République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Akli Mohand Oulhadj – Bouira

Faculté des sciences et des sciences appliquées

Département de Génie Electrique

Mémoire de Master

Option : Télécommunications

Thème :

Systèmes de Reconnaissance Automatique de la Parole

Réalisé Par :

- LARAK Ibtissam
- AFFOUN Saada

Dirigé par :

- OUAHABI Abdeldjalil.

Soutenue le : 23/09/2017 devant le jury composé de :

- SAOUDI	Président de jury
- SAIDI	Examinateur
- CHALABI	Examinateur

Année universitaire : 2016- 2017

Remerciements

Tout d'abord, nous tenons à remercier Allah, le clément et le miséricordieux de nous avoir donné la force et le courage de mener à bien ce travail.

Un grand merci à notre encadreur Mr Abdeldjalil OUAHABI pour son soutien et sa disponibilité.

Nous voudrions aussi remercier tous les professeurs qui ont contribué à notre formation.

Merci enfin à tous ceux qui m'ont soutenu dans ce travail et qui n'ont pas été cités ici.

Résumé

Le domaine de la reconnaissance de la parole est devenu l'un des thèmes d'application les plus fertiles. Il trouve des applications dans notre vie quotidienne, soit en téléphonie, soit en contrôle vocal de systèmes ou encore dans des applications spécifiques au handicap. Cet engouement a été possible grâce aux ressources actuelles de traitement numérique du signal. Un système de reconnaissance automatique de la parole doit d'abord analyser efficacement le signal de parole puis en extraire des caractéristiques en vue de les exploiter par un algorithme de reconnaissance/classification. Ce travail donne un bref aperçu du système de reconnaissance de la parole et ses différentes phases telles que l'analyse temps-fréquence et l'extraction des caractéristiques.

Mots clés : Parole, Analyse du signal de parole, Extraction des caractéristiques, Reconnaissance automatique de la parole.

Abstract

The field of speech recognition has become one of the most fertile application themes. It finds applications in our daily lives, either in telephony, voice control systems or in specific applications for disability. This enthusiasm has been possible using current resources of digital signal processing. An automatic speech recognition system must first effectively analyze the speech signal and then extract characteristics for use by a recognition / classification algorithm. This work gives a brief overview of the speech recognition system and its different phases such as time-frequency analysis and feature extraction.

Key words : Speech, Speech signal analysis, Features extraction, Automatic speech recognition.

SOMMAIRE

CHAPITRE I GENERALITES SUR LA PAROLE

CHAPITRE II : ANALYSE DU SIGNAL DE PAROLE

CHAPITRE III : EXTRACTION DES PARAMETRES

CHAPITRE IV LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Liste des figures

Liste des abréviations

RAP: Reconnaissance Automatique Parole

TF: Transformée de Fourier

TFD: Transformée de Fourier Discrète

FFT: Transformée de Fourier Rapide (Fast Fourier Transform)

TFCT: Transformée de Fourier à court terme

WT: Transformée en Ondelette (Wavelet Transform)

MFCC: *Mel-frequency cepstral coefficients*

LPC: Codage Prédictif Linéaire

DCT: Ttransformée en Cosinus Discrète

PLP: Perceptual Linear Prediction

DTW: Dynamic Time Warping

HMM: Hidden Markov Models

DAP: Décodage Acoustico Phonétique

GMM: Gaussian Mixture Models

ANN: Artificial Neural Network

Introduction générale

Introduction générale

Grace à la parole nous pouvons donner une expression à nos pensées, et l'utiliser pour exprimer des idées, des opinions, des sentiments. De nos jours, nous l'utilisons non seulement pour communiquer avec les êtres humains, mais aussi avec des machines.

Un système de Reconnaissance Automatique de la Parole (RAP) est un système qui a la capacité de détecter la parole et de l'analyser dans le but de générer une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. Cette analyse se fonde sur l'extraction des paramètres descriptifs de la parole.

Notre principal objectif dans ce travail est de comprendre les bases de la RAP, Plus précisément, nous nous intéressons, a l'analyse du signal parole, et sa représentation dans le domaine temps-fréquence par différentes méthodes, utilisant la programmation matlab, dont on a essayé de donner le mieux à travers ce travail. Ce mémoire se compose de quatre chapitres :

- Le premier c'est la production de parole et sa perception, ainsi que les caractéristiques générale de signal de la parole.
- Le deuxième chapitre comporte l'étude des outils de traitement du signal vocal, et les résultats obtenus à partir de la programmation matlab, prendront la base de données NOIZEUS.
- Le troisième chapitre représente l'extraction des paramètres utilisés dans le traitement de signal de la parole.
- Le quatrième chapitre c'est des généralités sur les méthodes utilisées dans la RAP.

Pour finir, nous présentons les conclusions et les perspectives de ce travail.

CHAPITRE I

GENERALITES SUR LA PAROLE

I.1 Introduction :

La parole est l'un des principaux moyens de communication entre êtres humains, sa simplicité en fait d'ailleurs le moyen de communication le plus populaire dans la société humaine (il est plus facile de parler à quelqu'un que de lui écrire ou de lui faire un schéma). Néanmoins, cette simplicité (pour l'être humain) renferme un traitement très complexe fait par notre cerveau, de la production de la parole jusqu'à sa perception et sa compréhension, ce qui rend la parole difficilement automatisable pour une machine [1].

Dans ce chapitre nous présenterons la production de la parole et son traitement, ainsi que l'appareil auditif de l'être humain et la perception de la parole.

I.2 Production et perception de la parole :

I.2.1 Description de l'appareil phonatoire :

La parole est un phénomène acoustique qui se distingue des autres sons par des caractéristiques liées aux mécanismes de sa production par l'appareil phonatoire. Ce dernier fait intervenir divers éléments : l'air, comme source d'énergie ; les cordes vocales, comme principal organe vibratoire ; la langue et les lèvres, comme organes vibratoires accessoires ; les cavités buccale et nasale, comme caisses de résonance ; et le système nerveux qui contrôle l'ensemble [2].

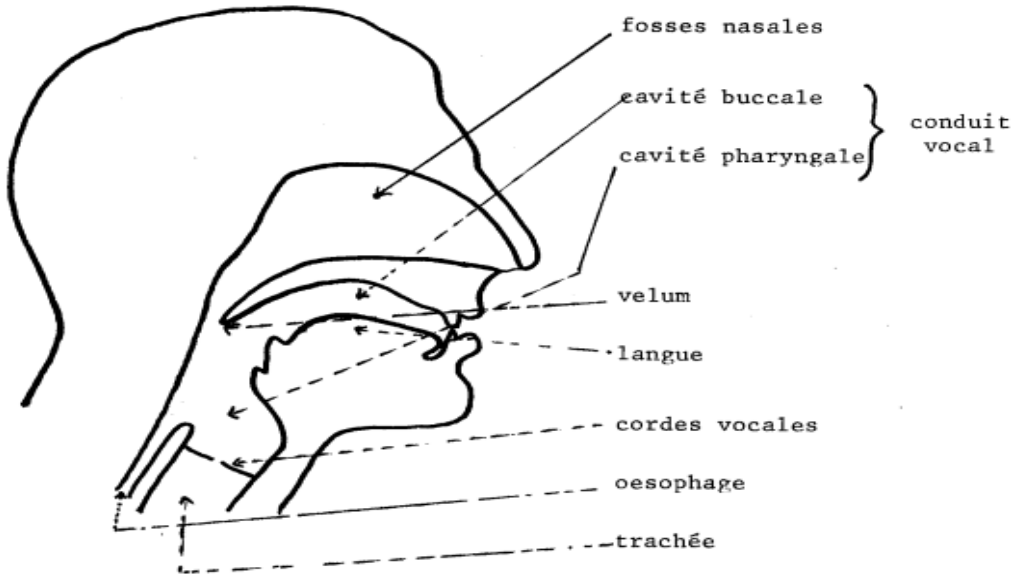


Fig I.1 l'appareil phonatoire [3].

Le figure I.1 représente schéma générale de l'appareil phonatoire humaine, On y voit, pour l'essentiel :

- **Les cordes vocales** : membranes musculeuses qui coiffent l'orifice de la trachée (larynx). Leur ouverture (glotte) présente une surface variable.
- **Le conduit vocal** : cavité de forme variable, allant des cordes vocales aux lèvres.sa longueur est d'environ 17.5 cm chez l'adulte. On y distingue souvent une cavité buccale et une cavité pharyngale.
- **La cavité nasale** : qui peut être mise en dérivation sur le conduit vocal, par l'abaissement du velum (comme c'est le cas sur la figure I.1).
- **Les articulateurs** : (langue, lèvres, mâchoires, …) qui permettent de modifier la forme du conduit vocal [3].

I.2.1.1 La production de la parole :

Produire de la parole, c'est tout d'abord expulser de l'air des poumons ; ce flux d'air va être mis en forme par la multitude de cavités, constrictions et orifices qui jalonneront son chemin depuis les poumons jusqu'aux lèvres et au nez. C'est au niveau du larynx que s'opère la première mise en forme de ce flux d'air [4].

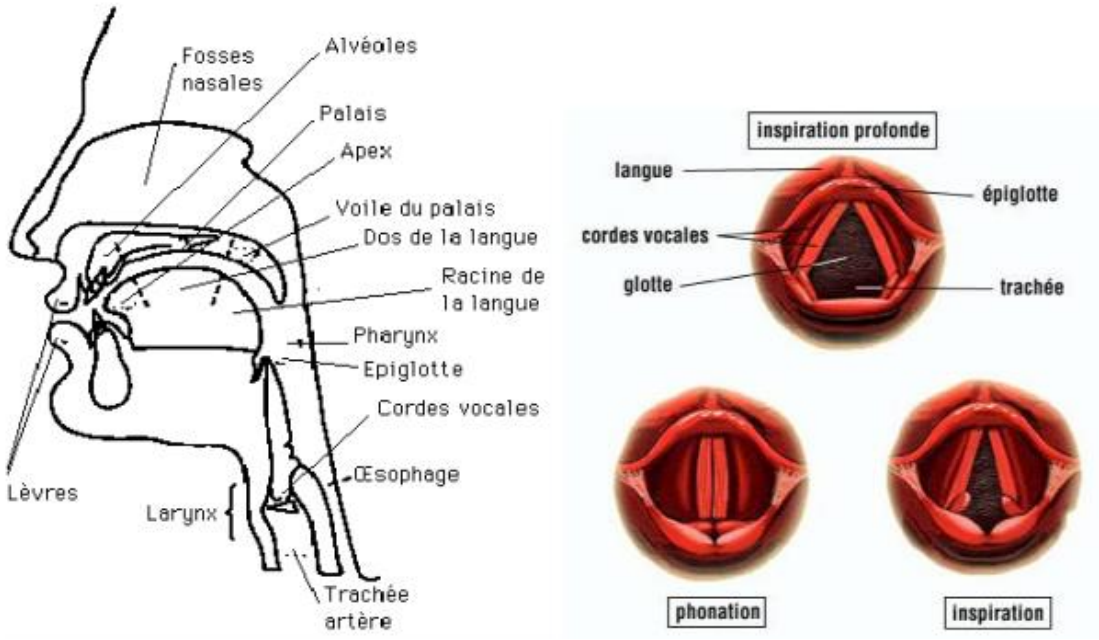


Fig I.2 Organes de l'appareil phonatoire [2].

Le signal de parole est produit par l'écoulement de l'air à travers le larynx et le conduit vocal (figure I.2).

L'appareil respiratoire est considéré comme un générateur d'air qui force l'écoulement de ce dernier à travers le larynx. Cette force est produite, pendant la phase d'expiration, par un effort musculaire au niveau de l'abdomen qui refoule le diaphragme et provoque la compression de l'air dans les poumons et la trachée artère.

