الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider – Biskra

Faculté des Sciences et de la technologie

Département : génie électrique

Ref :……

جامعة محمد خيضر بسكرة
كلية العلوم و التكنولوجيا
قسم:……
المرجع:……

Thèse présentée en vue de l'obtention
Du diplôme de

Doctorat en sciences en : Génie électrique

Option : Electronique

Modélisation neuro-prédictive pour La classification phonétique

Présentée par :

Mustapha Kamel Abderrahmane DIDICHE

Soutenue publiquement le 11 /12/2014

Devant le jury composé de :

GHODBANE Hatem	**Maitre Conférence A**	**Président**	**Université de Biskra**
TALEB AHMED a/Malik	**Professeur**	**Examinateur**	**Université de Valencienne France**
MELAAB Djamel	**Maitre Conférence A**	**Examinateur**	**Université de Batna**
ATHAMNIA Noureddine	**Maitre Conférence A**	**Examinateur**	**Université de Batna**
DEBILOU Abderazak	**Maitre conférence A**	**Examinateur**	**Université de Biskra**

SOMMAIRE

Remerciements

Merci Dieu le tout puissant pour nous avoir donné la foi et le courage pour accomplir ce travail.

Il nous serait impossible de nommer ici tous ceux qui ont de loin ou de prés aidé d'une manière ou d'une autre à la réalisation de ce modeste projet.

Tout d'abord, nos gratitudes vont à notre encadreur, Mr. A. DEBILOU ainsi qu'a MR BENYETTOU qui a bien voulu nous accueillir dans son laboratoire « SIMPA (Signal Image Parole)» de l'Université des Sciences et de la Technologie d'Oran - Mohamed BOUDIAF, Faculté des sciences - Département d'informatique.

Ensuite ma femme et mes enfants qui se sont sacrifiés pour 'établir un cadre agréable et serein afin de pouvoir travailler.

Nous remercions également les aimables membres du jury qui ont bien voulu juger ce travail et de nous avoir honoré de leur présence.

Nous remercions nos familles et particulièrement nos parents qui ont su installer en nous la soif du savoir et de la connaissance.

Merci également à mes confrères de l'INTTO qui ont bien voulu m'accorder cette chance et qui m'ont poussé jusqu'au bout telle que Mr Boutaleb ainsi que Melle imine et à travers eux toutes la famille d'enseignant. Et sans oublier mon ami Mokhtari Salim et à travers lui tous ceux qui m'ont soutenu et encouragé dans toutes ses longues années de travail et d'haleine.

INTRODUCTION GENERALE

Le travail réalisé lors de cette thèse s'inscrit dans le cadre général de la reconnaissance automatique de la parole, RAP. Les travaux entrepris jusqu'à présent ont permis de réaliser des systèmes qui , s'il permettent une reconnaissance de vocabulaires de plus en plus étendus, restaient jusqu'à ces dernières années , assez sensibles aux conditions sonores de l'environnement d'utilisation , les conditions parfaites rencontrées en laboratoire ayant longtemps masqués ces contraintes .cette sensibilité au bruit est au freins majeurs à l'emploi de la reconnaissance automatique de la parole dans des applications dites grand publics qui supposent que l'utilisation d'un système de dialogue oral homme machine, DOMH , permet de reconnaitre tout, quelque soit l'environnement sonore.

Aucun système, jusqu'à ce jour, n'a pu effectuer très efficacement la reconnaissance de la parole continue, en temps réel et dans un milieu bruité. Différents outils ont été développés pour permettre la reconnaissance de la parole continue. Les objectifs de ce travail sont d'une part d'étudier la pertinence de l'analyse par démodulation (utile en reconnaissance de parole continue), et d'autre part d'élaborer une architecture à réseaux de Neurones Multicouche pouvant servir pour la reconnaissance de parole continue.

La reconnaissance automatique de la parole peut être basée directement sur une comparaison de formes nouvelles avec des références de notes à connaitre, ou bien sur l'identification d'un ensemble d'unité élémentaire (phénomènes, diphones, syllabes). Dans le premier cas, il s'agit d'une reconnaissance dite globale (approche retenue dans ce travail), dans le second cas d'une reconnaissance dite analytique.

Notre intérêt porte sur la reconnaissance de notes isolées arabes, la reconnaissance globale est sans doute la plus adaptée dans le contexte actuel de notre problème .Le signal de la parole est l'extraction des paramètres nécessaires à ce type de reconnaissance sont abordés ainsi que le mode de fonctionnement des systèmes de reconnaissance. Les modèles mathématiques que nous avons choisi d'employer pour atteindre notre objectif sont les réseaux connexionnistes, neuromimétiques, qui présentent des capacités très intéressantes en reconnaissance et en classification des formes.

Notre travail est organisé de la manière suivante :
Le premier chapitre concerne une présentation de la complexité de traitement des informations par notre serveur qui n'est pas évidente pour une machine.

Le deuxième chapitre démontre toutes les contraintes et la problématique posée en reconnaissance de la parole. Surtout dans un milieu fortement bruité.

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

Le quatrième chapitre qui est considéré comme le noyau de cette thèse, traite la paramétrisation via les coefficients **MFCC** et **NPC** dont aucun besoin pour la classification par les réseaux de neurones **MLP**.

Le cinquième chapitre de cette thèse propose une méthode de coopération entre deux réseaux de neurones dont l'un est supervisé et l'autre non supervisé, ceci dont l'objectif d'améliorer les scores.

Le sixième chapitre cite une des méthodes de rassemblement de corpus et dans ce chapitre on parle de corpus **TIMIT** qui a été proposé aux Etat Unis.

La comparaison et discutions des résultats obtenus font l'objet du chapitre sept qui met en application un logiciel d'analyse de parole sous un environnement c++.

Cette interface est présentée par un oscillogramme, un spectrogramme et deux zones de texte qui affichent une transcription phonétique du signal de parole en caractère arabe et l'autre en caractère occidentaux, enfin un menu rempli de fonctionnalité. Les tests sont donnés pour le **MLP** et la coopération entre **MLP-LVQ** pour l'apprentissage et la phase de généralisation qui est la reconnaissance.

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

The work done in this thesis falls within the general framework of automatic speech recognition, RAP. The work undertaken so far have helped to realize systems which, if possible recognition vocabularies increasingly widespread, remained until recent years, quite sensitive to noise conditions of the use environment, perfect conditions encountered in the laboratory has long masked these constraints. This noise sensitivity is major obstacles to the use of automatic speech recognition in said large public applications that assume the use of a spoken dialog system man machine DOMH, allows recognizing all or some sound environment.

No system, so far, has been very effectively performing continuous speech recognition in real time and in a noisy environment. Various tools have been developed for the recognition of continuous speech. The objectives of this work are firstly to examine the relevance of the analysis by demodulation (useful in continuous speech recognition), and secondly to develop an architecture Multilayer Neural networks can be used for recognition A continuous speech.

Automatic speech recognition can be based directly on a comparison with new forms of references in footnotes to know, or the identification of a set of basic unit (phenomena, diaphones, and syllables). In the first case it is called a global acknowledgment (approach taken in this work); in the second case a so-called analytical recognition.

Our focus is on the recognition of notes Arab isolated global recognition is without doubt the most suitable in the current context of our problem. The speech signal is the extraction of parameters necessary for this type of recognition are discussed as well as the mode of recognition systems. Mathematical models we have chosen to use to achieve our goal are neural networks, neural, which have very interesting capabilities for recognition and classification of shapes.

Our work is organized as follows:

The first chapter deals with an overview of the complexity of information processing by our server that is not obvious to a machine.

The second chapter shows all the constraints and the issues raised in speech recognition. Especially in a very noisy environment.

The third chapter is to describe the physiological aspect of human organs used in the perceptual and articulator phonetics, the physical properties of sounds treated condition.

The fourth chapter is considered the core of this thesis addresses the parameterization via the **MFCC** coefficients and **NPC** with no need for classification by **MLP** neural networks.

The fifth chapter of this thesis proposes a method of cooperation between two neural networks, one of which is supervised and one unsupervised, it aimed to improve the scores.

The sixth chapter cites methods of gathering corpus and in this chapter we talk about **TIMIT** corpus was proposed to U.S. State

Comparing and discussing the results is the subject of Chapter Seven which implements speech analysis software developed entirely in a visual environment C + +6.0.

This interface is shown by a waveform, spectrogram and two text boxes that display a phonetic transcription of the speech signal in Arabic script and the other Western character, and finally full menu functionality. The tests were given to a **MLP** learning rate but a recognition rate, by cons for cooperation between **MLP-LVQ** learning rate and recognition rate.

In the end a conclusion was drawn from this thesis that the work should continue prospects in this field are very rich and important and should be prospected.

Résumé

Les applications en reconnaissance de la parole les plus évoluées se heurtent cependant à de grandes difficultés lorsqu'il s'agit de traiter des signaux en environnement fortement perturbé comme la téléphonie cellulaire par exemple. Des auteurs ont souligné récemment l'importance de revenir sur certain aspect de la chaine de traitement, l'extraction de caractéristiques en particulier, quitte à renoncer, à la course aux performances en termes de scores de reconnaissance.

Donc l'extraction de caractéristiques du signal de parole est un domaine de l'analyse du signal peu exploré par la communauté de recherche en parole. la raison principale en est que nous disposons aujourd'hui d'outils performants : des outils essentiellement fondés sur l'analyse fréquentielle des signaux pour leur paramétrisation et des outils d'analyse statistique pour leur classification.

Le travail présenté dans ce mémoire propose de reprendre les toutes premières étapes de traitement des systèmes de reconnaissance de la parole, à savoir l'extraction de caractéristiques et la classification phonétique **[16],[5]** . Une nouvelle modélisation permettant de prendre en compte les caractéristiques non linéaires du processus de production de la parole est proposée. Fondée sur l'utilisation d'un perceptron multicouche, elle permet de sursoir aux limites bien connues des systèmes connexionnistes appliqués à la modélisation non linéaire des signaux.

Nous nous intéresserons également à l'aspect discriminant de ces caractéristiques.

Mots clés : LPC, MFCCs. FFT, NPC, LVQ-NPC, MLP

Notations et Abréviations

x(n) : composante temporelle du signal de parole

X(k): composante fréquentielle du signal de parole

F_0 : Fréquence fondamentale ou pitch

F_1, H1 : premier Formant ou première Harmonique

F_2, H_2 : deuxième Formant ou deuxième Harmonique

F_1, H_1 : troisième Formant ou troisième Harmonique

s(n) : contribution de l'excitation

h(n) : conduit vocal

S(w) : spectre du signal

H(w) : filtre du signal

E(w) : spectre de l'excitation

m : fréquence en Mels

f : fréquence en Hertz

H(z) : filtre à réponse impulsionnelle finie

S_e : signal échantillonné

S_n : signal préaccentué

S_w: signal fenêtré

W (n): fenêtre

MFCC: scal Mel frequency cepstral coefficient

FFT: transformé de Fourier rapide

DFT: transformé de Fourier discrète

DCT: transformé en cosinus discrète

$\hat{c}_i$: pondération

$\Delta\hat{c}_i$: dérivation de la pondération

Θ : longueur de la fenêtre

E(m) : logarithme de spectre d'amplitude ou logarithme d'énergie

B(m) : bande passante des filtres triangulaires

LPC : coefficient de prédiction linéaire

$\hat{s}_n$: estimation du signal de parole

G : gain d'excitation

Z : transformé en Z

E_n : erreur quadratique moyenne

A : coefficient de prédiction

φ_s : fonction d'auto corrélation du signal s

R_s : matrice de covariance du signal s

Z_{cr} : taux de passage à zéro

Δ : dérivée du premier ordre

$\Delta\Delta$: dérivée du second ordre

$\hat{y}_k$: signal prédit

w : vecteur des poids de la couche cachée

a : vecteur des poids de la couche de sortie

G_w : fonction de la couche cachée

H_a : fonction de la couche de sortie

F : composition des fonctions des couches cachées et de sorties

k : numéro d'itération

μ : le pas d'apprentissage

Δw_{ij}^t : gradient de l'erreur par rapport aux poids

$d(a, mi,\tau)$: distance euclidienne

$m_{i,\tau}$: prototype de la classe C_i

μ_i : mesure de la mauvaise classification

DFE : extraction de caractéristique discriminante

MCE : critère de minimisation de l'erreur de classification

$C(a_0)$: classe d'apprentissage

GDP : descente de gradient probabilistique

$\alpha(t)$ et $\beta(t)$: pas d'apprentissage du classifieur LVQ

Δw^{mod} : rapprochement des modifications des premières couches

Δw^{disc} : éloignement des modifications des premières couches

Δa : modification résultante des vecteurs caractéristiques

Chapitre 1: Introduction de la Classification Phonétique

1-1- Introduction :

La reconnaissance automatique de la parole est un domaine de la science ayant toujours eu un grand attrait auprès des chercheurs comme auprès du grand public.

En effet, qui n'a jamais rêvé de pouvoir parler avec une machine ou, du moins, piloter un appareil ou un ordinateur par la voix.

L'homme étant par nature paresseux, une telle technologie à toujours suscité chez lui une part d'envie et d'intérêt, ce que peu d'autres technologies ont réussi à faire.

La parole est le principal moyen de communication dans toute société humaine. Son apparition peut être considérée comme concomitante à l'apparition des outils, l'homme ayant alors besoin de raisonner et de communiquer pour les façonner.

L'importance de la parole fait que toute interaction homme-machine devrait plus ou moins passer par elle. D'un point de vue humain, la parole permet de se dégager de toute obligation de contact physique avec la machine, libérant ainsi l'utilisateur qui peut alors effectuer d'autres tâches.

La parole a toujours été le meilleur moyen de communication entre les êtres humains, sa simplicité en fait d'ailleurs le moyen de communication le plus populaire dans notre monde.

Seulement, cette simplicité exige un traitement très complexe réalisé par notre cerveau, chose qui n'est pas évidente pour la machine.

Néanmoins, avec l'envahissement de l'informatique et le développement technologique (électronique, télécommunication , etc.,...) a suscité le besoin de nouveau moyen de dialogue homme machine (la parole), des moyens qui libéreraient l'homme d'un contact constant avec la machine limitant ainsi l'utilisation du clavier et des autres périphériques qui rendaient la communication avec la machine très difficile et très lente.

La parole est le principal support de la langue, comme pour l'écriture, le message doit donc être bien structuré selon des règles connues par tous ceux qui parlent la même langue **[2]** (grammaire, syntaxe, vocabulaire…).

Bien que l'évolution de la langue mue par les tendances actuelles rend de plus en plus difficile le fait de fixer des règles statiques une bonne fois pour toute, l'être humain grâce à l'ordinateur le plus sophistiqué au monde (son cerveau), pressé de faire passer son message, transgresse le plus souvent ces règles et arrive sans difficultés à le comprendre, ce qui montre le caractère dynamique du cerveau humain qui arrive très bien à s'adapter à de nouvelles situations,**[12],[17],[7]** caractère qui devrait caractériser tout système de traitement de la parole.

Le traitement automatique de la parole ouvre des perspectives nouvelles compte tenu de la différence considérable existant entre la commande manuelle et vocale. L'utilisation du langage naturel dans le dialogue personne/machine met la technologie à la portée de tous et entraine sa vulgarisation, en réduisant les contraintes de l'usage des claviers, souris et codes de commandes à maitriser.

En simplifiant le protocole de dialogue personne/machine, le traitement automatique de la parole vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse.

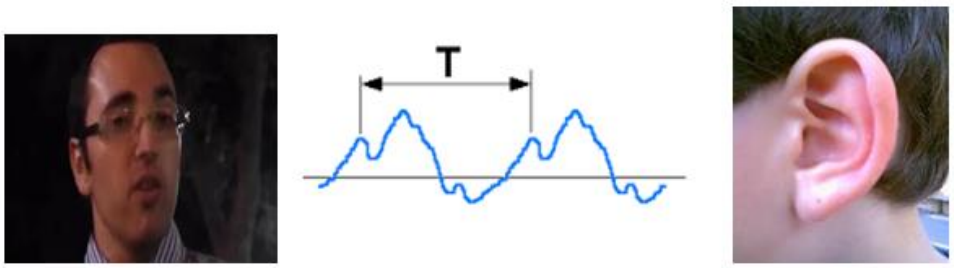


***Figure 1-1*: Exemple de dialogue personne-personne.**

Pour l'être humain comprendre ce que dit son interlocuteur est très simple, même s'il parle avec cette personne pour la première fois. L'interlocuteur arrive à générer grâce à son système articulatoire un signal acoustique (le son étant le support de la parole) continu et très complexe, perçu par notre système perceptif ce signal est traité par notre cerveau pour le filtrer de toute information inutile (bruit), en extraire des informations sur l'interlocuteur (voix, état mentale…), décoder le message en unités de base (phonèmes, syllabes…) qui seront réunies en mots choisis à partir d'un lexique qui permet de compenser d 'éventuelles erreurs de décodage, affecter des significations à ces mots (phrases) selon une sémantique et un contexte. Tout cela pour montrer le nombre colossal de connaissances nécessaires pour le traitement de la parole **[4], [11]**. En général, le traitement automatique de la parole utilise des sources de connaissances Phonétiques, Phonologiques, Prosodiques, Lexicales, Syntaxiques, Sémantiques et Pragmatiques.

La phonétique : Science qui étudie les caractéristiques physiques des sons sur trois Plans complémentaires (articulatoire, acoustique, perceptif).

La phonologie : Étudie l'aptitude linguistique en relation avec le son, en faisant abstraction de ses propriétés physiques. Elle définit un inventaire des unités de base (*phonèmes*) avec des contraintes de combinaison.

La prosodie : La prosodie peut être considérée comme une sorte de "ponctuation acoustique" de la parole.

Elle recouvre les aspects liés à la hauteur de la voix, à l'intensité et à la durée des segments syllabiques. Son rôle dans la langue est multiple.

Le lexique : Les performances d'un système de reconnaissance sont affectées par la taille du vocabulaire et aussi par le degré de confusion entre les mots. Le dictionnaire doit être étudié de telle sorte qu'il autorise économiquement la représentation de toutes les prononciations envisageables des mots, mais aussi pour qu'il permettent d'accéder directement à tous les mots contenant la même syllabe ou le même trait acoustique, de telle sorte qu'il soit possible de générer les hypothèses des mots à partir des caractéristiques du signal dans l'analyse ascendante.

La syntaxe : Du point de vue de la langue, la syntaxe est l'ensemble des règles contraignantes l'ordre des mots dans la phrase. Dans un système de compréhension, le but de la syntaxe est de réduire le nombre de phrases autorisées à partir du vocabulaire choisi.

La sémantique: La sémantique est définie d'un point de vue linguistique, comme la relation entre la forme des signes linguistiques, ou "signifiants", et ce qui est signifié, ou "signifiés". En reconnaissance de la parole la sémantique restreint la combinatoire syntaxique.

La pragmatique: La pragmatique peut être définie aussi comme l'étude des aspects du langage qui font référence aux relations entre locuteur et interlocuteurs, d'une part, et entre interlocuteurs et situation concrète, d'autre part . La pragmatique recouvre l'ensemble des relations entre le langage et le contexte d'énonciation. Dans les applications de communication homme-machine, la pragmatique joue un rôle très important dans l'interaction entre l'homme et l'application et pour résoudre les problèmes référentiels.

Cela dit les connaissances nécessaires pour la compréhension de la parole sont énormes, ainsi pour ce mémoire nous allons nous contenter d'étudier l'aspect de reconnaissance de la parole qui a pour objectif de décoder le signal de la parole en unités de bases (phonèmes, mots …) sans en donner une signification (sans comprendre le sens des phrases construites).

De nos jours, les systèmes de reconnaissance de la parole ont évolué et utilisent non seulement des connaissances en linguistique (Phonétiques, Phonologiques, Prosodiques, Lexicales…) mais aussi des connaissances dans les domaines : **[2]** Traitement du signal, Reconnaissance des formes…

1-2- Les différentes méthodes de reconnaissance de la parole :

Le principe général d'un système de reconnaissance de la parole peut être décrit comme une suite de mots prononcés **M** est convertie en un signal acoustique S par l'appareil phonatoire.

Ensuite le signal acoustique est transformé en une séquence de vecteurs acoustiques ou d'observation **O** (chaque vecteur est un ensemble de paramètre).
Finalement le module de décodage consiste à associer à la séquence d'observation O une séquence de mots reconnus **M'**.

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

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

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

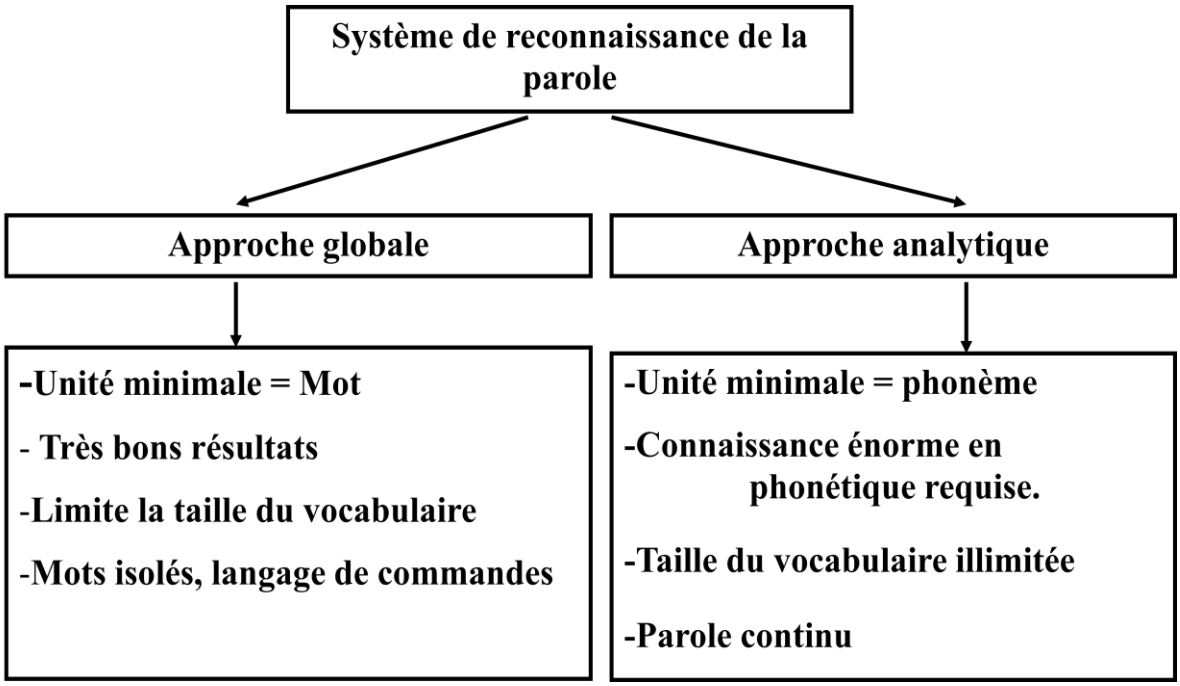


***Figure 1-2*: les approches des systèmes de RAP**

1-2-1- La méthode globale :

Cette méthode considère le plus souvent le mot comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans ce type de méthode, on compare globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique **[19]**. Cette méthode a pour avantage d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes. Ce type de méthode est utilisé dans les systèmes de reconnaissance de mots isolés, reconnaissance de parole dictée avec pauses entre les mots… et présente l'inconvénient de limiter la taille du dictionnaire.

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

1-2-2- La méthode analytique:

Cette méthode fait intervenir un modèle phonétique du langage. Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabe, demi-syllabe, diphonie, phonème, phone homogène, etc.).

Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance dans cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification de ces unités en utilisant des méthodes de reconnaissance des formes (classification statistique, réseau de neurones, etc.) ou des méthodes d'intelligence artificielle (systèmes experts par exemple). **[19]** Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation et d'étiquetage phonétique. C'est pourquoi le DAP (Décodage Acoustico-Phonétique) **[17], [12]** est fondamental dans une telle approche.

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

1- Représentation du message (signal vocal) sous la forme d'une suite de segments de parole, c'est la segmentation.

2- Interprétation de segments trouvés en termes d'unités phonétiques, c'est l'identification.

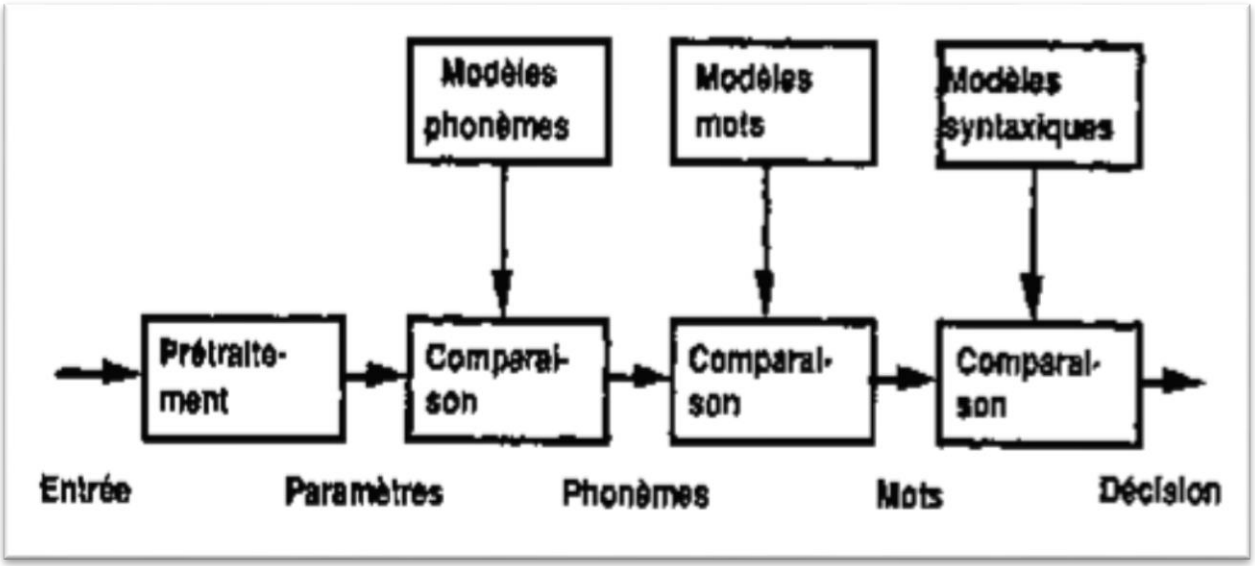


Figure 1-3: schéma synoptique d'un système de RAP selon l'approche analytique

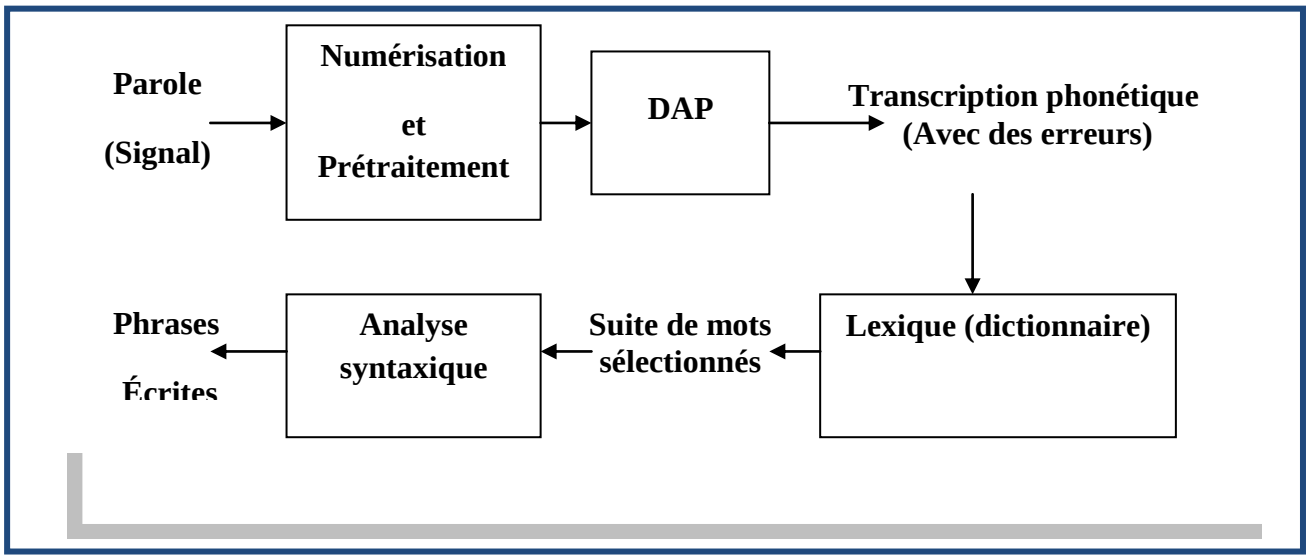


Figure 1-4: schéma général d'un système de reconnaissance de parole continu

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

- Dépendant du locuteur (mono locuteur)
- Multi locuteur
- Indépendant du locuteur

Chapitre 2:Problématique en Reconnaissance Automatique de la Parole

2-1- Introduction :

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

La parole est un phénomène à priori très simple à comprendre. Mais l'homme peut rencontrer des difficultés lorsqu'il essaie de suivre une conversation dans une langue qui lui est inconnue.

De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance ainsi que par le milieu ambiant.

L'obstacle principal dans la reconnaissance de parole, est la grande variabilité de ces caractéristiques.

Nous allons maintenant voir les problèmes directement liés à la parole.

2-2- Variabilité intra locuteur :

La variabilité interlocuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain **[10]**.

Une même personne ne prononce jamais un mot deux fois de façon identique.
La vitesse d'élocution en détermine la durée

L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie. Il existe un autre type de variabilité intra locuteur lié à la phase de production de parole ou de préparation à la production de parole **[2]**.

Il est possible de voir la phase de production de la parole comme un compromis entre une minimisation de l'énergie consommée pour produire des sons et une maximisation des scores d'atteinte des cibles que sont les phonèmes tels qu'ils sont théoriquement définis par la phonétique.

La coarticulation peut se produire à l'échelle d'un ou de plusieurs phonèmes adjacents.

2-3- Variabilité interlocuteur :

Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonnantes, modifient la production acoustique. En plus, il y a la hauteur de la voix, l'intonation et l'accent différent selon le sexe, l'origine sociale, régionale ou nationale.

La variabilité interlocuteur est un phénomène majeur en reconnaissance de la parole. Un locuteur reste identifiable par le timbre de sa voix malgré une variabilité qui peut parfois être importante. La contrepartie de cette possibilité d'identification à la voix d'un individu est l'obligation de donner aux différents sons de la parole une définition assez souple pour établir une classification phonétique commune à plusieurs personnes. La cause principale des différences interlocuteurs est de nature physiologique.

La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence de base, le fondamental.

Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents qui sont regroupés selon les classes que nous avons énoncées précédemment.

Or le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le genre et l'âge. Ainsi, le conduit vocal féminin adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte. Le conduit vocal d'un enfant en bas âge est bien sûr inférieur en longueur à celui d'un adulte.

Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes **[2], [10]**. La variabilité interlocuteurs trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

Chacun de ces facteurs détermine la situation de communication, et influe à sa manière sur la forme et le contenu du message.

2-4- Variabilité due à l'environnement:

La variabilité due à l'environnement (considérée comme du bruit) peut provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution **[2]**. Le bruit est défini au sens large comme étant tout signal perturbateur entachant à un degré ou un autre l'intégrité d'un signal utile (véhiculant l'information). **[10]** Le bruit ambiant peut ainsi provoquer une déformation du signal de parole en obligeant parfois le locuteur à accentuer son effort vocal.

2-5- Les différents types de bruit :

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

2-5-1 Les bruits additifs :

Les bruits additifs sont dus à la multiplicité des systèmes de communication dans un même environnement. Plusieurs émetteurs et plusieurs receveurs pouvant être confinés dans un même espace, les messages de tous les émetteurs peuvent donc se trouver en concurrence sur une même voie sans que les récepteurs possèdent un mécanisme infaillible pour isoler le message qui leur est destiné. L'émetteur et le récepteur peuvent aussi se trouver en présence d'un ou de plusieurs équipements générant un bruit de fond de force variable **[2]**.

2-5-2 Les bruits convolutionnels :

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

communication à longue distance tels que le téléphone, les moyens radiophoniques et, récemment, radiotéléphoniques. Ces moyens de communication à longue distance ont été élaborés à partir d'un compromis coût/efficacité. La parole, lorsqu'elle est transmise par un tel moyen, est forcément dégradée tout en gardant une grande intelligibilité **[2]**. Un des champs possibles d'application de la RAP sont les serveurs vocaux accessibles par les lignes téléphoniques. Mais la parole transmise par téléphone souffre de déformations variables induites par la qualité de la connexion.

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

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

2-5-3- Les bruits physiologiques :

D'autres bruits peuvent également être considérés dans le domaine de la RAP mais ils n'ont pas la généralité des bruits de type additif ou convolutionnel car ils sont spécifiques à l'être humain lors de sa phase de production de parole. Les plupart des systèmes de RAP fonctionnent mal en milieu bruité car les contraintes posées par de tels environnements n'ont pas été prises en compte dès le départ. L'homme essaie, lui même, de s'adapter aux conditions sonores rencontrées en modifiant sa méthode de production de parole. Un des phénomènes les plus remarquables de modification de production de la parole par l'homme est l'effet Lombard. Lorsqu'un locuteur est placé dans un environnement bruité, il modifie sa voix, et

son effort vocal, en “haussant le ton” de manière à ce que la parole produite conserve un bon RSSB par rapport à l’environnement.

Cette accentuation de la voix pose cependant un problème majeur aux systèmes de RAP car les spectres de tous les phonèmes peuvent être modifiés ce qui a pour effet de nettement amoindrir les taux de reconnaissance. **[2]** Certaines études montrent que l’homme arrive à avoir de meilleures capacités de compréhension dans le cas de la parole Lombard que pour la parole normale lorsqu’il lui est demandé de reconnaître des mots isolés ou de la parole continue masqués par du bruit

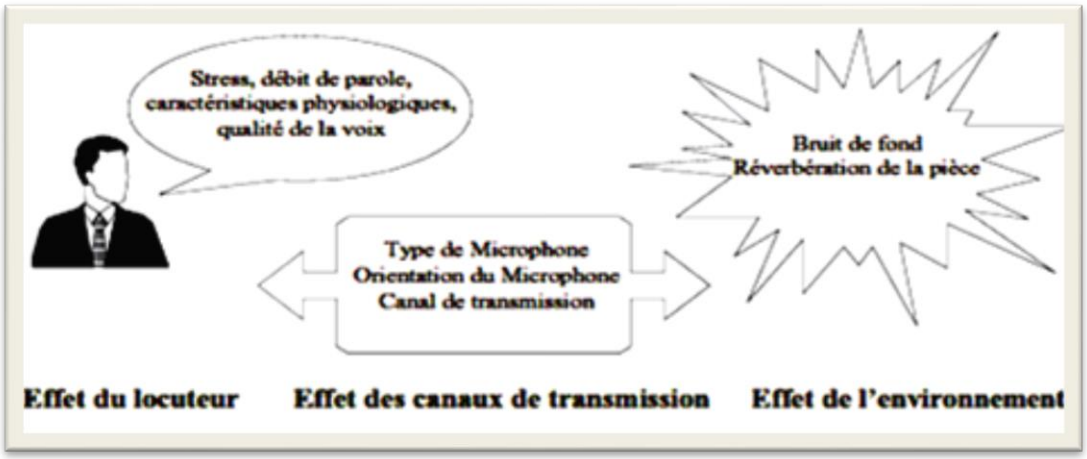


Figure 2-1: Représentation schématique de différentes sources de variabilité .

