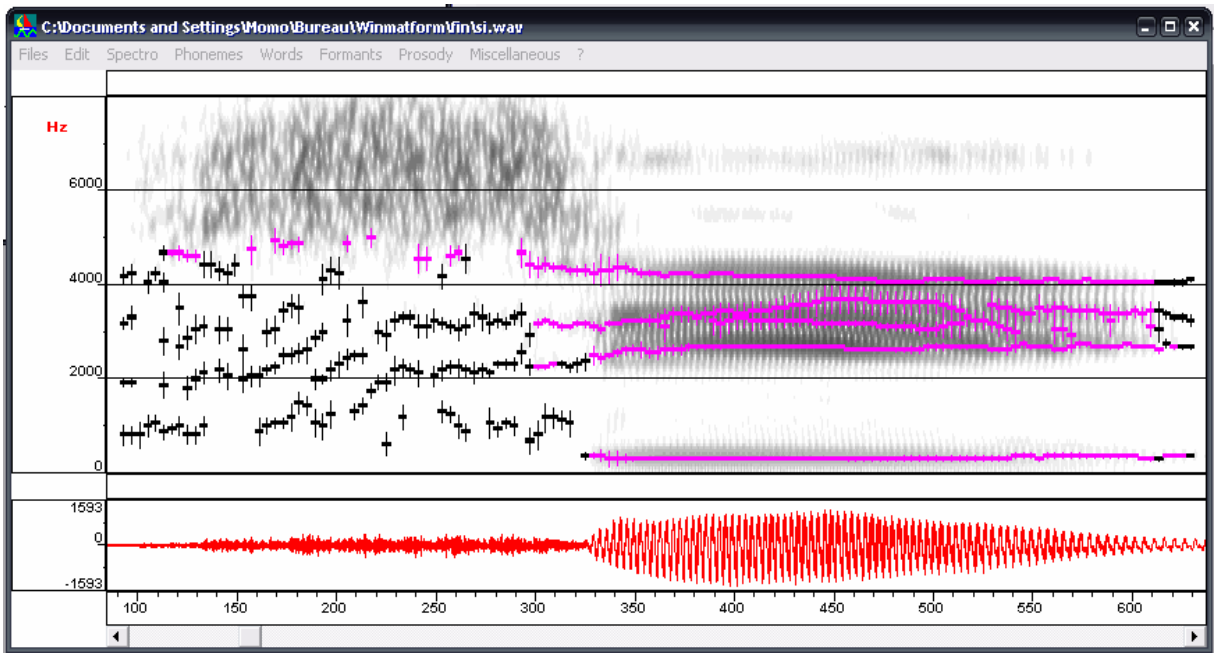
F1≈ 1320 Hz; F2≈ 2992 Hz ; F3≈ 4605 Hz ; F4≈ 5723 Hz ; F5≈ 6708 Hz ;



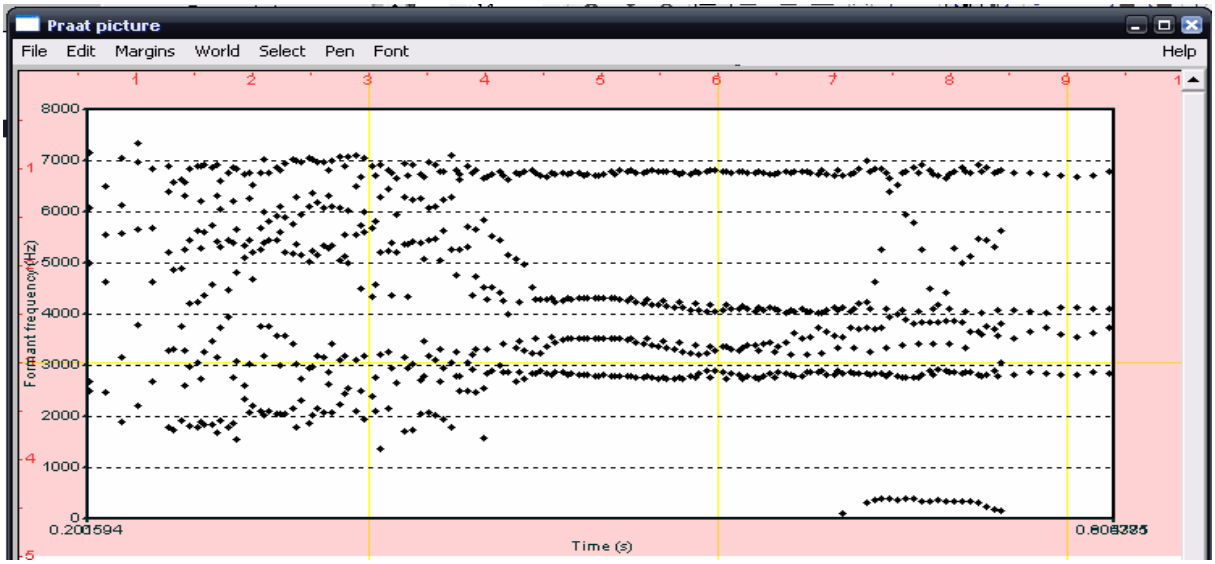
Avec SAFAS



Avec winsnoori



Avec praat



Nous remarquons encore une fois que les résultats obtenus sont similaires aux trois logiciels.