Au sommet de celle-ci se trouve le larynx, principal organe de la phonation, sur lequel se trouvent les cordes vocales qui, sous l'effet de la pression exercée par l'air, vont vibrer dans un régime quasi-périodique. La fréquence de vibration des cordes vocales est appelée fréquence laryngienne ou fréquence fondamentale de l'onde acoustique. L'ouverture triangulaire dessinée par les cordes vocales est appelée la glotte (figure I.2). Des muscles intrinsèques et extrinsèques au larynx agissent sur ce dernier pour régler l'ouverture de la glotte et la tension des cordes vocales. Ces réglages permettent de varier le flux d'air et la fréquence de vibration des cordes vocales, et par conséquent, l'intensité et la hauteur des sons [2].

Le flux d'air est ensuite modulé en fréquence en passant dans une série d'articulateurs du conduit vocal. Différents sons sont ainsi produits en fonction de la configuration géométrique de ces articulateurs. Parmi ces articulateurs, nous pouvons distinguer la langue, la mandibule, les lèvres et le voile du palais. Ce dernier permet, selon sa position, le couplage ou non du conduit vocal avec les fosses nasales.

Le mode de fonctionnement précédent décrit la production des sons voisés (sonores). Les sons non-voisés (sourds) sont quant à eux produits quand l'air passe librement à travers la glotte sans vibration des cordes vocales [2].

Ainsi, les fricatives non-voisées sont produites quand l'air passe à travers la glotte et à travers une constriction du conduit vocal, pour produire un flux d'air turbulent. Les fricatives voisées combinent en revanche des composantes turbulente et périodique : les cordes vocales vibrent dans ce cas.

Pour produire les occlusives, une forte pression d'air est créée en amont d'une occlusion en un point du conduit vocal (palais, dents ou lèvres). Cette occlusion est ensuite brusquement relâchée, causant un flux d'air impulsionnel et un son transitoire. On distingue aussi les occlusives voisées des occlusives non-voisées [2].

I.2.2 Audition et perception de la parole :

L'appareil phonatoire, émetteur d'informations, ne serait d'aucune utilité si l'information générée ne pouvait être captée et analysée par un récepteur. Parmi tous les récepteurs existants, l'homme a acquis la capacité de découvrir le sens caché sous les sons produits par son interlocuteur. Nous allons maintenant présenter l'anatomie de l'oreille, organe récepteur de l'information sonore, et les capacités de perception qui caractérisent cet organe lorsqu'il est en parfait état et n'a subi aucune atteinte venue amoindrir ses capacités intrinsèques [5].

I.2.2.1 Structure du système auditif

L'oreille est divisée en trois parties distinctes (figure I.3), cette division se faisant en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan. Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement [5].

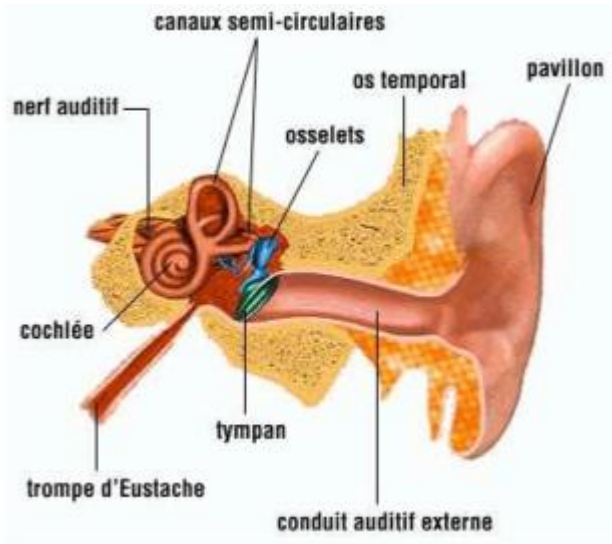


Fig I.3 Organes de l'appareil auditif humain [2].

I.2.2.2 Acoustique de l'audition

La psychoacoustique a pour objet l'étude des relations quantitatives entre les stimulus acoustiques et les réponses du système auditif de l'être humain [2].

Les résultats les plus marquants de cette science sont les suivants :

-Échelle d'intensité : Le système auditif ne présente pas une sensibilité à l'intensité sonore identique à toutes les fréquences. En effet, des sons d'intensité sonore égale n'auront pas la même sonie (l'intensité perçue) selon qu'ils soient de haute fréquence 10kHz, de basse fréquence 100Hz, ou de fréquence moyenne 1kHz. Ainsi, si ces trois sons ont une même intensité de 40dB, les sons de fréquence 100Hz et 10kHz seront plus faiblement perçus que le son de fréquence 1kHz. Les courbes d'isosonie représentent les niveaux d'intensité sonore générant une perception auditive d'égale intensité en fonction de la fréquence du son stimulant (figure I.4) [2].

-Échelle de hauteur : La tonie (la hauteur) d'un son est la qualification subjective de sa fréquence. Des études psychoacoustiques ont en effet montré que la perception humaine du contenu fréquentiel des sons ne suit pas une échelle linéaire mais une échelle fréquentielle de Mel. Cette échelle est approximativement linéaire de 20 Hz jusqu'à 1kHz et logarithmique de 1kHz jusqu'à 20kHz. La fréquence de Mel est calculée à partir de la fréquence acoustique comme suit :

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \quad \text{(I.1)}$$

L'unité de la tonie d'un son dans cette échelle est le Mel [2].

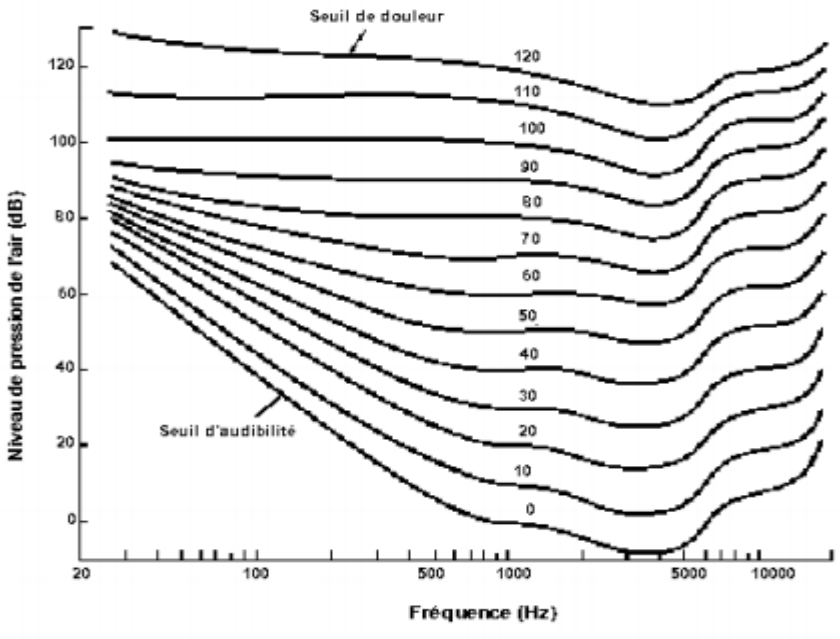


Fig I.4 Courbes d'isosonie [2].

I.2.2.3 Les effets de masque et les bandes critiques :

On parle de masquage lorsqu'un son devient totalement ou partiellement inaudible en présence d'un autre son. Il existe plusieurs sortes d'effet de masque :

-Le masquage simultané (fréquentiel) : il peut être total ou partiel et dépend des intensités et des fréquences des deux sons : masqué et masquant.

-Le masquage temporel proactif : il apparaît lorsqu'un son masqué est émis après un son masquant.

-Le masquage temporel rétroactif : il apparaît lorsqu'un son masqué est émis avant un son masquant [2].

Des observations psychoacoustiques ont par exemple montré la déformation du seuil d'audibilité d'un son de fréquence 2kHz en présence d'un son de fréquence 1kHz et d'intensité 80dB. Le son de 2kHz ne peut être perçu que si son intensité est supérieure à 50dB alors que, comme on peut le voir dans la figure I.4, son seuil d'audibilité est proche de 0dB lorsqu'il est émis seul sans un son masquant. L'étendue fréquentielle des effets de masque prouve qu'un son excite le système auditif sur une bande de fréquences plus large que son spectre fréquentiel. Ainsi, grâce à la décomposition fréquentielle opérant initialement au niveau de la membrane basilaire, le système auditif se comporte comme un banc de filtres passe-bande. Les largeurs de bande de ces filtres, appelée bandes critiques, ont été mis en évidence expérimentalement grâce au phénomène de masquage d'un

son pur (de fréquence f) par un bruit de largeur de bande fréquentielle variable. On appelle Bark la largeur d'une bande critique.

Une bande critique correspond aussi à l'écart fréquentiel minimal pour que deux harmoniques d'un son soient discriminés perceptivement [2].

I.3 Les caractéristiques de la parole :

I.3.1 Le signal de la parole :

La parole peut être vue comme un signal, la variation d'une valeur au cours du temps. Dans l'air, le signal de parole est une fluctuation locale de la pression [6], la parole est avant tout un signal acoustique plongé dans un espace de communication linguistique. La communication parlée n'est donc possible qu'entre deux locuteurs ayant un même espace de communication linguistique [7], Sur le plan physique, la parole est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur. Il s'agit d'une onde sonore créée par le passage de l'air expulsé des poumons dans les appareils phonatoires et articulatoires du locuteur, ce qui provoque une modification de cette onde puis elle se propage dans l'air [8].

Le signal de la parole ne peut être exploité directement. En effet, le signal contient de nombreux autres éléments que le message linguistique : des informations liées au locuteur, aux conditions d'enregistrement, etc. Toutes ces informations ne sont pas nécessaires lors du décodage de la parole et rajoutent même du bruit. De plus, la variabilité et la redondance du signal de la parole le rendent difficilement exploitable tel quel. Il est donc nécessaire d'en extraire uniquement les paramètres qui seront dépendants du message linguistique. Généralement, ces paramètres sont estimés via des fenêtres glissantes sur le signal. Cette analyse par fenêtre permet d'estimer le signal sur une portion du signal jugée stationnaire : généralement 10 à 30 ms en limitant les effets de bord et les discontinuités du signal via une fenêtre de Hamming [9].

I.3.2 Le phonème :

L'unité de codage linguistique de la phonologie est en effet le phonème. Selon la théorie des traits distinctifs, chaque phonème se caractérise par un ensemble de traits qui le différencie des autres phonèmes. Ces traits distinctifs reflètent des propriétés de nature acoustico-auditive ou articulatoire. Le phonème est une unité phonologique distinctive et commutable. Exemples : Pour le Français, la substitution de la consonne occlusive labiale sonore du mot bain par une consonne occlusive labiale sourde (pain) provoque un changement de sens. Les consonnes [b] et [p] sont donc

deux phonèmes différents. En revanche, la substitution d'une consonne [ʀ] grasseyée (dorsale) par une consonne [r] roulée (apicale) dans un mot comme renard ne change pas le sens de ce mot. Les consonnes [ʀ] et [r] sont donc deux variantes d'un unique phonème. Un phonème peut en effet avoir plusieurs variantes articulatoires (phonétiques), appelées les allophones. L'apparition de ces variantes est due au phénomène de coarticulation : comme chaque articulateur du conduit vocal évolue de façon continue, les mouvements articulatoires peuvent être modifiés, suivant le contexte et sous l'effet de l'inertie mécanique de ces articulateurs, de manière à minimiser l'effort à produire pour réaliser une séquence de phonèmes donnée. Les consonnes liquides par exemple, se caractérisent par une forte sensibilité à la coarticulation d'où une grande variabilité acoustique. Les réalisations physiques d'un phonème peuvent aussi varier en fonction d'autres facteurs tels que le locuteur et le dialecte [2].