2-6- Coarticulation :

La cause principale réside dans les contraintes que la dynamique des articulateurs impose à la production de la parole.

Le débit de la parole atteint facilement 12 phones/sec. Alors que les articulateurs ne peuvent prendre que quatre ou cinq positions différentes par seconde (**Pallier**, **1994**), les sons sont donc coarticulés.

Résoudre le problème de la segmentation de la parole signifie, comprendre comment les phonèmes peuvent être perceptiblement séparés si les informations acoustiques sont à ce point superposées. Un autre problème lié à la coarticulation est celui de l'invariance perceptive. Les sons **/t/** et **/k/** sont articulés différemment avant les voyelles **/u/** et **/i/**.

Cela fait que les spectrogrammes des syllabes **/tu/** et **/ku/** et de **/ti/** et **/ki/** sont plus proches que le sont ceux de **/tu/** et **/ti/**. Néanmoins pour l'oreille humaine les deux ***/k/*** sont très similaires et différents des deux **/t/**.

Ce problème de l'invariance perceptive est particulièrement vrai pour les consonnes plosives : **/q/**, **/t/**, **/k/**, **/b/**, **/d/**. Le phénomène affecte moins les segments vocaliques ou les consonnes continues (**/f/**, **/s/**, **/r/**).

Les voyelles sont en effet des segments plus stables acoustiquement (pendant leur production les articulateurs ne bougent pas) et d'assez longue durée pour être facilement reconnues dans les spectrogrammes, quel que soit le contexte consonantique. Sur un spectrogramme, il est facile de constater que la plupart des phonèmes n'on pas de partie stable.

En fait, seules les réalisations acoustiques des phonèmes apparaissent dans le signal de parole, et il est tout à fait clair que le contexte joue un rôle prépondérant.

Un son n'est pas influencé de la même façon par ses voisins, comme les articulateurs ont tendance à anticiper le geste articulatoire lié au son suivant, un son est plus influencé par son successeur que son prédécesseur.

La portée des effets de coarticulation est à priori limitée aux voisins directs d'un son puisque les articulateurs se déplacent de la position qu'ils occupaient pour le son précédent à celle qu'ils occuperont pour le son suivant.

Cela n'est pas toujours vrai, car il est possible qu'un geste articulatoire n'implique pas tous les articulateurs simultanément.

C'est notamment le cas, dans le contexte **VCV** (voyelle consonne voyelle) où la première voyelle peut être influencée par la seconde malgré la présence de la consonne **[2]**.

On constate souvent pour la séquence **/abi/** que F2 de **/a/** est « attiré » vers F2 de **/i/**, c'est-à-dire vers une valeur plus élevée que celui d'un **/a/** prononcé isolément.

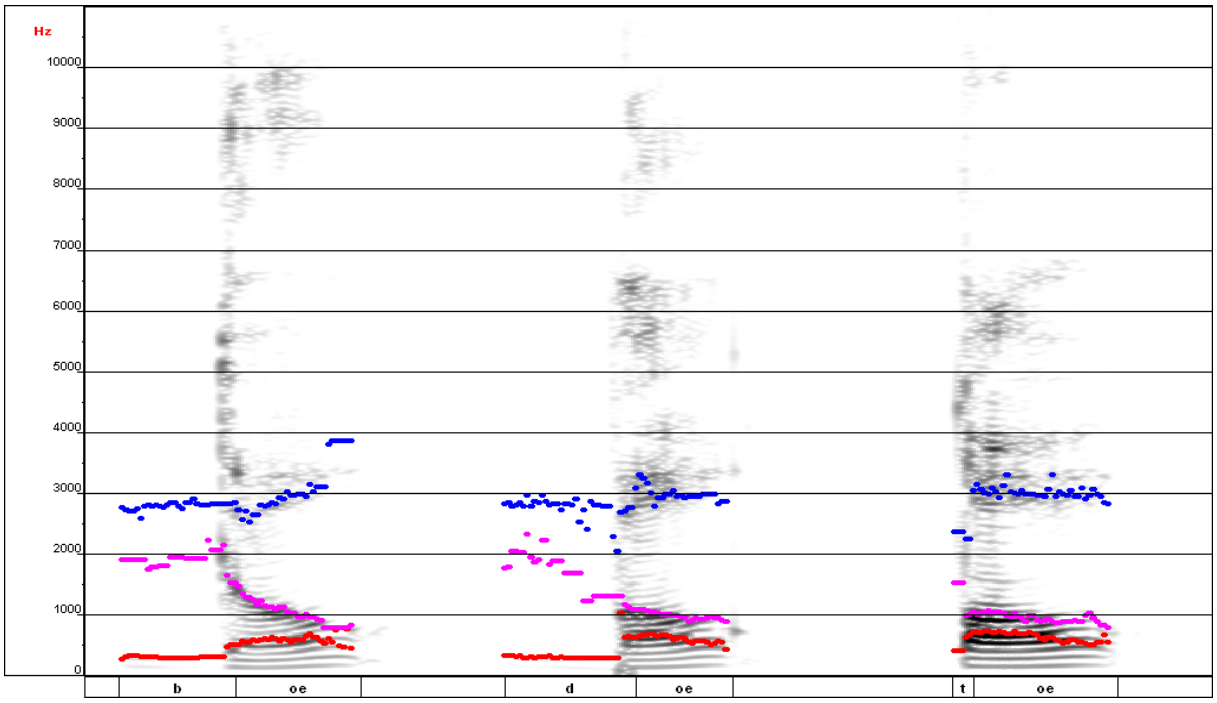


Figure 2-2: L'effet de la coarticulation dans quelques syllabes.

/boe/, **/doe/**, **/toe/** effet très visible dans la syllabe **/boe/** où les formants de **/oe/** s'éloignent de leurs trajectoires habituelles sous l'effet du phonème **/b/**.

Chapitre 3: Phonétique à partir de la modélisation Articulo-acoustico-perceptive

3-1- Phonétique articulatoire :

L'étude des sons du langage humain est envisagée, en phonétique articulatoire, sous l'angle de la production. Cette discipline comporte un volet sur la physiologie, consacré à la connaissance des organes de la phonation et un volet descriptif portant sur le rôle des différents organes dans la production des sons du langage. Il est possible d'opérer une classification des sons à partir de critères articulatoires. **[2]**Ces critères permettent également de décrire les sons de nombre d'autres langues et résument, en quelque sorte, les possibilités et les limites de l'appareil phonatoire. Bien entendu, les modalités d'exploitation des organes articulatoires peuvent varier selon la langue ou la famille de langues considérée. Ce chapitre est diffusé sous forme de cours en détails à l'université de Lausanne et dont le site est www.unil.ch/sli/page98709.html

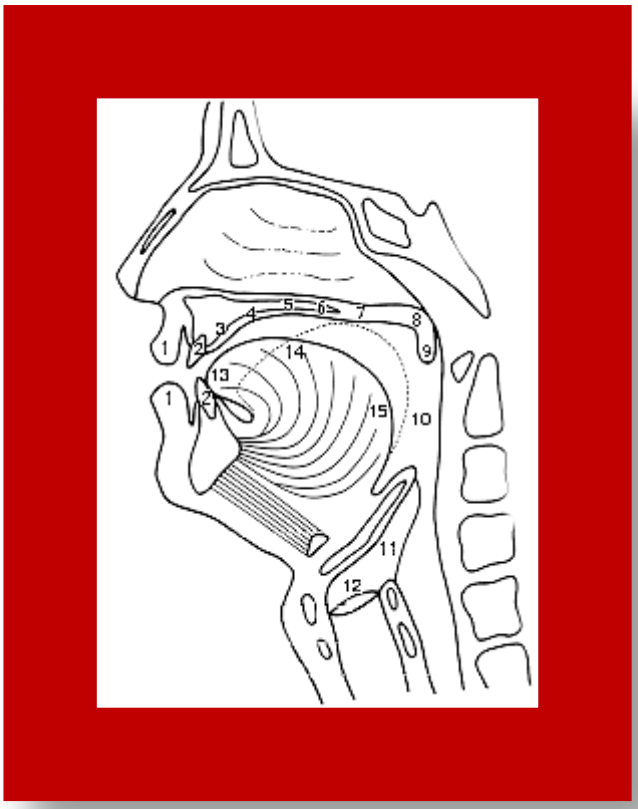
Système phonatoire		Organe anatomique		Nomenclature phonétique correspondante	
	1	Lèvres		Labiales	
	2	Dents		Dentales	
	3	Alveolus		Alvéolaires	
	4	palais dur		Pré-palatales	
	5			médio-palatales	
	6			post-palatales	
	7	voile du palais		Pré-vélaires	
	8			post-vélaires	
	9	luette (uvula)		Uvulaires	
	10	Pharynx		Pharyngales	
	11	Larynx		Laryngales	
	12	Glotte		Glottales	
	13	apex	de la langue	apicales (pré-dorsales)	Dorsales
	14	Dos		médio-dorsales	
	15	racine		radicales (post-dorsales)	

Figure 3-1: Système phonatoire de l'être humain

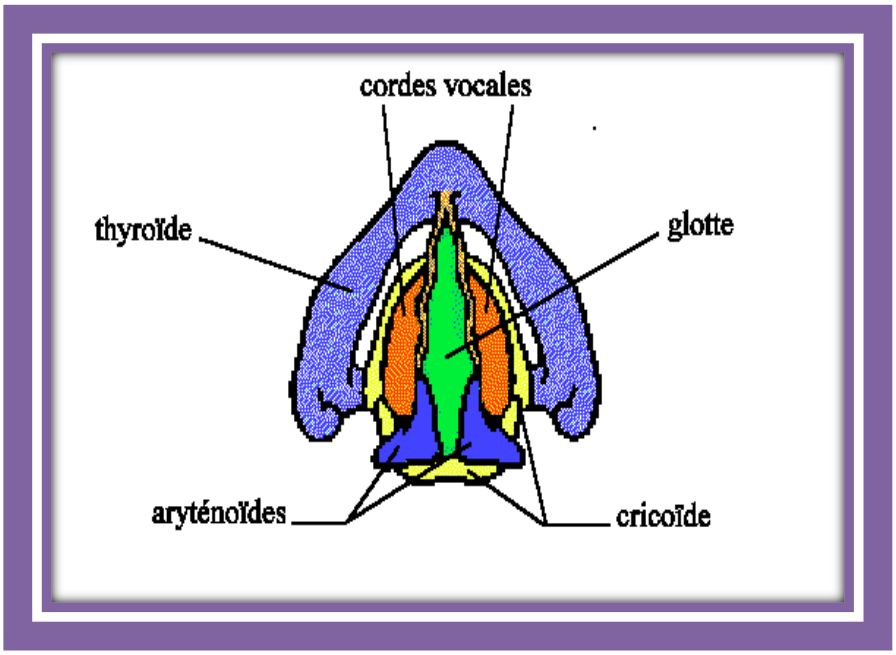


Figure 3-2: le larynx

3-1-1- Principe de production d'un son :

Le processus de production de parole est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique **[21]**. Il y a une grande quantité d'organes et de muscles qui entrent dans la production de sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre les poumons, le larynx, et les cavités supra-glottiques.

Les poumons, le larynx fournissent ce qui est essentiel pour la production de n'importe quel son, une source d'air et une source de bruit. Les cavités supra-glottiques renferment les organes qui permettent de modifier le son qui est émis par le travail conjoint des poumons, larynx.

Lorsque l'air est expulsé des poumons, il passe à travers un tube formé de plusieurs cartilages appelé le larynx **[21]**. Le larynx contient des muscles et des cartilages. Les cartilages les plus importants et les plus connus sont les cordes vocales qui peuvent s'ouvrir et se refermer très rapidement (jusqu'à 400 fois par seconde chez les enfants, par exemple), produisant ainsi des variations de pressions dans l'air. Ces variations de pression sont perçues comme du son par l'oreille humaine.

Les cordes vocales sont gardées ouvertes ou fermées par les aryténoïdes (cartilages en forme de pyramide situés à l'arrière des cordes vocales). Une voix typique d'un homme résulte de mouvements d'ouverture de 100 à 120 fois par seconde (un cycle d'ouverture est appelé un Hertz : Hz) alors que celle d'une femme est produite par entre 175 et 250 vibrations des cordes vocales par seconde. Ce bruit,

sera modifié par les divers organes de la parole qui font partie des cavités supra-glottiques.

Ces cavités servent à faire résonner le son et à lui donner une « couleur » particulière qui permettra de différencier les voyelles entre elles par exemple, ou les consonnes.

Donc 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. Les résonateurs principaux sont :

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

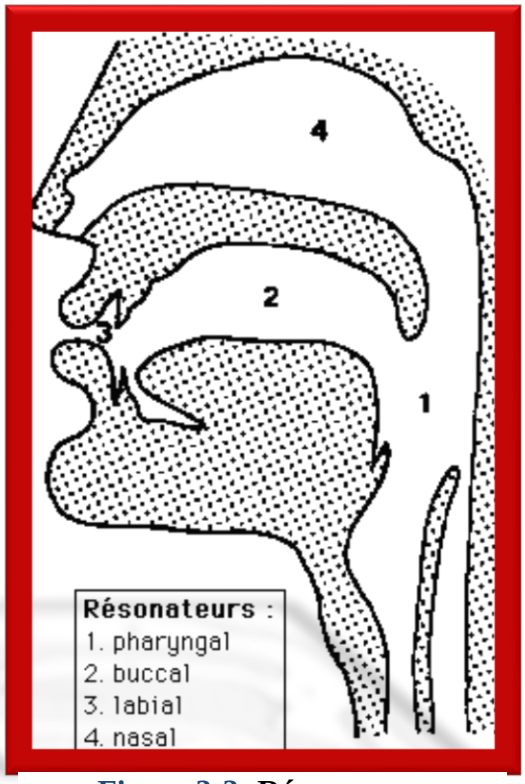


Figure 3-3: Résonateurs

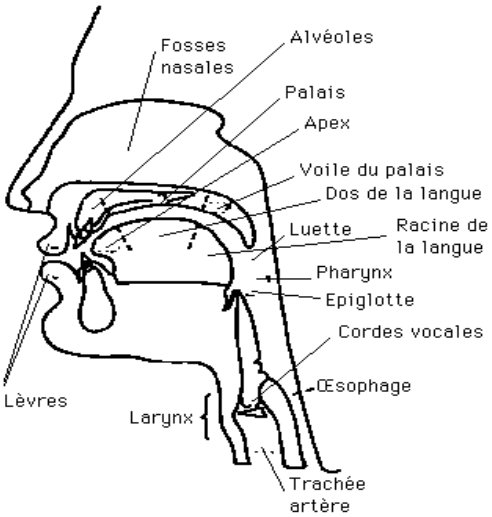
La présence ou l'absence d'obstacles sur le parcours de la colonne d'air modifie la nature du son produit. C'est, entre autres, en classant ces obstacles éventuels que la phonétique articulatoire dégage les différentes classes de sons décrites ci-dessous. Pour un petit nombre de réalisations, l'air ne 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, voire même par une variation de pression purement interne.

3-1-2- Consonnes et voyelles :

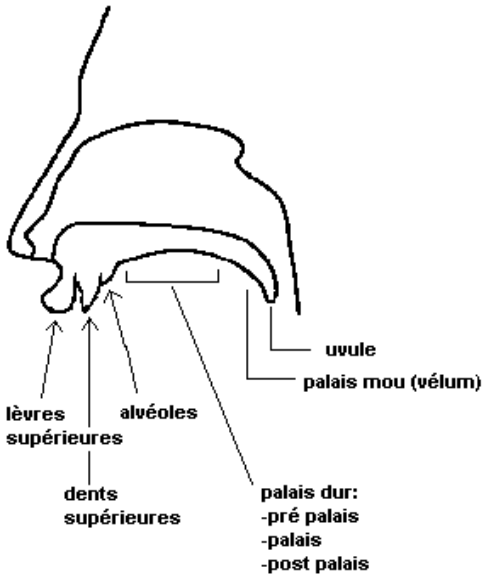
La distinction entre voyelles et consonnes s'effectue de la manière suivante :

- Si le passage de l'air se fait librement à partir de la glotte, on a affaire à une voyelle **[21]**.
- 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.



3-1-3- 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.

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-glottiques (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.

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 **[21]**. Le point d'articulation peut se situer aux endroits suivants:

- les lèvres (*labiales* ou *bilabiales*).
- les dents (articulations *dentales*).
- les lèvres et les dents (*labiodentales*).
- les alvéoles (articulations *alvéolaires*).
- le palais (*pré-palatal, médio-palatales et post-palatal*).
- le voile du palais (articulations *vélaires*).
- la luette (articulations dites *uvulaires*).
- le pharynx (articulations *pharyngales*).
- la glotte (articulations *glottales*).

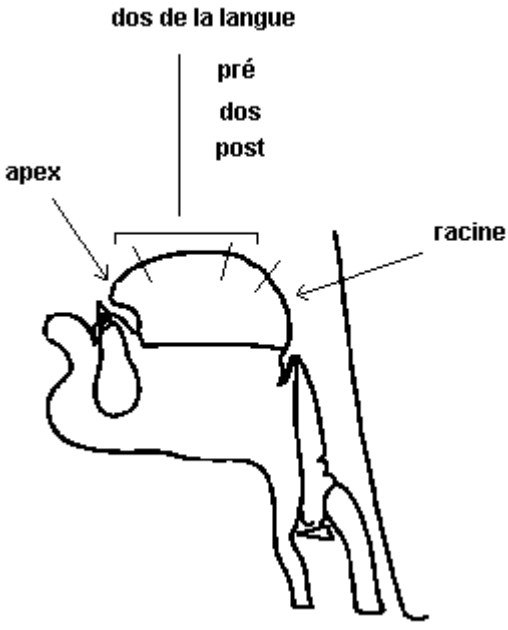


Figure 3.4 Points d'articulation

3-1-4- Description des voyelles:

La caractéristique majeure des voyelles est le libre passage de l'air à partir des cavités supra-glottiques. **[8]** Le seul traitement que l'air peut dès lors subir est la

résonance (c'est-à-dire le renforcement de certaines bandes de fréquences). Le timbre d'une voyelle dépendra de la variation des éléments suivants :

- le nombre des résonateurs (buccal, labial et nasal).
- la forme du résonateur buccal.
- le volume du résonateur buccal.

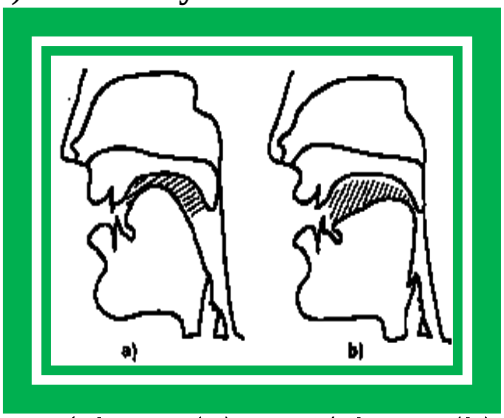
Si le voile du palais est relevé, l'air ne traverse pas le résonateur nasal, mais se répand exclusivement dans le résonateur buccal. Si le voile du palais est abaissé, l'air traverse simultanément les résonateurs buccal et nasal. Si les lèvres sont projetées vers l'avant et **arrondies**, il se forme un résonateur à la sortie du canal buccal, le résonateur labial. Si, au contraire, les lèvres sont appliquées contre les dents, le résonateur labial ne se forme pas.

D'après les critères ci-dessus, on oppose :

- Des **voyelles nasales** (présence du résonateur nasal) à des **voyelles orales** (absence).
- Des **voyelles arrondies** (résonateur labial) à des **voyelles non-arrondies** (absence).

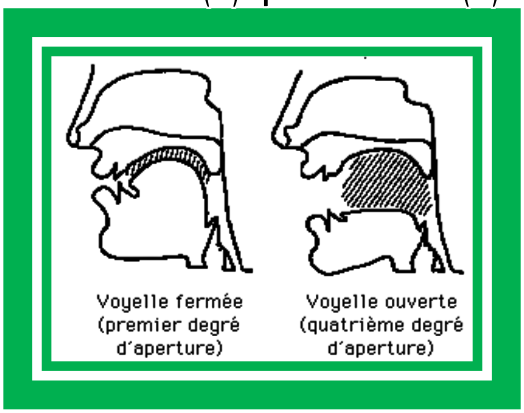
La **forme du résonateur buccal** est déterminée par l'emplacement de la masse de la langue dans la bouche. A partir de là, on envisage trois cas:

- des **voyelles antérieures** (le dos de la langue se trouve dans la région pré-palatale),
- des **voyelles postérieures** (la langue se trouve dans la région post-palatale ou vélaire).
- des **voyelles centrales** (la masse de la langue se trouve dans la région médio-palatale).



antérieures(a) postérieures(b)

Le **volume du résonateur buccal** est le dernier facteur pris en compte dans l'analyse articulatoire du timbre des voyelles. Ce volume dépend directement du **degré d'aperture**, c'est-à-dire de la distance séparant le point le plus élevé de la langue du palais.



On distingue arbitrairement quatre degrés d'aperture, du premier, le plus fermé, au quatrième, le plus ouvert.

Exemples :

 Voyelle mi-ouverte (de troisième degré d'aperture) antérieure arrondie.

 Voyelle fermée (de premier degré d'aperture) postérieure arrondie.

 Voyelle fermée (de premier degré d'aperture) antérieure non-arrondie.

3-1-5- Description des consonnes :

Il existe deux grands types d'articulations consonantiques **[8]**:

- soit le passage de l'air est fermé et le son résulte de son ouverture subite, on a alors affaire à des occlusives.
- soit le passage se rétrécit mais n'est pas interrompu; on parle dans ce cas de *continues*, dont les fricatives sont les plus représentatives.

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 **[21]***. 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.

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***. **[21]** La distinction entre mode d'articulation nasal et oral concerne aussi bien les consonnes que les voyelles.

3-1-5-1- Consonnes occlusives sourdes :

Lorsqu'elle est suivie d'une voyelle, l'évolution acoustique d'une occlusive sourde est la suivante:

- Arrêt de la colonne d'air par la fermeture soudaine du chenal expiratoire, ce qui cause un silence momentané dû à l'occlusion.
- libération de l'air interne par le relâchement brusque de l'occlusion (une explosion).
- un bruit de friction, produit au niveau de la constriction, le spectre est celui d'un bruit de bande aigu.
- un bruit d'écoulement glottal, le spectre est celui d'un bruit de bande plus aigu que le précédent, mais qui peut momentanément coexister avec lui.
- vibration de cordes vocales due à la voyelle qui suit l'articulation consonantique qui produit un spectre harmonique, l'intensité décroît régulièrement du grave vers l'aigu et ce signal apparaît avec un certain retard par rapport à l'explosion, un retard appelé en anglais « voice onset time » (V.O.T.).

3-1-5-2- Consonnes occlusives sonores :

Les étapes de la réalisation sont identiques à celles qui président à la production des occlusives sourdes, mais elles sont accompagnées d'une vibration des cordes vocales.

Le voisement a pour conséquences :

- une atténuation des bruits d'explosion et de friction.
- la disparition du bruit d'écoulement glottal, caractéristique des occlusives sourdes.

3-1-5-3- Fricatives :

Les consonnes fricatives sont produites par un resserrement du chenal expiratoire qui ne va pas, contrairement à ce qui se passe pour les occlusives, jusqu'à la fermeture complète. Ce sont essentiellement les lèvres et la langue qui, selon leur position et leur tension musculaire particulière, conditionnent le type de friction réalisée.

On distinguera ci-dessous des articulations fricatives *proprement dites* d'articulations spirantes qui leur sont apparentées.

Lors de la réalisation d'une fricative, le passage de l'air peut se faire de deux manières :

- la langue assure le passage de l'air expiré par un canal médian, ce sont les fricatives **dorsales** décrites dans la section des fricatives proprement dites.
- la langue forme un canal latéral pour le passage de l'air; ces articulations sont décrites dans la section Latérale fricative.
- finalement, il existe des articulations fricatives pour lesquelles la forme de la langue n'a pas d'importance : il s'agit des fricatives labiales et dentales (ce qui est logique puisque le point d'articulation de ces productions ne se situe pas à proprement parler dans la cavité buccale), ces articulations sont rangées dans les catégories fricatives proprement dites et spirantes.

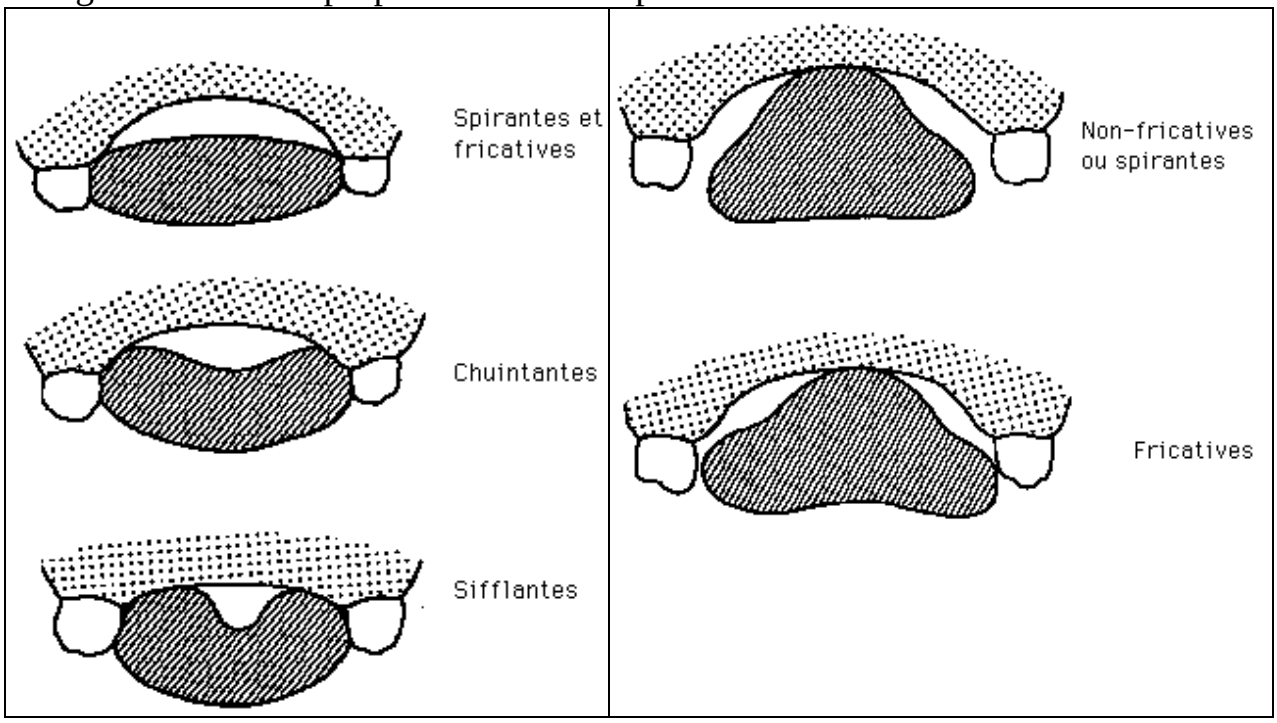


Figure 3.5- Consonnes fricatives dorsales et latérales

On trouve des articulations désignées **comme sifflantes ou comme chuintantes**. La production d'une sifflante implique une forte tension linguale: un canal se creuse sur toute la longueur de la langue, et en particulier au point d'articulation, où l'air passe par une petite ouverture ronde. Les chuintantes ressemblent aux sifflantes, mais le canal qui se creuse sur la langue est moins profond, et l'ouverture au point d'articulation est plus ovale. Les lèvres sont souvent arrondies ou projetées vers l'avant lors de la réalisation d'une chuintante.

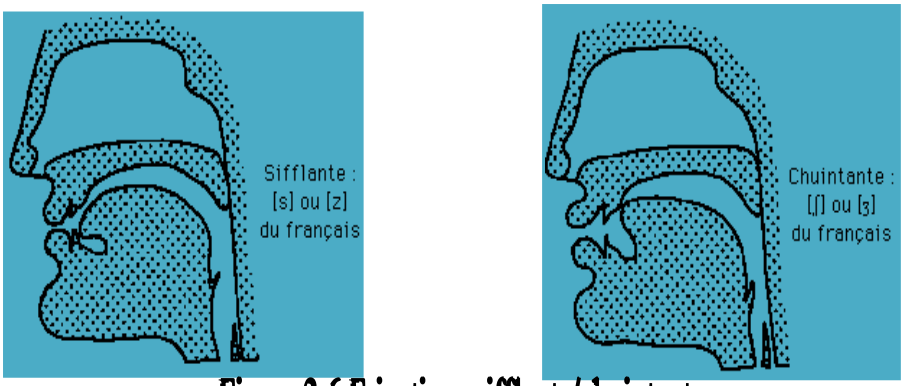


Figure 3.6 Fricatives sifflante/chuintante

3-1-5-4- Spirantes

Les spirantes présentent le même rétrécissement du chenal expiratoire que les fricatives, mais la tension des organes phonateurs lors de la réalisation d'une spirante est beaucoup plus faible, ce qui a pour conséquence, non plus de produire une friction, mais d'engendrer un effet de résonance au point d'articulation.

En gros, il y a friction lorsque l'articulation est tendue, ce qui engendre des fricatives; il y a résonance quand l'articulation est lâche, ce qui produit une spirante, tous les autres facteurs étant égaux par ailleurs.

Notons encore qu'on peut faire correspondre de nombreuses spirantes à des articulations occlusives relâchées. Ces correspondances seront signalées dans les descriptions des articulations.

3-1-5-5- Latérales :

On considère généralement les articulations latérales comme des articulations particulières, bien que, physiquement parlant, on puisse les classer parmi les fricatives et les spirantes.

On appelle ces articulations *latérales* car, lors de leur réalisation, le dos de la langue prend contact avec le palais, alors que l'avant de celle-ci s'affaisse pour laisser s'écouler l'air interne par un canal latéral ou parfois bilatéral. À l'inverse, pour les articulations dorsales, la langue prend appui sur les molaires, et l'air s'écoule par un canal médian, sur le dos de la langue.

On distingue donc deux types de latérales:

- les latérales fricatives, dont l'articulation, requérant une forte tension musculaire, ressemble fortement, à l'exception du point d'appui de la langue, à celle des fricatives.
- les latérales non-fricatives, parfois appelées *liquides*, dont l'articulation est très proche de celle des spirantes.

L'emplacement du canal latéral par lequel s'écoule l'air n'a pas d'importance : qu'il soit à gauche, à droite ou même bilatéral, la qualité du son n'est pas altérée.

3-1-5-6- Consonnes nasales :

Comme l'air peut s'écouler librement par les fosses nasales, l'occlusion peut être maintenue très longtemps. L'explosion est extrêmement brève, voire presque absente.

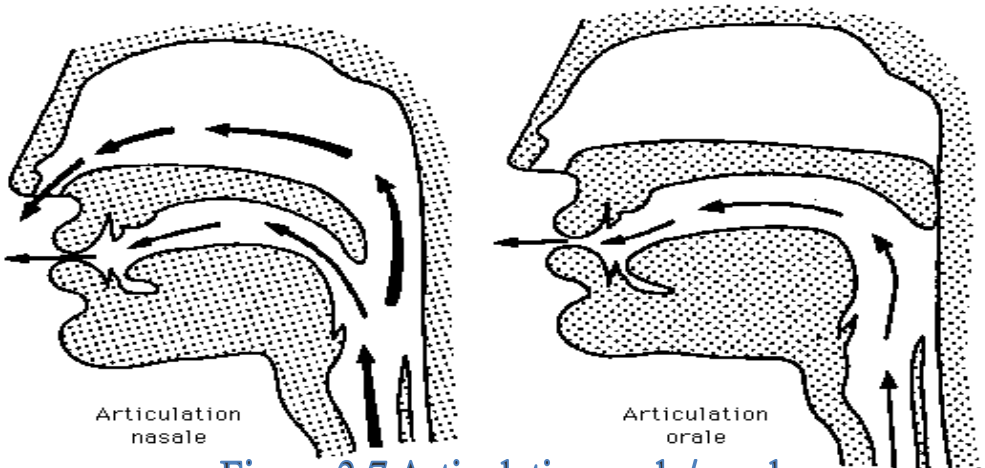


Figure 3.7 Articulation orale/nasale

3-1-5-7- Vibrantes:

Les consonnes vibrantes sont le produit d'un ou de plusieurs *battements*, c'est-à-dire de vibrations, sous la pression de l'air interne, d'un des organes de la parole: pointe de la langue, voile du palais ou luette. L'organe concerné prend contact avec un point fixe, opposé, du chenal expiratoire. Il en résulte une ou plusieurs occlusions successives, très rapides, accompagnées de résonances brèves. Les vibrantes sont généralement sonores.

On peut répartir les vibrantes en deux classes :

- les vibrantes à un seul battement, dites vibrantes battues.
- les vibrantes à plusieurs battements, dites vibrantes roulées.

3-1-5-8- La notion de semi-voyelle :

Pour les réalisations vocaliques les plus fermées, l'aperture buccale doit respecter une dimension minimale en dessous de laquelle on n'a plus affaire à des voyelles, mais à des consonnes spirantes ou fricatives, selon le degré de la tension musculaire.

D'autre part, la réalisation des voyelles les plus fermées exige également une durée articulatoire minimale, en dessous de laquelle on ne perçoit plus une voyelle mais une fricative ou une spirante.

On appelle *semi-voyelles* les sons produits par l'un comme l'autre des phénomènes décrits ci-dessus, bien que ces procédés articulatoires soient assez différents. Les

semi-voyelles sont produites par le passage de l'air à travers le conduit vocal, mais ce dernier fonctionne également et simultanément en mode résonnant.

3-1-5-9- Coarticulation:

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

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

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

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

Phénomènes de Coarticulation:

- **Assimilation** : phénomène par lequel un son tend, du fait de sa proximité par rapport à un autre, à devenir identique, ou à prendre certaines de ses caractéristiques (voisement ou dévoisement par exemple).

- **Dilation :** modification des caractéristiques d'un son due à l'anticipation d'un autre son qui ne lui est pas contigu.

- **Différenciation :** changement phonétique qui a pour but d'accentuer ou de créer une différence entre deux sons contigus.
- **Dissimilation :** changement phonétique qui a pour but d'accentuer ou de créer une différence entre deux sons voisins mais non contigus.

- **Interversion :** lorsque deux sons contigus changent de place dans la chaîne parlée.

- **Métathèse :** lorsque deux sons non contigus changent de place dans la chaîne parlée.

3-1-6- Description articulatoire des consonnes arabes :

3-1-6-1- Occlusives orales :

Occlusive bilabiale sonore.

b 

Les deux lèvres prennent fermement contact l'une contre l'autre, avec vibration des cordes vocales.