Notons que **SAFAS** offre des trajectoires formantiques plutôt stables par rapport à celles obtenues avec les deux autres.

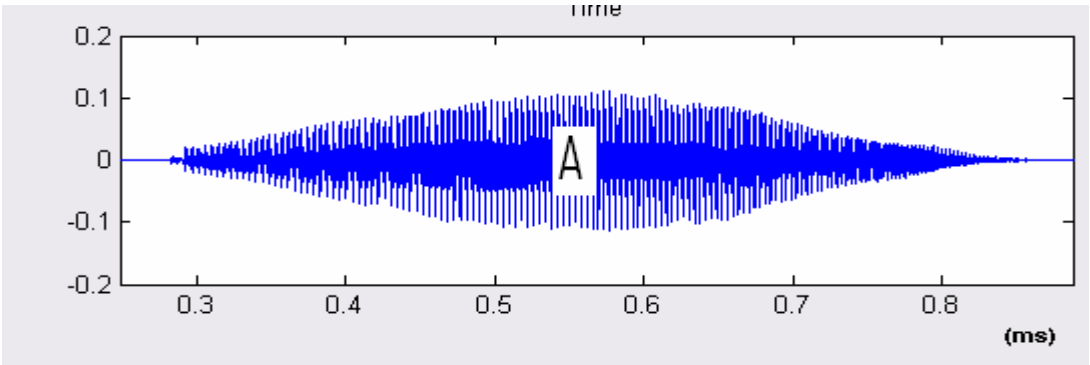
Les trajectoires formantiques se superposent bien sur le spectrogramme avec des valeurs très conformes à celles de la littérature.

Nous remarquons par contre que le logiciel winsnoori n'affiche pas les trajectoires formantiques au delà de 5000Hz pourtant la fréquence d'échantillonnage est de 16Khz. Ce qui ampute les fricatives surtout non voisées de leurs trajectoires formantiques.

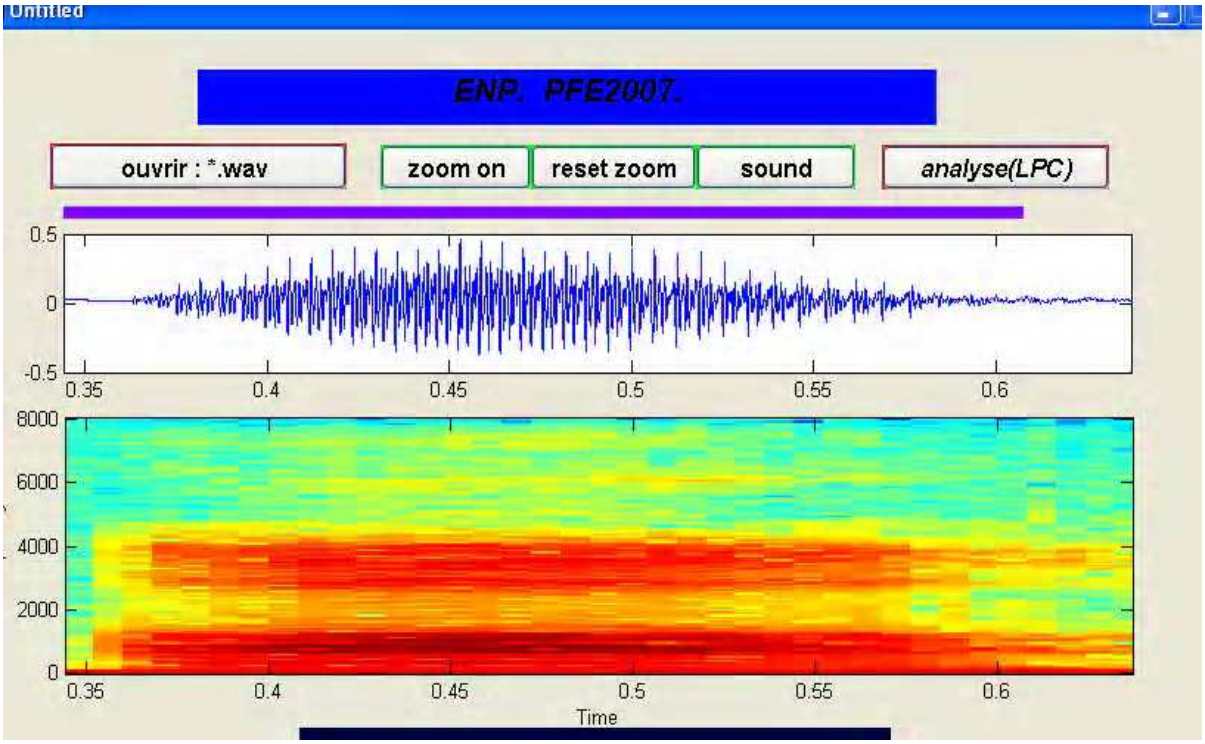
4.5.2. Résultats obtenus pour la Voyelle orale /A/