I.3.2.1 La classification des phonèmes :

Les deux grandes classes phonétiques en fonction de leur mode articulatoire : les voyelles et les consonnes :

- **les voyelles** : cette classe correspond, à quelques nuances supplémentaires près, aux voyelles de l'écrit. Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais ce sont principalement les formants en basses fréquences qui caractérisent les voyelles. Cette caractéristique permet d'ailleurs de distinguer grossièrement les voyelles en fonction de leur premier et deuxième formant [5].

- **Les consonnes** : Les consonnes se caractérisent d'un point de vue articulatoire par le passage complètement ou partiellement obstrué de l'air, en un ou plusieurs endroits du conduit vocal. Suivant leur mode d'articulation, on peut décomposer les consonnes du Français en trois classes : les occlusives, les fricatives et les sonantes [2].

Les consonnes occlusives sont constituées de la séquence d'événements acoustico-articulatoires suivante :

-Une occlusion complète du conduit vocal qui peut-être labiale pour [p] et [b], dentale pour [t] et [d], et palatale pour [k] et [g]. Cette phase d'occlusion correspond acoustiquement à un silence pour les occlusives sourdes ([p], [t], [k]) et à un son de basse fréquence, émis par la vibration des cordes vocales, pour les occlusives sonores ([b], [d], [g]) [2].

Les consonnes fricatives sont assimilables à du bruit qui est produit suite à une ou plusieurs constrictions de l'écoulement de l'air dans le conduit vocal. Ces constrictions se placent aux niveaux : labial pour [f] et [v], dental pour [s] et [z] et palatal pour [ʃ].

Les fricatives peuvent être sourdes ([f], [s], [ʃ]) ou sonores ([v], [z]). Ces dernières combinent des composantes d'excitation périodique et turbulente : la source vibratoire glottale est active et on observe une modulation du bruit de constriction par la fréquence fondamentale de la source glottale.

Les consonnes sonantes se caractérisent par une structure formantique qui ressemble à celle des voyelles. Plusieurs sous-classes de sonantes peuvent être distinguées :

-Les consonnes nasales [m] et [n] sont produites par le couplage du conduit nasal avec le conduit oral et par une occlusion, labiale pour [m] et dentale pour [n].

-Les consonnes liquides [l] et [ʀ] sont produites par une constriction au niveau de la langue (apicale pour [l] et dorsal pour [ʀ]). Les structures formantiques des liquides se caractérisent par une forte sensibilité à la coarticulation d'où une grande variabilité acoustique. Leur intensité sonore est en outre assez faible [2].

I.3.3 Le pitch (la fréquence du fondamentale) :

La fréquence fondamentale F0 (ou pitch [période des sons voisés]) joue un rôle important dans la parole. C'est elle qui véhicule une grande partie de l'information prosodique. L'intensité de la voix et les durées successives des syllabes complètent ces informations. D'une manière générale, la prosodie, qui peut être considérée comme l'effet des différentes variations de la fréquence fondamentale F0, 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. F0 moyen-homme 100 à 150 Hz, F0 moyen-femme 200 Hz à 300 Hz et F0 moyen-enfant 350 à 400 Hz [1].

La fréquence fondamentale est plus libre et varie en général plus rapidement dans le cas de la voix parlée [10].

La période du fondamental est par définition la fréquence de vibration des cordes vocales, elle est appelée aussi le pitch. La figure I.5 représente la fréquence fondamentale [11].

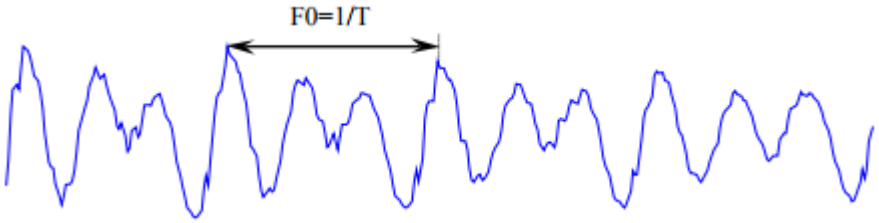


Figure I.5 fréquences fondamentales [11].

L'extraction du pitch est une tâche particulièrement difficile pour trois raisons :

- La vibration des cordes vocales n'a pas nécessairement une périodicité complète.

- Il est difficile de séparer le pitch des effets du trait vocal.

- La plage de dynamique de la fréquence fondamentale est très grande [11].

I.3.4 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 F1, F2, F3… 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 F1, F2 à travers un triangle acoustique [1].

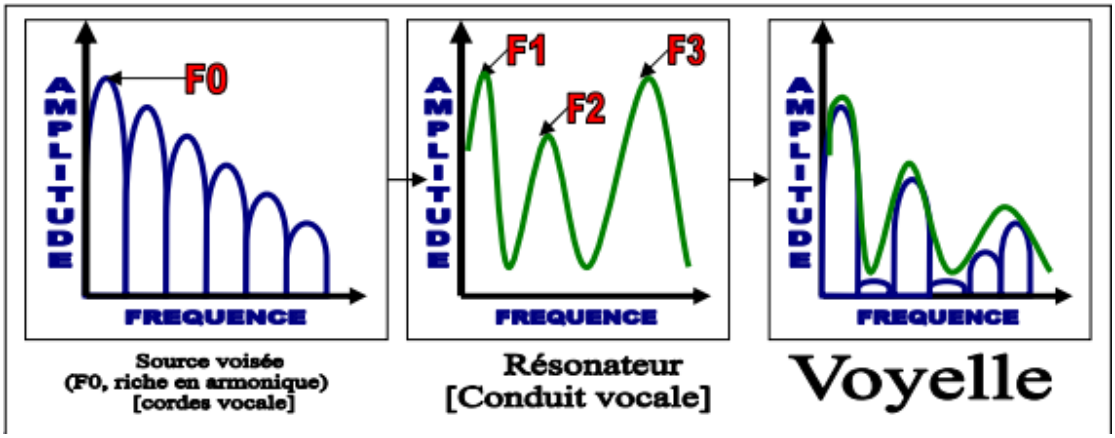


Fig I.6 la production d'une voyelle [1].

I.4 Numérisation :

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) [1].

I.4.1 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 a 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 » [12].

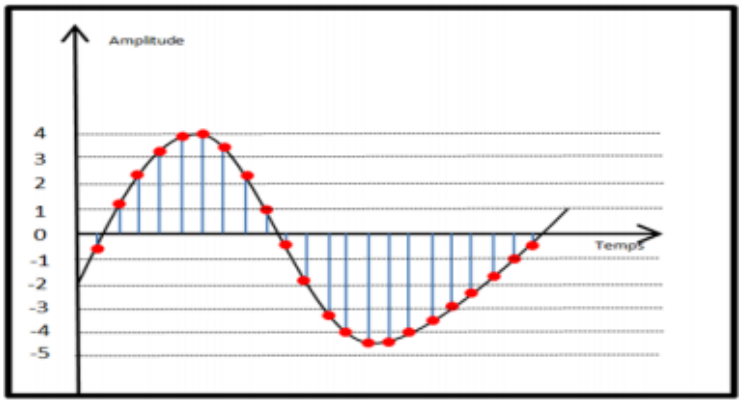


Fig I.7 signal échantillonné [12].

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

Pour les techniques d'analyse, de synthèse ou de reconnaissance de parole, la fréquence peut varier de 6000 à 16000 Hz. Puisque fm = 5 KHz alors fe = 10 KHz et donc il va falloir mesurer le signal tous le $\mathbf{1/10000}\ \boldsymbol{seconde} = \mathbf{0,1}\ \boldsymbol{msec}$ [12].

I.4.2 Quantification :

Quantifier un signal consiste à placer les amplitudes des échantillons sur une échelle de valeurs à intervalles fixes. Chaque impulsion correspond donc à un nombre binaire unique. -Une quantification à n bits permet d'utiliser 2n valeurs différentes. -Pour 8 bits, on a 256 valeurs et pour 16 bits, on a 65536 valeurs. La transformation d'une valeur physique (en volts) en une valeur binaire introduit donc une distorsion. De même lorsque l'impulsion dépasse la valeur maximale prévue [12].

I.4.3 Le codage :

C'est la représentation linéaire des valeurs quantifiés qui permet le traitement du signal sur machine [12].

I.5 Conclusion :

Dans ce chapitre nous avons passé de la production de la parole jusqu'à sa perception et ainsi que les caractéristiques générale de signal de la parole.

Dans le prochain chapitre nous étudierons les différentes méthodes d'analyse de signal parole et sans application sous matlab.

CHAPITRE II

ANALYSE DU SIGNAL DE PAROLE

II.1 Introduction :

L'étude de l'évolution temporelle et fréquentielle d'un signal de parole permet de mettre en évidence les caractéristiques de ce signal. Cet objectif est atteint grâce aux méthodes modernes de traitement du signal qui permettent de calculer par exemple, la transformée de Fourier d'un signal de parole pour déduire son spectre de puissance à court terme, et le spectrogramme qui représente l'évolution temporelle de ce spectre.

Dans ce chapitre nous décrirons les différents outils nécessaires à l'analyse de la parole.

Nous commencerons par une brève description du signal parole, ensuite nous exposerons les Techniques d'analyse à courte terme avec la Transformée de Fourier à court terme (TFCT) et la Transformée en Ondelette (WV: Wavelet Transform)

II.2 Description du signal de parole :

Une fois numérisé, le signal de parole peut être traité de différentes façons suivant les objectifs visés. Le nombre de technique possible étant très vaste, nous allons, dans ce qui suit, citer les outils relatifs au signal de parole

II.2.1 Description temporelle :

Le signal de parole est un signal quasi-stationnaire. Cependant, sur un horizon de temps supérieur, il est clair que les caractéristiques du signal évoluent significativement en fonction des sons prononcés comme illustré sur la figure ci-dessous. [13]

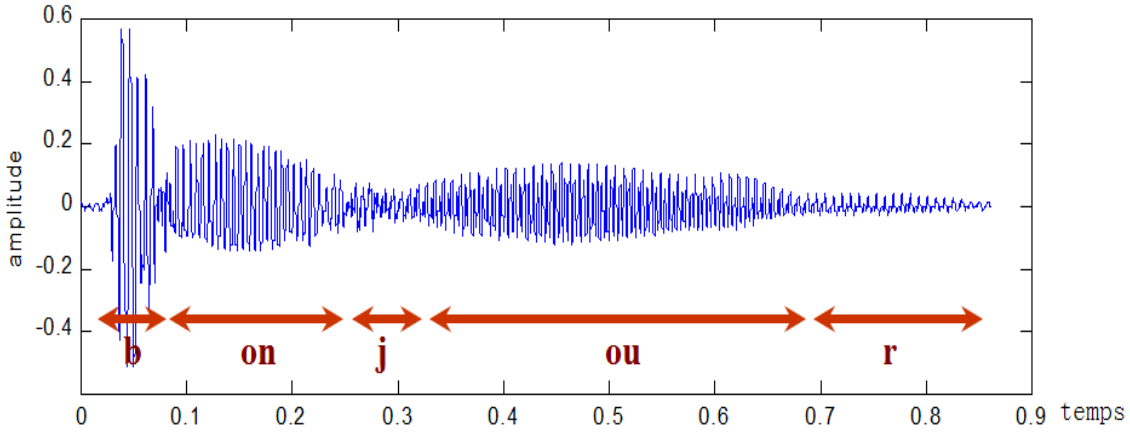


Fig II.1 Représentation temporelle du signal de parole

La première approche pour étudie le signal de parole consiste à observer la forme temporelle du signal. On peut à partir de cette forme temporelle en déduire un certain nombre de caractéristiques qui pourront être utilisées pour le traitement de la parole. Il est, par exemple, assez clair de distinguer les parties voisées, dans lesquelles on peut observer une forme d'onde quasi-périodique, des parties non voisées dans lesquelles un signal aléatoire de faible amplitude est observé. De même, on peut voir que les petites amplitudes sont beaucoup plus représentées que les grandes amplitudes ce qui pourra justifier des choix fait en codage de la parole. [13]