Occlusives alvéo-dentales

t 

Occlusive dentale ou alvéolaire sourde. La langue Prend contact avec le bourrelet formé par les alvéoles. A cette occlusive correspond une articulation relâchée, qui prend la forme d'une spirante.

Occlusive dentale ou alvéolaire sonore.

d

Même articulation que la précédente, mais avec Vibration des cordes vocales. A cette occlusive correspond une articulation relâchée, qui prend la forme d'une spirante.

Occlusive rétroflexe sourde.

ʈ 

La langue est retournée et sa pointe ou sa face intérieure prend appui sur un point de la partie antérieure du palais.

Occlusive rétroflexe sonore.

ɖ

Même articulation que la précédente, mais avec vibration des cordes vocales. L'articulation nasale rétroflexe est également très souvent sonore.

Occlusives vélaires

k 

Occlusive vélaire sourde. Alors que la pointe de la langue est appuyée contre la face interne des dents du bas, le dos de celle-ci prend contact avec le palais mou, appelé aussi voile du palais.

Occlusives uvulaires

q 

Occlusive uvulaire sourde. Pendant que la pointe de la langue demeure appliquée contre la face interne des dents du bas, le dos de celle-ci, relevé loin vers l'arrière prend contact avec le palais mou au niveau de la luette.

3-1-6-2- Occlusives nasales :

Occlusive bilabiale

Occlusive bilabiale nasale. La partie buccale de l'articulation est la même que pour l'orale correspondante : les deux lèvres prennent fermement contact l'une contre l'autre.

Occlusive alvéo-dentale

Occlusive dentale ou alvéolaire nasale. La partie buccale de l'articulation est la même que pour l'orale correspondante : la langue prend contact avec le bourrelet formé par les alvéoles.

3-1-6-3- Fricatives :

Fricatives sifflantes alvéolaires

Sifflante alvéolaire (fricative) sourde. Les sifflantes apico-alvéolaires sont produites par le rapprochement de la pointe de la langue vers la région alvéolaire. En ce qui concerne la forme de la langue, cette articulation suit les règles générales établies pour les sifflantes.

Sifflante alvéolaire (fricative) sonore

Même articulation que la précédente, mais avec vibration des cordes vocales. A ce détail près, les remarques faites pour la variante sourde s'appliquent à la sonore.

Fricatives sifflantes rétroflexes

Sifflante rétroflexe (fricative) sourde. La pointe de la langue est dirigée vers le haut et vers l'arrière ; la partie inférieure de la langue se rapproche de la partie antérieure du palais.

Fricatives spirante inter-dentale sonore rétroflexe. La pointe de la langue est proche des dents de la mâchoire supérieure avec vibration des cordes vocales. 

Fricatives chuintantes alvéolaires

Chuintante alvéolaire (fricative) sourde. La langue prend appui contre les alvéoles ; la forme de la langue est telle que décrite pour les chuintantes en général. 

Chuintante alvéolaire (fricative) sonore. Même articulation que la précédente, mais avec vibration des cordes vocales. 

Fricatives labiodentales

Fricative labiodentale sourde. La lèvre inférieure est rapprochée des dents du haut, et peut parfois les effleurer avec sa partie externe supérieure ou, parfois, avec sa partie interne). 

Fricative vélaire

Fricative vélaire sourde. La partie postérieure du dos de la langue se rétracte fortement vers l'arrière et vers le haut, au niveau du palais mou (ou voile du palais). 

Fricative vélaire sonore. Même articulation que la précédente, mais avec vibration des cordes vocales. 

Fricatives pharyngales

Fricative pharyngale sourde. La racine de la langue est fortement repoussée vers l'arrière et se rapproche de la paroi postérieure du pharynx. Le passage de l'air est alors considérablement rétréci et on perçoit une forte friction. La tension articulatoire est très forte. 

Fricative pharyngale sonore. Même articulation que la précédente, mais avec vibration des cordes vocales.

Fricatives glottales

Fricative glottale sourde. La glotte est presque entièrement close, à l'exception d'une étroite ouverture dans sa partie supérieure au niveau des cartilages aryténoïdes, On perçoit une forte friction quand l'air s'écoule par ce canal.

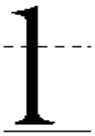
Fricatives spirantes inter-dentales

Fricatives Spirantes inter-dentale sonore. La pointe de la langue est proche des dents de la mâchoire supérieure; elle peut soit se placer derrière les dents, soit dépasser quelque peu à l'extérieur (on parle alors de *dentale*). Cette articulation peut être interprétée comme la réalisation relâchée de l'occlusive [t].

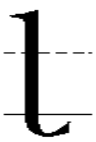
Fricatives spirantes inter-dentale sonore. Même articulation que la précédente, mais avec vibration des cordes vocales. Cette articulation peut être interprétée comme la réalisation relâchée de l'occlusive [d].

3-1-6-4- Latérales, sonnante :

Latérale, sonnante alvéolaire. La pointe de la langue se pose sur le bourrelet gingival. L'air s'écoule sur les côtés de la langue.



Latérale, sonnante alvéolaire rétroflexe. La pointe de la langue se retourne vers l'arrière, et la face inférieure de celle-ci prend contact avec la partie antérieur du palais dur. Les côtés de la langue s'abaissent pour permettre le passage de l'air.



3-1-6-5- Vibrante :

Vibrante roulée

Vibrante roulée alvéolaire. La région alvéolaire sert de point d'appui à la pointe de la langue qui entre en vibration sous la poussée de l'air interne. Cette vibration produit de petites occlusions successives, entrecoupées de résonances de type vocalique.

Vibrante battue

Vibrante battue alvéolaire rétroflexe. La pointe de la langue, dirigée vers l'arrière, vibre brièvement contre le palais antérieur. Elle prend contact avec celui-ci par sa face intérieure, puis va se rabattre en position de repos, plate, la pointe contre les dents du bas.

3-1-6-6- Description des articulations semi-vocaliques :

Sonnante, centrale palatale sonore.

Le son produit provient soit de l'articulation très fermée d'une voyelle antérieure non-arrondie du premier degré d'aperture (un [i]), soit d'une même articulation vocalique, normalement fermée, mais dont la réalisation est brève Le dos de la langue se creuse en canal et se rapproche de la partie antérieure ou centrale du palais dur, avec vibration des cordes vocales.

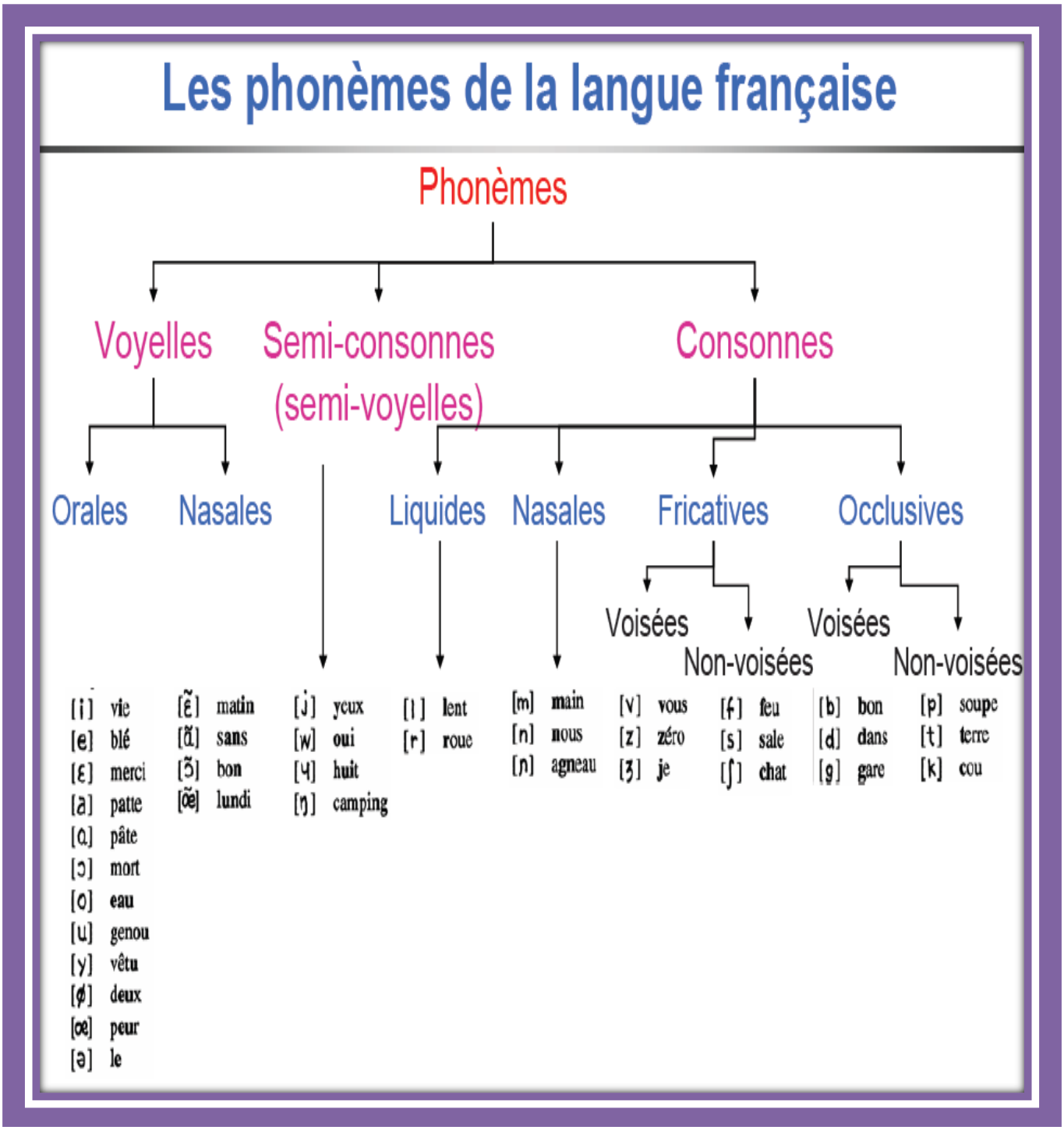
Sonnante, centrale labio-vélaire sonore.

Le son produit provient soit de l'articulation très Fermée d'une voyelle postérieure arrondie de premier degré d'aperture (un [u]), soit d'une même articulation vocalique, normalement fermée, mais dont la réalisation est brève.

Figure 3-8 ***:Schéma de l'API***

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

Figure 3-9 : ***Schéma de l'arbre phonétique***



3-2- Phonétique acoustique:

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire.la phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone(lui-même associé a un préamplificateur).le signal électrique résultant est le plus souvent numérisé.il peut alors être soumis à un ensemble de traitement statistique qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre.

3-2-1- Le son : Qu'est-ce que c'est ?

Une vibration mécanique de la matière et de l'air qui met en alternance le tympan ou le micro ne constitue pas en elle-même un son. Car c'est dans le cerveau que naît et se forme le son. L'oreille recueille les vibrations de l'air, les transforme en impulsion électrique au moyen des cellules nerveuses, impulsion qui est perçue et interprétée en son par le cerveau. Le son est donc une perception.

Par définition, le son est ce que l'oreille perçoit de la vibration d'un corps. Généralement la vibration se propage dans l'air jusqu'à l'oreille, mais le son se propage aussi dans l'eau et même dans les corps solides. Le son est donc une vibration, une sorte d'onde (produite par un objet) qui se propage par et à travers des corps physiques (air, eau, métal, bois...). **[4]** Cet ébranlement de la matière peut se caractériser par une variation de pression que l'on mesurerait dans une pièce avec un instrument de mesure adéquat (par exemple une oreille). On parle alors de pression acoustique.

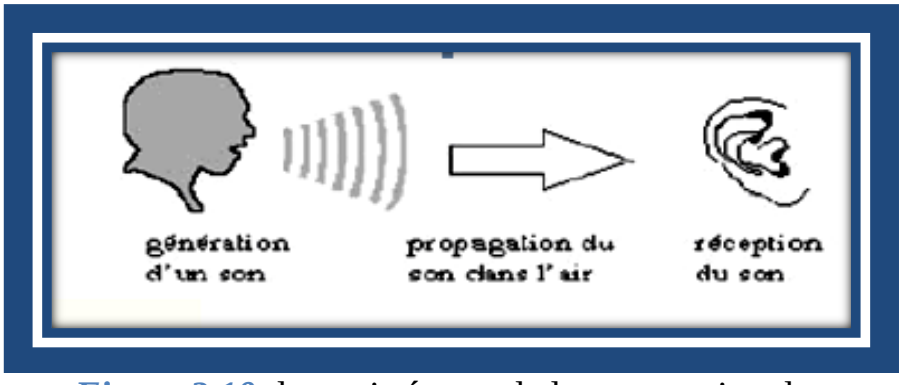


Figure 3.10- les trois étapes de la propagation du son

Au final, les molécules du corps physique restent à leur place. Elles oscillent, vibrent, autour d'une position. C'est en oscillant de cette façon qu'elles rencontrent et bousculent les molécules voisines qui, de proche en proche, vont se mettre à

osciller et transporter le son à leur tour, et ce jusqu'à l'oreille, si l'excitation primitive des premières molécules était assez puissante pour faire vibrer les molécules jusque là. Empiriquement, la propagation du son diminue avec la distance.

Ceci est dû à **l'amortissement** du système : l'air oppose une résistance au son, comme il oppose une résistance aux corps qui le traversent. Il en est de même d'autres fluides comme l'eau. C'est cette résistance qui entraîne la diminution du volume sonore avec la distance, puis la disparition complète du son.

C'est parce que cette propagation se fait de proche en proche qu'elle n'a pas une vitesse très grande d'une part, et qu'elle s'atténue relativement vite, d'autre part (heureusement). Par conséquent la vitesse d'un son, appelée aussi «célérité», est fonction de certaines caractéristiques du corps qui le propage.

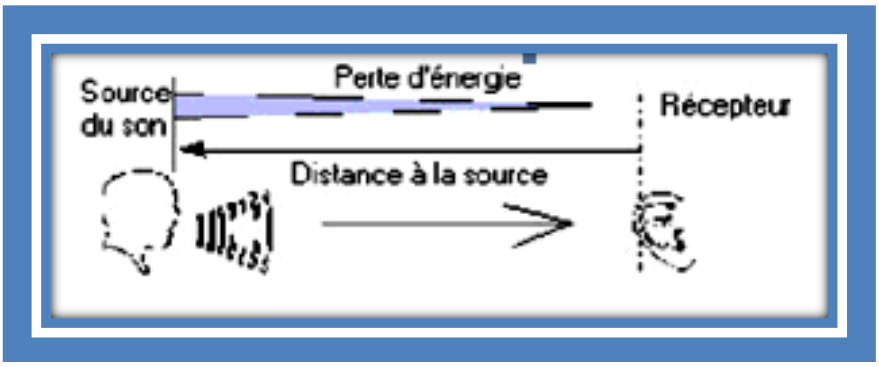


Figure 3.11- Propagation du son atténuée

3-2-2- Du son au signal :

Le signal correspond à la mesure d'une grandeur physique. Mesure de grandeur physique : signal sismique, mesure du pouls, déplacement, intensité, etc.… la plupart des grandeurs physiques sont aujourd'hui converties en signaux électriques plus codées en signal numérique binaires. Donc Le signal est une variation (dans le temps de préférence) d'une grandeur physique de nature quelconque porteuse d'information. L'opération de numérisation du signal audio se réalise en théorie en trois étapes (échantillonnage, quantification, codage).

3-2-2-1- Échantillonnage :

Avant tout traitement, il est nécessaire de numériser le signal continu sortant du microphone ou d'un appareil d'enregistrement. Cette opération s'appelle échantillonnage du signal. L'échantillonnage procède à un découpage dans le temps du signal X(t). Sachant que l'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50Hz-8KHz], la fréquence

d'échantillonnage devrait donc au moins être égale à 16 KHz, selon le théorème de Shannon.

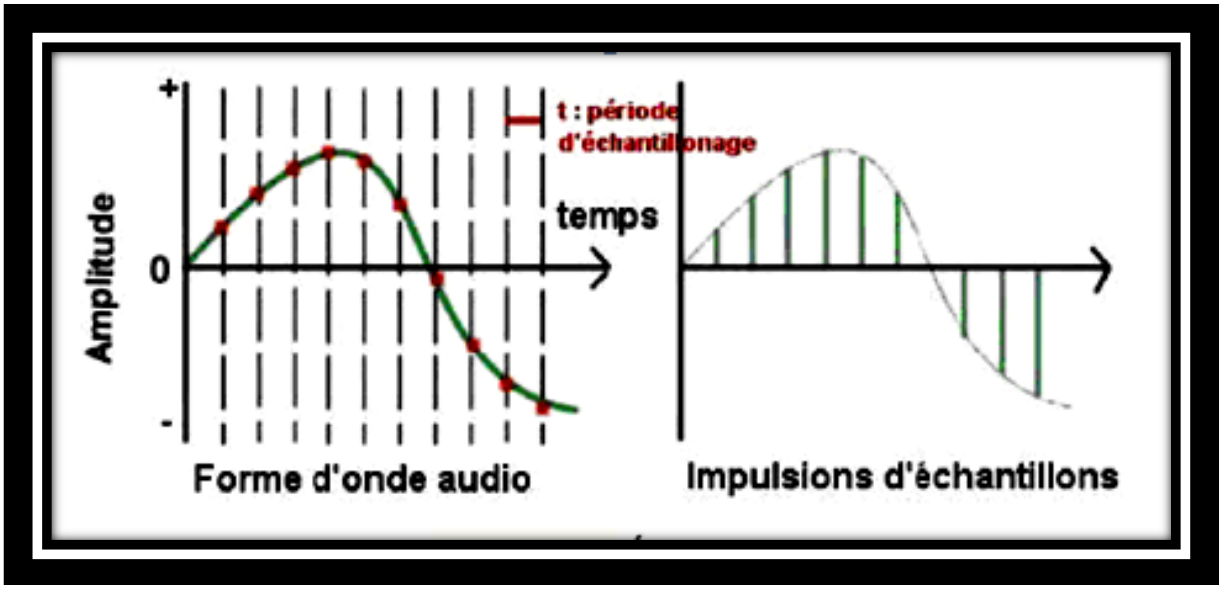


Figure 3.12- Échantillonnage

3-2-2-2- Quantification: [2], [4], [21]

Cette étape consiste à approximer les valeurs réelles des échantillons selon une échelle de n niveaux appelée échelle de quantification. Chaque impulsion correspond donc à un nombre binaire unique.

- Une quantification à n bits permet d'utiliser 2^n valeurs différentes.
- Pour 8 bits, on a 256 valeurs et pour 16 bits, on a 65536 valeurs.

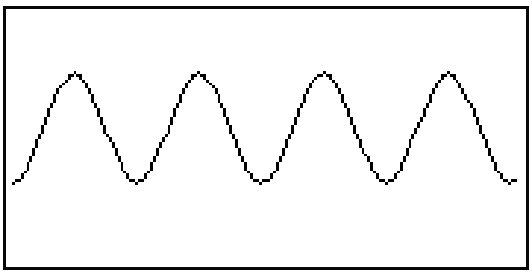
La transformation d'une valeur physique (en volts) en une valeur binaire introduit donc une distorsion. L'erreur systématique que l'on commet en assimilant les valeurs réelle de l'écart au niveau du quantifiant le plus proche est appelé bruit de quantification.

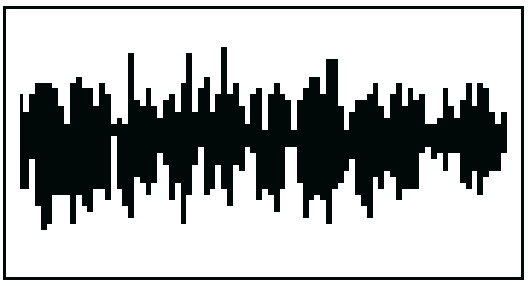
3-2-2-3- Le codage: [2], [4], [21]

C'est la représentation binaire des valeurs quantifiées qui permet le traitement du signal sur machine. Donc le codage désigne le type de correspondance que l'on souhaite établir entre chaque valeur du signal analogique et le nombre binaire qui représentera cette valeur.

3-2-3- Caractéristiques de l'onde sonore :

Une vibration est dite ***périodique*** lorsqu'elle se répète à des intervalles de temps égaux, au contraire les ondes ***apériodiques*** sont irrégulières, anarchiques, elles sont à l'origine de ce que nous percevons comme des " bruits ".





L'***amplitude*** d'un son correspond à la variation de pression maximale de l'air engendrée par les oscillations, et donc au volume sonore.

L'amplitude d'une vibration peut être exprimée objectivement par le calcul des variations de pression d'air (exprimée en Micron Bar et convertie en watt/cm^2). On utilise toutefois plus fréquemment une unité de mesure relative, le décibel (dB), pour rendre compte de *l'intensité* d'un son.

La vitesse des mouvements d'aller et de retour des vibrations est responsable de la sensation de hauteur. Plus les mouvements vibratoires sont rapides, plus le son sera aigu. À l'inverse, un mouvement plus lent engendre un son plus grave.

De façon objective, la hauteur d'un son correspond à sa ***fréquence*** qui est exprimée en cycles par secondes ou Hertz.

Un son comportant 100 cycles par seconde, soit 100 mouvements complets d'aller et de retour par rapport au point de repos, aura une fréquence de 100 Hertz.

Le ***timbre*** est un paramètre beaucoup plus subjectif: il s'agit de ce qui différencie deux sons de même hauteur et de même amplitude.

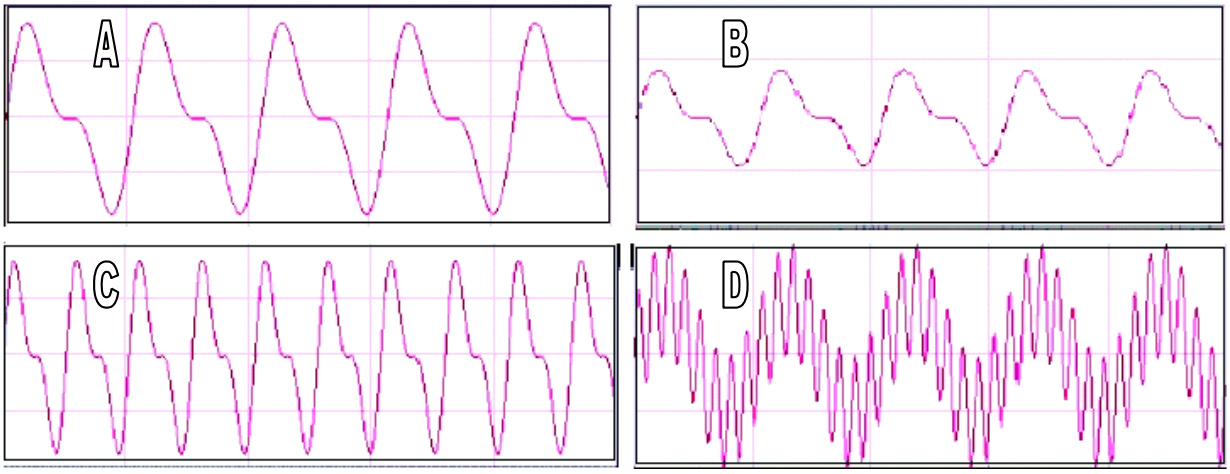


Figure 3.13- Exemples d'ondes périodiques

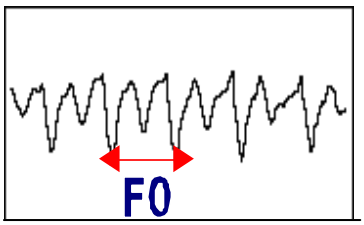
Les 4 figures ci-dessus montrent :

- A) Une onde A périodique.
- B) L'onde A avec une amplitude 2 fois plus faible.
- C) L'onde A avec une fréquence plus élevée (10 au lieu de 5).
- D) Une onde D avec les mêmes (fréquence/amplitude) que A mais avec un timbre différent.

L'onde apériodique complexe est constituée d'une multitude de mouvements vibratoires anarchiques. Les sons tels que " **s** " ou " **ch** " sont formés d'ondes apériodiques complexes. L'onde périodique complexe est constituée par la sommation d'ondes périodiques simples (sinusoïdales), c'est-à-dire une somme de sons purs. Il est donc possible de décomposer l'onde périodique complexe en ondes simples et de calculer la fréquence de chacune d'entre elles.

3-2-4- Propriétés de l'onde périodique complexe :

L'onde périodique complexe est constituée d'une première onde que l'on appelle le " fondamental ou F0" qui correspond à la période de l'onde. C'est la fréquence de cette onde qui nous permet d'évaluer, de façon globale, la hauteur du son. Les ondes qui accompagnent le fondamental sont appelées les *harmoniques*. La fréquence de chacune des harmoniques est un multiple entier de la fréquence du fondamental.



Onde périodique complexe

Les harmoniques possèdent ainsi des fréquences propres qui sont plus élevées que celle du fondamental. Leur intensité est, par contre, moins importante que l'intensité du fondamental. Ces harmoniques sont générées par des vibrations des cordes vocales, Donc la fourniture laryngée ou, si l'on préfère, le voisement produit donc un son riche, composé de la fréquence du fondamental et d'harmoniques.

3-2-5- Visualisation des sons :

3-2-5-1- L'oscilloscope:

L'oscilloscope, l'une des plus anciennes de ces représentations, montre l'évolution temporelle de l'amplitude du signal. C'est une simple fonction du temps qui ne dévoile pas la structure interne du son (sa composition fréquentielle) et qui se révèle peu intéressante pour des objets sonores complexes et notamment pour l'étude de la parole.

3-2-5-2- Spectrogramme:[2]

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 niveau dans un diagramme en deux dimensions temps-fréquence. Alors 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.

On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (10ms), ils mettent en évidence l'enveloppe spectral du signal, et permettant par conséquent de visualiser l'évolution temporelle des formants.

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné, une transformée de Fourier rapide étant régulièrement calculée à des intervalles de temps rapprochés. Avant le calcul des transformées successives, le signal doit d'abord être préaccentué par un filtre du premier ordre pour égaliser les hautes fréquences dont l'énergie est toujours plus faible que celle des basses fréquences.

Cette phase de préaccentuation du signal est suivie par une phase de fenêtrage, nécessaire du fait de la théorie qui sous-tend la transformée de Fourier. Dans cette

méthode d'analyse, le signal est considéré comme indéfiniment stable et constitué d'une somme invariable de fonctions sinusoïdales de fréquences différentes. Pour contourner cette contrainte théorique d'invariabilité du signal, il faut convoluer le signal avec une fenêtre temporelle qualifiée de glissante puisque chaque calcul de spectre nécessite de convoluer le signal avec la fenêtre temporelle à un instant particulier.

Le choix de la taille de la fenêtre, en nombre de points de convolution, est également important vis-à-vis de la qualité de l'analyse fréquentielle obtenue. Ainsi, une fenêtre de petite taille (avec un nombre de 128 points, par exemple) permettra d'obtenir une bonne analyse dans le domaine temporel, du fait de son étroitesse, mais ne permettra pas d'obtenir une bonne information fréquentielle, la taille de la fenêtre étant alors trop petite pour ne pas tronquer les phénomènes de basses fréquences. À l'inverse, une fenêtre de grande taille (plus de 512 points) permettra d'obtenir une bonne information fréquentielle mais ne permettra pas d'obtenir une bonne information temporelle car tout événement, même de courte durée, est jugé présent sur l'ensemble du pas de temps analysé puisque la théorie de la transformée de Fourier considère les signaux indéfiniment stables. Une fois la convolution effectuée, la transformée de Fourier est calculée sur la totalité de la fenêtre, le reste du "signal" étant alors égal à 0.

Ceci permet d'obtenir un spectre qui correspond à une trame, et l'ensemble de trames calculées permet d'obtenir le spectrogramme désiré.

$$S[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \qquad avec \quad 0 \le k < N \qquad \textbf{3-1}$$

Où **X**(n) est la composante temporelle et S(K) est la composante fréquentielle

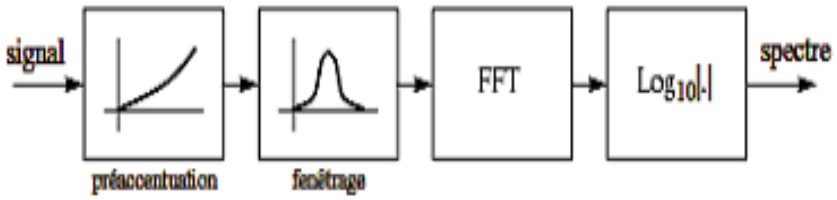


Figure 3.14- Méthode de calcul d'une transformée de Fourier rapide

3-2-5-3- Problèmes posés par la transformée de Fourier : [2],[4]

La transformée de Fourier et l'implantation algorithmique efficace qui y a été associée, la transformée de Fourier rapide, présente de nombreux avantages en tant que méthode d'analyse temps-fréquence.

La rapidité de sa mise en œuvre l'a propulsé au rang d'élément incontournable des systèmes de traitement de signal. Mais, après la naissance de la notion de représentation temps-fréquence, qui fait suite à l'utilisation de représentations spectrographiques, les études théoriques du domaine ont permis de mettre à jour quelques désavantages qui sont impossibles à éliminer et qui constituent ainsi les limites d'exploitation de la transformée de Fourier.

Au rang de ces problèmes se trouve le compromis entre finesse d'analyse en fréquence et en temps. Le fait que la transformée de Fourier ne prenne pas en compte les dépendances temporelles implique, lorsque cette méthode est adaptée aux signaux non stationnaires, de considérer l'inégalité **d'Heisenberg-Gabor**. Cette inégalité postule qu'un signal ne peut être concentré sur des supports temps et fréquence qui soient, simultanément, arbitrairement petits.

Une autre constatation exhibe une limitation qui dépasse le cadre de l'inégalité d'Heisenberg Gabor et qui nous amène à nous demander ce que les transformées de tous types permettent de représenter. La théorie de **Slepian-Pollack-Landau** prouve en effet qu'un signal ne peut pas parfaitement confiner son énergie sur des supports finis, même s'ils sont arbitrairement grands.

La transformée de Fourier et les autres transformées existantes ne permettent donc pas de représenter correctement un signal temporel discret, qui est déjà une approximation de la réalité. Ainsi, bien que la transformée de Fourier permette d'extraire d'un signal des connaissances a priori inaccessibles, l'information obtenue ne peut pas, théoriquement, être correcte. Ce qui pousse certains chercheurs du domaine à dire que nous serons toujours à la recherche d'une inaccessible fréquence instantanée.

Mais ces limites théoriques relatives aux représentations temps-fréquence ne sont pas les seuls problèmes existants. Le défaut majeur de la transformée de Fourier pour l'étude de la parole vient de l'inévitable intermodulation source/conduit présente dans le spectre qui ne permet pas de connaître précisément la hauteur du fondamental. Cette intermodulation est due à la convolution qui est réalisée par le conduit vocal sur la fréquence fondamentale produite par les cordes vocales.

La déconvolution ne pouvant pas être réalisée par une simple transformée, il a donc fallu développer une technique particulière capable de la réaliser pour fournir ces deux informations utiles à l'analyse de la parole.

L'étude des représentations temps-fréquence et les limites de la transformée de Fourier ont donc poussé à créer des méthodes de traitement de signal plus adaptées à la parole, que ces méthodes soient spécifiques à la recherche ou qu'elles soient créées pour des applications plus industrielles avec une volonté de compression maximale du signal agrémentée d'une conservation de sa qualité subjective.

Le spectrogramme, fondé sur la transformée de Fourier, n'est cependant pas la seule méthode d'analyse existante, Les méthodes de représentation temps-fréquence ont grandement évolué ces dernières années avec l'apparition des méthodes de transformation comme la transformer en ondelettes, mais la transformée de Fourier reste, aujourd'hui encore, la principale méthode d'analyse du signal du fait du nombre de connaissances accumulées au cours des recherches passées.

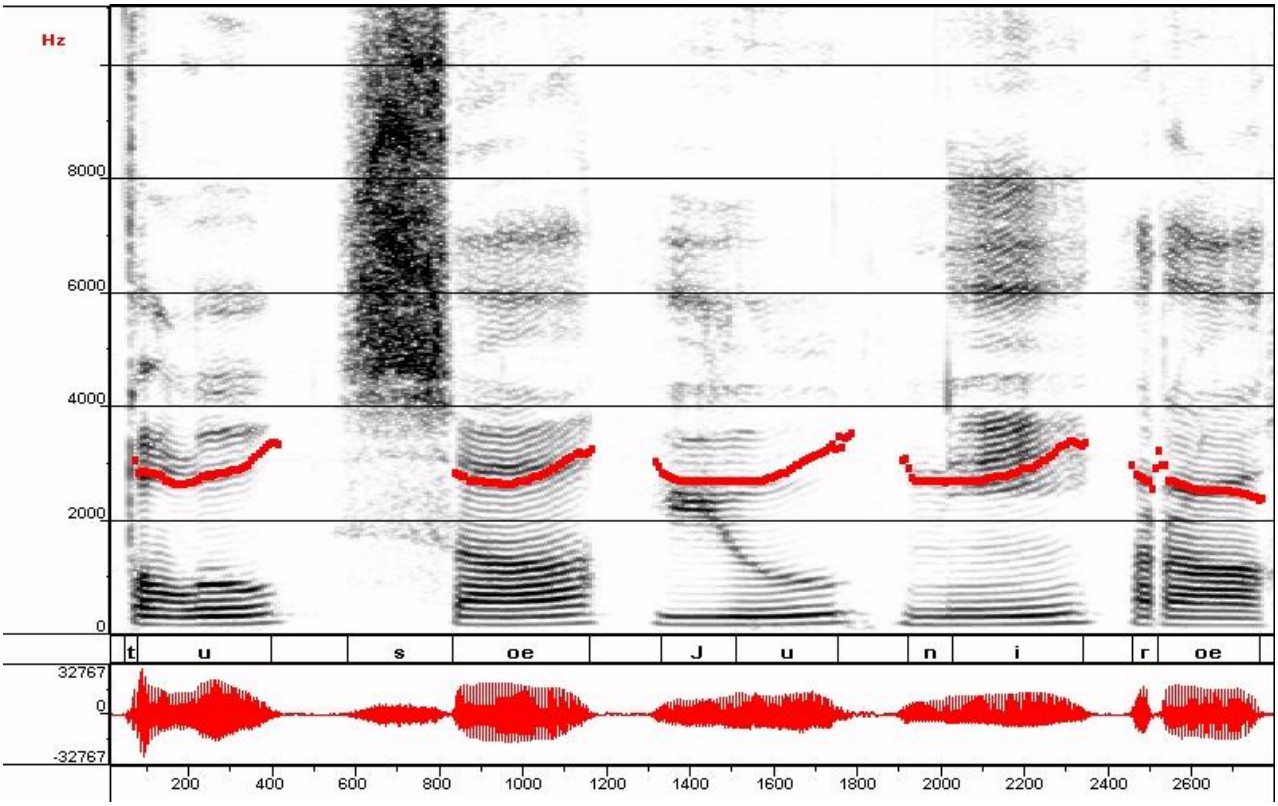


Figure 3.15-Exemple d'un spectrogramme bande étroite avec F0 marqué en rouge

Dans la figure ci-dessus une représentation temporelle (en bas oscilloscope) et fréquentielle (en haut spectrogramme), et la fréquence fondamentale marquée en haut en rouge de 5 syllabes CV (consonne voyelle) **tu**, **soe**, **ju**, **ni**, **roe**.