Notons que dans la littérature pour le /a/ les cinq premiers formants sont aux alentours des valeurs suivantes :

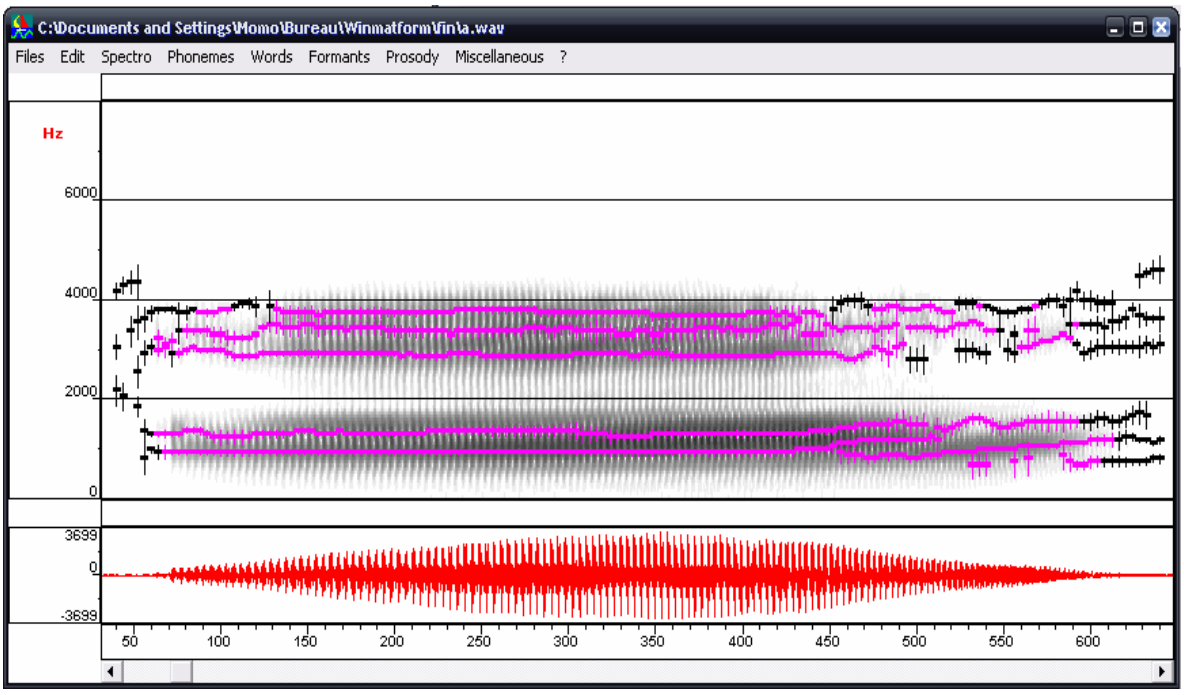
F1≈800Hz ; F2≈ 1500Hz ; F3≈2700Hz F4≈3030Hz F5≈3900Hz



Avec SAFAS



Avec winsnoori



Avec praat

Notons que les trois logiciels donnent presque les mêmes résultats.

Pour SAFAS sur la figure les trajectoires formantiques avoisinent les valeurs suivantes :

F1= 890.Hz
F2= 1300.Hz
F3= 3050.Hz
F4= 3600.Hz
F5= 6200.Hz

Pour winsnoori les trajectoires formantiques avoisinent les valeurs suivantes :

F1= 860.Hz
F2= 1251.Hz
F3= 2900.Hz
F4= 3330.Hz
F5= 3750.Hz

Pour praat les trajectoires formantiques avoisinent les valeurs suivantes :

F1= 1100.Hz
F2= 1420.Hz
F3= 3140.Hz
F4= 3650.Hz
F5= 6740.Hz

Nous remarquons une grande similarité entre les quatre trajectoires formantiques F1, F2, F3 et F4 obtenues à l'aide des trois outils.