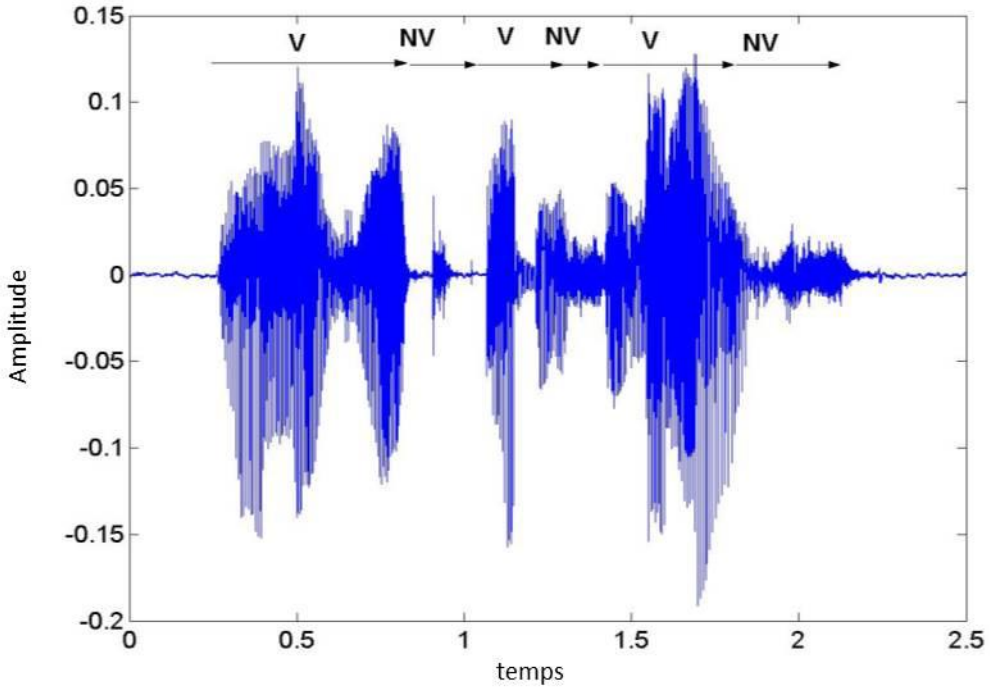


Fig II.2 Signal temporel de la phrase "La musique adoucit les mœurs" :

(V=partie voisée ; NV = partie non voisée)

II.2.2 Description fréquentielle :

Une seconde approche pour caractériser et représenter le signal de parole est d'utiliser une représentation spectrale. Les méthodes spectrales occupent une place prépondérante en analyse de la parole, l'oreille effectue, entre autres, une analyse fréquentielle du signal qu'elle perçoit; de plus les sons de la parole peuvent être assez bien décrits en termes de fréquence. [13]

La transformé de Fourier permet d'obtenir le spectre d'un signal, en particulier son spectre fréquentielle, c'est-à-dire sa représentation amplitude-fréquence.

- L'analyse De Fourier :

L'analyse de Fourier fournit une manière d'analyser les propriétés spectrales d'un signal donné dans le domaine fréquentiel. Les outils d'analyse de Fourier considèrent un signal comme superposition des fonctions sinusoïdales de base de différentes fréquences, phases et amplitudes.

L'analyse de Fourier fournit un outil pour trouver les paramètres des sinusoïdes fondamentaux (Transformée directe) ou pour synthétiser le signal original domaine - tems d'après la présentation du domaine – fréquence (Transformée inverse). [14]

La Transformée de Fourier (TF) [15] d'une fonction *x(t)* est définie par :

$$X(f) = \int x(t) e^{-2i\pi ft}\, dt \qquad \text{(II.1)}$$

La Transformée de Fourier Discrète (TFD) [16] d'un signal discret *x(m)* observé sur une durée de *N* échantillons est périodique de période *N* et définie par :

$$X(n) = \frac{1}{N} \sum_{m=0}^{N-1} x(m) e^{-2i\pi \frac{mn}{N}} \qquad \text{(II.2)}$$

II.2.3 Description temps/fréquence :

La représentation la plus répandue est le spectrogramme.

- Le spectrogramme :

Le spectrogramme permet de donner une représentation tridimensionnelle d'un son dans laquelle l'énergie par bande de fréquences est donnée en fonction du temps. Plus précisément, le spectrogramme représente le module de la transformée de Fourier discrète calculé sur une fenêtre temporelle plus ou moins longue. [13]

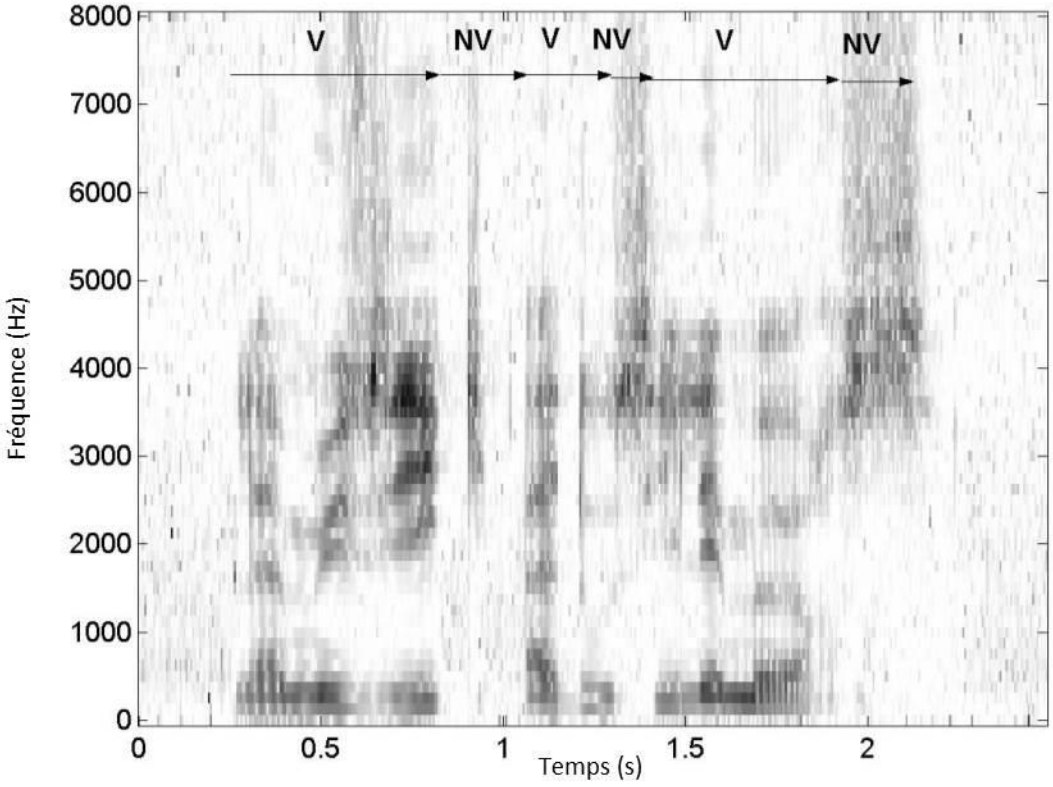


Fig II.3 Spectrogramme de la phrase "La musique adoucit les moeurs" [13]

(Le spectrogramme représente le module de la transformée de Fourier où cours du temps avec les Fréquences en ordonnée, le temps en abscisse et l'énergie en niveau de gris. Ainsi une zone sombre, indique une forte énergie à la fréquence et au temps correspondants)

La taille de la fenêtre d'analyse est un paramètre important pour cette représentation. Pour de petites fenêtres (typiquement de l'ordre de 3 à 10 ms), on obtiendra une représentation avec une très bonne localisation temporelle mais avec une précision fréquentielle moins précise. On aura dans ce cas un spectrogramme à bande large. Dans le cas contraire où l'on choisit des fenêtres d'analyse de plus grande taille (typiquement supérieures à 20 ms), on obtient une plus grande précision fréquentielle au prix d'une localisation temporelle plus approximative. On parlera dans ce cas de spectrogramme à bande étroite. Pour la parole, les deux types de représentations sont utilisés suivant que l'on souhaite observer la structure fine du contenu fréquentiel (qui est clairement visible sur le spectrogramme à bande étroite) ou que l'on souhaite observer l'enveloppe spectrale ou les formants (qui sont plus clairement visible sur un spectrogramme à bande large). [13]

La figure II.4 propose les spectrogrammes à bande étroite et à bande large d'une voyelle /a/ prononcée avec une fréquence fondamentale augmentant avec le temps. Les harmoniques sont alors très clairement identifiées sur le spectrogramme à bande étroite. [13]

(a) (b)

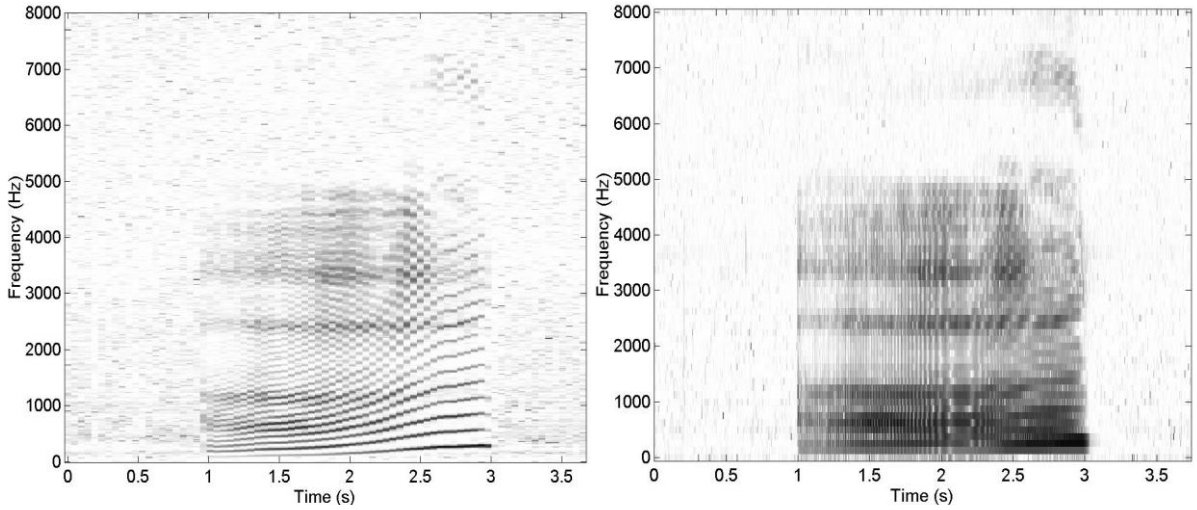


Fig II.4 Spectrogramme bande étroite (a) et spectrogramme large bande (b) d'une voyelle /a/ produite avec une élévation progressive de la fréquence fondamentale"

II.3 Les techniques d'analyse temps/fréquence : (l'analyse à court-terme)

Puisque le son de la parole change en continu en raison des mouvements articulatoires des organes vocaux de production, le signal doit être traité avec des petits segments, dans lesquels les paramètres demeurent quasi stationnaires (Fig.II.6). Le calcul de la DFT du signal entier jetterait les propriétés spectrales locales qui présentent des réalisations de différents phonèmes. Au lieu d'exécuter la DFT pour le signal entier, une fenêtre DFT est calculée. Un segment en général autour de 10-30 millisecondes, est multipliée par une fonction fenêtre et la DFT du segment fenêtré est alors calculée Ce processus est répété jusqu' à la fin du son articulé de sorte que le segment soit décalé en avant par une quantité fixe des points, en général autour 30 à 75 % de la longueur du segment. [14]