Remarque :

Structure harmonique claire pour les voyelles **[oe]**, **[u]**, **[i]**, pour **[u]** c'est surtout les harmoniques en basses fréquences qui sont renforcées (0-1000)Hz, pour **[i]** les harmoniques entre (1000-3000)Hz sont affaiblies, pour **[oe]** les harmoniques sont maintenues à peu prés partout entre (0-4000) Hz, ce qui laisse penser que les résonateurs du conduit vocale agissent comme un filtre sur une source (riche en harmoniques) pour la production des voyelles. Certes, mais les voyelles ne sont pas les seules à présenter un spectre riche en harmoniques, les consonnes sonores tel que **[j]**, **[n]**, **[r]** qui montrent que leur production est accompagnée d'une vibration des cordes vocales, source de l'onde périodique de période $\mathbf{f_0}$ et riche en harmoniques, mais leur durée en générale est moins importante. Mais ce n'est certainement pas le cas pour le **[s]** qui ne présente pas de structure harmonique et une dispersion d'énergie instable située dans les hautes fréquences comme du bruit. Enfin pour **[t]** ainsi pour la plupart des occlusives sourdes, une barre d'explosion de bruit dont la durée dans le temps est vraiment très petite qui rend son étude vraiment difficile avec un spectrogramme de bande étroite.

3-2-5-4- Bandes étroites versus bandes larges :

Puisque la largeur de la fenêtre détermine la résolution spectrale de l'analyse, il apparaît donc un **conflit** entre la résolution **temporelle** et la résolution **fréquentielle**,

Il s'agit essentiellement d'effectuer un choix entre les paramètres qui nous intéressent :

- Un spectrogramme à bandes étroites (10-32 ms) offre une bonne résolution au niveau fréquentiel permettant une bonne représentation de la structure harmonique du signal, mais l'analyse temporelle est moins fine (difficile d'étudier des phénomènes de courtes durées telles que le phonème t dans l'exemple ci-dessus).
- Inversement, un spectrogramme à bandes larges (3-6 ms) offre une meilleure résolution temporelle et permet de dégager les **formants** vocaliques, mais l'analyse fréquentielle est moins fine avec la disparition de la structure harmonique du signal.

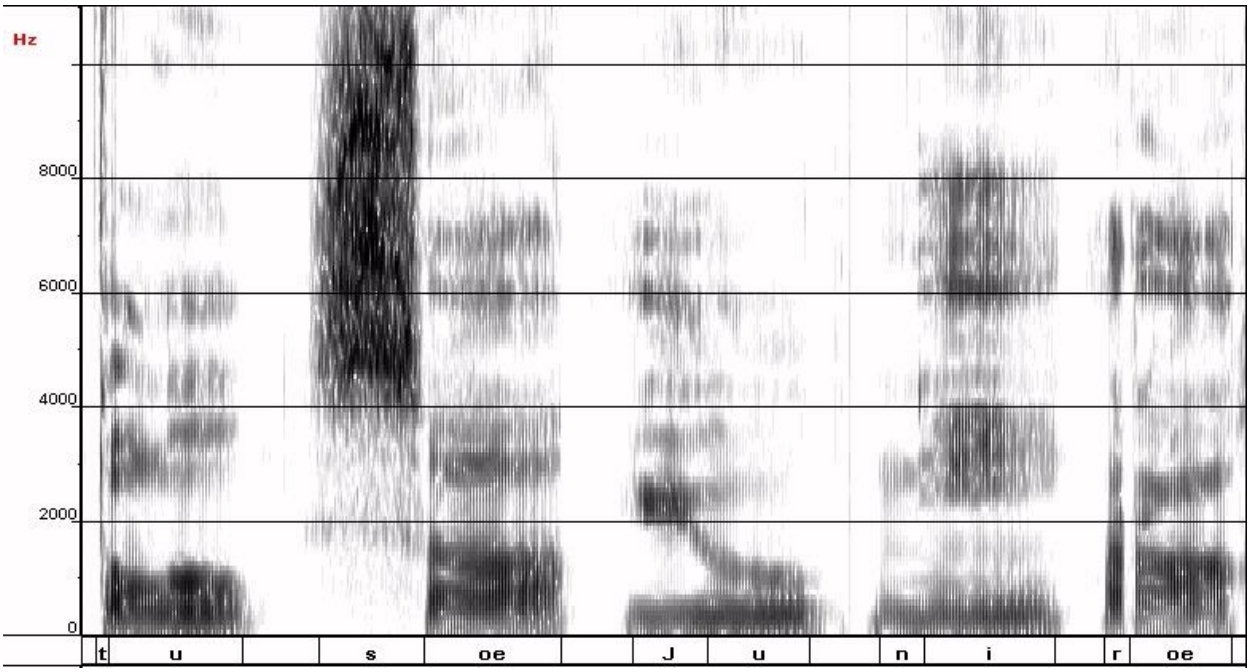


Figure 3.16- Exemple d'un spectrogramme bande large

Remarque :

On indique ici brièvement certaines des analyses particulières qui peuvent être menées sur un signal sonore numérisé en entrée d'un logiciel dédié à l'étude de la parole, elles s'adressent souvent aux voyelles mais peuvent s'appliquer à tout message parlé.

3-2-5-5- La fréquence fondamentale:

La fréquence fondamentale $\mathbf{F_0}$ (ou pitch [période des sons voisés]) joue un rôle important dans la parole **[4]**. C'est elle qui véhicule une grande partie de l'information prosodique. L'intensité de la voix et les durées successives des syllabes complètent ces informations. D'une manière générale, la prosodie, qui peut être considérée comme l'effet des différentes variations de la fréquence fondamentale $\mathbf{F_0}$, de l'intensité et de la durée, peut faire ressortir bien des caractéristiques du locuteur, comme son genre, ses origines géographiques et culturelles, ses émotions, etc. mais participe aussi à la caractérisation de la langue elle-même, par la manière dont elle est utilisée pour différencier les divers éléments syntaxiques comme les énoncés (interrogatifs, exclamatifs ou déclaratifs), l'importance de certains mots, ou bien même pour caractériser les différences lexicales entre les mots. $\mathbf{F_0}$ moyen-homme 100 à 150 Hz, F0 moyen-femme 200 Hz à 300 Hz et $\mathbf{F_0}$ moyen-enfant 350 à 400 Hz.

3-2-5-6- Les formants:

Le voisement que nous percevons est différent de celui produit à la source par les cordes vocales. Ce que nous entendons est le fruit d'un phénomène de filtrage sur une onde complexe. La capacité d'amplifier ou, au contraire, d'atténuer certaines fréquences est la propriété de tout résonateur. Dans le cas précis de la parole, le stimulus est fourni par l'onde périodique complexe provenant du mouvement des cordes vocales et ce sont les cavités supra-glottiques qui assurent la fonction de résonateur. La forme, la dimension ainsi que la matière qui compose ces cavités sont autant de particularités qui détermineront les fréquences qui seront mises en évidences et celles qui seront atténuées. Les cavités supra-glottiques ont la capacité de neutraliser certaines harmoniques et d'en mettre d'autres en évidence par un simple changement de configuration. Lorsque l'on prononce, sur une note constante ou à une hauteur de voix constante, des voyelles aussi différentes que " **oe i u** ", c'est le procédé d'atténuation et de renforcement qui entre en jeu et qui est responsable de l'apparition du timbre propre à chacune des voyelles. Donc le conduit vocal possède une fonction de transfert (filtre) appliquée à une source produit un phonème, on appelle formant les maxima locaux (pics) de cette fonction de transfert notés $\mathbf{F_1}$, $\mathbf{F_2}$, $\mathbf{F_3}$… et correspondent aux zones de renforcement maximal. Néanmoins, du point de vue perceptif, seuls quelques formants jouent un rôle central au niveau de la parole. En théorie on décrit souvent les voyelles grâce à leurs 2 premiers formant $\mathbf{F_1}$, $\mathbf{F_2}$ a travers un triangle acoustique.

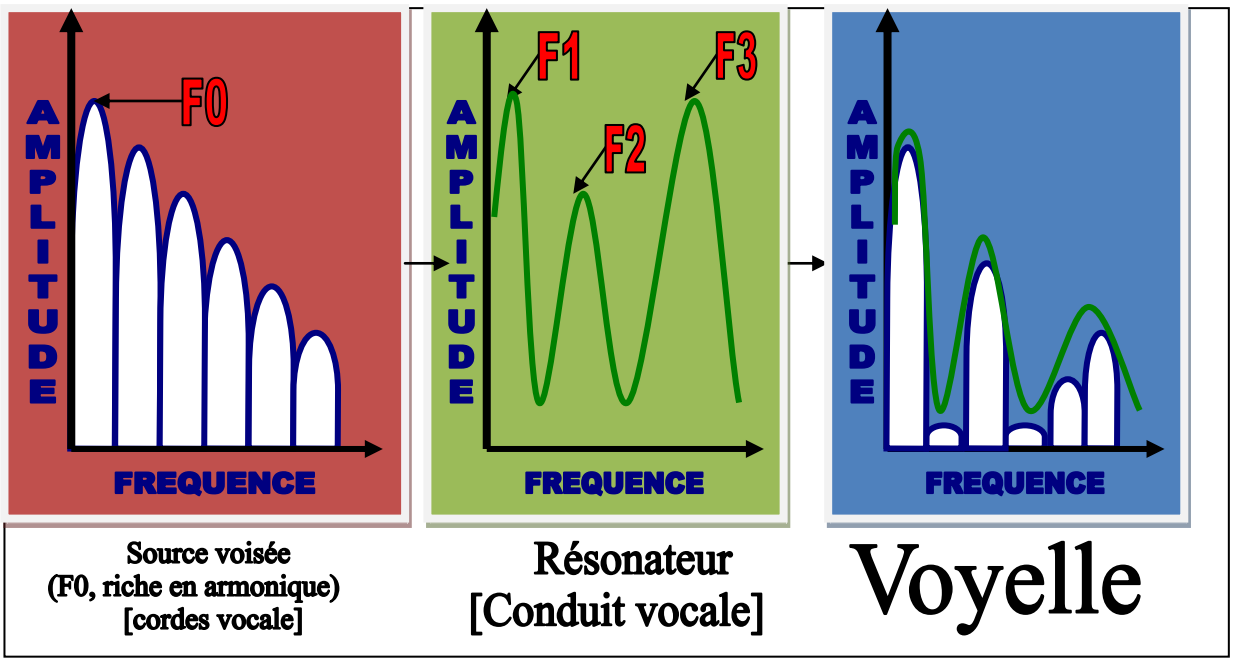


Figure 3.17- Production d'une voyelle

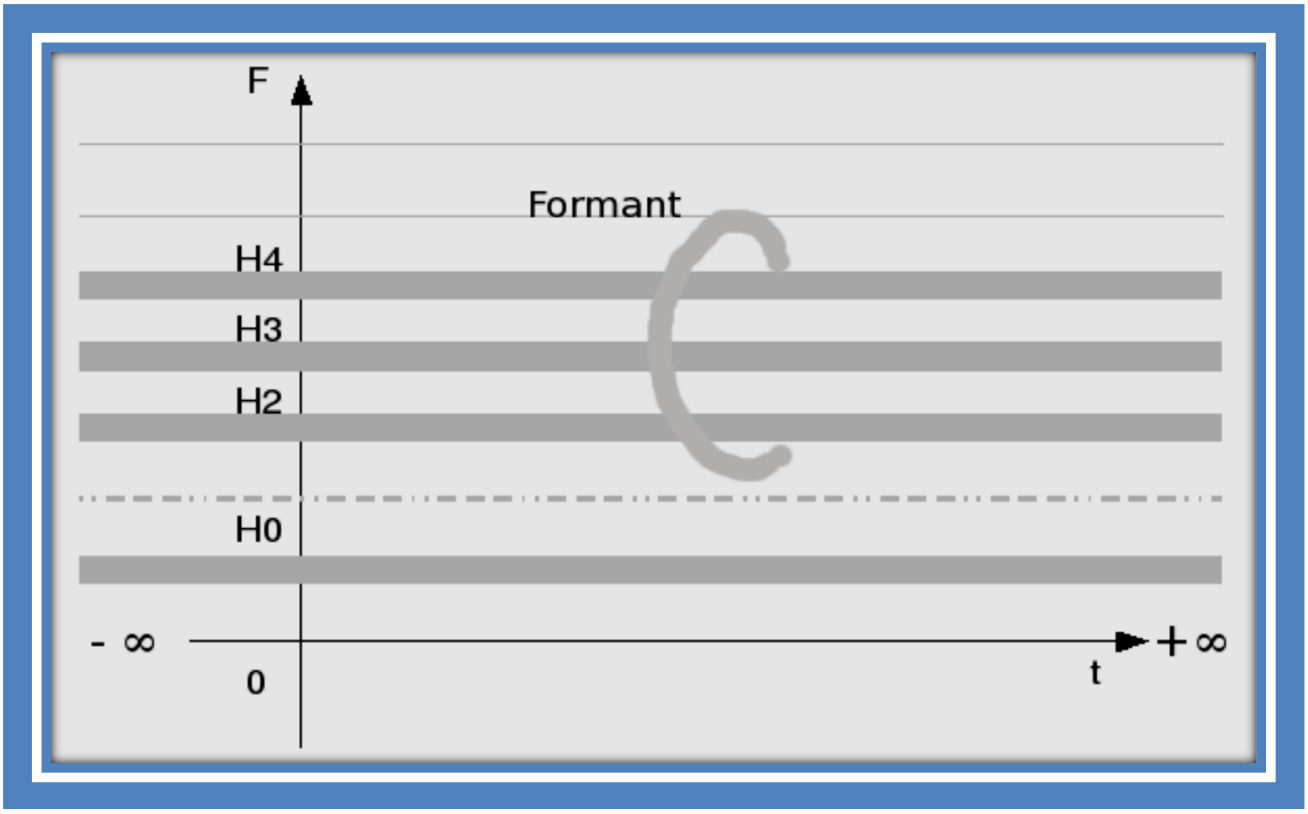


Figure 3.18- exemple d'un son continu de fréquence fondamentale $\mathbf{H_0}$ avec trois harmoniques $\mathbf{H_2}$, $\mathbf{H_3}$, $\mathbf{H_4}$

3-2-5-7- Acoustique des voyelles et des consonnes :

La description acoustique des voyelles prend principalement en compte les valeurs des deux premiers formants. 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. La labialisation (ou arrondissement) des articulations vocaliques agit sur la valeur du troisième formant. Ce lien ne vaut toutefois que pour les voyelles antérieures. La production des (voyelles/consonnes) nasales implique la participation d'un résonateur supplémentaire: les fosses nasales. Cela a pour conséquence d'affaiblir l'intensité des formants 1 et 2 et de générer un formant supplémentaire dont la valeur se situe entre 300Hz et 600Hz.

La caractérisation acoustique des consonnes doit rendre compte de certaines propriétés acoustiques des ondes non périodiques dont elles sont constituées. On peut effectuer une certaine classification à partir des bruits d'explosion ou de friction de même qu'à l'aide des mouvements de transitions des voyelles vers les

consonnes puis des consonnes vers les voyelles. Tout d'abord, on doit retenir que les consonnes sont constituées d'ondes apériodiques complexes, c'est-à-dire de " bruits ".

C'est pourquoi, tant du point de vue acoustique que perceptif, la classification des consonnes reposera sur la nature du bruit. C'est en modifiant le parcours du courant d'air provenant des poumons par un rétrécissement ou une fermeture temporaire du canal buccal suivi d'une ouverture brusque que sont produits les sons consonantiques. Le bruit consonantique sera continu ou discontinu, et couvrira une certaine zone de fréquence de telle sorte qu'on le percevra plus grave ou plus aigu. Un bruit où prédominent les fréquences basses sera perçu comme grave alors qu'un bruit où prédominent les fréquences hautes sera perçu comme aigu. Le bruit d'explosion du " **t** " s'oppose à celui du " **b** " par sa hauteur. De même, le bruit du " **s** " est plus aigu que celui du " **ch** ". Lorsque la consonne est voisée, il y a aussi production d'ondes périodiques. Certaines consonnes (**m**, **n**, **l**, **r** en arabe) possèdent une structure acoustique qui ressemble à celle des voyelles. La faible tension qui accompagne les phases d'occlusion (**m**, **n**) ou de constriction (**l**, **r**) permettent la mise en évidence des ondes périodiques produites, ce qui se traduit par l'apparition de formants. C'est la raison pour laquelle on utilise fréquemment le terme " sonantes " pour désigner ces consonnes.

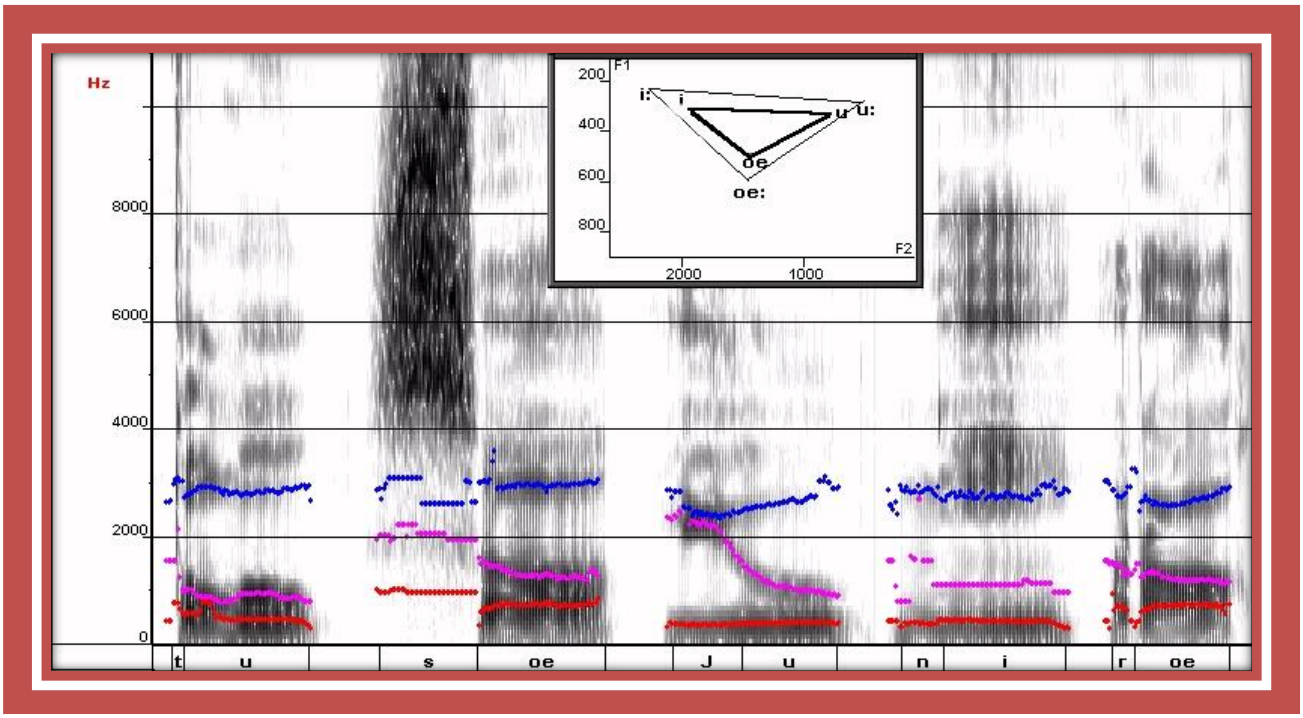


Figure 3.19- Les transitions formantiques des 3 premiers formant $\mathbf{F_1}$, $\mathbf{F_2}$, $\mathbf{F_3}$ des syllabes [**tu**], [**soe**], [**ju**], [**ni**], [**roe**], et le triangle acoustique $\mathbf{F_1}$, $\mathbf{F_2}$ des voyelles

3-2-5-8- Lissage Cepstral :

On utilisant un lissage Cepstral on peut séparer dans le signal de parole $\boldsymbol{s}_{(n)}$ les contributions de l'excitation $\boldsymbol{e}_{(n)}$ et du conduit vocal $\boldsymbol{h}_{(n)}$.

Le but du lissage est d'éliminer les segments issus de l'assemblage qui sont non significatifs.
Le lissage s'effectue en deux parties. Il y a en fait deux lissages :

- Le premier est un pré lissage qui s'effectue à 20 ms pour éliminer les segments trop petits.
- Le second a lieu entre 300 et 700 ms pour la parole et 1500 et 2500 ms pour la musique. Ce deuxième lissage consiste à ne garder que les segments représentatifs de la parole(ou de la musique).

Donc son utilisation est appliquée pour la détermination :
- Les formants
- La fréquence fondamentale

Principe

Le spectre du signal de parole $\boldsymbol{S}_{(w)}$ est obtenu grâce à un filtre $\boldsymbol{H}_{(w)}$ (transformation, supposée linéaire, effectuée par le conduit vocal) appliqué au spectre de l'excitation $\boldsymbol{E}_{(w)}$.
Pour séparer les deux informations présentes dans le signal de parole s(n), il est nécessaire d'effectuer une déconvolution (homomorphisme) du signal.

3-3- Phonétique perceptive:

3-3-1- L'appareil auditif :

L'appareil auditive remplit un grand nombre de fonctions diverses :

-information sur l'environnement, les objets qui nous entourent.
-alerte, détection, information sur la proximité et la direction des sources sonores.
La manière dont nous percevons ces sources, elles sont perçues différemment suivant le contexte de l'auditeur, comme le montrent les deux exemples suivants :

1-un coup de klaxon automobile sera généralement ignoré ou perçu comme gênant si l'on est chez soi.

2-dans une conversation que nous n'écoutions pas jusqu'alors, la simple prononciation de notre nom peut focaliser soudain notre attention.

-la reconnaissance des sources
-l'appréhension d'espace clos
-la notion de confort acoustique est essentielle dans les lieux voués à l'audition
-la communication enfin est un des rôles essentiels de l'audition, la communication sonore passe par :

1-l'intelligibilité proprement dite de la parole, mais aussi,
2-le timbre qui donne des indications sur le locuteur (âge, sexe, état de fatigue,)
3-l'intonation, qui contribue au sens (interrogation, exclamation.)Mais aussi qui exprime l'humeur ou les sentiments.

Nous allons présenter l'anatomie des organes de l'oreille.

3-3-2- L'appareil auditif humain :

L'appareil auditif humain est subdivisé en trois parties distinctes**, [2]** :l'oreille externe, moyenne et interne :

L'oreille externe :

Celle-ci se compose du pavillon et du conduit auditif externe :
-le pavillon réalise un filtrage sélectif suivant la direction d'incidence du son
-le conduit auditif externe c'est un cylindre fermé à une extrémité par le tympan

L'ensemble de l'oreille externe à pour effet une augmentation de l'intensité sonore au niveau du tympan

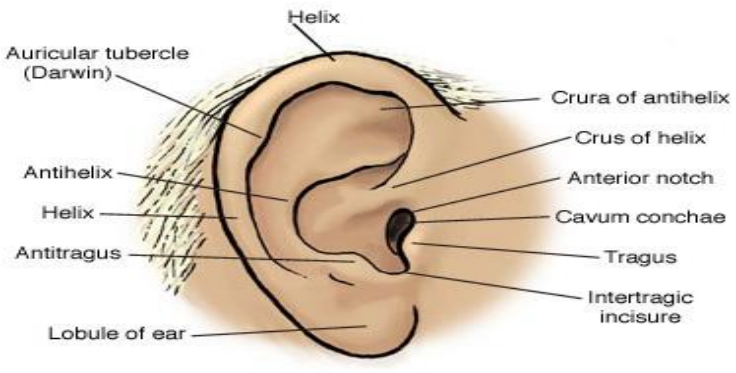


Figure 3-20 **:** oreille externe

L'oreille moyenne :

L'oreille moyenne est une cavité d'air dans un os, le **rocher,** qui renferme un système articulé de trois osselets : le **marteau**, l'**enclume**, et l'**étrier**.

L'oreille moyenne joue un rôle d'amplification et d'adaptation d'impédance. Elle joue également un rôle essentiel de protection de deux manières :

-la limitation mécanique naturelle des mouvements des osselets.
-une limitation par contraction d'un muscle lié à l'étrier.

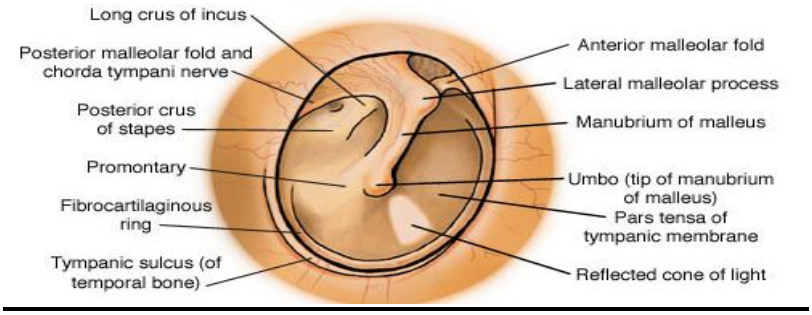


***Figure 3-21*:** oreille moyenne

L'oreille interne :

C'est dans l'oreille interne que l'énergie mécanique est transformée en énergie bioélectrique, câd en potentiels d'action nerveux. L'oreille interne se compose :
-de l'appareil vestibulaire, comprenant les trois canaux semi-circulaire
-de la cochlée, qui à globalement la forme d'un canal en colimaçon attacher à deux membranes basilaires et la membrane tectorielle.
La membrane basilaire sépare le canal de la cochlée en deux rampes remplies de liquide et formé par des cellules ciliées : Les cellules ciliées internes et externes.

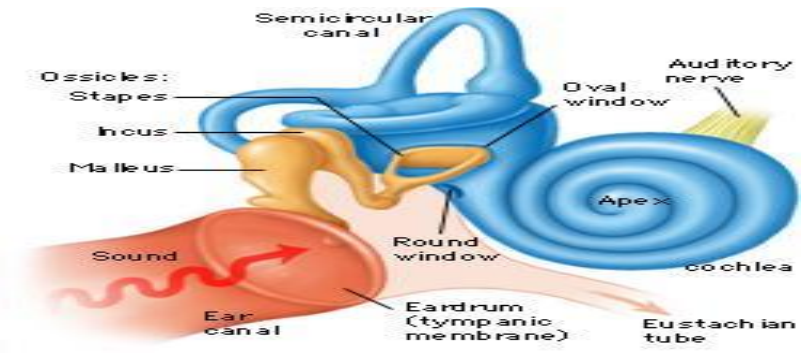


Figure 3-22 : oreille interne

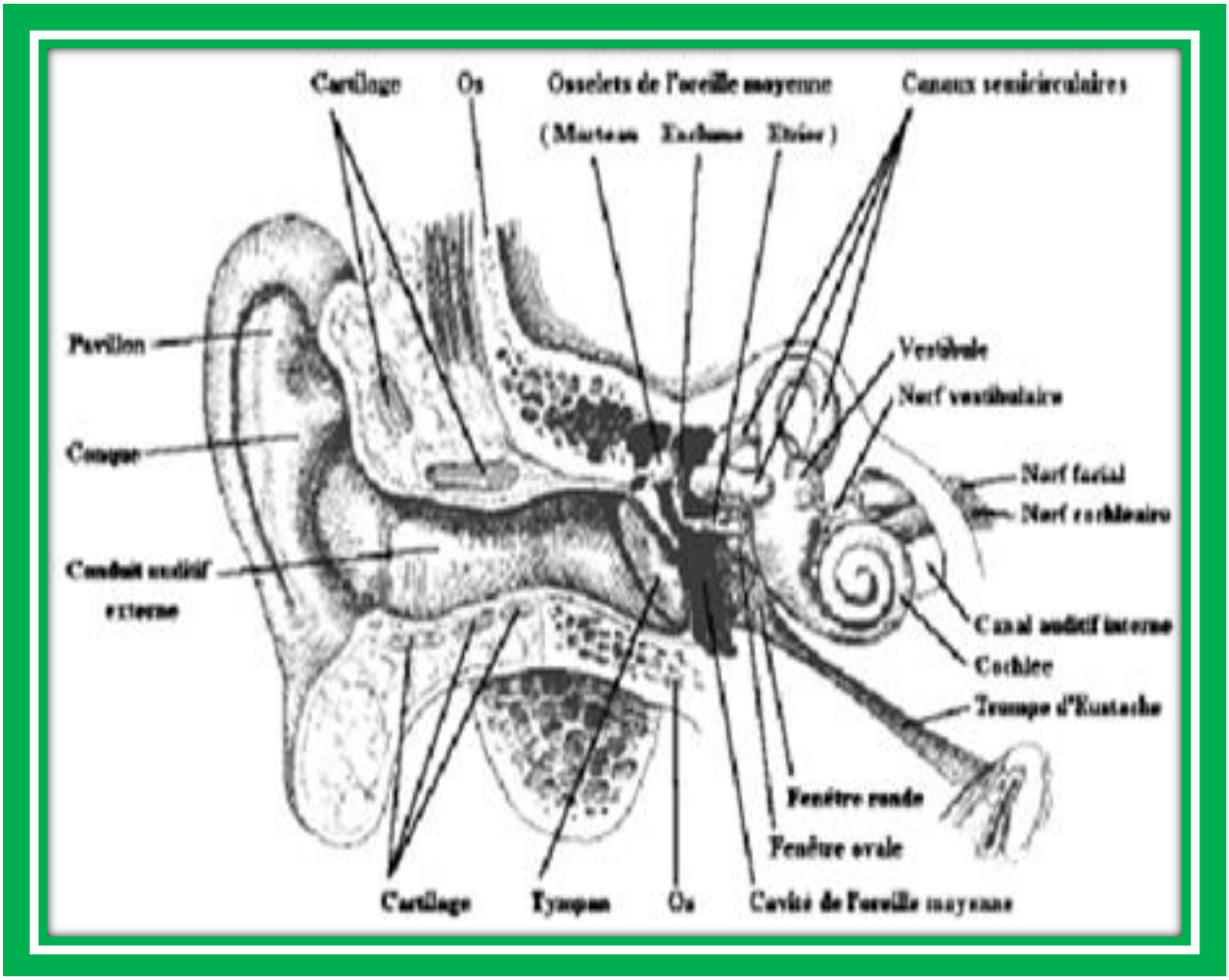


Figure 3.23 : Coupe de l'appareil auditif humain

3-3-3- Les deux organes sensoriels de l'oreille interne:

L'oreille interne regroupe 2 organes sensoriels distincts : le vestibule, organe de l'équilibration et la cochlée, organe de l'audition. D'une même origine embryologique (la vésicule otique) ces deux organes partagent aussi d'autres propriétés morphologiques et physiologiques comme le liquide endolymphatique, les cellules ciliées et leurs propriétés de transduction. Ci-dessous, le schéma principal représente, par transparence, le labyrinthe membraneux contenant l'endolymphe; en haut à gauche : le labyrinthe osseux.

3-3-4-Transfert des pressions acoustiques (ondes sonores) du milieu aérien aux fluides et aux structures de l'oreille interne (cochlée) :

Les vibrations mobilisent le tympan et la chaîne des osselets. L'étrier, plaqué sur la fenêtre ovale transfère la vibration au compartiment périlymphatique de la rampe vestibulaire et aux structures de l'oreille interne. En fonction de sa fréquence, la vibration a un effet maximal (résonance) en un point différent de la membrane basilaire : c'est la tonotopie passive.

Figure 3.24- Cas d'un son aigu

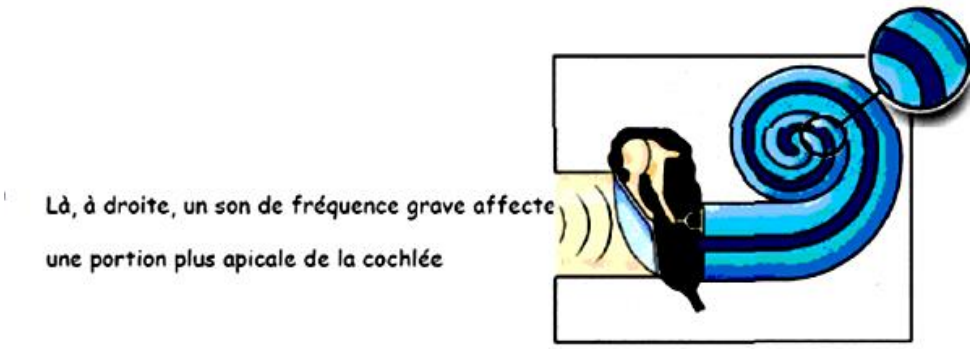


Figure 3.25- Cas d'un son grave

3-3-5- Courbes psycho-acoustiques:

Plusieurs échelles essaient de rendre compte de la réalité perceptive de l'oreille. Elles peuvent toutes être rapprochées des échelles de la membrane basilaire et du rang des cellules ciliées. Ces échelles ne présentent pas toutes la même morphologie. **[2]** En effet, celles qui essaient de restituer le plus correctement possible les échelles de la perception humaine sont non linéaires, telles que les

échelles Mel ou Bark. Les échelles qui peuvent être qualifiées de plus mathématiques sont en revanche linéaires, telle que l'échelle des fréquences.

Ces différentes échelles essaient de rendre compte du mode de perception de l'homme en permettant de distinguer les plages de plus ou moins grande importance.

Ainsi les basses fréquences sont-elles perçues de manière plus fine par l'homme que les hautes fréquences. Cette différence dans la finesse de perception permet de comprendre plus facilement certaines courbes, en particulier les courbes situant l'utilisation du spectre sonore par l'homme.

L'homme est en effet très limité dans ses capacités de perception auditive vis-à-vis d'autres membres du règne animal. Il lui est ainsi impossible de distinguer des sons de plus de 20 kilohertz, les ultrasons, alors que certains animaux qui lui sont familiers peuvent percevoir des sons allant jusqu'à 50 kilohertz. De même lui est-il impossible de distinguer des sons d'une fréquence inférieure à 20-25 hertz, les infrasons.

À l'intérieur de cet espace fréquentiel existe un sous-espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie en dessous de laquelle l'homme ne percevra pas un son d'une fréquence appartenant pourtant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition et il est variable en fonction de la fréquence.

Inversement, il existe une limite d'énergie maximale. Cette limite ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être irrémédiablement endommagées.

Cette limite s'appelle le seuil de douleur et elle aussi est variable en fonction de la fréquence. Il est intéressant de noter qu'il existe dans l'oreille deux muscles qui permettent à l'homme de débrayer le transfert des vibrations du tympan à la cochlée pour limiter les dégradations qui peuvent survenir dans le cas où un bruit dépassant le seuil de douleur est perçu.

L'espace de fréquences et d'énergies ainsi défini (figure ci-dessous) constitue la zone d'audition à l'intérieur de laquelle l'homme peut recevoir des informations de son environnement.

C'est bien sûr à l'intérieur de cet espace que se trouve le champ de la musique qui circonscrit lui-même le champ de la parole.