Cependant, la trajectoire formantique du cinquième formant obtenu à l'aide de winnsnoori semble différer complètement. En effet, la trajectoire formantique du F5 est aux alentours des 3700Hz avec winsnoori alors qu'avec praat et SAFAS, elle est aux alentours des 6000Hz. Notons que d'après la littérature, la valeur du cinquième formant pour le phonème

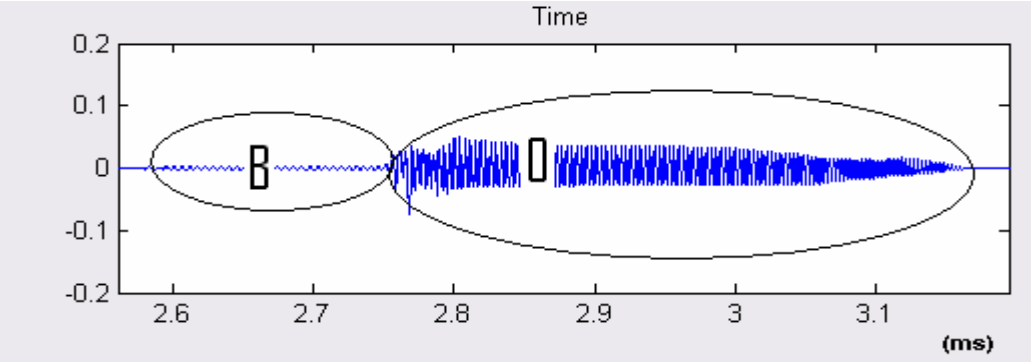
/a/ es t aux alentours de 4000Hz. Donc, la valeur du F5 donnée par SAFAS et praat n'est pas la bonne.

Avec SAFAS et praat, le cinquième formant n'apparaît pas du tout.

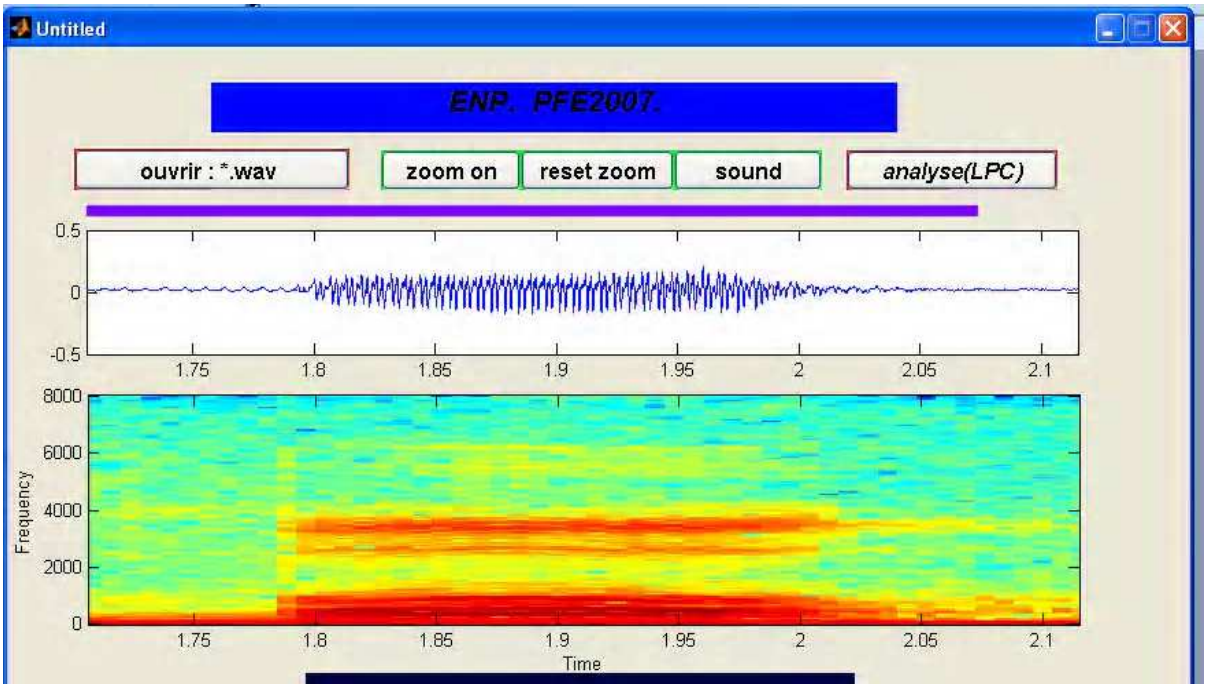
Ce qui montre que les résultats obtenus avec SAFAS sont tout à fait comparables à ceux obtenus avec les autres logiciels utilisant la détection de trajectoire formantiques à l'aide de la LPC, ce qui montre aussi les grandes performances (les 4 premiers formants sont similaires) et les limites (quant à la précision du cinquième formant) du modèle LPC