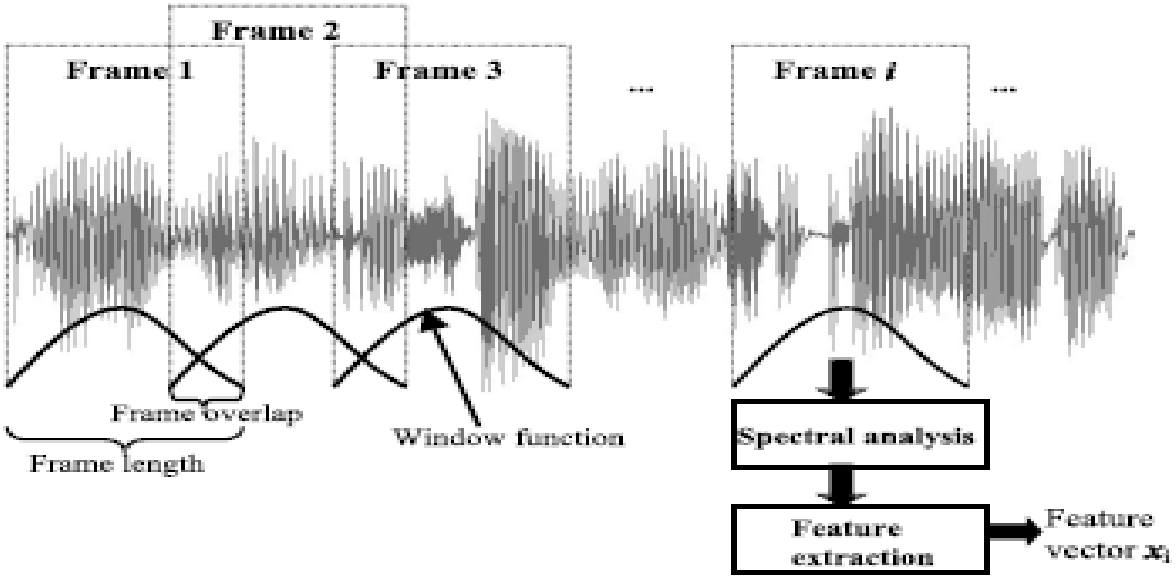


Fig II.5 l'analyse spectrale à courte terme [14]

II.3.1 La transformée de Fourier à court terme (TFCT) :

Pour donner un pouvoir de localisation aux fonctions analysantes de la transformée de Fourier, qui oscillent avec la même amplitude sur tout l'axe des réels, on pondère ces fonctions par une fonction fenêtre de manière à sélectionner uniquement la partie utile du signal. La fenêtre est bien sûr translatée de manière observer toutes les parties utiles du signal. Concrètement, la transformée de Fourier à fenêtre glissante s'exprime par:

$$F_x(t,f) = \int_{-\infty}^{+\infty} x(u)\, h^*(u-t)\, e^{-2i\pi ft}\, du \qquad \text{(II.3)}$$

Le signal est caractérisé par x(t), h est une fonction de fenêtrage centrée en t. Pour obtenir la représentation spectrale autour de t, il suffit de déplacer par translation la fenêtre h et d'effectuer une transformation de Fourier sur le signal ainsi fenêtré.

On notera que la transformée dépend maintenant de deux variables: une variable de fréquence et une variable de localisation temporelle du contenu fréquentiel. Cette transformée nous permet donc bien d'atteindre le but recherché qui était d'avoir des informations sur le signal en temps et en fréquence à partir de la transformation réalisée. [17]

Ce traitement fait l'hypothèse de stationnarité durant la durée de la fenêtre, quelle que soit la partie du signal considéré. La longueur de la fenêtre est cependant choisie pour respecter cette hypothèse. Ce choix influence directement les propriétés de résolution de la composition ; plus la fenêtre g est petite, plus la résolution temporelle est meilleure mais plus la résolution fréquentielle

est mauvaise. Si une haute résolution fréquentielle est nécessaire alors une longue fenêtre temporelle g sera utilisée et il sera difficile de respecter les hypothèses de stationnarité. La forme, la longueur de cette fenêtre ainsi que le pas d'incrémentation sont des paramètres fixés avant l'analyse. Ils présupposent une bonne connaissance a priori du signal à analyser. [18]

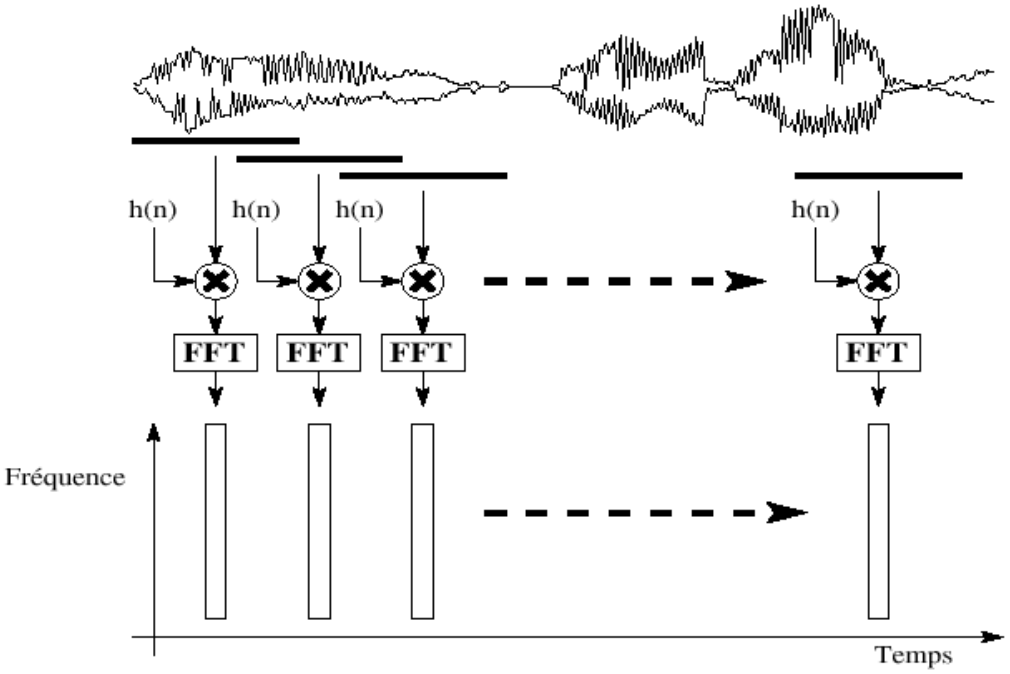


Fig II.6 Description schématique de l'analyse temps/fréquence par la FFT

(On notera bien que la fenêtre que l'on fait glisser est toujours la même et a donc la même résolution en temps et en fréquence sur tout le signal).

- Les fenêtres les plus utilisées sont :

Si nous définissons h(n) fenêtre où $0<n<N-1$ et N représente le nombre d'échantillon dans chacune des trames, alors le résultat du fenêtrage est le signal x_h, donné par la formule [12]:

$x_h = x(n)\, h(n), \qquad 0 < n < N-1$

Les fenêtres les plus utilisées sont :

- Fenêtre de Hamming :

$$h(n) = \begin{cases} 0.54 - 0.46\, cos\frac{2\pi n}{N-1} & si \quad 0 \le n \le N-1 \\ 0 & sinon \end{cases} \qquad \text{(II.4)}$$

- Fenêtre rectangulaire :

$$h(n) = \begin{cases} 1 & si \quad 0 \le n \le N-1 \\ 0 & sinon \end{cases} \tag{II.5}$$

- Fenêtre triangulaire (ou Bartlett) :

$$h(n) = \begin{cases} \frac{2n}{N-1} & si \quad 0 \le n \le \frac{N-1}{2} \\ \frac{2-2n}{N-1} & si \quad \frac{N-1}{2} < n \le N-1 \\ 0 & sinon \end{cases} \tag{II.6}$$

- Fenêtre Hanning :

$$h(n) = \begin{cases} 0.5 - 0.5\, cos\frac{2\pi n}{N-1} & si \quad 0 \le n \le N-1 \\ 0 & sinon \end{cases} \tag{II.7}$$

- Fenêtre Blackman :

$$h(n) = \begin{cases} 0.42 - 0.5\, cos\frac{2\pi n}{N-1} + 0.08\, cos\ \frac{4\pi n}{N-1} & si \quad 0 \le n \le N-1 \\ 0 & sinon \end{cases} \tag{II.8}$$

La figure II.7 illustre la forme que prennent les fonctions définies ci-dessus.

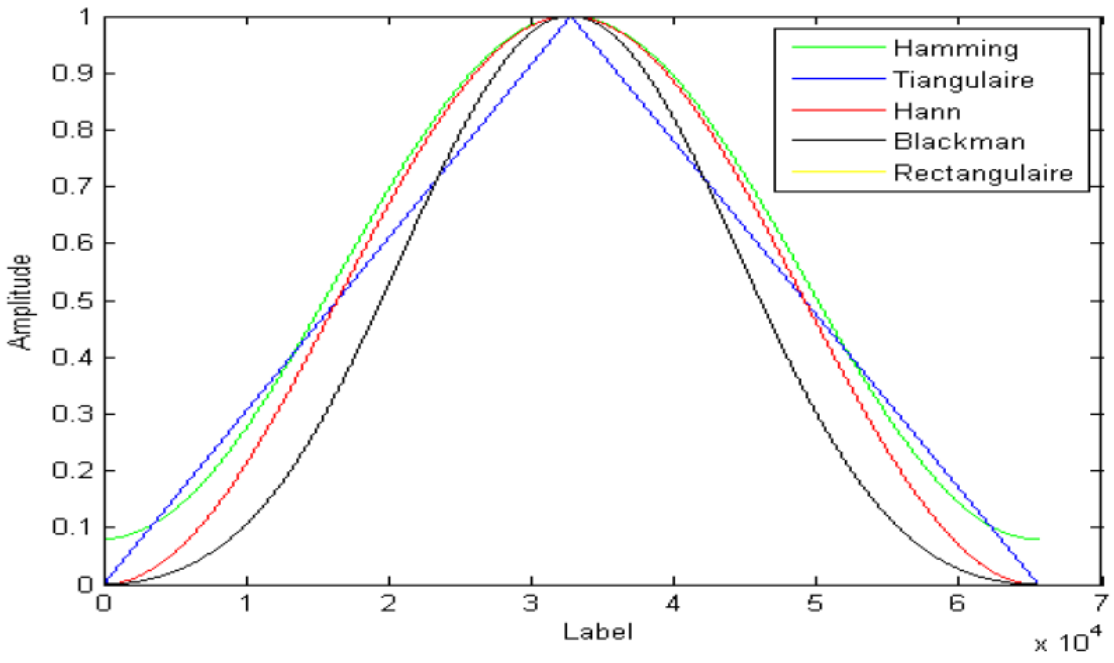


Fig II.7 fonction de fenêtrage

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%).

- **Inconvénients de la TFCT :** [19]

✓ Une fois la fenêtre choisie, la résolution fréquentielle et temporelle sont à leur tour définitivement fixée.

✓ Le signal doit être stationnaire dans la fenêtre utilisée par la TFCT.

✓ Une fenêtre étroite permet d'obtenir une bonne localisation du signal en temps par contre la localisation fréquentielle sera mauvaise.

✓ Une fenêtre temporel large conduit à une à une mauvaise résolution temporelle et une bonne résolution fréquentielle.

✓ Du moment où la longueur de la fenêtre est fixée une fois pour toute, l'analyse simultanée

Des phénomènes dont les échelles du temps sont différentes est impossible. Pour palier à tous ces inconvénients, un autre outil de traitement est utilisé, c'est la transformée en Ondelettes. L'apport principal de ce nouvel outil est la représentation conjointe Temps-échelle de signaux non-stationnaires et c'est l'objet du paragraphe Suivant. [12].