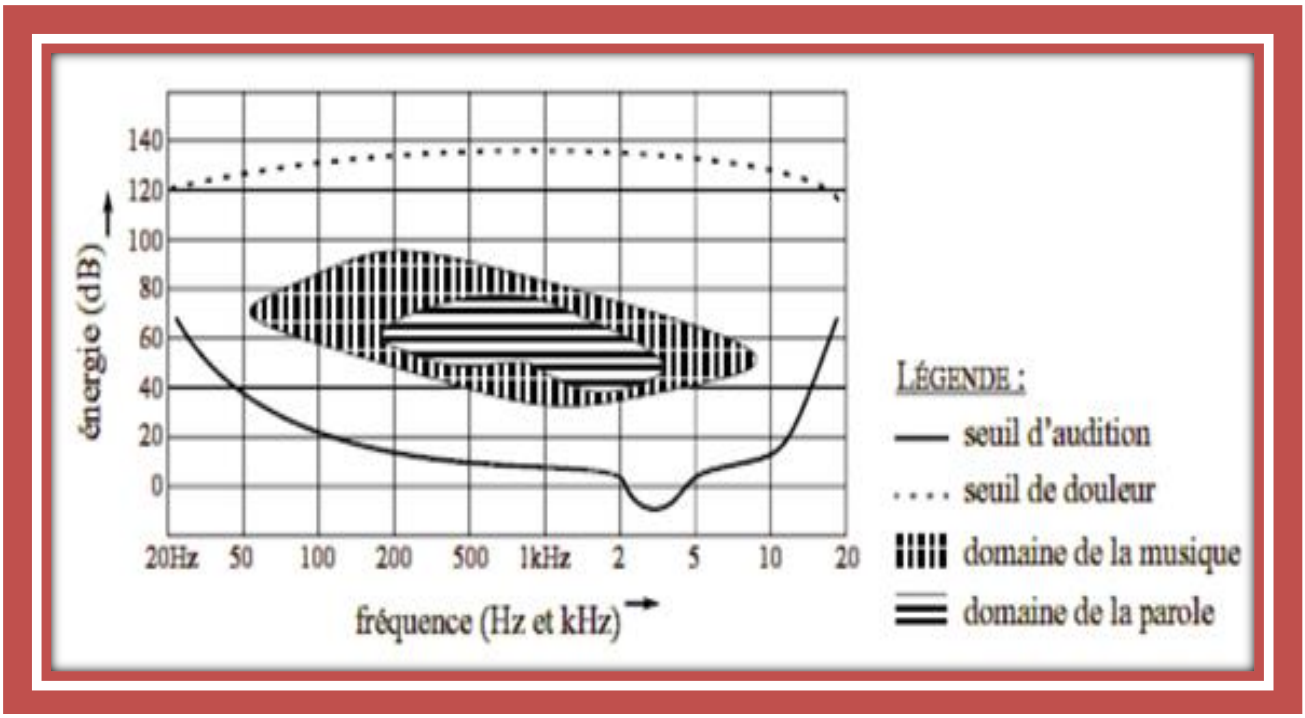


Figure 3.24- L'aire d'audition

Il est important de parler de l'oreille pour comprendre comment nous percevons le son. D'abord l'oreille n'a pas une sensibilité linéaire, mais une sensibilité logarithmique (par rapport à la pression acoustique). Ensuite sa sensibilité est différente selon la fréquence et le niveau sonore général.

3-3-6- L'échelle Mel :

C'est une modélisation de l'oreille humaine. A noter que le cerveau effectue en quelque sorte une reconnaissance vocale complexe avec filtrage des sons. Prenons l'exemple suivant où vous êtes en compagnie de nombreuses personnes, l'ensemble de ses personnes parle en même temps et vous discutez avec votre voisin.

Malgré le bruit, vous arrivez à discerner clairement ce que vous dit votre voisin, vous ignorez de façon naturelle le bruit de fond et vous amplifier le son qui vous paraît le plus important.

Le cerveau ne se contente non pas seulement de filtrer les sons et de les amplifier mais aussi de prédire. Prenons l'exemple suivant où une personne discute avec vous avec un volume sonore très bas, vous n'avez pas entendue une certaine partie de la phrase mais vous arrivez à la reconstituer et à la comprendre. A partir

de l'étude du cerveau nous pouvons nous faire une idée de la complexité de la reconnaissance de la parole et nous pouvons nous rapprocher d'un modèle de plus en plus puissant et parfait.

On considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence. L'échelle Mel modélise assez fidèlement la perception de l'oreille :

Linéairement jusqu'à 1000 Hz, puis logarithmiquement au dessus.

La formule donnant la fréquence *m* en Mels, *m* à partir de celle de *f* en Hz, est :

$$m = \frac{1000 \cdot \ln\left(1 + \frac{f}{700}\right)}{\ln\left(1 + \frac{1000}{700}\right)} \approx 1127 \cdot \ln\left(1 + \frac{f}{700}\right) \approx 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad \textbf{3-2}$$

L'échelle Mel permet donc de modéliser une perception de l'oreille linéairement.

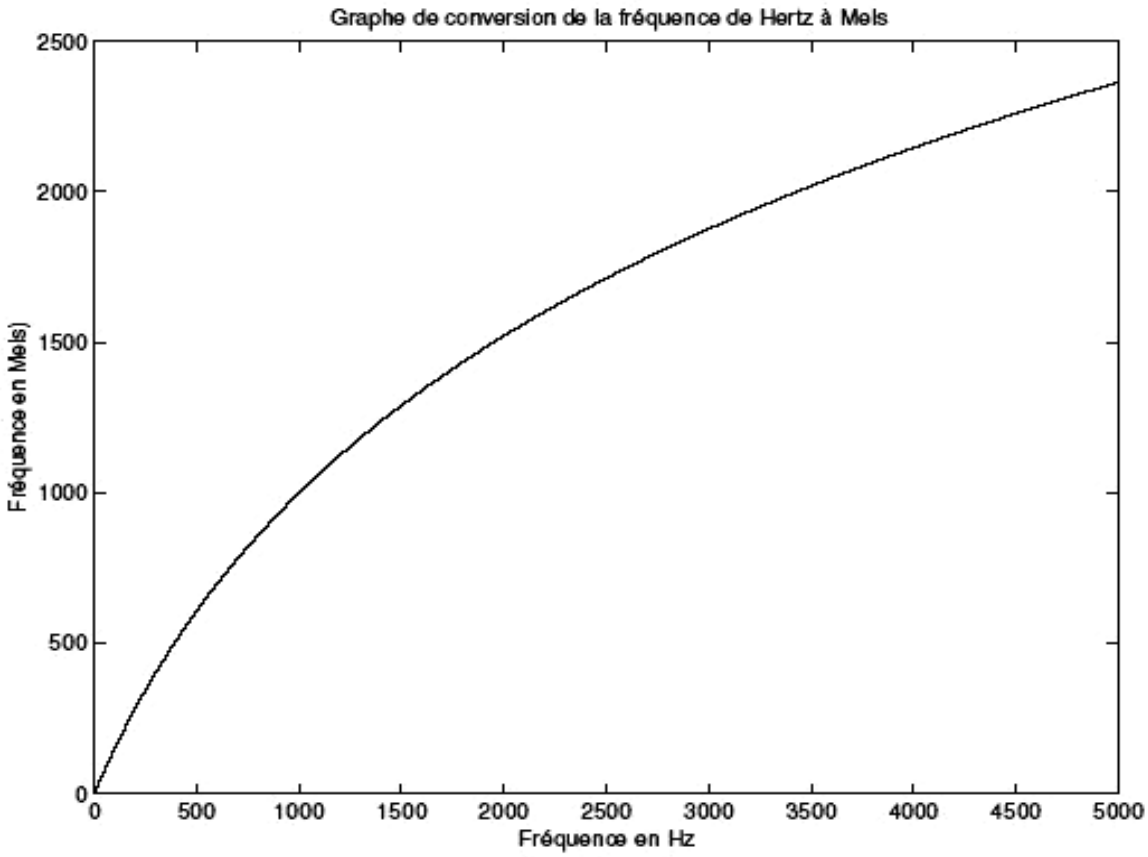


Figure 3.27- graphe de conversion

On remarque qu'avant 1000 Hz, la courbe est à peu près droite, ce qui traduit bien l'équivalence entre Hz et Mels à ces fréquences.

3-4- Relations articulo-acoustico-perceptives :

Le signal de parole est le résultat de la propagation d'une onde acoustique dans un tuyau de forme variable **[6], [9]**. Ce tuyau est mis en forme par un ensemble de muscles agissant sur des structures rigides, semi-rigides ou des hydrostats comme la langue ou les lèvres. De même, les sources d'excitation sont produites par des compressions exercées sur les poumons, les affecteurs 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. Comme nous l'avons mentionné, ce qui fait toute la difficulté du traitement automatique de la parole est la phénoménale variabilité des signaux que l'on peut enregistrer dans différents points dans cette chaîne de communication (signaux articulatoires, acoustiques, activité neuronale).

Cette variabilité puise ses origines dans différentes sources indépendantes : les caractéristiques anatomiques, l'influence du contexte phonétique, les stratégies articulatoires propres à chaque locuteur, les états mentaux, les origines socioculturelles et géolinguistiques, etc.

Chaque auditeur sait, d'une part, s'affranchir de ces sources de variabilité afin d'associer au signal un contenu conventionnel (au sens du code linguistique). D'autre part, chaque auditeur sait identifier ces sources de variabilité afin d'acquérir l'habilité motrice nécessaire à une communication riche et robuste. Au contraire, les machines ne disposent pas aujourd'hui de moyens efficaces pour gérer cette variabilité.

La principale raison provient d'un nombre de degré de liberté en excès dans le passage des commandes musculaires du locuteur au percept de l'auditeur. Nous essayons d'illustrer ce passage sous la forme d'une pyramide. Cette non-unicité se manifeste aussi entre niveaux intermédiaires :

Le passage d'un niveau de caractérisation à un autre est une relation de type "plusieurs-à-un".

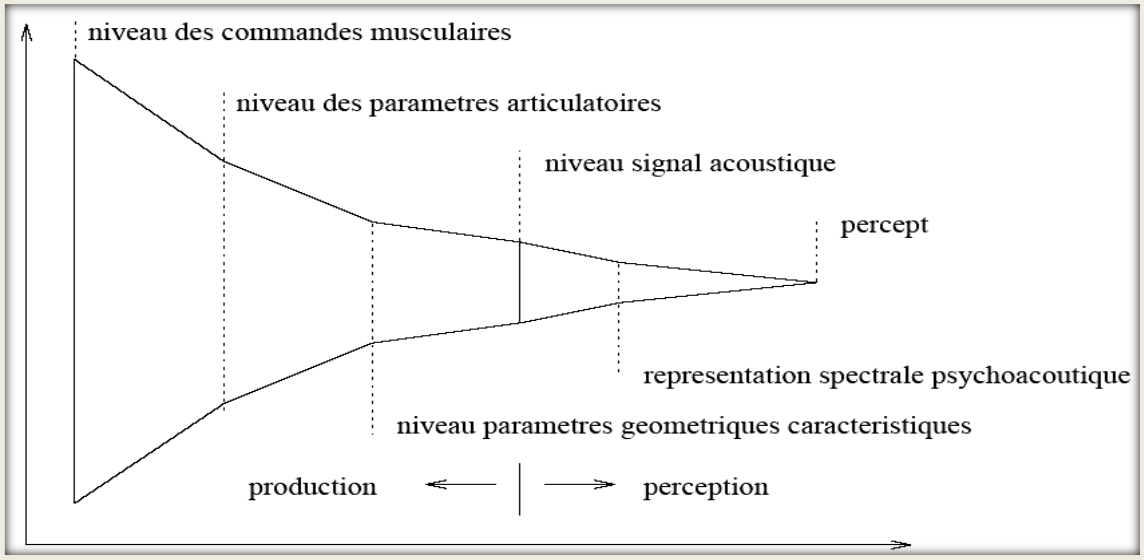


Figure. 3.28- Niveaux de représentation d'un son donné de sa production à sa perception

3-5- La théorie quantique des traits distinctifs (Stevens)

3-5-1- Les traits distinctifs :

L'analyse phonologique réduit les multiples différences acoustiques entre les mots d'une langue à quelques oppositions élémentaires, ou traits distinctifs. Les traits sont des marques discriminatives sans contenu sémantique propre **[6], [9]**. Ils ne sont pas définis en termes d'oppositions sémantiques mais bien en termes d'oppositions phonétiques **[43].** La mise en évidence des traits distinctifs comporte deux étapes. La première étape consiste à découper les mots, ou plus précisément les mots et les portions de mots qui restent porteurs de signification (les morphèmes), en segments linguistiques élémentaires, ou phonèmes. La seconde étape consiste à rechercher des oppositions phonétiques élémentaires entre les phonèmes pour spécifier les traits distinctifs.

Le point de départ de la procédure de détermination des phonèmes repose sur le jugement perceptif des locuteurs de la langue considérée. Les morphèmes sont subdivisés en des segments qui peuvent être perçus de manière distincte.

Les traits distinctifs sont des classes d'oppositions phonétiques minimales entre phonèmes. A chaque trait correspond une opposition phonétique élémentaire qui ne

peut pas être scindée en d'autres oppositions, plus fines et dotées d'un pouvoir distinctif indépendant au sein de la langue considérée.

Les traits sont des propriétés relationnelles: à chaque trait correspond un rapport constant qui, selon **[38]** peut être mis en évidence avec des mesures acoustiques ou articulatoires.

Pour mettre les traits en évidence, on recherche des oppositions phonétiques minimales entre paires de mots. En français, '**bon - son**' n'est pas une paire minimale car il y a 3 différences articulatoires entre les phones en position initiale et chaque différence peut jouer un rôle distinctif indépendant: le mode d'articulation, le lieu d'articulation et le voisement.

Le trait est défini par une propriété phonétique invariante dans l'ensemble des paires minimales correspondantes. En termes acoustiques ces invariants sont:

Pour le mode d'articulation, la durée du segment de bruit, plus long pour les fricatives que pour les occlusives; pour la distinction de lieu labiodentale, la répartition de l'énergie, sur des fréquences plus grave pour les labiales; pour le voisement, le point de départ des vibrations périodiques, anticipé pour les voisées

3-5-2- La théorie quantique :

Cette théorie n'est pas un modèle de perception de parole en tant que tel, et n'est ni un modèle actif ni passif, car aucune référence explicite du rôle actif ni passif de l'auditeur dans le traitement de la perception de la parole n'est donné.

En se basant sur les propositions de la théorie quantique acoustique de la parole (**Fant, 1960**), **Stevens** explique que, pour un paramètre articulatoire précis, des changements articulatoires mineurs produits dans certaines régions conduiront à des changements acoustiques importantes et l'inverse et tout aussi vrai. **[41], [42]** pour cette théorie, le rôle attribué à l'auditeur est tout simplement d'arriver à dériver les traits constituant le signal acoustique présentent dans des régions précises dans le conduit vocal et les mouvements de différents paramètres acoustiques constituant ces traits. L'hypothèse de base est que le langage est adapté à des systèmes biologiques humains spécialisés pour la communication orale. Il s'agit du système articulatoire (pour la production des sons) et du système perceptif (perception humaine des sons). Il est donc supposé que la sélection des sons dans une langue est gouvernée par des contraintes physiques et physiologiques. Ces sons sont sélectionnés par des contraintes et forment ce qui est appelé l'espace des sons possibles qui est parfaitement délimitable. La notion

quantique du langage traduit la notion de discrétisation ou de catégorisation des sons dans cet espace.

La forme sigmoïdale de la courbe reflète la relation non-linéarité entre les mouvements articulatoires et leurs corrélats acoustiques (figure 3.27).D 'où l'existence de régions de l'espace articulatoire pour lesquelles:

- Le résultat acoustique est stable (régions I et III sur la figure),
- Une petite variation du paramètre articulatoire implique une variation abrupte du paramètre acoustique (région II).

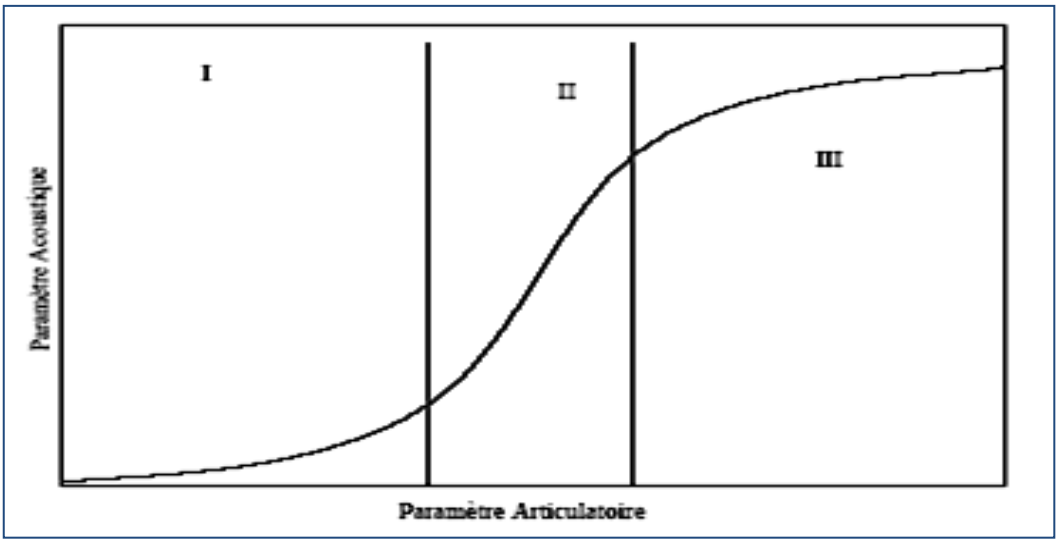


Figure 3.29- courbe représentant schématiquement les principes de base de la Théorie Quantique (d'après **Stevens** (1972)).

Les différences entre les trois zones s »expliquent par le fait que dans la région I, les changements de paramètres articulatoires produiront des changements acoustiques négligeables. Dans la région II, des changements importants des paramètres articulatoires produiront des modifications importantes.

Les régions I et III sont des régions de plateaux acoustiquement stables. La région II en revanche est une région de transition abrupte du paramètre acoustique traduisant une zone d'instabilité avec les régions I et III. Ce modèle apporte quelques explications quant à la relation entre les mécanismes articulatoire et acoustique.

3-5-3- Théorie Quantique appliquée aux voyelles :

Les relations entre les fréquences des formants vocaliques et les caractéristiques résonnantes des cavités sont relativement complexes. La fréquence de chaque formant dépend de la forme globale du conduit vocal supra- laryngé **[39]** et il n'y a, en outre, pas de relations biunivoques entre les configurations articulatoires et les formants puisque différentes formes de conduit vocal peuvent éventuellement produire le même pattern formantique **[34]**.

Pour étudier la production des voyelles, on considère le conduit vocal comme étant un tube acoustique de section non uniforme, fermé d'un côté (côté de la glotte) et ouvert de l'autre (côté des lèvres). Tout modèle articulatoire tient compte de la variation de la constriction formée au niveau de la langue **[6], [9]**. Ceci se traduit par la présence d'au moins deux cavités acoustiques dans le conduit vocal. Les fréquences des formants sont interprétées comme étant les résonances des deux cavités.

Les paramètres de contrôle sont la position de la constriction et l'aire aux lèvres. Ces deux paramètres sont des bons candidats car, en variant ces deux paramètres, on peut simuler l'ensemble des voyelles orales avec un modèle élémentaire. **[36], [37]** et **[6]** a montré que les configurations articulatoires requises pour la production des voyelles cardinales [**a,i,u**] , sont optimales dans le positionnement des articulateurs n'entraînent que de faibles changements dans les propriétés du signal acoustique. Pour la production de la voyelle [**i**], le modèle du conduit vocal utilisé peut être rapproché par un résonateur d'un tube montré à la Figure 3.26, le tube est divisé en deux parts par une section étroite. Toute la longueur $l_1 + l_2 + lc$ = 16 cm, longueur de constriction LC = 2cm, et sections A1 et $A2$ sont 3cm^2.

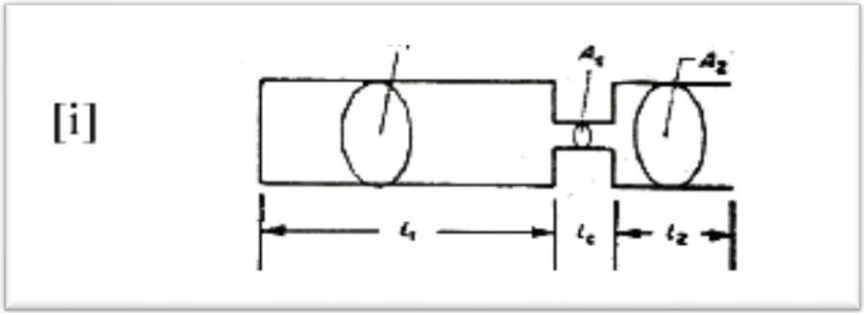


Figure 3.30 **:** modèle du conduit vocal

On varie la position de la constriction, donc la longueur du premier tube et du deuxième changent tout en conservant la longueur totale constante. L'axe des abscisses représente la position de la constriction et l'axe des ordonnées représente les fréquences de résonance (les formants).

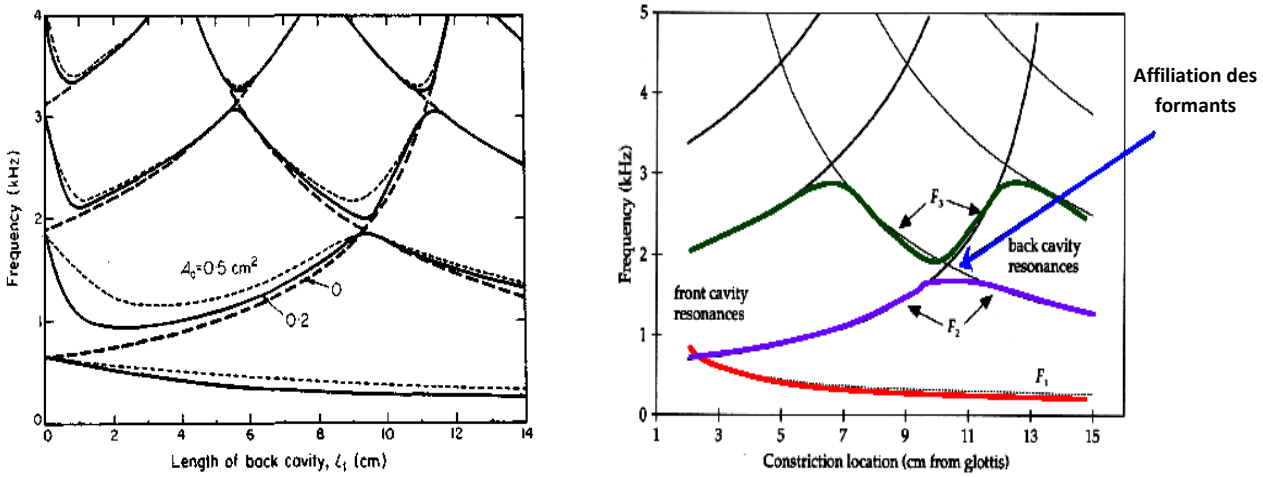


Figure 3.31 : constriction au niveau du conduit vocal

Nous observons sur cette représentation (figure 3.29) que les fréquences de résonance de l'ensemble du conduit (notés **F_1**, **F_2**, **F_3**, etc.) balayent chacune un domaine de fréquence. Il en résulte ainsi des trajectoires qui se rencontrent. Ces points de convergence des trajectoires formantiques montrent les lieux de changement d'affiliation des formants appelés aussi points focaux. Un tel point est marqué sur le graphique de gauche.

Stevens a remarquée que les fréquences normales du système réalisent des valeurs maximum ou minimum, et ces fréquences sont relativement peu sensibles aux petits changements dans L1.

Ces régions sont à proximité de *L1* = 5.5, 9.3, et 11.2 centimètres dans la figure. **Stevens** a démontré qu'il y a des régions de stabilité acoustique (régions quantiques) ou deux formants convergents (voir la figure 3.31).

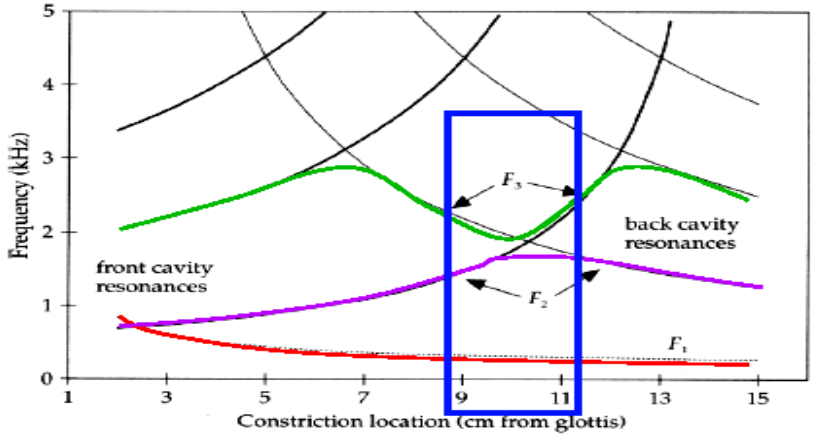


Figure 3.32 : localisation de la constriction

Pour la production de la voyelle [**a**], le modèle du conduit vocal utilisé peut être rapproché par un résonateur de deux tubes montrés à la Figure .3.31. La partie postérieure de la région voca1 est resserrée, et la cavité buccale est relativement grande. On fait varier les longueurs des 2 tubes en sachant que la somme des longueurs des deux tubes, L1 et L2, est toujours égale à 16 cm.

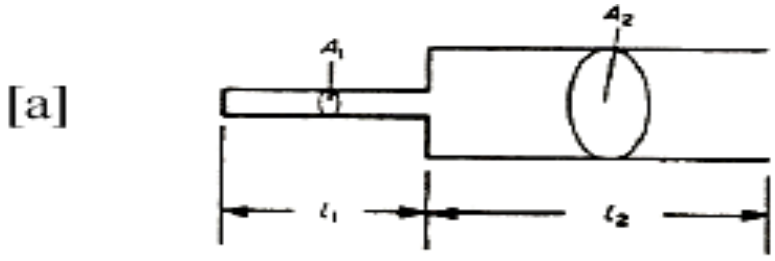


Figure 3.33 : modèle de conduit vocal avec variation des longueurs

La représentation graphique (Figure 3.34) montre la variation des fréquences en fonction du d1.

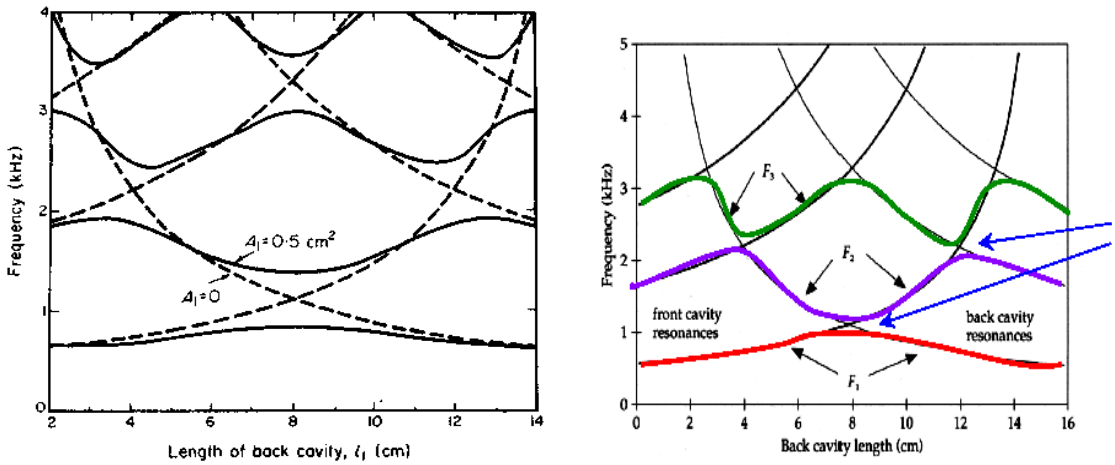


Figure 3.34 : affiliation des formants focaux

Les relations entre les fréquences de formant et la dimension L1 montre encore des régions où les fréquences réalisent des valeurs maximum ou minimum. Quand L1 est dans une de ces régions, comme à proximité de L1 = 8 centimètres dans la figure, les fréquences de formant sont relativement peu sensibles aux changements de L1, tandis qu'entre ces régions (par exemple, près de L1 = 6cm), de petites perturbations dans L1 peut provoquer les changements substantiels des fréquences des formants.

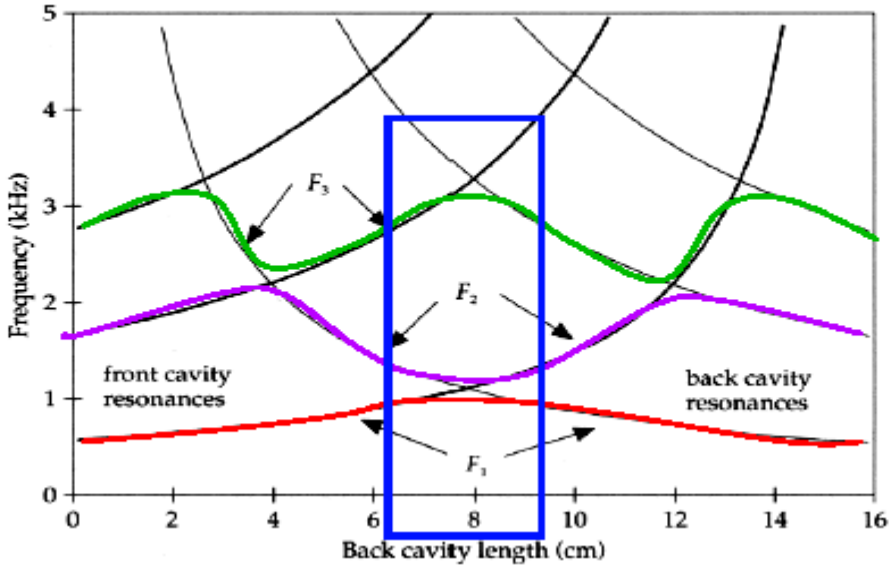


Figure 3.35: la longueur de la cavité buccal

Et pour la production de la voyelle [**u**], le modèle du conduit vocal utilisé peut être rapproché par un résonateur de quatre tubes montré à la Figure 3.36.

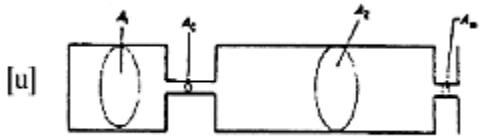


Figure 3.36 : résonateur représentant un modèle du conduit vocal

Cette configuration est semblable à celle dans la figure 3.27, sauf qu'il y a le rétrécissement du tube à l'extrémité ouverte. La variation des fréquences normales de cette configuration par rapport à la position de constriction est donnée dans la figure suivante (figure 3.37).

Cette figure montre un large minimum de $\mathbf{F_2}$ dans une gamme des longueurs de la cavité arrière. F_2 est à moins de 100 hertz de sa valeur minimum pour L1 entre 2 et 7.5cm. Dans cette marge de *L1*, l'espacement entre $\mathbf{F_1}$ et $\mathbf{F_2}$ est de 400-500 hertz, alors que $\mathbf{F_1}$ ne réalise pas une valeur maximum, il change par seulement environ 80 hertz. La position exacte de la constriction pour laquelle un minimum de $\mathbf{F_2}$ est atteint dépend de la taille de l'ouverture à l'extrémité de rayonnement du tube et sur la longueur et la taille de la constriction. Il y a une gamme des positions de constriction qui rapporteront une valeur de $\mathbf{F_2}$ qui est bas, relativement stable, et près de $\mathbf{F_1}$.

Ces positions sont réalisées en manœuvrant le corps de langue dans une position soutenue pour former une constriction dans la région du palais ou du pharynx supérieur.

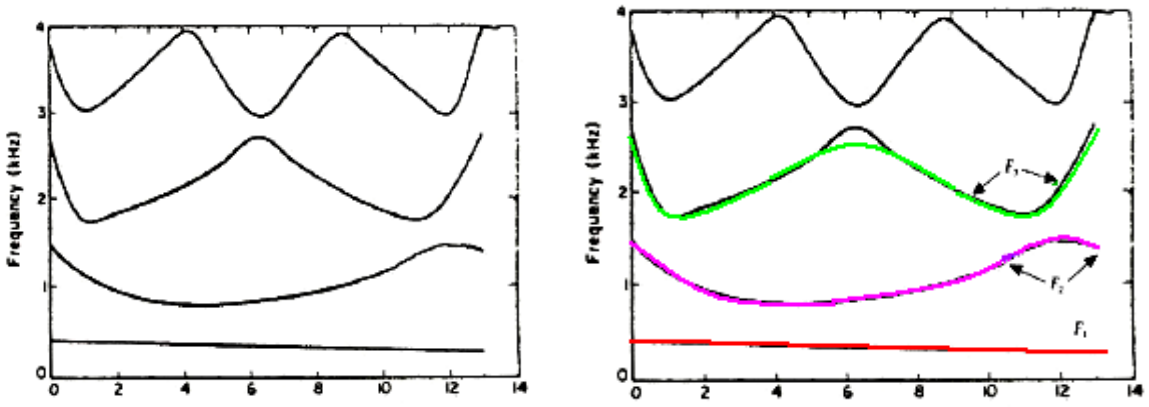


Figure 3.37 : position de la langue sur le palais ou le pharynx supérieur

Nous pouvons conclure, que ces trois voyelles sont des voyelles quantiques. Car il y a des régions de stabilité acoustique (régions quantiques) où deux formants convergent.

• [**i**] est produit à la convergence du $\mathbf{F_2}$ et du $\mathbf{F_3}$, crée par une constriction dans la région palatale.

• [**a**] est produit à la convergence de $\mathbf{F_1}$ et de $\mathbf{F_2}$, où la cavité avant et la cavités arrières sont des longueurs approximativement égales.

• [**u**] est produit à un large minimum $\mathbf{F_2}$, où $\mathbf{F_2}$ est près d'un $\mathbf{F_1}$ stable.

Le $\mathbf{F_1}$ a un intérêt particulier pour la production du voisement parce que la présence de ce formant dépend de l'état de la glotte. La fréquence du $\mathbf{F_1}$ dépend essentiellement du degré de constriction du conduit vocal **[39], [35]**. La fréquence du $\mathbf{F_1}$ augmente lorsque la constriction buccale diminue, ce qui est le cas pour la voyelle [**a**], dite ouverte, par rapport aux voyelles [**i**, **y**, **u**] dites fermées. Pour les occlusives voisées, la fréquence du $\mathbf{F_1}$ durant l'occlusion atteint la limite inférieure de 200 Hz, contre 750 Hz environ pour [**a**] et 250-300 Hz pour les voyelles fermées.

3-5-4- Approche globale :

Stevens introduit la notion de stabilité dont la fonction serait accrue dans le processus auditif. Un phonème est d'autant plus fréquent dans les langues du

monde qu'il est stable. Ce phonème sera qualifié de quantique. Stevens considère qu'un phonème acoustiquement stable possède une forme perceptive plus riche de par sa stabilité auditive et est, par conséquent, plus aisément discriminable.