4.5.2. Résultats obtenus pour la consonne plosive /B/

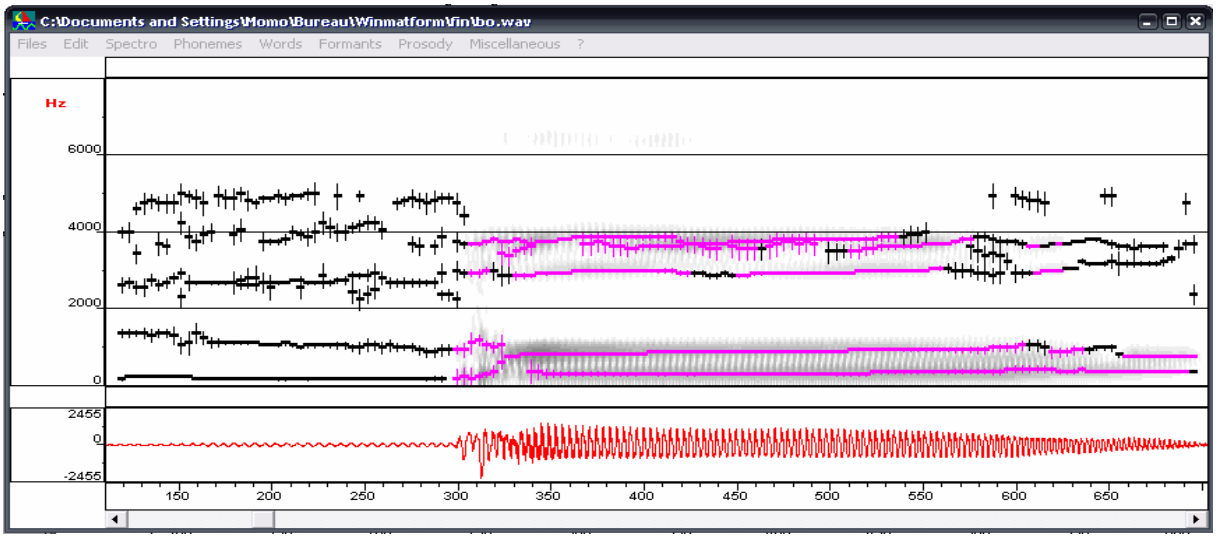
Dans la syllabe /BO/



Avec SAFAS



Avec winsnoori



Avec praat :

Les trajectoires formantiques obtenues pour winsnoori et SAFAS sont identiques.

Après avoir comparé nos résultats avec ceux de Winsnoori et de praat nous pouvons dire que notre outil est fiable puisque il nous procure les mêmes performances que ces deux autres qui sont des outils utilisés par beaucoup de chercheurs. Cependant, ces résultats sont très dépendants des paramètres d'entrées du logiciel (tel que le choix de la fenêtre d'analyse, la durée de chevauchement des fenêtres, la valeur de l'ordre du filtre etc…)

Il serait donc intéressant de voir le protocole d'évaluation sur la Reconnaissance automatique des consonnes fricatives étudié, afin de pouvoir effectuer des ajustements si le résultat obtenu n'est pas satisfaisant.

4.6. Algorithme de reconnaissance

- Début ;
- La première chose à faire avant de procéder à une analyse est d'obtenir un son. Pour cela Pour ouvrir le fichier audio (.wav) dont vous voulez faire la transcription, Cliquez sur [Ouvrir : *.wav] dans l'interface.
- Une fois le fichier est chargé, le logiciel affiche le signal acoustique ainsi que son spectrogramme sur la figure et à partir de bouton [**Zoom on]** on peut faire la segmentation manuellement ;
- Cliquez sur le bouton **ANALYSE (LPC)** pour démarrer l'analyse suivant les étaps :
  - pré-accentation de signal, création des fenêtres de 25 ms ;
  - Multiplication par une fenêtre de hamming ;
  - Calcul des paramètres a(i) du filtre LPC ;
  - en utilisant la commande "ROOTS" on détermine les pôles dans le plan "Z" ;
  - la détection des formants et calculée les moyennes formants ;
- si le signal né pas encore terminé répéter l'étape précédente ;
- calcul la distance euclidienne entre les moyennes formants calculés et les références dans le dictionnaire ;
- On prend la distance minimum ; Si la distance minimal inférieur ≤ seuil, afficher la réponse (le phonème), sino afficher « Le phonème entré ne correspond pas » ;
- Fin ;