II.3.2 La transformée en ondelettes :

Le besoin d'améliorer l'analyse classique des méthodes du type transformation de Fourier à court terme se fit alors assez vite sentir. L'idée principale fut de définir une analyse du même type mais en faisant dépendre la largeur de la fenêtre d'analyse de sa position.
On pouvait ainsi régler la finesse de l'analyse en temps ou en fréquence indépendamment l'une de l'autre. Une des solutions proposées dans les années 80 fut la transformation en ondelettes. [20]

L'analyse est réalisée au moyen d'une fonction d'analyse spécifique appelée ondelette de base. Durant l'analyse, cette ondelette est positionnée dans le domaine temporel pour sélectionner la partie du signal à traiter. Puis, elle est dilatée ou contractée par l'utilisation d'un facteur d'échelle a permettant de concentrer l'analyse sur une gamme donnée d'oscillations. Quand l'ondelette est dilatée, l'analyse regarde les composants du signal qui oscille lentement; quand elle est contractée, l'analyse observe les oscillations rapides comme celle contenues dans une discontinuité de signal. Par ce traitement d'échelle (contraction - dilatation d'une ondelette), la transformée en ondelettes amène à une décomposition temporelle du signal. [18]

- Définition :

Partant d'une fonction mère h dépendant de t et possédant de bonnes propriétés ("assez" localisable, "assez" régulière, . . .), il est possible de générer, par l'action d'une déformation dite du groupe affine sur le signal, une famille de fonctions $h_{t;a}$ appelée famille d'ondelettes :

$$h_{t,a}(u) = \frac{1}{\sqrt{a}}\, h\left(\frac{u-t}{a}\right) \tag{II.9}$$

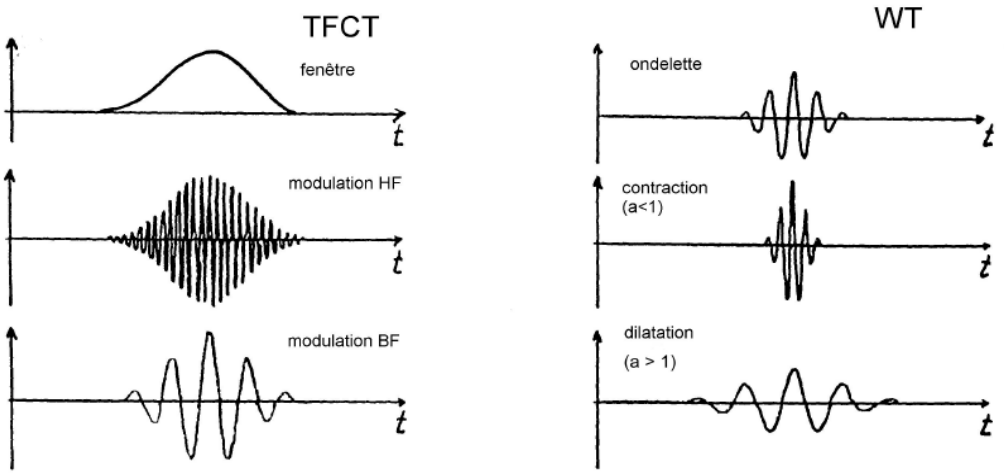
Où a > 0 est un paramètre d'échelle de contraction (a < 1) ou de dilatation (a > 1) de la fenêtre et t une translation de la fenêtre. [20]

Une fois cette famille générée, on décompose classiquement le signal x(t) sur cette famille selon le produit scalaire usuel dans l'espace des signaux. On obtient ainsi des coefficients d'ondelettes $T_x(t, a)$ qui caractérisent le coefficient de la décomposition du signal x(t) dans cette base [20]:

$$T_x(t,a) = \int_{-\infty}^{+\infty} x(u)\, h^*_{t,a}(u)\, du \tag{II.10}$$

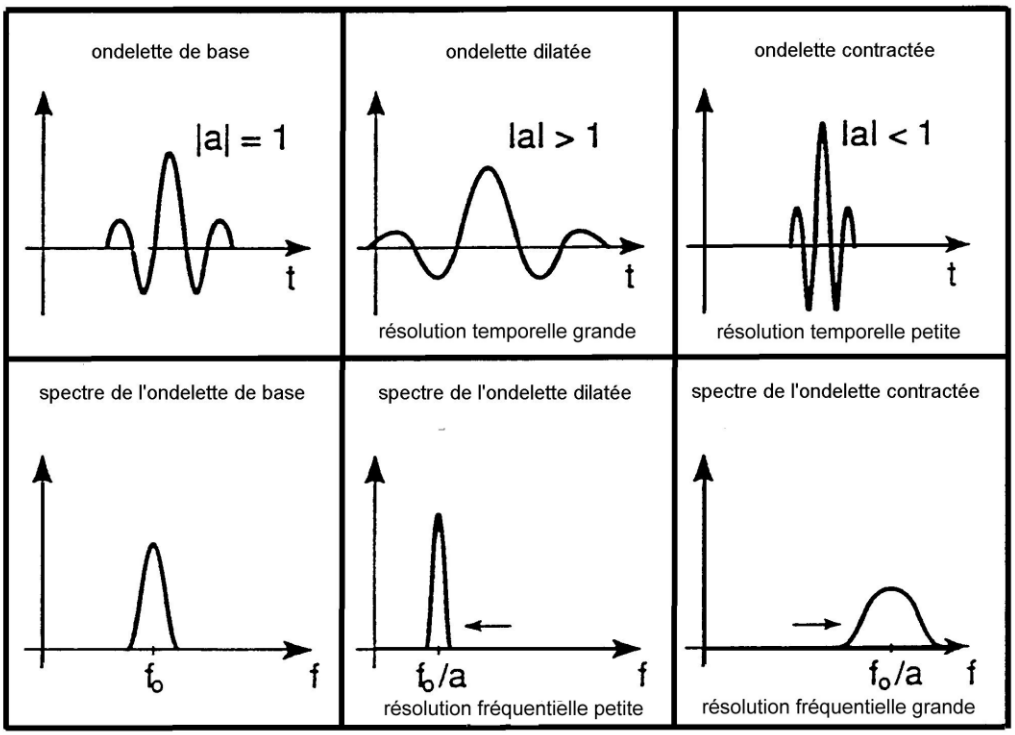
$$= \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(u)\, h^*\left(\frac{u-t}{a}\right) du$$

Pour mieux comprendre, l'analogie avec la TFCT est nécessaire. Les deux transformations peuvent être interprétées comme des projections du signal sur des fonctions qui sont simultanément localisées en temps et en fréquence/échelle. [18]



La TFCT définit un outil d'analyse temps fréquence tandis que la WT définit un outil temps échelle. Cependant, une relation entre l'échelle et la fréquence peut être établie supposant que l'ondelette de

base est positionnée dans le domaine fréquentielle autour d'une fréquence f_0. Il est important que f_0 soit reliée de manière franche à l'ondelette *h*, c'est à dire que f_0 soit le centre de gravité ou simplement la valeur maximale du spectre de l'ondelette de base. La figure ci-après le montre [18] :



Une fois l'ondelette de base localisée en fréquence, il est facile de démontrer que l'ondelette est positionnée autour de la fréquence f_0/a. Ainsi, la WT devient un outil temps fréquence.

Le spectre de l'ondelette de base montre que l'ondelette correspond à un filtre de bande passante autour de f_0. Le dimensionnement de l'ondelette dans le domaine temporel correspond à une translation dans le domaine des fréquences : Le spectre de l'ondelette dilatée est localisé vers les basses fréquences tandis que celui de l'ondelette contractée vers les hautes fréquences. [18]

Une autre caractéristique de l'ondelette dilatée est d'être plus diffuse dans le temps et donc d'avoir un spectre plus concentré autour de sa fréquence centrale. L'inverse est constaté pour l'ondelette contractée. Ceci est la conséquence du principe d'incertitude puisque une fonction rencontre quelques limitations dans sa résolution à la fois dans les domaines temporel et fréquentiel.

Il est facile de conclure que la WT favorise la résolution temporelle lors de l'analyse des composantes hautes fréquences, et privilégie la résolution fréquentielle lors de l'analyse des composantes basses fréquences. De plus, la WT donne une analyse à largeur de fréquence et de

temps relative constante, tandis que la TFCT correspond à une analyse à résolution temporelle et fréquentielle constante. [18]

II.4 Analyse du signal acoustique avec Matlab :

Notre objectif est de réaliser des algorithmes d'analyse de signal de parole sous MATLAB on prend un signal de la base de données NOIZEUS.

Utilisons les deux outils suivants :

- Le programme Matlab pour l'analyse et le traitement numériques des signaux.

Le logiciel Matlab servira pour traiter les signaux par tranches successives, extraire leurs caractéristiques et mettre en évidence les résultats obtenus. [21]

- NOIZEUS:

Un corpus de parole bruitée pour l'évaluation des algorithmes d'amélioration de la parole a été développé un corpus de parole bruyante (NOIZEUS) pour faciliter la comparaison des algorithmes d'amélioration de la parole entre les groupes de recherche. La base de données contient 30 bruyantes phrases IEEE (produites par trois hommes et trois femmes) haut-parleurs corrompus par huit différents bruits du monde réel à différents RSB. Ce corpus est disponible aux chercheurs gratuitement. [22]

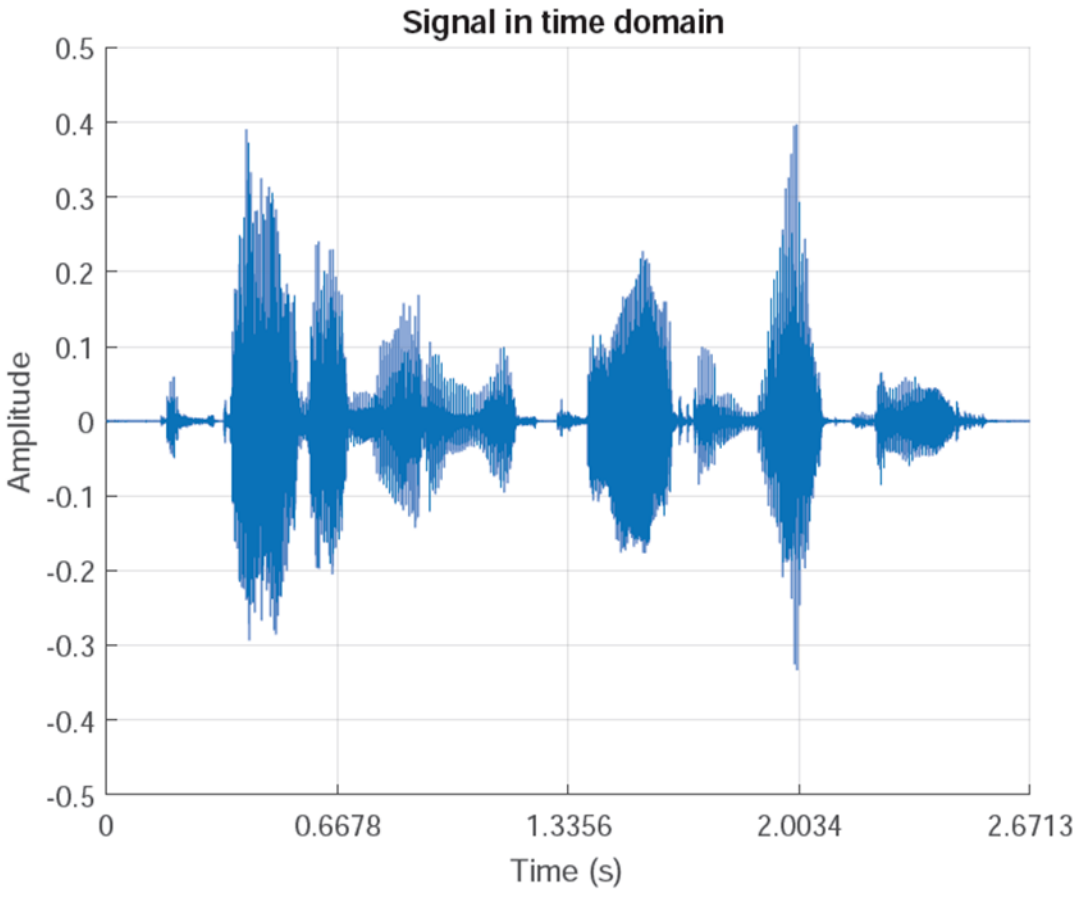
- **Programmation MATLAB :**

```
close all,

%% Signal in time domain

[x,Fs]=audioread('sp10.wav');
N=length(x);
t=(0:N-1)/Fs;
figure;
h=gcf;clf(h); grid on; hold on;
set(h,'Name','Time Domain Representation');
plot(t,x);
title('Signal in time domain')
ylabel('Amplitude')
xlabel('Time (s)')
axis([0 t(end) -0.5 0.5])
set(gca,...
'Units','normalized',...
'YTick',-0.5:.1:0.5,...
'XTick',0:t(end)/4:t(end))
 print -r600 -dpdf figure_1_T.pdf
```

```
%% Signal in frequency domain

NFFT = length(x);
X = fft(x,NFFT)/NFFT;
f = Fs/2*linspace(0,1,NFFT/2+1);
figure;
h=gcf;clf(h); grid on; hold on;
set(h,'Name','Frequency Domain Representation');
plot(f,2*abs(X(1:NFFT/2+1)))
title('Single-Sided Amplitude Spectrum')
xlabel('Frequency (Hz)')
ylabel('|X(f)|')
print -r600 -dpdf figure_2_F.pdf
```