3-5-5- Analyse critique :

Les arguments les plus convaincants fournis par la théorie Quantique portent sur l'articulation des structures sonores. A Ce propos, **Lindblom & Engstrand** (1989) énoncent: «the essence of an intuition that underlies the Q T is that continuous Articulatory motion produces clear discrete acoustic segments». Cependant, les arguments avancés par Stevens concernant le lieu d'articulation sont plus controversés (**Abry, Boë et Schwartz 1989**).

D'autre part, il apparaît que la règle selon laquelle les voyelles quantiques sont les plus rencontrées dans les langues du monde n'est pas vérifiée.

Vallée (1994) remarque que la voyelle [**y**] devrait être, au même titre que la voyelle [**i**], très fréquente dans les systèmes, alors qu'en réalité, [**y**] est rencontrée approximativement dix fois moins que [**i**].

3-5-6- Rapprochement F'_2 : [12] et [17]

La plupart des études en traitement de la parole et en phonétique ont montré que presque toutes les voyelles peuvent être simulées perceptiblement par des modèles à deux-formants, dans lesquels, l'un ou l'autre des deux formants constitue une moyenne pondérée des fréquences de deux ou plusieurs formants. Dans de telles études, des auditeurs ont été invités à apparier des configurations de deux-formants aux voyelles synthétiques standards formées de quatre formants ayant des fréquences appariant ceux des voyelles naturelles.

Le premier formant de ce type de configuration à deux formants était fixé à la fréquence du $\mathbf{F_1}$ standard, et les sujets ont ajusté le deuxième formant $\mathbf{F'_2}$ jusqu'à ce qu'ils aient entendu une voyelle qui est perçue comme équivalente à la voyelle originale composée de quatre formants **[14]**. Dans le cas des voyelles postérieures, le $\mathbf{F'_2}$ a été ajusté sur une intermédiaire de fréquence entre le $\mathbf{F_2}$ et le $\mathbf{F_3}$ de la voyelle originale. Dans le cas des voyelles antérieures, le $\mathbf{F'_2}$ a été ajusté sur une fréquence près du $\mathbf{F_2}$. Pour expliquer ces observations.

Delattre suggère que quand les formants sont étroitement proches en fréquence comme le cas de $\mathbf{F_1}$ et $\mathbf{F_2}$ pour les voyelles antérieures et le $\mathbf{F_2}$ et $\mathbf{F_3}$ pour les voyelles postérieures, elles sont intégrées d'une façon perceptive telle que le

formant pertinent $\mathbf{F'_2}$ est équivalent à une moyenne des formants standards. **Chistovich**, a prouvé que quand deux crêtes spectrales ou plus se produisent dans une même bande critique dans l'échelle de Bark, la qualité perçue de la voyelle est équivalente à une configuration avec une crête spectrale unique située au centre de gravité des fréquences des formants. La fréquence perceptive du formant $\mathbf{F'_2}$ est une moyenne pondérée en fréquence et en amplitude des crêtes spectrales dans une marge de 3-3.5 Barks.

Dans cette marge, la fréquence perceptive du formant $\mathbf{F'_2}$ est décalée vers la fréquence de la crête la plus élevée en amplitude. Quand la distance de fréquence entre les crêtes spectrales excèdent 3.5 Barks, les formants sont perceptiblement éloignés et les changements de leur amplitude relative n'affectent pas la qualité perçue des voyelles.

3-5-7- Technique d'estimation de F'2 : *[12] et [17]*

Les méthodes d'estimation des formants perceptifs peuvent être classées dans deux catégories. La première calcule $\mathbf{F'_2}$ comme moyenne de l'ensemble des fréquences des quatre premiers formants. La deuxième estime $\mathbf{F'_2}$ sans utiliser la fréquence standard des formants. Ainsi, **[31]** **[22]**, **[23]** et **[24]**, proposent des formules pour estimer le $\mathbf{F'_2}$ comme moyenne pondérée des quatre premiers formants. **[28]** estiment $\mathbf{F'_2}$ par un spectre LPC d'ordre élevé modifié en échelle Mel et appliquant un seuillage d'égale sonie. **[29]** a utilisé une technique de prédiction linéaire perceptive (PLP) basé sur un modèle LPC de 5ème ordre dans un contexte auditif.

Chapitre 4:Paramétrisation via les Coefficients MFCC et les coefficients NPC

4-1 : INTRODUCTION :

La parole est produite par l'articulation des membres phonatoires de l'homme et prend une forme analogique apériodique, ce qui est impossible pour que la machine puisse l'interpréter ou le prédire, car elle ne comprend que du numérique. Pour cela on doit faire un traitement de paramétrisation sur ce signal.

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes [2], un tel système prend un signal en entrée et retourne un vecteur de paramètre (appelé indifféremment vecteur acoustique ou encore vecteur d'observation).Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

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

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

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

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

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

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

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

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

- Mise en forme du signal
- Détermination des coefficients MFCC
- Détermination des coefficients LPC
- Détermination des coefficients NPC par extension des LPC

4-2- Mise en forme du signal de la parole :

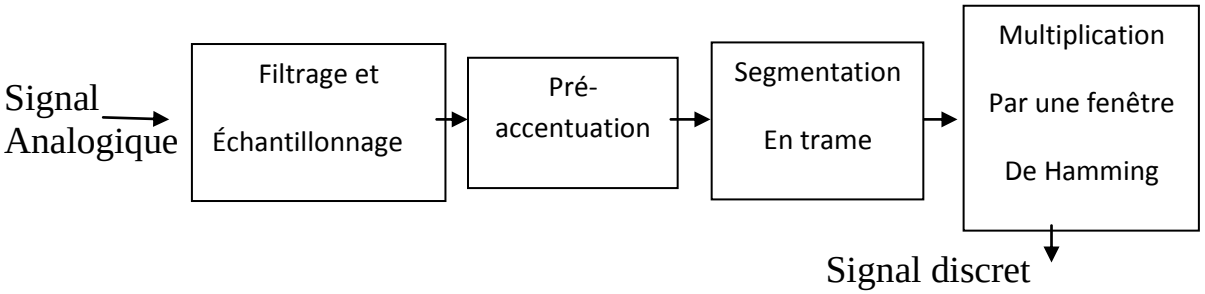


Figure 4-1 : *Mise en forme du signal*

4-2-1 Echantillonnage :

L'échantillonnage consiste à transformé une fonction **s(t)** à valeur continues en une fonction **ŝ(t)** discrète constituée par la suite des valeurs **s(t)** aux instants d'échantillonnage **t=KT** avec **K** un entier naturel. Le choix de la fréquence d'échantillonnage n'est pas aléatoire car une petite fréquence nous donne une présentation pauvre du signal. Par contre une très grande fréquence nous donne des mêmes valeurs, redondance, de certains échantillons voisins, donc il faut prélever suffisamment de valeurs pour ne pas perdre l'information contenue dans **s(t).**

Cette problématique à été résolver par le théorème de Shannon « la fréquence d'échantillonnage assurant un non repliement du spectre doit être supérieure à 2 fois la fréquence haute du spectre du signal analogique »

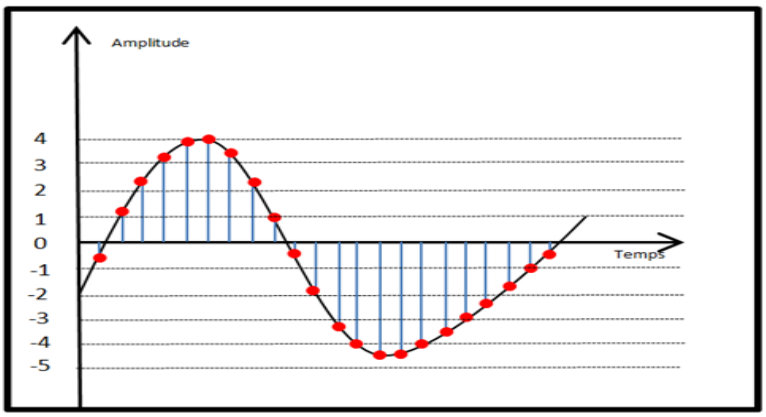


Figure 4.2- signal échantillonné

Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limitée à 3400 Hz et l'on choisit $\mathbf{f_e = 8000}$ Hz.

Pour les techniques d'analyse, de synthèse ou de reconnaissance de parole, la fréquence peut varier de 6000 à 16000 Hz.
Puisque $\mathbf{f_m = 5}$ KHz alors $\mathbf{f_e = 10\ KHz}$ et donc il va falloir mesurer le signal tous le $\frac{\mathbf{1}}{\mathbf{10000}}\ \boldsymbol{seconde} = \mathbf{0,1}\ \boldsymbol{msec}$.

4-2-2 Quantification :

Cette étape consiste à approximer les valeurs réelles des échantillons selon une échelle de **n** niveau appelée échelle de quantification. Il y'a donc $\mathbf{2^n}$ valeurs possibles comprises entre **-2n-1** et **2n-1** pour les échantillons quantifiés.
L'erreur systématique que l'on commet en assimilant les valeurs réelles de l'écart au niveau de quantifiant le plus proche est appelé bruit de quantification.

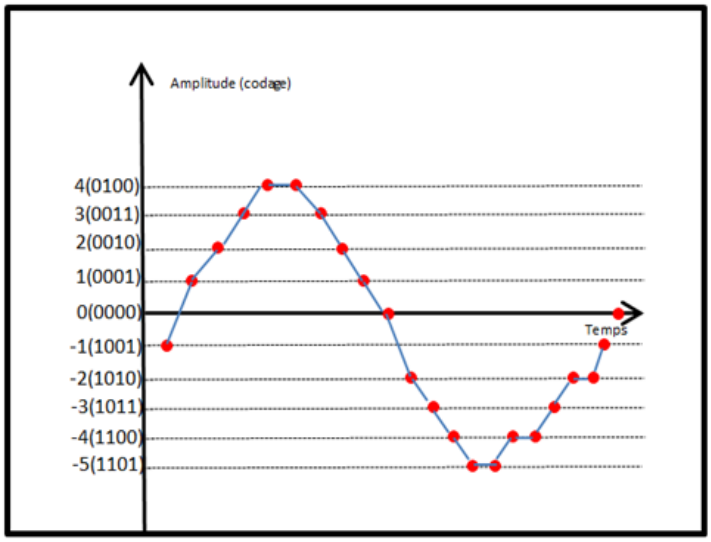


Figure 4-3- signal quantifié

4-2-3 : Codage

C'est la représentation linéaire des valeurs quantifiés qui permet le traitement du signal sur machine.

4-2-4 : préaccentuation

Le signal échantillonné $\mathbf{s_e}$ est ensuite préaccentué afin de relever les hautes fréquences qui sont moins énergétiques que les basses fréquences. Cette étape

consiste à faire passer le signal $\mathbf{s_n}$ dans un filtre numérique à réponse impulsionnelle finie de premier ordre donné comme suit :

$$\boldsymbol{H(z) = 1 - \propto z^{-1}} \qquad \text{Avec } \mathbf{0,9 \leq \propto \leq 1} \qquad \textbf{4-1}$$

Ainsi, le signal préaccentué $\mathbf{s_a}$ est lié au signal $\mathbf{s_e}$ par la formule suivante :

$$\boldsymbol{s_a(n) = s_e(n) - \propto s_e(n-1)} \qquad \textbf{4-2}$$

4-2-5 : fenêtrage

Rappelons que la parole est un signal réel, continu, d'énergie finie, et surtout non stationnaire, mais qui présente une évolution lente au cours du temps des propriétés du signal de la parole, ceci nous permet d'isoler de court segments du signal et de les traiter successivement comme s'ils résultent d'un son stationnaire avec des propriétés invariables, par ailleurs on garde une région commune de 5 à 12 ms entre intervalles successives, on définit donc la fenêtre d'analyse qui n'a des valeurs non nulles qu'à l'intérieur de l'intervalle **[2]** , **[3]**.

Donc, si nous définissons **w(n)** fenêtre où $\mathbf{0} < n < N - 1$ et **N** représente le nombre d'échantillon dans chacune des trames, alors le résultat du fenêtrage est le signal $\mathbf{s_w}$, donné par la formule :

$$\boldsymbol{s_w = s(n)w(n)} \qquad \mathbf{0} < n < N - 1 \qquad \textbf{4-3}$$

Les fenêtres les plus utilisées sont :

- Fenêtre de Hamming

$$\boldsymbol{w(n) = \begin{cases} 0,54 - 0,46cos\frac{2\pi n}{N-1} & 0 \leq n \leq N-1 \\ 0 & sinon \end{cases}} \qquad \textbf{4-4}$$

- Fenêtre rectangulaire

$$\boldsymbol{w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & Sinon \end{cases}} \qquad \textbf{4-5}$$

- Fenêtre triangulaire

$$\boldsymbol{w(n) = \begin{cases} \frac{2n}{N-1} & si \;\; 0 \leq n \leq \frac{N-1}{2} \\ \frac{2(N-n-1}{N-1} & si \;\; \frac{N-1}{2} < n \leq N-1 \\ 0 & sinon \end{cases}} \qquad \textbf{4-6}$$

- Fenêtre de Hanning

$$\boldsymbol{w(n) = \begin{cases} 0,5 - 0,5cos\frac{2\pi n}{N-1} & Si \quad 0 \leq n \leq N-1 \\ 0 & sinon \end{cases}} \qquad \textbf{4-7}$$

- Fenêtre de Blakman :

$$w(n) = \begin{cases} 0,42 - 0,5cos\frac{2\pi n}{N-1} + 0,08cos\frac{4\pi n}{N-1} & si \quad 0 \leq n \leq N-1 \\ 0 & sinon \end{cases} \qquad \textbf{4-8}$$

La figure 4-4 illustre la forme que prennent les fonctions définies ci-dessus.

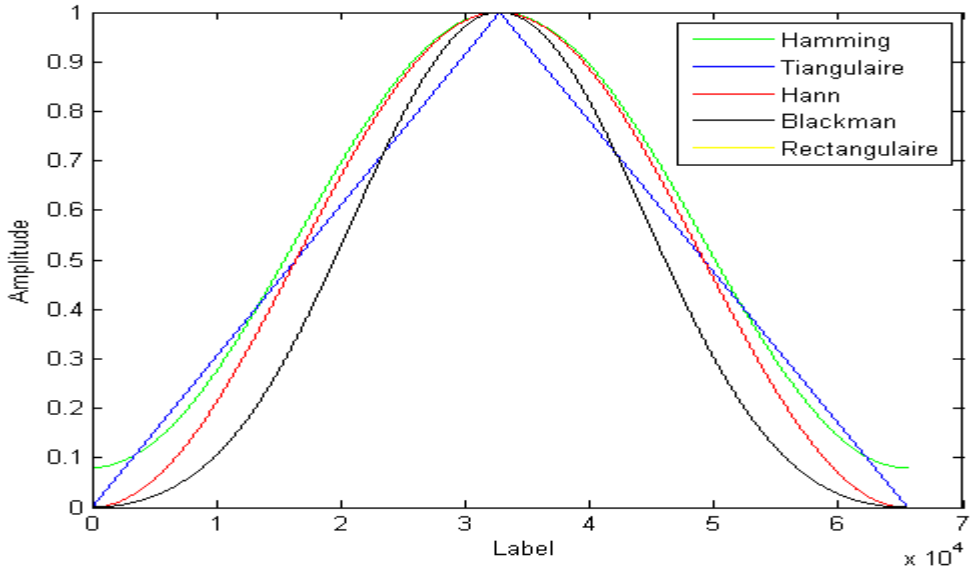


Figure 4-4 : fonction de fenêtrage

Donc parmi les types de fenêtres citée ci-dessus, on choisit celle de Hamming car elle n'introduit pas de grande perturbation sur le signal (atténuation ou rapport du lobe principal sur le lobe secondaire = 41 dB, avec concentration de l'énergie dans le lobe centrale = 99,96%).

4-3 : paramétrisation via les MFCC (Scal-Mel Frequency Cepstral Coefficient) :

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

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

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

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

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

$$\boldsymbol{Mel(f) = blog_{10}(1 + \frac{f}{c})} \qquad \textbf{4-9}$$

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

Mel	**500**	**600**	**700**	**800**	**900**	**1000**	**1100**	**1200**	**1300**	**1400**	**1500**	**1600**	**1700**	**1800**	**1900**	**2000**	**2100**
Fréqu.	**510**	**630**	**770**	**920**	**1080**	**1270**	**1480**	**1720**	**2000**	**2320**	**2700**	**3150**	**3700**	**4400**	**5300**	**7700**	**9500**

Tableau 1 : Transformation des échelles

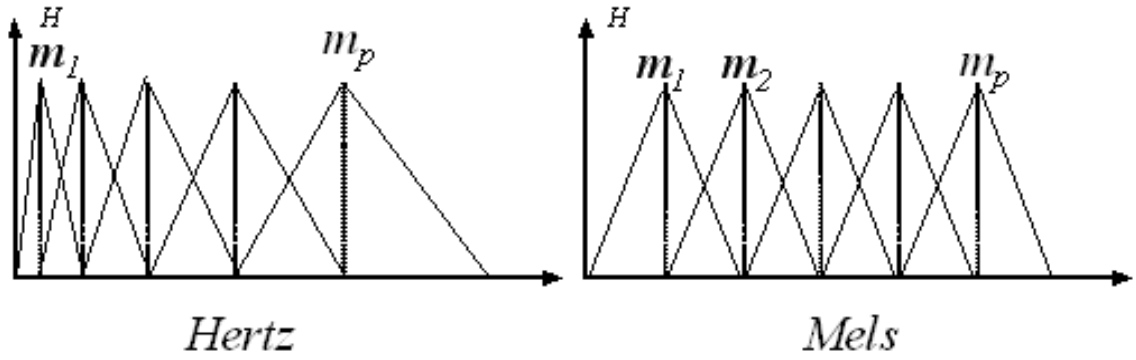


Figure 4.5- Répartition des filtres triangulaires sur les échelles

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

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

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

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

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

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

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

L'algorithme peut être décrit comme suit :

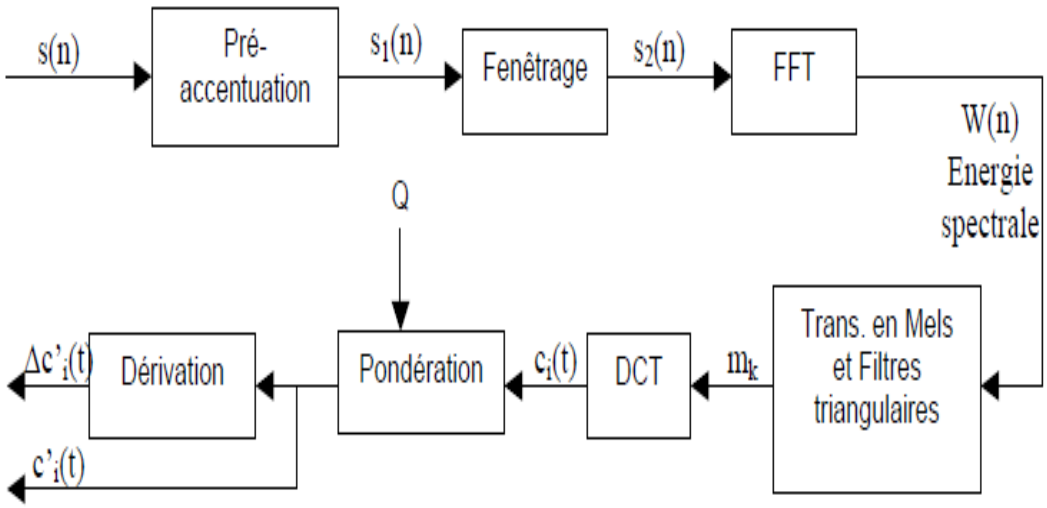


Figure 4-6: Processus d'extraction des MFCCs

a- Calcul de la transformée de Fourier (FFT) :

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

$$x[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-2j\pi}{N}kn} \quad , \quad 0 \leq k \leq N-1 \qquad \textbf{4-10}$$

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

b- Transformation en échelle Mels et Filtres triangulaires :

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

$$\mathbf{W_n = |s(f_n)|^2} \qquad \textbf{4-11}$$

La chaîne des coefficients $\mathbf{m_k}$ (k = 1,2, …, K) de **K** filtres obtenue par la somme accumulée après avoir transformé $\mathbf{W_n}$ en échelle Mels et après avoir passé à travers de chaque filtre de bande.
L'ensemble des filtres de bande comprend usuellement plus de 20 filtres triangulaires.

c- Transformation en Cosinus Discret : DCT (Discret Cosinus Transform) :

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

$$\mathbf{C_i = \sqrt{2/k} \sum_{j=1}^{k} \ln(m_j) \cos\left(\frac{\pi i}{K}(j-0.5)\right)} \qquad \textbf{4-12}$$

i = 1, 2, ..., K

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

Nous allons vérifier si cette combinaison de 13 coefficients plus leurs variations en $\mathbf{1^{ere}}$ ordre, en $\mathbf{2^{eme}}$ sont les meilleures pour la reconnaissance.

d- Pondération

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

$$\hat{c}_i=\left[1+\frac{Q}{2}\sin\left(\frac{\pi i}{Q}\right)\right]c_i \quad 1\leq i\leq Q \qquad 4\text{-}13$$

e- Dérivation

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

$$\Delta\hat{C}_i = \frac{\sum_{\theta=1}^{\Theta}\theta(\hat{C}_{t+\theta}-\hat{C}_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta_2} \qquad 4\text{-}14$$

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

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

- Signal découpé en trame de **N** échantillon :

  Le Signal échantillonné doit subir une préaccentuation puis un fenêtrage.

- On fait passer un filtre de Hamming. On obtient des trames pondérées

- Celles-ci subissent une transformé de Fourier discret **DFT.**

- Puis les spectres de signal obtenus par **DFT** sont multipliés par des filtres triangulaires **B(m).**

- Ces derniers doivent passer par une étape ou on fait sortir des logarithmes d'énergie appelés aussi des logarithmes de spectre d'amplitudes. **E(m)= log |.|**

- Enfin les coefficients cepstraux des **MFCC**, peuvent être obtenus par une transformé de cosinus discrète de **E(m).** Appelé $\mathbf{C_{(n)}}$ ou **iDCT.**

Les coefficients **MFCC [17]** sont un type de coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole. Le codage **MFCC** utilise une échelle fréquentielle non-linéaire. L'échelle Mel est définie par :

$$B(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad \textbf{4-15}$$

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

Soit un signal discret **{*x[n]*}** avec $\mathbf{0 \le n \le N-1}$, ***N*** est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète ***S[k]*** est obtenue par 4-2:

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

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

$$B[m] = B(f_l) + m\frac{B(f_k) - B(f_l)}{M+1} \qquad 0 \le m \le M+1 \qquad \textbf{4-16}$$

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

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

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_k) - B(f_l)}{M+1}\right)$$

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

$$B^{-1}(b) = 700 * (10^{b/2595} - 1)$$

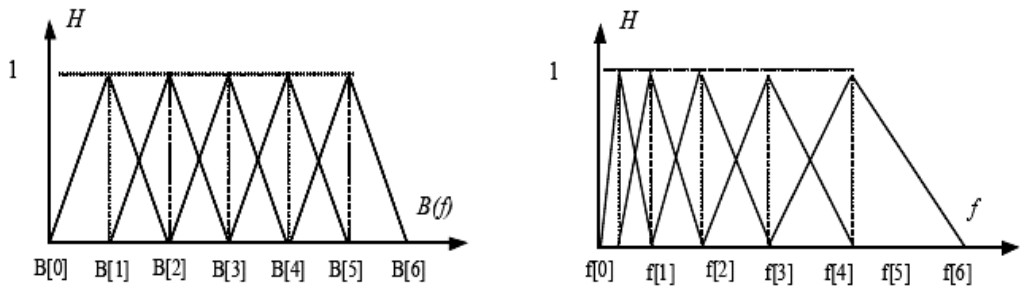


Figure 4.7- Les filtres triangulaires passe-bande en Mel-Fréquence (**B(f)**) et en fréquence **(f)**

Le coefficient $H_m[k]$ de chaque filtre est déterminé par le système suivant :

$$H_m[k]=\begin{cases}0 & si \quad k \le f[m-1] \\ \dfrac{k-f[m-1]}{f[m]-f[m-1]} & si \quad f[m-1] \le k \le f[m] \\ \dfrac{f[m+1]-k}{f[m+1]-f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1]\end{cases} \qquad \textbf{4-17}$$

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

$$E[m]=\log\left[\sum_{k=0}^{N-1}|S[k]|^2 H_m[k]\right] \quad 0 \le m < M \qquad \textbf{4-18}$$

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

$$c[n]=\sum_{m=0}^{M-1} E[m]\cos\left(\frac{\pi n(m+\frac{1}{2})}{M}\right) \quad 0 \le n < M \qquad \textbf{4-19}$$

En résumer :

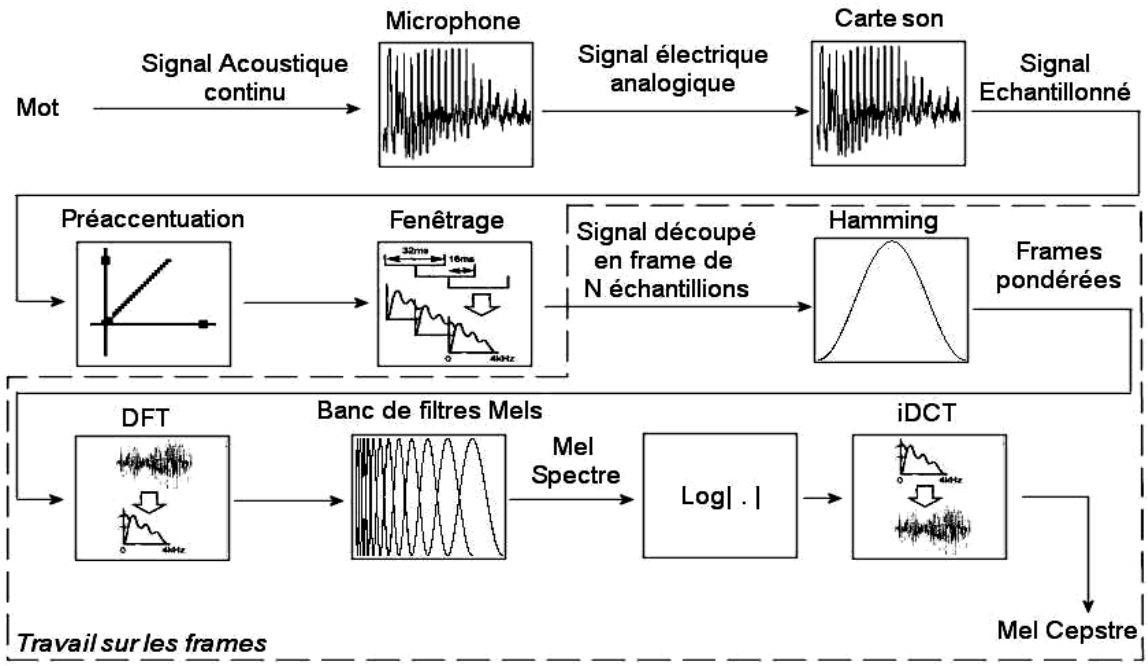


Figure 4.8- Etapes du calcul des coefficients MFCC

4–4- Paramétrisation basée sur un modèle de production de la parole (LPC) :

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

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

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

$$\hat{s}_n = \sum_{i=1}^{p} a_i \, x_{n-i} \qquad \textbf{4-20}$$

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

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + G e_n \qquad \textbf{4-21}$$

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

$$GE(Z) = \left(1 - \sum_{i=1}^{p} a_i Z^{-i}\right) X(Z) \qquad \textbf{4-22}$$

D'où:

$$H(Z) = \frac{X(Z)}{E(Z)} = \frac{G}{(1-\sum_{i=1}^{p} a_i Z^{-i})} = \frac{G}{A(Z)} \qquad \textbf{4-23}$$

Cette équation peut être interprétée comme suit :

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

Les coefficients $\mathbf{a_i}$ sont les coefficients qui minimisent l'erreur quadratique moyenne :

$$E_n = \sum_m (G - e_{n+m})^2 = [\sum_m (x_{n+m} - \sum_{i=1}^{p} a_i x_{n+m-i})]^2 \qquad \textbf{4-24}$$

Calcul des coefficients a_i : La prédiction de $\mathbf{s_n}$ est

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} \qquad \textbf{4-25}$$

Et l'erreur est donc

$$\varepsilon_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k\, s_{n-k} \qquad \textbf{4-26}$$

La fonction d'énergie associée (minimisation de l'erreur au sens des moindres carrés est

$$E = \sum_{n=1}^{N} \varepsilon_n^2 = \sum_{n=1}^{N} (s_n - \hat{s}_n)^2 \qquad \textbf{4-27}$$

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

$$\forall i = 1, \ldots\ldots, p \qquad \frac{\partial E}{\partial a_i} = 0 \qquad \textbf{4-28}$$

C'est-à-dire

$$\forall i = 1, \ldots\ldots, p \qquad \sum_{n=1}^{N} s_{n-i}\, s_n = \sum_{n=1}^{N} \sum_{k=1}^{p} a_k s_{n-i} s_{n-k} \qquad \textbf{4-29}$$

Ou encore

$$\forall i = 1, \ldots\ldots, p \qquad \sum_{k=0}^{p} a_k \left(\sum_{n=1}^{N} s_{n-i}\, s_{n-k} \right) = 0 \qquad \textbf{4-30}$$

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

$$\varphi_s(i) = \sum_{n=1}^{N} s_{n-i}\, s_n, \qquad i = 1, \ldots\ldots\ldots, p. \qquad \textbf{4-31}$$

On note

$$\phi(i,k) = \sum_{n=1}^{N} s_{n-i}\, s_{n-k} = \varphi_s(i-k) \qquad \textbf{4-32}$$

(Par prolongement périodique du signal échantillonné) et on obtient

$$\forall i = 1, \ldots\ldots, p \qquad \sum_{k=1}^{p} a_k \varphi_s\,(i-k) = \varphi_s(i) \qquad \textbf{4-33}$$

Finalement

$$\begin{Bmatrix} \varphi_s(0) & \varphi_s(-1) & \dots & \varphi_s(1-p) \\ \varphi_s(1) & \varphi_s(0) & \dots. & \varphi_s(2-p) \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \varphi_s(p-1) & \varphi_s(p-2) & \dots & \varphi_s(0) \end{Bmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{Bmatrix} = \begin{Bmatrix} \varphi_s(1) \\ \varphi_s(2) \\ \vdots \\ \varphi_s(p \end{Bmatrix} \qquad \textbf{4-34}$$

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

R_s : est la matrice d'auto covariance du signal **s**, c'est une matrice de **Toeplitz**. Cette équation mène au système d'équation appelée équation normales :

$$R_s \, \mathbf{a} = [\sigma_e^2, 0, \dots \dots \dots .]^{\mathbf{T}} \qquad \textbf{4-35}$$

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

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

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

- Il présente également une grande variance pour les différentes classes en reconnaissances.

- Il est optimisé pour la tâche de reconnaissance.

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

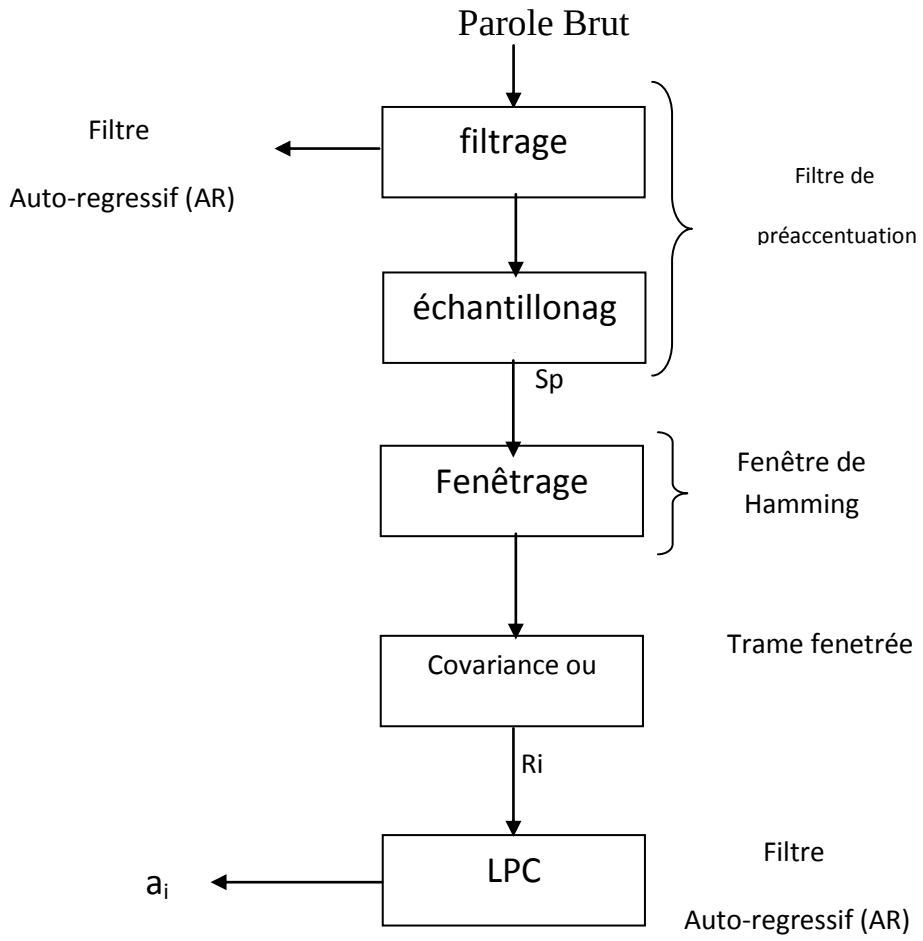


Figure 4-9 : étapes de calcul des coefficients des LPC (a_i)

4 – 5- Autres paramétrisations :

❖ Energie : généralement, l'énergie du signal est utilisée en compliment des coefficients issus d'une paramétrisation sur une analyse dans le domaine Cepstral. L'énergie correspond à la puissance du signale

$$\mathbf{E_n} = \sum_{n=0}^{n-1} S_n^2 \qquad 4\text{-}36$$

Le calcul de l'énergie se fait généralement sur les fenêtres glissantes de 25 ms avec décalage de 10 ms (soit une valeur toutes les 10 ms du signal).

❖ Taux de passage par zéro : (Z.C.R) zéro crossing rate. le ZCR est un bon complément de l'énergie. un taux de passage par zéro faible et une énergie forte sont un bon indice d'un son voisé, alors qu'un taux de passage par zéro élevé et une énergie plus faible caractérisent plutôt une zone non voisée :

$$\mathbf{Z_{cr}} = \mathbf{0.5} * \sum_{n=0}^{n-1} | signe(x_n) - signe(x_{n-1})| \quad 4\text{-}37$$

Alors le $\mathbf{Z_{cr}}$ propose une bande d'amplitude autour de 0 pour limiter un certain nombre de phénomène parasites qui provoquent de faibles oscillations aux alentours de 0.