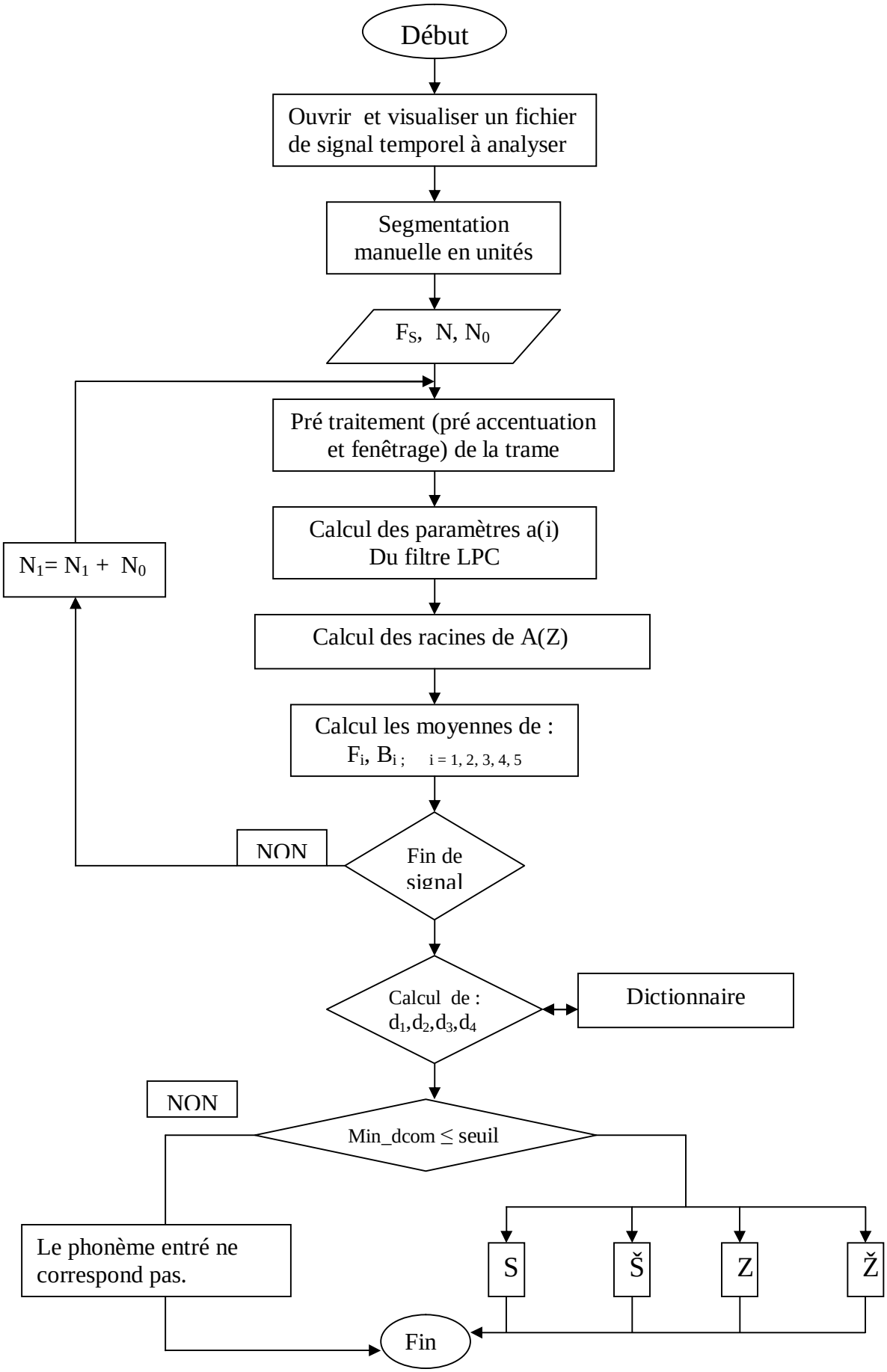


Fig. 4.2 : Organigramme de reconnaissance de phonème spécifiques de l'AS

4.7. Protocole d'évaluation

Le taux de reconnaissance correcte de chaque phonème présents sur les tableau suivant, et nous avons calculer le taux de reconnaissance par la relation suivant :

$$Ir = \frac{\text{Nombre de segments de test correctement reconnu}}{\text{Nombre total de segments de test}} \times 100$$

Et les tableaux suivant présent les différant valeurs de taux de reconnaissance de chaque consonne étudiés (tableau 4.2, 4.3)