- La visualisation du signal temporel et de son spectre :

(a)



(b)

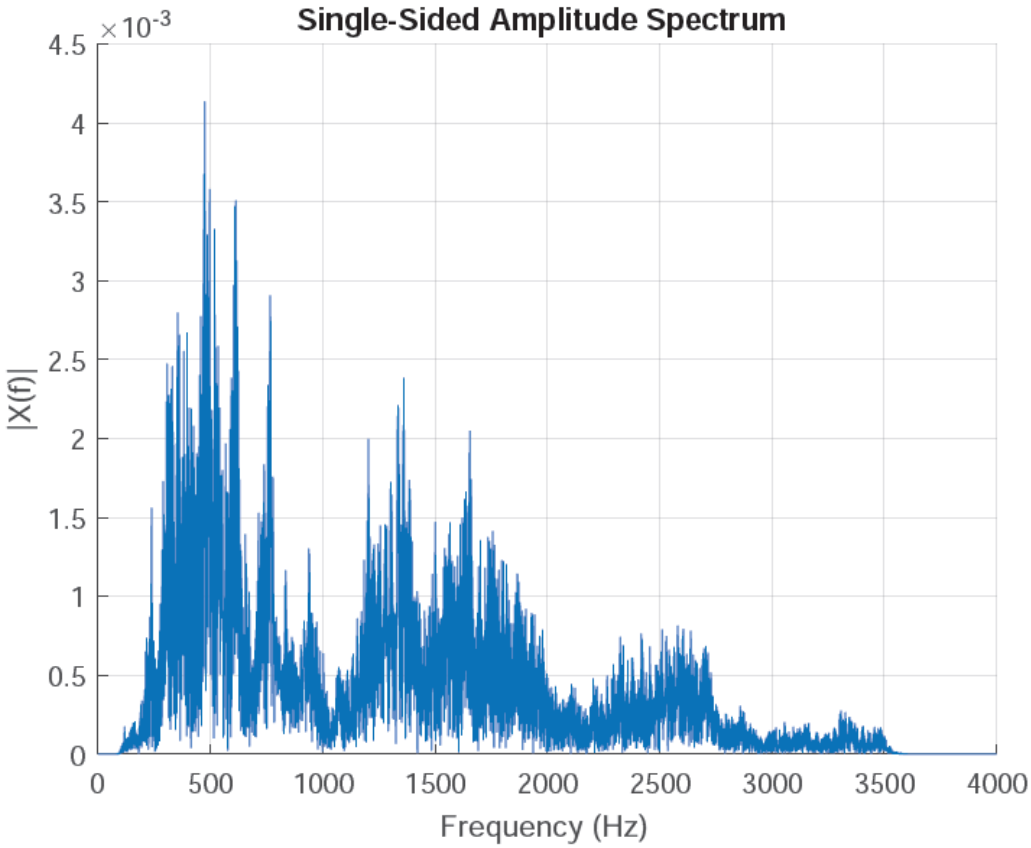


Fig II.8 Signal de parole, (a) représentation temporelle, (b) représentation spectrale.

```
%% Short time Fourier transform (STFT)


x = x(:);        % represent x as column-vector if it is not
wlen = 128;      % length of the hamming window
h =64;           % hop size
nfft =256;       % number of FFT points
fs =Fs;          % sampling frequency, Hz
xlen = length(x);% length of the signal

% form a periodic hamming window
win = hamming(wlen, 'periodic');

% form the stft matrix
rown = ceil((1+nfft)/2);            % calculate the total number of rows
coln = 1+fix((xlen-wlen)/h);        % calculate the total number of columns
x_STFT = zeros(rown, coln);           % form the stft matrix

% initialize the indexes
indx = 0;
col = 1;
```

```
% perform STFT

while indx + wlen <= xlen
    % windowing
    xw = x(indx+1:indx+wlen).*win;

    % FFT
    X = fft(xw, nfft);

    % update the stft matrix
    x_STFT(:, col) = X(1:rown);

    % update the indexes
    indx = indx + h;
    col = col + 1;
end

% calculate the time and frequency vectors
t = (wlen/2:h:wlen/2+(coln-1)*h)/fs;
f = (0:rown-1)*fs/nfft;

% Plot the spectrogram
figure;
h=gcf;clf(h); grid on;
set(h,'Name','Time-Frequency Domain Representation');
imagesc(t,f,abs(x_STFT)); axis xy;xlabel('Time (s)','Fontsize',12);
ylabel('Frequency (Hz)','Fontsize',12);
title('Spectrogram','fontsize',13);colormap((1-gray).^2)
print -r600 -dpdf figure_3_TF.pdf
```

- La visualisation du Spectrogramme :

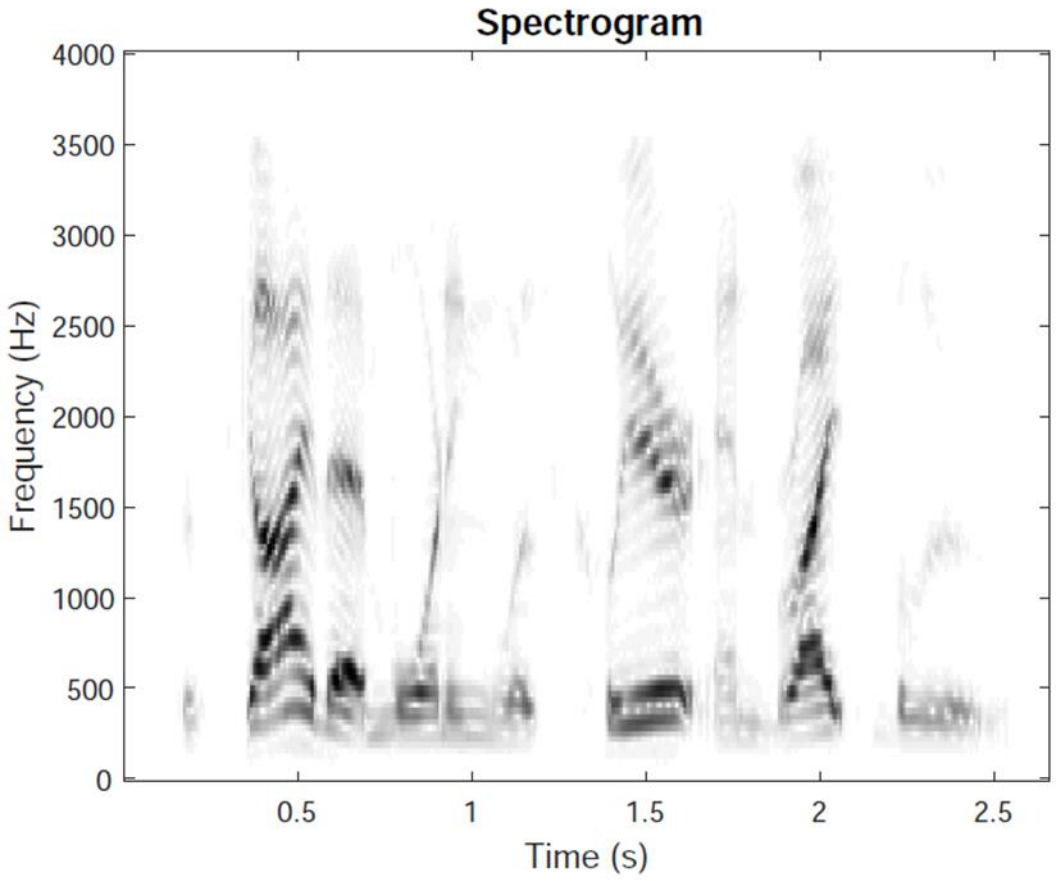


Fig II.9 Spectrogramme de signal de parole

II.5 Conclusion :

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

Dans le chapitre suivant nous allons présenter les méthodes d'analyses et d'extraction les plus utilisés pour le signal de parole dont le but de la reconnaissance automatique de la parole.

CHAPITRE III

EXTRACTION DES PARAMETRES

III.1 Introduction :

Le signal de parole ne peut directement être transformé en hypothèses de séquences de mots. L'extraction de ses paramètres est une étape importante puisqu'elle doit déterminer les caractéristiques pertinentes du signal. Il est nécessaire de découper le signal audio par trame, en prenant généralement une taille fixe définie aux alentours de 25 ms, afin de rendre le signal quasi-stationnaire (chapitre précédent). Un vecteur de paramètres est ensuite extrait pour chaque trame. Cette extraction peut se faire au moyen de multiples techniques. Ces différentes méthodes permettent d'extraire des coefficients caractéristiques pour chaque trame.

Parmi les coefficients les plus utilisés et qui représentent au mieux le signal de la parole en reconnaissance de la parole, nous trouvons les coefficients cepstraux, appelés également cepstres. Les deux méthodes les plus connues pour l'extraction de ces cepstres sont : l'analyse spectrale et l'analyse paramétrique. Pour l'analyse spectrale (par exemple, Mel-Scale Frequency Cepstral Coefficients (MFCC)) comme pour l'analyse paramétrique (par exemple, le codage prédictif linéaire (LPC)), le signal de parole est transformé en une série de vecteurs calculés pour chaque trame. Ces coefcients jouent un rôle capital dans les approches utilisées pour la reconnaissance de la parole. [23]

Dans ce chapitre, nous allons présenter les méthodes les plus utilisées d'extraction des paramètres pertinents du signal parole pour la reconnaissance automatique de la parole.

III.2 Représentation cepstrale :

La parole peut être représentée sous la forme d'un modèle source-filtre. Cette représentation permet ainsi de représenter le signal de parole s(t) sous la forme de la convolution du signal source g (t) par la réponse impulsionnelle du filtre h(t) représentant le conduit vocal [13]:

$$s(t) = g(t) * h(t) \qquad \text{(III.1)}$$

L'étude de ce signal à l'aide de la FFT présente un défaut particulier liée à cette convolution qui rend difficile l'observation de la seule contribution du conduit vocal. Le cepstre (parfois appelé lissage cepstral) permet de séparer les contributions respectives de la source et du conduit vocal.

En effet, l'équation précédente se récrit dans le domaine spectral sous la forme :

$$S(\omega) = G(\omega)\ H(\omega) \qquad \text{(III.2)}$$

où $S(\omega)$,$G(\omega)$ et $H(\omega)$ représentent respectivement les transformées de Fourier de s(t), g(t) et h(t).