❖ Paramétrisation dynamiques : dérivée première et secondes Δ et ΔΔ :

Pour enrichir la paramétrisation, les dérivées de première et seconde ordre sont souvent utilises. Cela permet d'ajouter de l'information concernant la dynamique du signal.

Les coefficients Δ (dérivées du premier ordre) sont souvent estimés grâce au développement limite d'ordre 2 :

$$C_i'(\mathbf{t}) = \frac{C_i(t+1) - C_i(t-1) + 2[C_i(t+2) - C_i(t-2)]}{10} \quad \mathbf{4\text{-}38}$$

$\mathbf{C_i(t)}$ correspond en $i^{ième}$ coefficient pour la trame t

$C_i'(\mathbf{t})$ sa dérivée

ΔΔ (Coefficients du second ordre) sont estimés de la même manière à partir des coefficients du premier ordre.

4 – 6 - Le Codage Neuro-Prédictif (Neural Predictive Coding ou NPC)

Au cours des deux dernières décennies, nous avons pu constater un développement fulgurant des réseaux de neurones. Cet intérêt à démarré avec l'application réussie de cette technique puissante pour des problématique très différentes, et dans des domaines aussi divers que la médecine, la production industrielle, la géologie ou encore la robotique et traitement du signal de parole.
Le succès croissant des réseaux de neurones sur la plupart des autres statistiques peut s'attribuer à leur puissance, leur polyvalence et à leur simplicité d'utilisation. Les réseaux de neurones sont des techniques extrêmement sophistiquées de modélisation et de prévision, en mesure de modéliser des relations entre les données ou des fonctions particulièrement complexes.

La neurophysiologie moderne a identifié des classes de simplifications remarquables qui témoignent du caractère hautement prédictif du fonctionnement du cerveau et permettent de résoudre des problèmes complexes avec une grande rapidité.

La possibilité d'apprendre sur la base d'exemples constitue l'une des nombreuses fonctionnalités des réseaux de neurones qui permettent à l'utilisateur de modéliser ses données et établir des règles précises qui vont guider les relations sous-jacentes entre différents attributs de données. L'utilisateur des réseaux de neurones collecte des données représentatives puis il fait appel aux algorithmes d'apprentissage, qui vont apprendre automatiquement la structure des données.
Bien qu'il existe différente manière de réaliser l'apprentissage des réseaux de neurones, la plupart d'entre elles utilisent des algorithmes numériques qui sont en mesure d'effectuer la tâche au cours d'un nombre fini d'itérations. Nous avons recours a ces algorithmes itératifs essentiellement en raison de la nature fortement non-linéaire des modèles. Un algorithme d'apprentissage itératif va ajuster graduellement les poids du réseau de neurones de telle sorte que toute donnée d'entrée **x**, le réseau de neurone est en mesure de produire une sortie aussi proche que possible de **t.**

La performance des réseaux de neurones se mesure par la manière dont ils savent prévoir des données inconnues. Ce processus est connu comme la généralisation. La question de généralisation est en fait une question majeure qui se pose lors de l'apprentissage. Ce phénomène s'exprime par une tendance au ***surajustement*** des données d'apprentissage s'accompagnant d'une difficulté à prévoir de nouvelles données. Par conséquent, lors de l'apprentissage des réseaux de neurones, nous devons toujours tenir compte des questions de performance et de généralisation

Le codage neuro-prédictif est une extension du codage **LPC [1], [12], [17]**, donc une méthode de codage temporelle mais contrairement à la méthode **LPC**, le codeur **NPC** extrait les caractéristiques non linéaires d'un phonème. Il est basé sur un **MLP** à une couche cachée suivi d'une couche de sortie à 1 neurone appelé cellule de prédiction. L'étape d'apprentissage consiste à prédire un échantillon (extrait du signal acoustique d'un phonème) à partir des n échantillons précédents grâce à un **MLP**.

$$\mathbf{y}_k = \left[y_{k-1}, y_{k-2}, \cdots y_{k-L} \right]^T \qquad \textbf{4-39}$$

Une séquence d'échantillons (**L** est la longueur de la séquence dite largeur de la fenêtre de prédiction qui est égale à la dimension de la couche d'entrée de notre **MLP**). Le prédicateur calcule $\hat{\mathbf{y}}_k$ à partir du vecteur $\mathbf{y}_k$:

$$\hat{y}_k = \mathrm{F}(\mathbf{y}_k) \qquad \textbf{4-40}$$

F Peut être vu comme la composition de deux fonctions $\mathbf{G_w}$ (correspond à la couche cachée) et $\mathbf{H_a}$ (correspond à la couche de sortie).

$$F = H_a \circ G_w \qquad 4\text{-}41$$

Avec $\hat{y}_k = H_a(z_k)$ et $z_k = G_w(y_k)$ 4-42

W indique le vecteur des poids de la couche cachée et **a** le vecteur des poids de la couche de sortie.

Tous les poids du réseau sont ajustés de façon à minimiser l'erreur quadratique :

$$L = \sum_k (y_k - \hat{y}_k)^2 \qquad 4\text{-}43$$

Ce sont ces poids qui vont constituer le vecteur acoustique caractéristique de chaque phonème, mais si on les utilise tous, on aura trop de paramètres (ex : un **MLP** 20 x 8 x 1 =321 poids) , le codeur **NPC** permet un nombre réduit de paramètres, l'idée est de ne prendre comme coefficients que les poids de la couche de sortie (ex : un **MLP** 20 x 8 x 1 = 8 poids pour couche de sortie = 8 coefficients) tant que les poids de la première couche sont utilisés seulement comme un filtre.

Cela est fait en créant une couche de sortie différente pour chaque phonème, quand à la première couche, elle reste constante pour tous les phonèmes. Considérant maintenant un ensemble de phonèmes i=1..N.

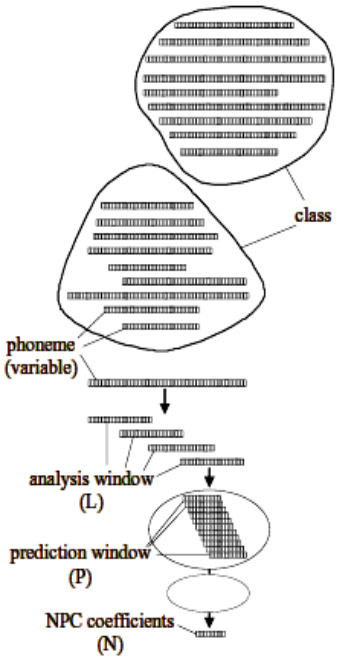


Figure 4.10- Calcul des coefficients NPC

Les poids de la couche de sortie sont propres à chaque phonème, et constitue notre vecteur de coefficients, Quand aux poids de la première couche, sont partagés par tous les phonèmes, et constituent la partie fixe du système. Le processus d'apprentissage se divise en deux phases **[12]**.

La première phase dite paramétrisation, faite par l'ajustement des poids de la première couche (calcul de la partie fixe du codeur), la deuxième phase dite codage faite par l'ajustement des poids de la couche de sortie. La phase de paramétrisation est faite en entraînant le **MLP** à prédire tous les échantillons de tous les phonèmes en minimisant l'erreur de prédiction.

Dans la deuxième phase on garde les poids de la première couche et on réinitialise les poids de la couche de sortie par des valeurs fixes, chaque phonème étant décomposé en fenêtre de taille choisie. On essaie de minimiser l'erreur de prédiction sur une fenêtre **(NPC1)** ou pour toute les fenêtres d'un phonème **(NPCII)** du MLP en ajustant que les poids de couche de sortie (avec un nombre choisi d'itérations, de préférence pas trop grand pour ne pas ralentir le système), sans toucher aux poids de la première couche qui resteront fixes pour tous les phonèmes **[12],[17]**.

Enfin on récupère les poids de la couche de sortie, qui constitueront notre vecteur de paramètres.

Présentation de la rétropropagation du gradient :

La règle du gradient de l'erreur (delta rule) est l'une des règles les plus utilisées pour l'apprentissage de réseaux de neurones. Cette règle, initialement développée pour résoudre des problèmes de traitements adaptatifs du signal **[WIDROW et HOFF, 1960]** à ensuite été exploitée pour obtenir le très populaire algorithme de rétro propagation du gradient de l'erreur (backpropagation) **[RUMELHART et al. 1986]** pour réseaux de neurones multicouche. L'objectif de cet algorithme est de minimiser une fonction de coût '**E**'.

L'équation exprime cette fonction de coût à partir de l'erreur quadratique, pour un couple entrée-sortie, avec $\mathbf{d_k}$ la sortie désirée pour le neurone d'indice **k** et $\mathbf{s_k}$ la sortie obtenue par le réseau.

Sachant que les réseaux de neurones multicouches se basent essentiellement sur l'outil le plus utilisé dans le domaine et qui est la rétropropagation du gradient.
C'est une technique de calcul des dérivées qui peut être appliquée à n'importe quelle structure de fonction dérivables.

Mathématiquement, cette méthode est basée sur l'algorithme de descente du gradient et utilise les règles de dérivation des fonctions dérivables **[17]**. Dans cette

méthode, l'erreur commise en sortie du réseau sera rétropropagée vers les couches cachées d'où le nom de rétropropagation.

$$\mathbf{E} = \sum_i (\mathbf{d_k} - \mathbf{s_k})^2 \qquad \textbf{4-44}$$

L'apprentissage comporte une première phase de calcul dans le sens direct où chaque neurone effectue la somme pondérée de ses entrées et applique ensuite la fonction d'activation **f** (fonction dérivable) pour obtenir la mise à jour du neurone.

$$s_i = f(p_i) \qquad où \qquad p_i = \sum_{j=0}^{n} w_{ij}\, x_j, \qquad \textbf{4-45}$$

Il faut faire la mise à jour avec $\mathbf{p_i}$ le potentiel post-synaptique du neurone **i**, $\mathbf{x_j}$ l'état du neurone de la couche précédente et $\mathbf{w_{ij}}$ le poids de la connexion entre les deux neurones.

Cette phase, dite de propagation, permet de calculer la sortie du réseau en fonction de l'entrée. L'algorithme de rétropropagation consiste à effectuer une descente de gradient sur le critère « **E** ».

Le gradient de **E** est calculé par tous les poids de la manière suivante :

$$\frac{\partial \mathbf{E}}{\partial \mathbf{w_{ij}}} = \frac{\partial \mathbf{E}}{\partial \mathbf{p_i}} \frac{\partial \mathbf{p_i}}{\partial \mathbf{w_{ij}}} = \frac{\partial \mathbf{E}}{\partial \mathbf{p_i}}\, \mathbf{x_j} \qquad \textbf{4-46}$$

Le gradient sera ensuite noté : $\mathbf{C_j}$ et $\mathbf{C_j} = -\frac{\partial \mathbf{E}}{\partial \mathbf{p_i}}$. **4-47**

On distingue alors deux cas, suivant que le neurone d'indice : est un neurone de sortie ou non.

Dans le cas de la couche de sortie :

Le gradient attaché aux cellules de sortie est obtenu par l'équation suivante :

$$C_i = -\frac{\partial E}{\partial p_i} = -\frac{\partial(\sum_k (d_k - s_k)^2)}{\partial p_i} = 2(d_i - s_i) f'(p_i) \qquad \textbf{4-48}$$

Car seul $\mathbf{s_i}$ dépend de $\mathbf{p_i}$ et $\mathbf{s_i = f(p_i)}$

Pour les neurones de la couche cachée :

L'ordre de calcul du gradient est l'inverse de celui utilisé pour la mise des états dans le réseau. Il s'effectue de la couche de sortie vers l'entrée, on parle alors de rétropropagation.

L'expression du gradient est obtenue comme indiqué dans l'équation :

$$C_i = -\frac{\partial E}{\partial p_i} = -\sum_{k=0}^{n} \frac{\partial E}{\partial p_k} \frac{\partial p_k}{\partial p_i} = \sum_{k=0}^{n} C_k \frac{\partial p_k}{\partial p_i} = \sum_{k=0}^{n} C_k \frac{\partial p_k}{\partial s_i} \frac{\partial s_i}{\partial p_i} \qquad \textbf{4-49}$$

Soit encore :

$$C_i = f'(p_i) \sum_{k=0}^{n} w_{ki}\, C_k \qquad \textbf{4-50}$$

Avec $\mathbf{C_k}$ le gradient du neurone **k** de la couche suivante (dans le sens de la propagation).

Dans le cas de l'algorithme de gradient total, les exemples de la base d'apprentissage sont présentés successivement au réseau, les gradients accumulés au fur et à mesure et la modification des poids n'intervient qu'après présentation de tous les exemples (par opposition au gradient stochastique où la modification des poids est effectuée pour chaque exemple présenté).

La modification des poids est obtenu par :

$$\boldsymbol{w_{ij}^{t+1} = w_{ij}^{t} + \alpha C_i S_j} \qquad \textbf{4-51}$$

Où α est un petit nombre positif qui représente le pas de déplacement en direction du minimum le plus proche.

Précisions que la phase d'apprentissage est souvent arrêtée lorsque l'erreur calculée sur l'ensemble de la base d'apprentissage est inférieure à un seuil déterminé par l'utilisateur.

Chapitre 5: Modélisation selon LVQ-NPC

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

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

Chaque classe est caractérisée par un ensemble fixé d'unité calculant des distances. Leurs poids w_i, constituent des vecteurs de référence de même dimensions que les entrées, lorsqu'on désire classifier un vecteur **x**, on sélectionne le vecteur de référence le plus proche w^* , et on regarde la classe **Φ (w*)** qui lui est associée.

Si l'on construit le diagramme de ***VONOROÎ*** associé aux divers vecteurs de référence, les frontières de séparation entre classe sont constituées des portions des frontières du diagramme séparant deux vecteurs de référence associés à deux classes différentes.

Ce principe permet donc de réaliser des frontières non-linéaires, et même non convexes, s'il y a plusieurs références par classes.

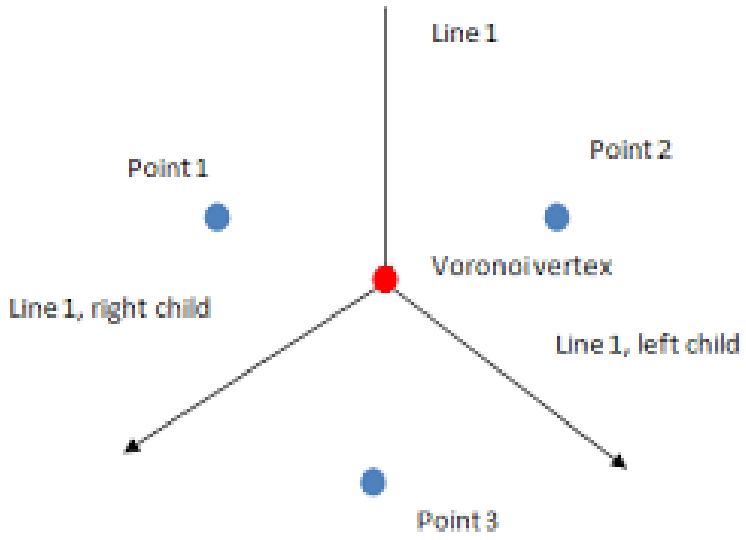


Figure 5.1 – diagramme de ***VONOROÎ*** montrant les frontières de séparations entre classes

Les algorithmes **LVQ** sont réputés efficaces pour des tâches de classification, mais sont peu adaptés aux données bruités.

Un système composé d'un perceptron multicouches et d'un **LVQ** permet d'obtenir de meilleurs résultat qui chacun de ces systèmes utilisés, isolément. Elle peut être interprétés comme la coopération de deux étages, possédant chacun leur propre algorithme d'apprentissage.

Elle permet d'aborder l'aspect structurel des systèmes d'apprentissage et de concevoir des systèmes modulaires, ce qui peut avoir un impact considérable sur le développement et l'implémentation de système d'apprentissage complexes.

Chaque système était entrainé séquentiellement : dans un premier temps, on entraine un **MLP** pour notre tâche de classification. Ensuite, on utilise les premiers étages du perceptron, et on classe les sorties obtenues au moyen de l'algorithme **LVQ**.

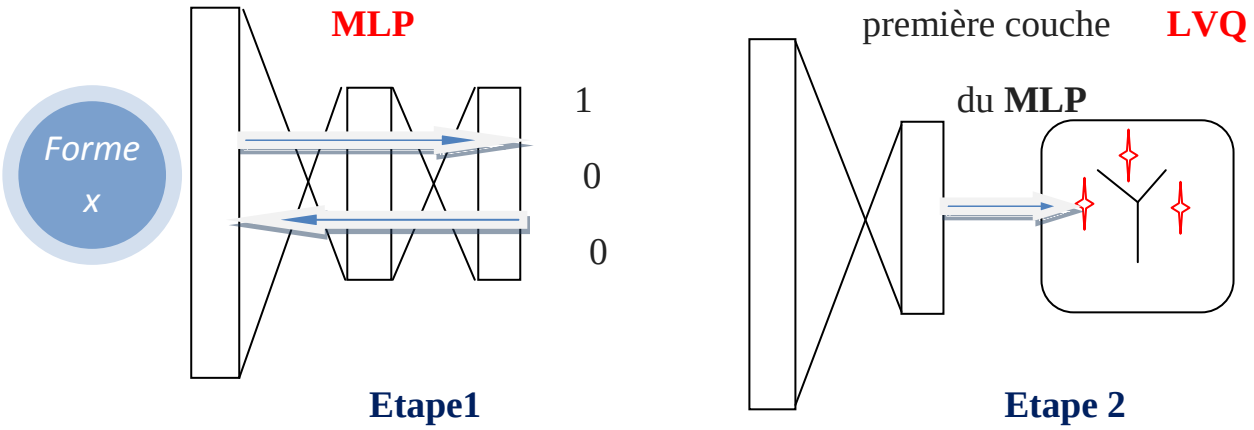


Figure 5.2- Apprentissage **LVQ-NPC**

Dans la première étape, les premières couches du perceptron ont été entrainées à produire des représentations qu'un séparateur linéaire pouvait décoder. Ce séparateur linéaire n'est pas optimal : utiliser sur ces données une méthode de plus proches voisin, comme **LVQ** permet d'améliorer les performances.

Or, ces données n'ont pas été optimisées pour être décodées à l'aide d'un tel algorithme.

On peut alors espérer de meilleur résultat en entrainant coopérativement nos systèmes et non séquentiellement.

L'objectif de l'apprentissage simultané du modèle **NPC** et du classifieur à prototypes est l'introduction de contraintes discriminante. Ce classifieur a été appliqué avec succès dans différents domaines comme la reconnaissance de l'écriture ou bien de reconnaissance de la parole.

L'apprentissage simultané des modèles **NPC** et **LVQ** peut se faire par le biais de la méthode **DFE** proposée par **Biem**. Cette méthode permet d'introduire les discriminations imposée par le **LVQ** dans le modèle **NPC**. La méthode **DFE** considère l'extracteur de caractéristiques **NPC** et le classifieur **LVQ** comme un seul module **[5], [16]** décrit par **Φ= (a,m)**.

Alors la mise en œuvre de la méthode d'extraction de caractéristique discriminante **DFE** (discriminative feature extraction) s'effectue en plusieurs étapes. Dans un premier temps, il faut définir une fonction discriminante décrivant le processus d'extracteur de caractéristiques et le classifieur. Dans le cas du **LVQ**, la fonction discriminante est la valeur négative du minimum de la distance entre le vecteur d'entrée (le vecteur caractéristique) et le prototype correct.

$$g_i = -\min_{\tau} \left\| a - m_{i,\tau} \right\|^2 = -\min_{\tau} d\left(a, m_{i,t}\right) \qquad \textbf{5-1}$$

Où **d(a, $\mathbf{m_{i,\tau}}$)** est la distance euclidienne entre le vecteur d'entrée **a** (vecteur caractéristique) et un prototype $\mathbf{m_{i,\tau}}$ de la classe $\mathbf{C_i}$

Par la suite, nous définissons une mesure de mauvaise classification :

$$\mu_i(a) = g_i(a) + \left[\frac{1}{M\ 1} \sum_{j \neq i} g_j(a)^{\psi}\right]^{\frac{1}{\psi}} \qquad \textbf{5-2}$$

Où $\boldsymbol{\psi}$ est un nombre positif. Pour une grande valeur $\boldsymbol{\psi}$, la mesure de mauvaise classification devient :

$$\mu_i(a) = g_i(a) + \bar{g}_i(a) \qquad \textbf{5-3}$$

$\bar{g}_i$ **(a)** est la fonction discriminante concurrente. Ceci revient à ne considérer que le premier prototype concurrent de la classe incorrecte :

$$\bar{g}_i(a) = \max_{j \neq i} g_i(a) \qquad \textbf{5-4}$$

La mesure de mauvaise classification $\mu_i(a)$ doit être positive quand ‘’**a’’** n’est pas bien classé et négative si ce n’est pas le cas :

$$\mu_r(a) = d\left(a - m_{i,\tau}\right)\ d\left(a, m_{i,t}\right) \qquad \textbf{5-5}$$

Où $m_{i,\tau}$ est le plus proche prototype de la véritable classe alors que $\boldsymbol{m_{j,\vartheta}}$ est le plus proche prototype de la classe incorrecte.

La prochaine étape consiste en la définition du critère de minimisation de l’erreur de classification **MCE** reflétant les erreurs de classification :

$$l_i(a) = l_i(\mu_i) = \frac{1}{1+ e^{c\mu_i}} \qquad \textbf{5-6}$$

Le risque empirique total du critère **MCE** est le suivant :

$$L(a,\mu) = \sum_{n=1}^{N}\sum_{i=1}^{M} l_i(a_n)\delta C(a_n)_i \qquad \textbf{5-7}$$

Où **C** $\mathbf{(a_n)}$ est la classe d’appartenance du vecteur caractéristique $\mathbf{a_n}$ et **δ** est le symbole de Kronecker qui vaut **1** lorsque **C** $\mathbf{(a_n)}$ **= i.** ***N*** est le nombre de fenêtre d’analyse et **M** le nombre de classes.

La descente du gradient probabilistique généralisé **GPD** est appliquée pour la mise à jour des paramètres **Φ = (a,m)**.

$$a_n = a_n - \beta(t)\frac{\partial l_{i(a_n)}}{\partial a_n} \qquad \textbf{5-8}$$

$$m_{i,\tau} = m_{i,\tau} - \propto (t)\frac{\partial l_{i(a_n)}}{\partial m_{i,\tau}} \qquad \textbf{5-9}$$

$$m_{j,v} = m_{j,v,\tau} - \propto (t)\frac{\partial l_{i(a_n)}}{\partial m_{j,v}} \qquad \textbf{5-10}$$

Où ***α(t)*** et ***β(t)*** sont les pas d’apprentissage du classifieur **LVQ** et du modèle **NPC**. Les pas d’apprentissage sont des fonctions décroissantes du nombre d’itération ***t***.

Les lois d’adaptation des prototypes du **LVQ** sont données par :

$$m_{i,\tau} = m_{i,\tau} + 2\alpha\,(t) l_i(a_n)\left(1 - l_i(a_n)\right)\left(a_n - m_{i,t}\right) \qquad \textbf{5-11}$$

$$m_{j,v} = m_{j,v} + 2\alpha\,(t) l_i(a_n)\left(1 - l_i(a_n)\right)\left(a_n - m_{j,v}\right) \qquad \textbf{5-12}$$

Et pour le modèle **NPC**, les vecteurs caractéristiques a_n évoluent selon la loi suivante :

$$\Delta a_n^{MCE} = 2\beta\,(t) l_i(a_n)\left(1 - l_i(a_n)\right)\left(m_{i,t} - m_{j,v}\right) \qquad \textbf{5-13}$$

On peut remarquer que les caractéristiques évoluent en fonction de la distance entre les deux prototypes les plus proches, l'un de la véritable classe l'autre de la classe incorrecte. Elles évoluent donc dans le sens de maximisation de la séparabilité entre ces deux classes.

Ces deux modifications doivent être associées aux modifications classiques du modèle NPC qui sont en fonction de l'erreur de prédiction :

$$\Delta w_{1.2}^{Pred} \; et \; \Delta a^{Pred}$$

Donc l'objectif de la coopération entre les deux modèles **NPC** et **LVQ** est l'introduction de contraintes discriminantes optimales dans la phase de paramétrisation du **NPC**. Par exemple, on pourrait opter pour une minimisation sous contrainte comme dans le cas où l'apprentissage simultané des classifieurs est effectué par le biais du formalisme **lagrangien**.

Ici, nous choisissons une autre approche. Les deux procédés de minimisation sont modérés par le biais d'un coefficient **θ**.

La modification résultante des vecteurs caractéristiques $\mathbf{a_n}$ est :

$$\Delta a = \theta \Delta a^{Pred} + (1 - \theta)\Delta a^{MCE} \qquad \textbf{5-14}$$

La seconde étape de la coopération consiste à modifier les poids des premières couches pour maximiser la séparabilité des classes.

Cependant, la relation entre les poids des premières couches $\mathbf{w_{1,2}}$ et le critère **MCE** n'est pas aussi directe que dans le cas des poids de la couche de sortie, c'est-à-dire les vecteurs caractéristiques.

En considérant l'objectif du modèle **NPC** dans la coopération, cela revient à rapprocher les caractéristiques de leurs prototypes adéquates et de les éloigner des prototypes incorrects **[16]**.

En d'autres mots, le vecteur caractéristique $\mathbf{a_{i,n}}$ produit par le modèle **NPC** pour la fenêtre d'analyse $\mathbf{y_{i,n}}$ (appartenant à la classe $\mathbf{C_i}$) doit être proche de l'un des prototypes $\mathbf{m_{i,\tau}}$.

Pour cela, on introduit une nouvelle étape dans le modèle **NPC**. Pour la fenêtre $\boldsymbol{y_{i,n}}$, on détermine les deux modifications nécessaires pour :

- Rapprochement des caractéristiques vers le prototype $\mathbf{m_{i,\tau}}$: minimisation de l'erreur de prédiction sous la contrainte que la couche de sortie soit $\mathbf{m_{i,\tau}}$. on obtient la modification des premières couches $\Delta w_{1,2}^{mod}$.
- Eloignement des caractéristiques du prototype $m_{j,\vartheta}$: maximisation de l'erreur de prédiction sous la contrainte que la couche de sortie soit $m_{j,\vartheta}$. On obtient la modification des premières couches $\Delta w_{1,2}^{disc}$.

Lors de modification des premières couches est une modération de ces deux effets :

$$\Delta w_{1.2} = \theta \Delta w_{1.2}^{mod} + (1-\theta)\Delta w_{1.2}^{disc} \qquad \textbf{5-15}$$

On peut remarquer que cette loi de modification n'intégré pas la modification du modèle **NPC** $\Delta w_{1,2}^{pred}$. En effet, cette modification n'est plus utile car la contribution $\Delta w_{1,2}^{mod}$ permet de tenir compte de la partie modélisation nécessaire au processus **LVQ-NPC**.

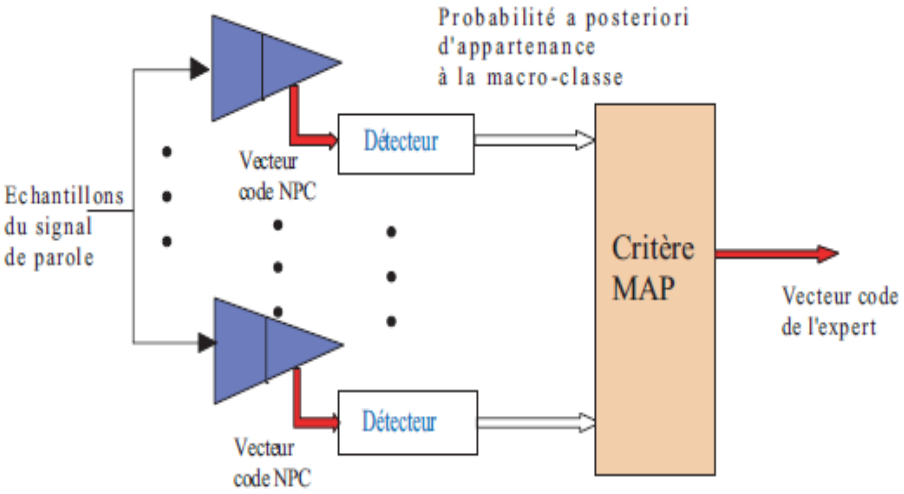


Figure 5-3 :coopération entre LVQ-NPC

Chapitre 6: Classification phonétique de TIMIT

6-1- Base de données TIMIT:

Tout travail s'appuyant sur l'apprentissage nécessite une base de données pour en faire apprendre le système et ensuite de l'évaluer. Ils existent plusieurs base de données internationales dans le domaine de la parole tels que TIMIT qui à été développée par la commission DARPA pour l'anglais américain.

La base TIMIT est une base de données acoustique et phonétique dédiée à la reconnaissance de la parole indépendamment du locuteur. Elle contient les enregistrements de 630 locuteurs américains, répartis en 8 « dialectes régionaux » (''dr1'' à ''dr8'') et prononçant chacun 10 phrases.

Le vocabulaire total de la base est de 6300 mots. Le texte est lu, et les conditions d'enregistrement sont bonnes. Les 630 locuteurs de la base (438 hommes et 192 femmes) sont répartis entre l'ensemble d'apprentissage (462 locuteurs dont 326 hommes et 136 femmes) et l'ensemble de test (168 locuteurs dont 112 hommes et 56 femmes).

Chaque locuteur est identifié par une lettre indiquant son genre (''m'' pour les hommes et '' f '' pour les femmes), ses 3 initiales et un chiffre. Un sous-ensemble, appelé noyau, a été extrait de l'ensemble de test original. Il contient 24 locuteurs : deux hommes et une femme par ''dialecte''. Les phrases de calibration étant exclues, ces 24 locuteurs prononcent 192 phrases. Le noyau de test comprend 7215 phonèmes. Sa taille réduite a permis de multiplier les expériences tout en conservant une estimation assez fiable des taux de reconnaissance réels.

Pour chaque phrase, nous disposons du texte en anglais, du signal échantillonné à 16 kHz sur 16 bits, de la segmentation phonétique en 61 classes, et de la segmentation en mots. La segmentation phonétique est très fine ; en particulier, l'occlusion précédant l'explosion des occlusives et des Afrique est étiquetée individuellement. Ce qui donne la liste des étiquetés de la base TIMIT, le phonème correspondant dans l'Alphabet Phonétique International (API) et une mise en contexte de ce phonème dans un mot anglais.

Et nous trouvons aussi d'autres bases de données de différentes langues connus, et d'autres inconnus.

Pour la langue arabe nous n'avons pas découvert une base de données standard, mais nous avons repérer deux dont l'une développé a l'université Houari Boumediene a Bab Ezzouar appelé ALGERIAN ARABIC SPEECH DATABASE (ALGASD) pour le traitement de la parole.

6-2- Base de données NTIMIT :

La base NTIMIT est obtenue par le passage artificiel de la base TIMIT à travers le réseau téléphonique. La bande passante est donc de 330 à 3400Hz avec un signal toujours échantillonné à 16kHz. Les lignes téléphoniques varient. La moitié des appels sont des appels locaux et donc de la même région. Les autres sont des appels longs distances.

Comme TIMIT, la base NTIMIT est divisée en régions et chaque région est divisée en deux sous-ensembles (TRAIN et TEST). La base fournit également une segmentation phonétique.

L'intérêt principal de cette base est l'extraction de caractéristiques de signaux de parole de qualité téléphonique. En effet, les performances des systèmes de reconnaissance de la parole et du locuteur se dégradent lors du traitement de cette base.

6-3- Résumé :

La base ALGASD développé Afin de pouvoir procéder au traitement de la parole arabe en prenant en compte les différents accents de différentes régions du pays.

Et vu la non disponibilité et le manque de moyens pour avoir une base de données audio nous a poussé à construire notre propre base de données destinée à faire la reconnaissance de la langue arabe.

Cette procédure suit les étapes suivantes :

- Introduction des fichiers sons
- Etiquetage manuel des données
- Etiquetage pour la reconnaissance de mots connectés
- Etiquetage pour la reconnaissance de mots continus
- Paramétrisation
- Algorithme des MLP et LVQ
- Initialisation
- Apprentissage
- Généralisation
- Evaluation
- Analyse des résultats

Chapitre 7:Comparaison et discussion des résultats obtenus

7-1- But du projet :

Le but de notre projet est la « Modélisation Neuro-Prédictive pour la Classification Phonétique», notre contribution s'effectuera essentiellement pour la langue arabe et par conséquent nous allons surtout essayer de régler certains problèmes spécifiques tels que les consonnes emphatiques et certaines plosives difficiles à séparer.

Le problème majeure rencontré dans notre projet est le manque de base de donnée pour la langue arabe (corpus) afin d'évaluer et tester les performances de notre approche.

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

Le problème ne réside donc pas dans les machines, il réside dans les hommes, ou plus précisément dans ce que l'on pourrait appeler « les exigences du dialogue homme-machines ».

Pour cela, nous avons décidé de construire une petite base de données pour la langue arabe. Sachant que notre système utilise les fonctions d'édition et de traitement de la parole et qui est structuré en module.

Mais, cela n'était pas chose facile, nous nous sommes confrontés à de nombreux problèmes :

- Incapacité de bien prononcé certain phonème pour la plupart de nos locuteurs.
- Mauvaise qualité de l'environnement et du matériel
- L'étiquetage nécessite une énorme connaissance en phonétique

Notre système reçoit en entrée le signal de parole et renvoie comme résultat un ensemble phonétique. Autour des modules de reconnaissance, nous avons développé des procédures pour l'analyse phonétique et l'affichage graphique ainsi que des modules d'évaluations des performances du système.

L'évaluation nécessite un corpus de phrases équilibrées par plusieurs locuteurs et étiquetées manuellement.

Nous avons aussi intégrer un système d'apprentissage et de décodage acoustico-phonétique basé sur un corpus composé de fichiers de format wav qui doivent être soigneusement étiquetés. Nous avons eu le libre choix des méthodes afin de résoudre le problème du décodage.

Notre choix s'est porté pour la phase de paramétrisation sur une comparaison entre le codage **MFCC** très réputé par les bons résultats qu'il a obtenu jusqu'ici avec une méthode de codage assez récente **NPC** qui apparemment arrive à obtenir de meilleurs résultats que le codage **MFCC**.

Pour la phase de classification, nous avons opté pour une architecture neuronal assez récente qui est le **MLP**, approprié pour la classification de phénomènes, ainsi qui une architecture de classification neuronal supervisé **LVQ** « Learning vectoriel quantification ».

7-2-Présentation de l'interface:

Le logiciel permet une étude du signal de parole grâce à une interface assez simple constituée par :

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

Acquisition et présentation paramétrique :

Le menu est le principal moyen d'interaction avec le logiciel, il présente de nombreuses fonctionnalités :

a- Le module d'acquisition :

Ce module est composé d'un ensemble de fonctions :

- Acquérir de la parole, le système demande le temps d'acquisition qui est limité à 4secondes. Le signal est stocké sur fichier disque contenant les échantillons de taille 16 bits.