[s]	résultats	[ž]	résultats
سأل المعلم التلميذ	1	ألم يجدك يتيما فآوى	1
السلطة القضائية	1	جلس يستمع الراديو	0
الحسنة تذهب السيئة	0	في ما جاء في البرنامج	1
خمسة أسر	1	بصوت جميل	1
جلس يستمع إلى	1	شجرة الأرز	1
مارس الرياضة	1	المجاهد	0
لبس ثوباً جديدا	1	جيجل مدينة سياحية	1
سبحان الله	1	ما أعجب الحياة	1
من سور القرآن	1	ذهب ليعالج مرضه	1
سعاد سلوكها غريب	0	و تعوذ جذوره إلى	1
تقع سوريا في الشام	1	الجوع يقتلني	1
taux	81		72

Tableau 4.2 : les taux de reconnaissance de [s] et [ž]

[z]	résultats	[š]	résultats
زعيم القبيلة	0	تنتشر بسرعة	0
زار المري	1	شعر بالخوف	1
مزارع	1	كان شابا قويا	1
لابد لي من الزواج بهذه المرأة	1	المشاهد	1
فاز في السباق	1	عاش لمدة طويلة	1
زبير بن الحارث	1	الشعب الجزائري	1
يزور جاره	1	الشعوب الحرة	1
بدون منازع	1	الشروق	1
الاعتزاز بالهوية	0	شوهد أمام النزل	1
الرموز الوطنية	1	فراش الموت	0
زيارة الأقارب	0	يناقش الأستاذ الموضوع	1
taux	90	taux	100

Tableau 4.3 : les taux de reconnaissance de [z] et [š]

4.8. Conclusion

Les résultats obtenus avec SAFAS sont pratiquement de même qualité que ceux obtenus à l'aide de winsnoori ou Praat, Cependant, notre but n'est pas de reproduire exactement les fonctions de praat ou winsnoori mais de réaliser un outil répondant à nos besoins.

SAFAS offre en plus des fonctions classiques (telles que découper zoomer enregistrer, détermination des trajectoires formantiques etc…) :

Notons que les meilleurs résultats avec SAFAS peuvent être réaliser avec une configuration des paramètres telle que :

*fenêtre d'analyse: hamming

*ordre du filtre: 13

*durée de la fenêtre: 25 ms.

*durée de chevauchement : 5ms.

Il serait donc nous intéressant de notre étude à la Reconnaissance Automatique des consonnes fricatives de l'Arabe Standard et a partir de protocole d'évaluation, on peut dire que ce système peut généralisé sur tout les phonème.

CONCLUSIONS GÉNÉRALES ET PERSPECTIVES

Conclusions générales et perspectives

Nous avons réalisé un outil de la reconnaissance automatique d'une classe des sons fricatives non emphatiques de l'AS [s], [š], [z], et [ž]. à partir d'un logiciel que nous avons baptisé SAFAS, dans l'environnement Matlab 7.

Cet outil permet de réaliser des traitements et nous pouvons citer en particulier des traitements sur les formants, et de la reconnaissance automatique des consonnes étudiées à savoir :

- La détermination des paramètres des formants (temps, bandes passantes, fréquences) ;

- le tracé de ces dernières directement sur un spectrogramme afin de s'assurer de la justesse de leurs formes ;
- la reconnaissance automatique des consonnes étudiées ;

Les résultats obtenus ne sont pas toujours très performants mais répondent cependant à nos attentes. En les comparant à ceux obtenus à l'aide d'autres outils de grandes renommées (comme winsnoori et praat), nous pouvons dire que SAFAS est du même niveau de performances.

Les résultats des paramètres des formants, ont été introduits dans un système de reconnaissance par une méthode de comparaison traditionnelle avec les références enregistrées.

En effet, la détection des formants n'est pas toujours évidente. Le spectre du signal de parole peut être très perturbé sur tout le signal ou sur une partie du signal par cause de plusieurs facteurs : naturels (bruits et autres) ou facteurs dũs aux condition d'enregistrement, de prétraitement ou au mauvais choix des paramètres d'analyse. C'est pour cela, qu'il est nécessaire que l'utilisateur ait un minimum de connaissances dans le domaine à traiter pour pouvoir intervenir sur le logiciel pour obtenir de bons résultats et ce avec n'importe quel outil de traitement.

Enfin, ce travail nous a permis de découvrir:

- l'immensité du langage de programmation MATLAB dont l'impressionnante bibliothèque nous a été d'une grande utilité ;
- un domaine très intéressant qui est le traitement du signal de parole ;
- les outils de traitement automatiques de la parole ;

RÉFÉRENCES BIBLIOGRAPHIQ

Références bibliographiques

[1] Le monde des sons, Dossier Pour la Science, Hors-Série n°32, (La voix humaine, Robert Sataloff, p 10-15), juillet-octobre 2001.