- Ecouter une partie ou l'ensemble du signal, le système demande le début et la fin de la zone à restituer dans le signal temporel, l'utilisation choisit le nombre de restitutions.
- Choisir la fréquence d'échantillonnage lors de l'acquisition et la restitution. Cette fréquence est utilisée par plusieurs modules et elle est fixée par défaut à 16 KHz.
- Lire un fichier de parole dans l'un des corpus existants.
- Sauvegarder une partie ou l'ensemble du signal temporel.
- Afficher le signal temporel sur une fenêtre du menu
- Faire un zoom sur le signal temporel.

Le module acoustique :

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

- Calculer et afficher un spectrogramme à bande large avec une fenêtre de Hamming.
- Spectrogramme calculé sur les coefficients LPC qui renforce les fréquences formantiques.
- Spectrogramme à bande étroite pour séparer les harmoniques de la fréquence fondamentale

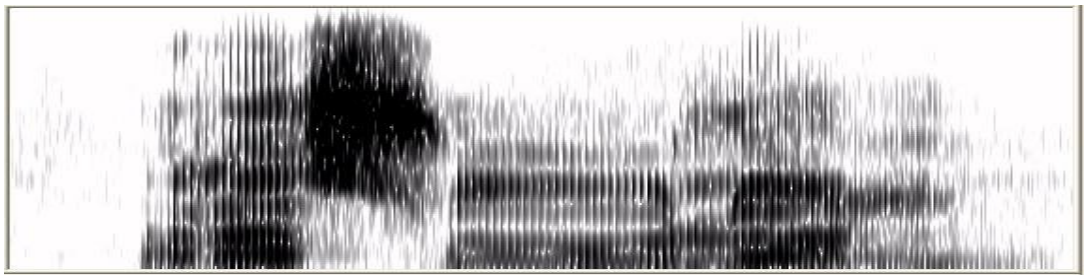


Figure 7.1 : Bande étroite

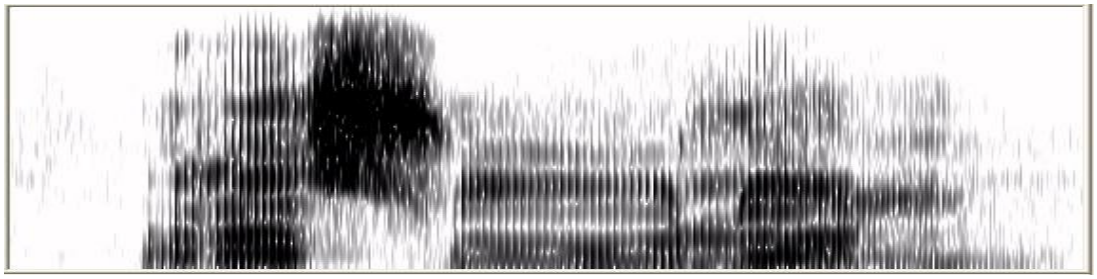


Figure 7.2 : Bande large

- Spectrogramme lissé cepstralement.

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

- Réafficher un spectrogramme déjà calculé en modifiant ses présentations (lissage).
- Calculer l'amplitude du signal
- Calculer le nombre de passage par zéro du signal temporel. Ce paramètre est utilisé pour distinguée entre parole et non parole et permet de différencier entre les sons voisés des sons non voisés.
- Calculer la fréquence fondamentale ou pitch, qui correspond aux vibrations des cordes vocales. La méthode utilisée est celle de l'auto corrélation.

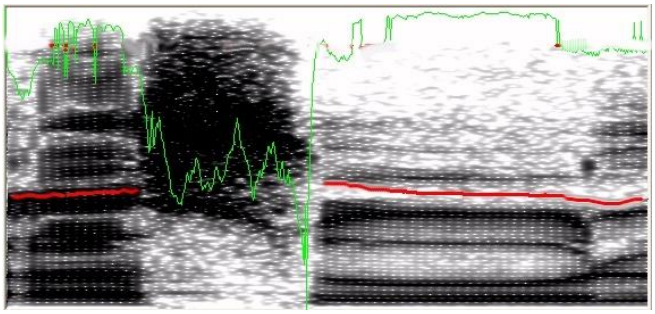


Figure 7.3 : pitch (F0 en rouge et le taux de voisement en vert)

c- le module de décodage :

- Le module de segmentation :

Il consiste à segmenter le signal de parole en grande classes phonétiques en utilisant des algorithmes non contextuels et reposant sur des critères simples. Nous avons retenu trois grandes classes : les voyelles, les plosives et les fricatives.

Les fonctions prévues à ce niveau sont : le calcul, l'affichage et la sauvegarde de la segmentation.

- Le calcul d'indice :

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

- La durée d'un segment
- Le degré de voisement

- La barre d'explosion et ses paramètres
- Les valeurs des formants
- Les transitions formantiques
- Le centre de gravité énergétique
- La limite inférieure du bruit de friction

d- le module d'étiquetage :

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

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

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

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

Notre base de données se divise en deux parties :

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

واحد ,ذئب ,طائرة ,ازرق, بني ,ارض,كبير

Remarque : l'acquisition s'est faite sur une trentaine de locuteurs mais beaucoup d'échantillons ont été rejetés à cause des problèmes cités précédemment. Nous avons fait appel aussi à la Radio local pour la réalisation de corpus.

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

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

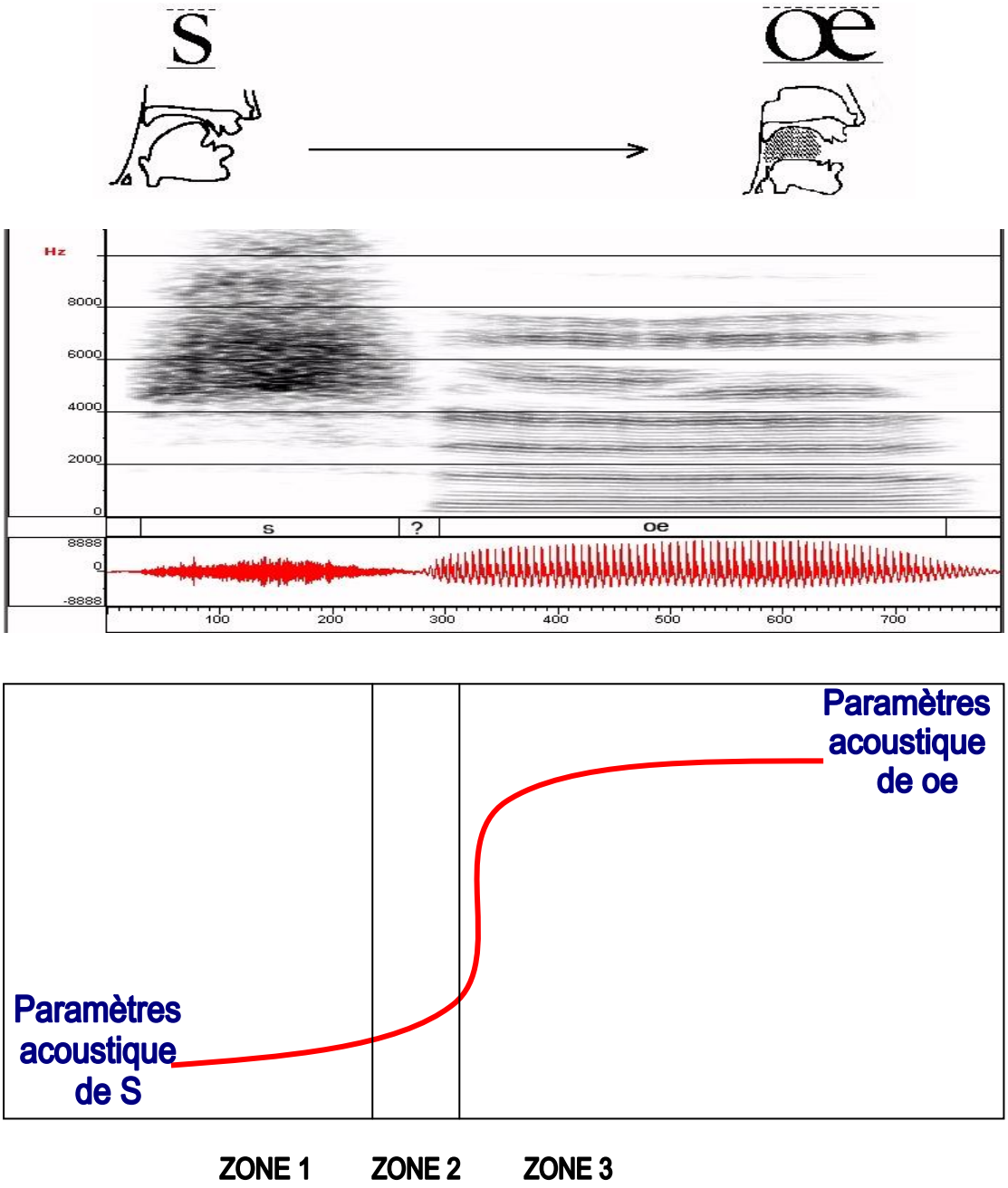


Figure 7.4- montrant le changement des paramètres acoustiques de la transition phonétique

/s/->/oe/ en fonction du changement des paramètres articulatoires

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

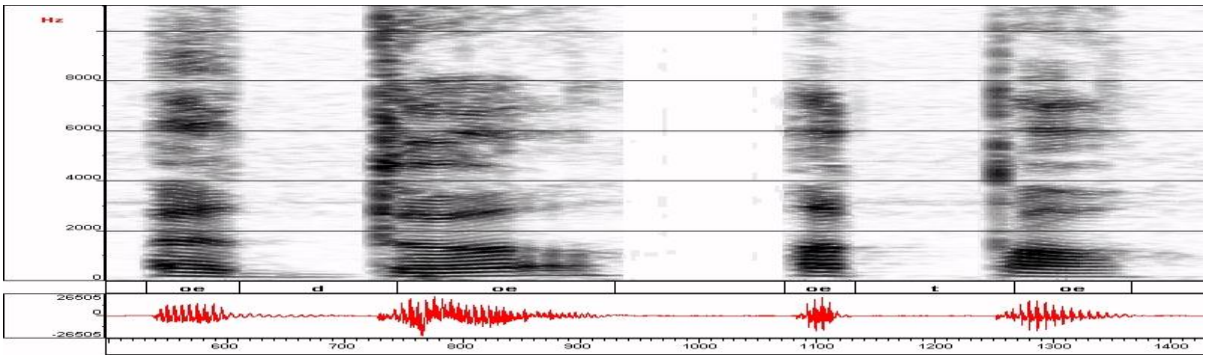


Figure 7.5- montrant le spectrogramme et l'oscillogramme de deux plosives /d/, / t / en contexte /oe/

Dans la figure ci-dessus sont présentées deux plosives en contexte /oe/, il est claire qu'elles sont composées de deux parties, une première partie énergétiquement faible, puis une explosion brutale de bruit (burst) qui a pour le cas de /d/, /t/ les mêmes caractéristiques vues leurs point d'articulation commun (alvéolaire) où la différence entre eux consiste dans la partie énergétiquement faible de /d/ où on perçoit des harmoniques (voisement) de faible énergie qui donne son caractère sonore à la consonne /d / contrairement au /t/ qui est une consonne sourde.

Il est clair qu'il est impossible de percevoir ces phonèmes en n'entendant que leurs débuts comme c'était le cas pour des voyelles et des fricatives.

Ce n'est qu'à la fin du phonème, et au début du suivant que l'on arrivait à percevoir ces consonnes, et le voisement qui apparaît dans /d/ est plutôt perçu

comme un /oe/, donc il est claire qu'un seul vecteur acoustique, surtout pris dans la partie silence de ces plosives, même pris dans le burst, on a vu que le burst de /d/ et celui de /t/ avez les même caractéristiques donc probablement le même vecteur acoustique donc impossible de les séparer grâce à un seule vecteur.

Même le voisement qui apparaît dans la partie silence du /d/ apparaît aussi dans toutes les consonnes sonores, donc pour arriver à reconnaître une plosive, il est essentiel d'étudier à la fois sa partie silence et sa partie burst, d'où l'utilisation d'un classificateur à mémoire ne fait pas de doute. Mais il est essentiel de ne lancer l'apprentissage qu'après avoir chargé (mémorisé) toute la plosive (afin d'avoir accès à toute les informations nécessaires sur la plosive pour la reconnaître).

Pour une fricative ou une voyelle, on peut lancer l'apprentissage après en avoir chargé qu'une partie (supposée suffisante pour les reconnaître). Pour notre étiquetage, nous avons pris la contrainte de ne représenter un phonème qu'avec un seul caractère, donc nous allons changer les symboles pris précédemment :

Voyelles :

Contexte non emphatique أَ	Contexte emphatique أَ ـً ـُ ـٌ	Contexte non emphatique أُ	Contexte emphatique أُ	Contexte non emphatique إِ	Contexte emphatique إِ
E	**a**	**u**	**O**	**i**	**E**

Tableau 7-1 : les voyelles

***Consonnes* :**

ب	د ض	ت ط	ك	ق	ذ ظ	ز	ج	ع	غ	ف	ث	س ص	ش	ح	خ	ه	م	ن	و	ي	ل	ر
B	**D**	**T**	**k**	**q**	**v**	**z**	**J**	**3**	**r**	**f**	**C**	**s**	**h**	**g**	**x**	**&**	**m**	**n**	**W**	**y**	**L**	**p**

Tableau 7-2 : les consonnes

e- module du Prétraitement:

Afin d'utiliser les fichiers wav. qui vont nous servir comme base d'apprentissage où base de test, on va d'abord leurs faire subir une série de prétraitements qui vise à préparer les entrées de notre réseau MLP. Après le chargement d'un fichier (22 kHz, 16bits, mono), les données sont sous forme d'un signal temporel avec des valeurs qui varient entre – 32768 et + 32768.

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

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

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

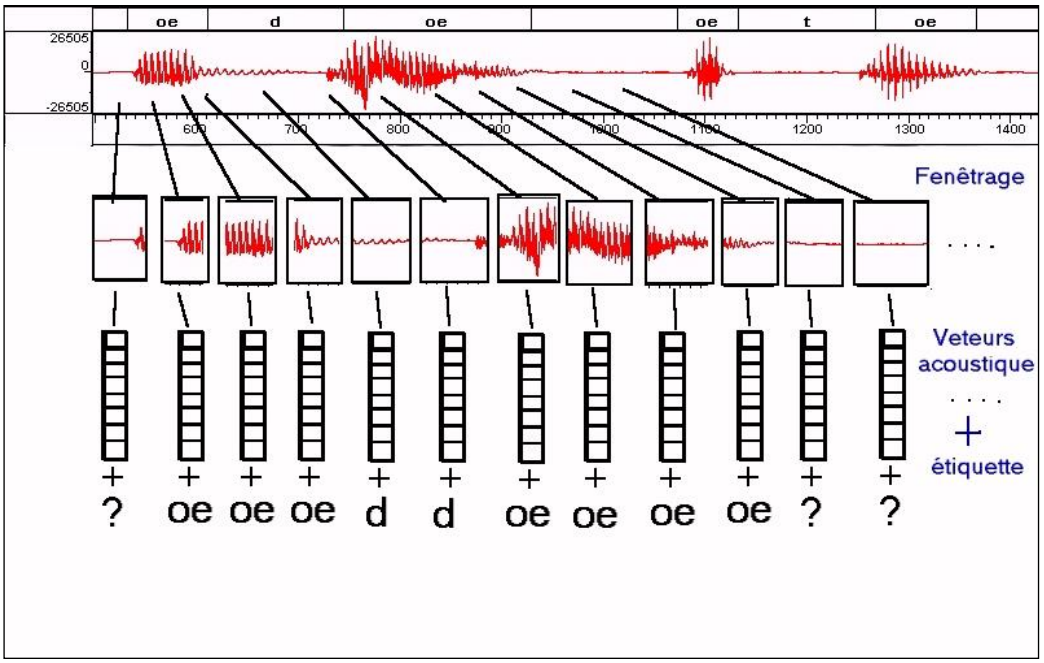


Figure 7.6- montrant l'extraction de vecteurs acoustiques à partir d'un signal avec étiquette

7-3- Mise en œuvre du réseau MLP:

Pour notre test, nous allons utiliser un MLP avec 3 couches :

- couche entrée (nombre de neurones= nombre de coefficients)
- couche cachée (nombre de neurones = nombre de coefficients)
- couche de sortie (nombre de neurones = nombre de classes phonétiques).

7-3-1-Apprentissage:

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

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

Ces entrées seront intégrées dans l'algorithme de la rétropropagation du gradient citée au chapitre 5.

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

7-3-2- Adaptation de l'étiquetage:

A- <u>Base des voyelles :</u>

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

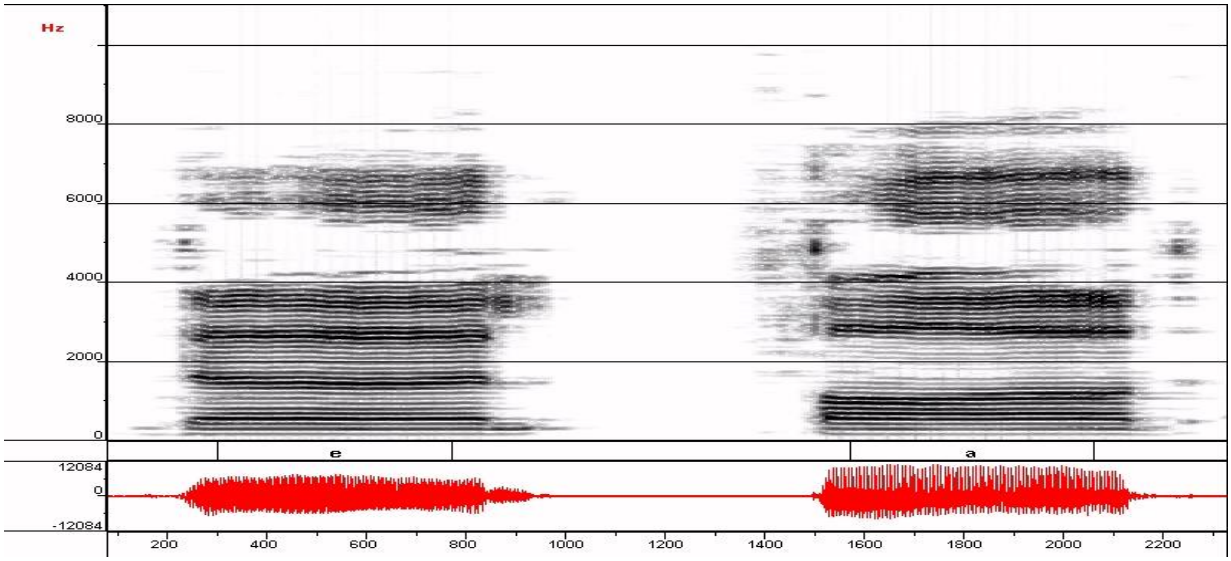


Figure 7.7- montrant un exemple d'étiquetage pour les voyelles /e/, /a/

B- Base de mots:

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

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

7-3-3- Généralisation:

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

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

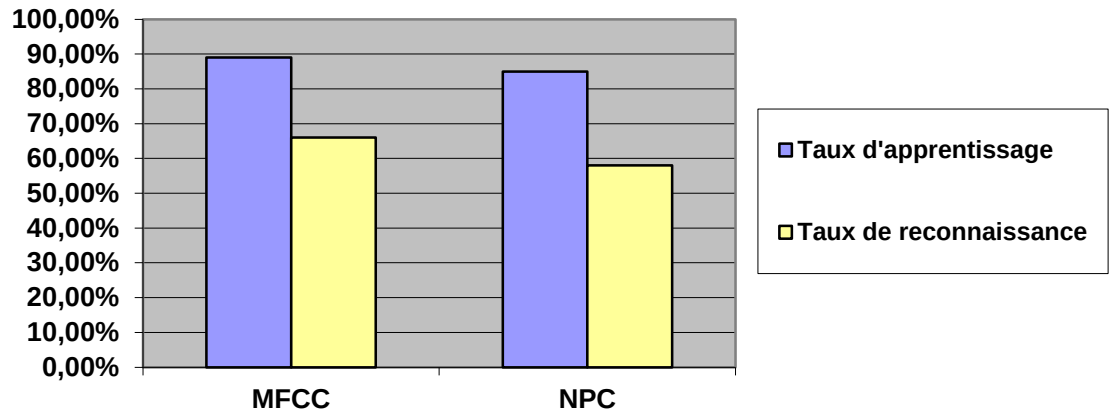
Ensuite, on prend soin de regrouper les segments similaires en notant leurs durées, pour des segments de durée inférieur à un seuil (exemple 50 ms), ce segment est ignoré (supposé bruit), pour des segments dépassant un seuil (exemple 150 ms) on rajoutera la marque / : / ce qui signifie pour une voyelle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

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

7-3-4- Test et résultats en reconnaissance phonétique:

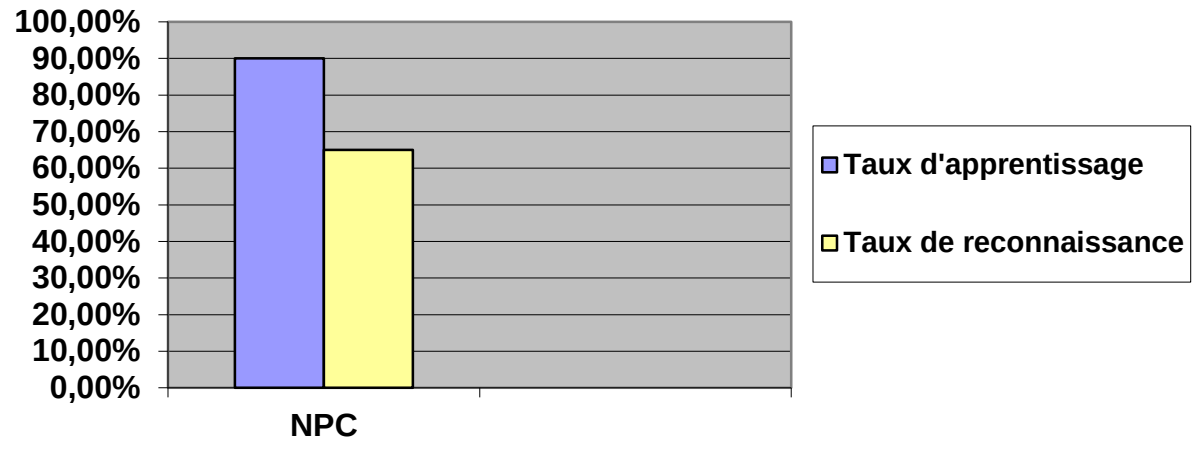
Afin de tester les performances de notre système, nous allons l'appliquer sur notre base de données. Pour cela, nous allons extraire une suite de vecteurs de 16 coefficients (MFCC/ NPC) de nos fichiers wav, et lancer le mécanisme (apprentissage/ test) comme évoqué précédemment.

Base de voyelles avec un MLP



Pour ce test, on a utilisé un apprentissage sur un MLP simple, après 1000 itérations en utilisant 12 coefficients NPC, et 16 coefficients MFCC. Les résultats montrent une supériorité des coefficients MFCC que se soit en taux d'apprentissage (89% MFCC, 85% NPC) ou en taux de reconnaissance (66% MFCC, 58% NPC). Le codage MFCC étant une méthode fréquentielle arrive à mieux reconnaître les voyelles que le codage NPC. Notre corpus étant bruité, les coefficients MFCC montrent une meilleure résistance au bruit que les coefficients NPC.

Base de voyelles avec un LVQ-NPC



On a essayé le même test avec un réseau LVQ. On constate une très bonne amélioration, en taux d'apprentissage (90% NPC) qu'en taux de reconnaissance (65% NPC).

Base de mot avec un MLP

Tableau 7-3 : Taux de reconnaissance sur les consonnes de la base des mots avec un codage MFCC

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

Tableau 7-4 : Taux de reconnaissance sur les consonnes de la base des mots avec un codage NPC

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

***Commentaire* :**

On remarque une large avance en taux de reconnaissance pour le codage NPC en ce qui concerne les plosives, on dirait que le codage NPC arrive à mieux prendre en compte les phénomènes rapide dans le temps, contrairement au codage MFCC qui arrivent plutôt à prendre en compte les consonnes de longue durée comme les nasale et les sonnâtes.

Tableau 7-5:Taux de reconnaissance sur les consonnes de la base des mots avec un codage LVQ-NPC

ب	د ض	ك	ق	ع	س ص	م	ن	ل	ر
74%	68%	64%	73%	45%	99.5%	74%	64%	66%	70%

Commentaire :

Là aussi, on peut remarquer certainement d'autres améliorations dans les scores de reconnaissances par rapport a ce qui a été cité plus haut. Mais les occlusifs restes encore loin de nos aspirations

Chapitre 8:Conclusion et Perspective

CONCLUSION

Les recherches effectuées dans le domaine de la reconnaissance automatique de la parole permettent d'envisager un éventail toujours plus large d'applications industrielles ou grand public.

Cependant, la compréhension des mécanismes de production et de reconnaissance de la parole pour l'homme ne suffit pas en elle-même pour élaborer effectivement les dites applications.

Les conditions de laboratoire qui ont prévalues lors de l'enregistrement des premiers corpus de parole utilisés à des fins de recherches sont en effet très différentes des conditions réelles que l'on rencontre généralement dans la vie.

La reconnaissance automatique de la parole pose un certain nombre de problèmes, les caractéristiques phonétiques et linguistiques de la langue également impliquées dans le processus.

Le travail de notre thèse consiste à présenter une contribution à la reconnaissance automatique de la langue arabe.

Nous avons d'abord effectué une étude phonétique de la langue basée essentiellement sur l'examen de spectrogrammes de mots et des phonèmes en tenant compte des différents contextes de production.

Cette étude nous a permis aussi de définir les caractéristiques acoustiques nécessaires au système de reconnaissance.

Ensuite, nous avons réalisé un système de décodage Acoustico-Phonétique qui reçoit en entrée le signal de parole et retourne comme résultat un treillis de phonèmes.

Les principales étapes du système utilisée dans notre travail sont : la segmentation du signal de parole en classe phonétiques (voyelles, plosives, fricatives et sonnantes) ; l'extraction des indices phonétiques utilisées en reconnaissance ; l'étiquetage des segments manuellement et qui nous a pris énormément de notre temps et aussi beaucoup de connaissance en phonétique.

Cette phase d'analyse du signal de parole montre la présence de plusieurs informations 'linguistiques, paralinguistiques et extralinguistiques) qui Entraînent une grande variabilité du signal qui nous ont permis d'extraire les caractéristiques appelés coefficients LPC et MFCC.

Mais les phénomènes non-linéaires présents lors de la production de la parole ont pour effet de limiter la validité de codage LPC.

Par ailleurs, les différentes expériences réalisées au niveau du signal de parole confirment l'intérêt d'une modélisation non-linéaire pour l'extraction des caractéristiques.

Pour mener à bien cette étude, différentes architecture connexionnistes ont été proposées.

L'utilisation de modèles connexionnistes nous à permis de mettre au point, grâce au mécanisme d »apprentissage, des systèmes qui sont immédiatement adaptés à différents condition de bruit.

Cette architecture en première phase, basée sur des perceptrons multicouches MLP, à prouvé être de bonne qualité mais l'étape de segmentation s'est révélée être de moindre qualité.

Ceci, nous à poussé à étudiés d'autres possibilités et dans ce cadre nous avons proposé un modèle qui consiste à optimiser simultanément un classifieur, en l'occurrence le classifieur à prototype LVQ, et du modèle NPC. Le résultat est nommé LVQ-NPC.

Elle justifie l'importance de l'utilisation de méthodes discriminantes en extraction de caractéristiques.

Cette approche s'avère intéressante pour l'amélioration des systèmes de reconnaissance dans le cadre du traitement non-linéaire.

PERSPECTIVE

Des perspectives propre a chaque sujet abordé dans cette thèse peuvent être envisagées. Dans un premier temps, les pistes de recherche concernant le décodage Acoustico-Phonétique sont nombreuses.

On peut aussi enrichir les apprentissages par un élargissement des corpus d'entrainement, acoustique et de langage.

Dans le même cadre, La classification peut être revue par d'autres algorithmes connexionnistes ou bien d'autres approches telles que la logique floue, etc.
L'objectif dans tous ceci est d'améliorer les scores de reconnaissance et contribué a l'avancée technologique

Listes des figures

Liste des tableaux

Bibliographies

[1]-**Steve Lawrence** «The Gamma MLP for Speech Phoneme recognition 2001»

[2]-Laurent Buniet « Traitement automatique de la parole en milieu bruité : étude de Modèles connexionnistes statiques et dynamiques 2000 »

[3]-Stromboni Jean Paul « un cours d'introduction au traitement du signal pour l'ordinateur multimédia 1995 »

[4]-Nguyen Quoc Cuong «Reconnaissance de la parole en langue Vietnamienne 2000»

[5]-Bruno GAS « Méthodes neuronales pour l'extraction de caractéristiques non linéaire et discriminantes : application aux signaux de parole. 23 Novembre 2005»

[6]-Kenneth N. Stevens « On the quantal nature of speech Journal of Phonetics (1989) 17, 3-45 Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge MA 02139, U.S.A»

[7]-Rodolphe Battault « La reconnaissance vocale »

[8]-Willy Serniclaes « Etude expérimentale de la perception du trait de voisement des Occlusives du français »

[9]-Stevens, Liljencrants & Lindblom « Theories of Vowel Systems: Quantal Theory and Adaptive Dispersion L 105/205 – Phonetics Scarborough 2005»

[10]-Neagu Adrian « Représentations phonétiques et identification des syllabes occlusive-voyelle en français 30 avril 1998 Institut de la Communication Parlée»

[11]- D.E. Kouloughli « Grammaire de l'arabe d'aujourd'hui presses Pocket 1994»

[12]-Benyettou. Abdelkader « Analyse et paramétrisation des phonèmes arabe en vue de la reconnaissance automatique de la parole 1986» université d'Oran

[13]-Tahar Saidane, Mounir Zrigui et Mohamed Ben Ahmed
«La transcription orthographique phonétique de la langue arabe 2004»Fès, Maroc

[14]-Kaïs Ouni et Noureddine Ellouze « Sur l'évaluation du second formant F'2 par une technique d'estimation spectrale basée sur une modélisation du filtrage auditif, journées d'étude sur la parole, Nancy 24-27 juin 2002»

[15]-Kenneth N. Stevens « The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data 1972 »

[16]-Chetouane Mohamed « codage neuro prédictif pour l'extraction de caractéristiques de signaux de parole, Paris, 2004 »

[17]-Louni Abderrahmane et Benyettou Abdelkader « un codage neuro prédictif pour l'extraction des traits distinctifs appliqué à la reconnaissance des phonèmes arabe, April, 2008, Oran, Algeria »

[18]-Amrouche A., Debieche M., Taleb-Ahmed « An efficient speech recognition system in adverse conditions using the nonparametric regression. Engineering application of artificial intelligence, 23(1), pp85-94.2010 »

[19]-Bendahmane Abderrahmane « reconnaissance de la parole par distance DTW exemple d'application pour la reconnaissance de chiffre isolés dans la langue arabe, laboratoire SIMPA, Oran, Algeria »

[20]-Amrouche A., Taleb-Ahmed A., Rouvaen J.M., Yagoub M., « improvement of the speech recognition in voisy environments using a nonparametric regression International journal of parallel, emergent and distributed system,vol 34,issue 1,pp.49-67…,2009, valencienne »

[21]-Odile Mella «contribution à l'identification automatique du locuteur sur des critères acoustiques et phonétiques 1993, INRIA Lorraine, Nancy »

[22]-Bladon A., Fant.G «A Two-formant model and the cardinal vowel, » STL-QPRS, no 1, pp.1-8, 1978.

[23]-Paliwal K.K., Ainsworth W.A., Lindsay D., «A study of two-formant models of vowel identification», speech communication...Vol, 2.no.4, December, pp.305-313

[24]-Carlson R., Ganstrôm B., Pauli S., «Perceptive evaluation of segmental cues» STL/QPSR1/1972, Stockholm, Roy.Inst.Technol. 18-24

[25]-Lieberman P., «Phonetic features and physiology: a reappraisal»J.of phonetics 4, 91-112

[26]-Schwartz J.L., Bôe L.J., Vallée N., Abry C., «The dispersion-focalization theory of vowel systems» 1997, France

[27]-Hermansky H., Morgan N., Hirsch H.G., «recognition of speech in additive and convolutional noise based on RASTA spectral processing. »IN: International conference on acoustics, speech, and signal processing, pp83-86, 1993

[28]-Itahashi S., Yokoyama S., «A formant extraction method utilizing and scale and equal loudness contour» STL-QPRS, 104, pp 17-29, 1978

[29]-Hermansky H., «perceptual linear predictive (PLP) analysis of speech», J.acoust.soc.vol 87, nà4, April 1990, pp.1738-1752

[30]-Lieberman A.M,, Cooper F.,Delathe P.C.,Gerstman L.J., «An experimental study of the acoustical determinants of vowels colour»,world8:195-210,1952

[31]-Bladon A., «two-formant models of vowels perception: shortcomings and enhancements», speech communication, vol.2, December, pp.305-313

[32]-Garofolo J.S., Fischer M., Fiscus G.J., Pallett S.D., Dahlgren N.L., «Acoustic-phonetic continuous speech corpus»Gaithersburg, 1993

[33]-Kammoun M.A., Ben-Hamida A., «Synthèse de la parole par corpus: utilization du modèle BI-Grams», Hammamet, 2008, Tunisia

[**34]-Liberman,p** «phonetic features and physiology reappraisal»J.of phonetics,91-112.1976.

[35]-Stevens, K.N and House, A.S«An acoustical theory of vowel production and some of its implications»J.speech Hrng.Res.4, 303-320, also in lehiste (1967).1961.

[36]-Stevens, K.N. «Perception of phonetic segments: evidence from phonology, acoustics and psychoacoustics» in the perception of language, D.L.Horton and J.J.Henkins eds., ch.Merrill publ.cy: colombus, 216-235.1971.

[37]-Stevens.K.N «The quantal nature of speech; evidence from Articulatory-acoustic data» in human communication: A unified view, E.E.David and P.B.Denes eds, MC Graw-Hill: New York.1973.

[38]-Jakobson, R.and Hall «phonétique et phonologie» In essays de linguistique générale, Jakobson(1963), N.Ruwet Trad., Ed.de minuit: paris;103-149.1956.

[39]-Fant.G «Acoustic Theory of Speech Production Mouton» S-Gravenhagen.1960.

[40]-Carlson, R.Granstrôm, B and Pauli, S. «perceptive evaluation of segmental cues»STL/QPSR 1/1972,Stockhôlm,Roy.Inst.Technol.,18-24.1972.

[41]-Stevens, K.N «On the quantal nature of speech» Journal of phonetics 17, 3-46.1989

[42]-Stevens,K.N «The quantal nature of speech, Evidence from Articulary-Acoustic Data» In P.B Denes et E.E.jr David (eds),Human communication, A unified view,51-66.NY :McGraw-Hill.1972.

[43]-Jacobson, R., G.Fant and M.Hall «Preliminairies to speech Analysis» Cambridge, MA: MIT Press.1952

[44]-Alexander K.S: probability inequality for empirical processes and a law of the iterated logarithm-Ann.Prob.Vol 12, pp 1041-1067, 1984