[2] J.Y. Antoine ,*'Outils informatiques d'analyse de corpus'* –Université Rabelais de Tours (France) , 2003

[3] http://www.info.fundp.ac.be/~gde/dec.html

[4] M. Kunt et R. Boite, *'Traitement de la parole'* - presses polytechniques Romandes (Suisse), 1987.

[5] K . Bouchefra, *'Contribution à la reconnaissance automatique de la parole Continue'*- Magistèr électronique ENP 1995.

[6] H. Sakoe et S. Chiba, *'Dynamic programming Algorithm Optmisation for spoken Word Recognition'*– IEEE Int. Conf. On Acoust. , Speech and Signal Proc . ICASSP , Vol . 26, pp. 43-49,1978.

[7] B. Flacon et PH. Lockwood, *'Systeme de reconnaissance de mots isolés multilocuteurs pour un vocabulaire de 130 mots . Intégration dans un poste de travail'* – 13 éme journnés d'etudes sur la parole GALF, BRUXELLES , 28-30, mai 1984.

[8] L. MYARA, *' la reconnaissance automatique de la parole '*- IEEE Int , Elève-Ingénieur Supinfo Paris Promotion SUPINFO, 2005.

[9] M. Kabache, *'Application des Réseaux de Neurones à la Reconnaissance Automatique des phonémes Spécifiques en Arabe Standard'*-magistèr –ENSLSH, alger-algérie, 2006.

[10] Z. Bendraoua et F. Bandou, *' Elaboration d'une Base de Données des Sons de l'Arabe Standard'*- PFE informatique USDB, blida-algérie 2006.

[11] www.isacolondecarvajale.perso.cegetel.net/maitrise_isacolon2004.pdf

[12] M. Chouai, *'Utilisation de la technique LPC pour la synthése de la parole'*- PFE électronique USTHB, alger-algérie, 1999.

[13] Calliope, *'La parole et son traitement automatique'*- collection technique et scientifique des télècommunications–édition Masson 1989.

[14] M.Basseville et I.Nikiforov, 'Détection of Abrupt Changes - theory and application' précédemment édité par Prentice-Hall, Inc, 2004.

[15] A.D. passos, 'Méthodes mathématique du traitement numérique du signal'edit Eyrolles - Paris ,1989.

[16] A. Hocini et H. Medjendjel, *'Localisation automatique des zones stables dans la parole continue'* - PFE électronique USTHB alger-algérie , 2000.

[17] L. Ali benali, *'Modélisation autoregressive et application du traitement du signal de parole'*-magitèr USTHB, alger-algérie, 1993.

[18] S. McCandless, An algorithm for automatic formant extraction using linear prediction analysis. *IEEE Transaction on ASSP*, p135–p141, 1974.

[19] John D. Markel and Augustine H. Gray. *Linear prediction of speech*. Springer-Verlag, Berlin–Germany, 1976.

[20] G. Duncan, B. Yegnanarayana, et Hema A. Murthy. A non parametric method of formant estimation using group delay spectra. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 572–575, 1989.

[21] A. S. Crowe. Generalised centroids, a new perspective on peak picking and formant. In *Proceedings of 7th FASE Symposium*, pages 683–689, Edinburgh–Scotland, 1988.

[22] A. Potamianos et P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of the Acoustical Societyof America*, 1996.

[23] A. Soquet. A cooperative approach to formant extraction. In *Proceedings of the International Congress of Phonetic Sciences*, pages 448–451, Stockholm - Sweden, 1995.

[24] A. Soquet. *Etude comparée de représentations acoustiques et articulatoires du signal de parole pour le décodage acoustico-phonétique*. PhD thesis, Université Libre de Bruxelles, Bruxelles – Belgique, 1995.

[25] Y. Laprie, Formant tracking adapted to acoustic-phonetic decoding. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 669–672, Paris - France, 1989.

[26] G. E. Kopec, Formant tracking using hidden Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1113–1116, 1985.

[27] R. Gerhard , A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1229–1232, 1986.

[28] R. Gerhard , Formant tracking with quasilinearization. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1 :307–310, 1988.

[29] M. J. Hunt. A robust formant-based speech spectrum comparison measure. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1117–1120, 1985.

[30] R. Battault, 'La reconnaissance vocale, techniques utilisées, applications actuelles et futures.' -IEEE,Int. -pfe électronique du C.N.A.M, juin 1998.

[31] J. Mariani, *Traitement du signal vol.7 N°4 spécial 1990 : Reconnais. de la parole. « Reconnaissance automatique de la parole : progrès et tendances ». , décembre 1990.*