Le cepstre qui est défini par le logarithme de la transformée de Fourier inverse du module de $S(\omega)$ s'écrit donc sous la forme :

$$c(\tau) = FFT^{-1}\, log|S(\omega)| = FFT^{-1}\, log|G(\omega)| + FFT^{-1}\, log|H(\omega)| \qquad \text{(III.3)}$$

On peut alors noter que le spectre s'exprime comme la somme de deux termes. Le premier terme FFT^{-1} log $|G(\omega)|$ est caractéristique de la source et représente ainsi la structure fine, tandis que le second terme est caractéristique de l'enveloppe spectrale et représente la contribution du conduit vocal. Le paramètre homogène à un temps est appelé quéfrence. A l'aide de cette représentation, il est possible d'isoler soit le pic (qui correspond au pitch) qui se trouve dans la région des hautes quéfrences (on a ici une méthode d'estimation de la fréquence fondamentale) soit d'isoler la partie correspondant aux basses quéfrences qui représente une version lissée de l'enveloppe spectrale. Ce procédé de séparation des éléments cepstraux est appelé un liftrage (par dérivation de l'appellation filtrage).

Lorsque le cepstre est obtenu en calculant la transformée de Fourier discrète, on obtient la forme suivante [13]:

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} log|X(k)|\, e^{\frac{2j(\pi)kn}{N}} \qquad pour \qquad 0 \leq n \leq N-1 \qquad \text{(III.4)}$$

III.3 Les coefficients MFCC (Mel-Frequency Cepstral Coefficients) :

La paramétrisation MFCC est probablement la paramétrisation la plus répandue dans les systèmes de reconnaissance actuels. Nous donnons ci-dessous les principales étapes de cette paramétrisation [13]:

1. Fenêtrage du signal Le signal de parole est séparé en trames de N échantillons, chaque trame étant séparée de M échantillons. Dans le cas courant où M < N on dira qu'il y a recouvrement (overlap en anglais) entre les trames. Pour M = 1/3 N. En pratique, la longueur N d'une trame est couramment choisie de façon à avoir des trames dont la durée est de l'ordre de 20ms associé à un recouvrement entre trames de 50% correspondant à une valeur de M = N/2 . L'opération précédente consiste ainsi à appliquer une fenêtre rectangulaire de durée finie sur l'ensemble du signal. Pour réduire les effets dus aux discontinuités aux bords de la fenêtre, il est fréquent de pondérer une trame de longueur N par une fenêtre de pondération. L'une des fenêtres les plus utilisée est la fenêtre de Hamming. Cette opération donne la trame fenêtrée :

$$S_h(n) = S'(n) * h(n) \qquad \text{(III.5)}$$

2. Calcul de la transformée de Fourier rapide (FFT) pour chaque trame du signal de parole
3. Filtrage par un banc de filtres MEL. Cette opération permet d'obtenir à partir du spectre S(k) de chaque trame, un spectre modifié qui est en fait une suite de coefficients, noté S'(k), représentant l'énergie dans chaque bande fréquentielle k (définies sur l'échelle Mel), pour k = 1 ……K. En pratique, on utilise des filtres triangulaires de largeur de bande constante et régulièrement espacées sur l'échelle Mel (On peut par exemple choisir un espacement entre filtres de 150 mels et une largeur des filtres triangulaire prise à leur base de 300 mels).
4. Calcul des coefficients MFCC : Les coefficients MFCC sont alors obtenus en effectuant une transformée en cosinus discrète inverse du logarithme des coefficients S'(k) :

$$\bar{c}_n = \sum_{k=1}^{K} (\log \bar{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \qquad pour \quad n = 1,2,\dots,L \qquad \text{(III.6)}$$

Où L est le nombre de coefficients cepstraux désirés.

5. Pondération : En raison de la grande sensibilité des premiers coefficients cepstraux sur la pente spectrale générale et de la sensibilité au bruit des coefficients cepstraux d'ordre élevé, il est courant de pondérer ces coefficients pour minimiser cette sensibilité. Cette pondération pourra s'écrire sous la forme :

$$\hat{c}_m = w(m)\, c_m \qquad pour \quad 1 \leq m \leq Q \tag{III.7}$$

Où Q est le nombre de coefficients cepstraux.

La fenêtre de pondération cepstrale est en fait un liftre passe bande dont un choix approprié peut être

$$w(m) = \left[1 + \frac{Q}{2} \sin\left(\frac{\pi m}{Q}\right)\right] \qquad pour \quad 1 \leq m \leq Q \tag{III.8}$$

Cette fenêtre tronque le nombre de coefficients et diminue le poids des premiers et derniers coefficients.

6. Calcul des dérivées temporelles Δ, Δ^2 : La représentation cepstrale donne une bonne représentation des propriétés fréquentielles locales du signal (pour une fenêtre de signal donnée). Une représentation améliorée peut être obtenue en incluant de l'information liée à l'évolution temporelle des coefficients cepstraux. Celle-ci peut être obtenue par exemple à l'aide des dérivées premières et secondes des coefficients cepstraux. Soit cm(t) les coefficients cepstraux obtenus à l'instant t (ou plus précisément à la fenêtre d'indice t). Cette suite est obtenue à des instants discrets et ainsi il est bien connu qu'un simple moyennage aux différences ne permet pas d'obtenir des estimations non bruitées. Ainsi, la dérivée est souvent obtenue en effectuant une moyenne sur un plus grand horizon temporelle sous la forme :

$$\Delta c_m(t) \approx \mu \sum_{k=-K}^{K} k c_m(t+k) \tag{III.9}$$

Où μ est une constante de normalisation et (2K +1) est le nombre de trames utilisées pour ce calcul.

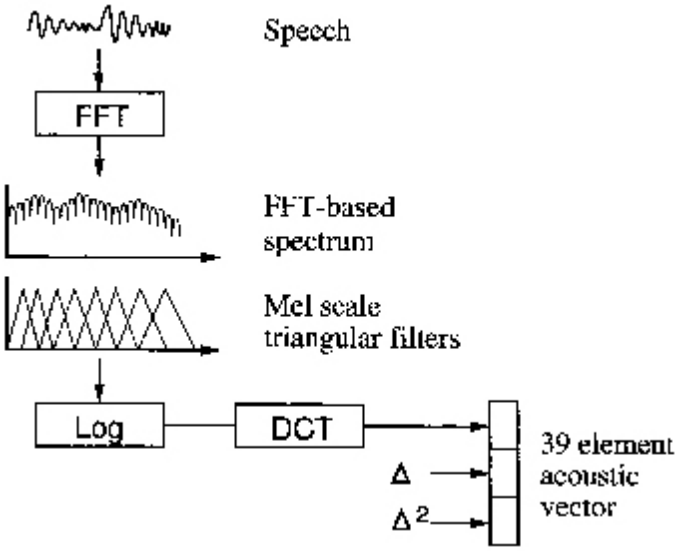


Fig III.1 Schéma bloc de la paramétrisation MFCC [13]

III.4 Linear Predictive Coding (LPC) :

Il consiste à synthétiser des échantillons de signal de parole à partir d'un modèle de système de production vocale et d'excitation. Il permet de prédire une valeur future du signal à partir d'une combinaison des valeurs précédentes. Le codage LPC est l'une des techniques les plus puissantes d'analyse de la parole qui a gagné en popularité en tant que technique d'estimation de formants. Les coefficients LPC sont calculés en découpant le signal de la parole en de petites fenêtres de courte durée. La fenêtre de Hamming est ensuite appliquée sur les différentes portions de signal obtenues. L'application de la fenêtre de Hamming permet de diminuer la distorsion spectrale. Avec l'équation :

$$s(n) = \sum_{k=1}^{p} a_k \times s(n-k) + e(n) \qquad \text{(III.10)}$$

Le signal à l'instant n est prédit à partir des p échantillons précédents.

La moyenne que constitue la somme pondérée du signal sur p pas de temps introduit une erreur car la parole ne constitue pas un processus parfaitement linéaire. Cette erreur est corrigée par l'introduction du terme e(n). Le codage par prédiction linéaire consiste donc à déterminer les coefficients a_k qui minimisent l'erreur e(n), ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage [24].

III.5 Perceptual Linear prediction (PLP) :

La prédiction linéaire perceptuelle PLP modélise la parole humaine en se basant sur le concept psychophysique de l'audition. Elle est développée par Hermansky. La paramétrisation du signal en coefficients PLP est identique à celle du codage LPC, sauf que les caractéristiques spectrales du PLP sont transformées pour correspondre aux caractéristiques du système auditif de l'homme [24].

III.6 Autres paramètres :

- **Energie du signal :**

Elle est évaluée sur plusieurs trames de signal successives mettant en évidence les différentes variations du signal. Elle correspond à la puissance du signal et se calcule comme suit [24]:

$$E(fenêtre) = \sum_{n \in fenêtre} |n|^2 \quad \text{(III.11)}$$

- **Taux de passage par zéro (zero crossing rate en anglais) :**

Il correspond au nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la valeur zéro. Pour un segment de parole donné, il est lié à la fréquence moyenne et est nul pour un segment silencieux. Il est calculé à partir de l'équation suivante [24]:

$$Z_m = \sum_n |sign[x(n)] - sign[x(n-1)]|\, w(m-n) \quad \text{(III.12)}$$

La fonction sign est définie comme suit :

$$sign[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \quad \text{(III.13)}$$

$$w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \geq N-1 \\ 0 & autrement \end{cases} \quad \text{(III.14)}$$

x(n) est le signal de la fenêtre n.

III.7 Conclusion :

Dans ce chapitre nous avons cité, les techniques d'extraction de caractéristique pertinente de signal de parole les plus utilisées dans les systèmes de reconnaissance automatique de la parole.

Le chapitre suivant sera consacré au sujet de la reconnaissance automatique de la parole.

CHAPITRE IV

LA RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

IV.1 Introduction :

La Reconnaissance Automatique de la Parole est un processus qui permet de passer d'un signal acoustique de parole à la transcription de ce signal en version écrite.

Dans ce chapitre nous décrirons les principes de la reconnaissance automatique de la parole, nous présentons quelques approches possibles sur les méthodes et algorithme utilisée ainsi que l'apprentissage dans la reconnaissance automatique de la parole.

IV.2 Les méthodes utilisées pour la reconnaissance de la parole :

On peut classer les méthodes de reconnaissance en deux classes :

-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. 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 [1]. Dans cette catégorie on trouve principalement la méthode DTW (Dynamic Time Warping), et la méthode HMM (Hidden Markov Models) [25].

- 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, diphone, 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). 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) est fondamental dans une telle approche [1].

IV.2.1 La programmation dynamique :

Lorsqu'un locuteur, même entraîné, répète plusieurs fois une phrase ou un mot, il ne peut éviter les variations du rythme de prononciation ou de la vitesse d'élocution, ces variations entraînent des transformations non linéaires dans le temps du signal acoustique, ce qui fait qu'on ne pourra comparer directement point à point (matching) deux formes acoustiques sans correction temporelle au préalable. Pour établir une meilleure correspondance entre les axes temporels des deux mots, en même temps que leurs comparaisons, on utilise une technique appelée technique d'alignement temporel dynamique ou DTW. C'est une technique basée sur la programmation dynamique qui consiste à trouver la trajectoire optimale entre le mot de référence et le mot inconnu [25].

IV.2.2 Les modèles acoustiques :

un décodage acoustico-phonétique (DAP) est défini généralement comme la transformation de l'onde vocale, en unités phonétiques - une sorte de transcodage qui fait passer d'un code acoustique à un code phonétique - ou plus exactement comme la mise en correspondance du signal et d'unités phonétiques prédéfinies (opération de couplage / identification) dans lequel le niveau de représentation passe du continu au discret [26].

Les modèles acoustiques sont des modèles stochastiques qui sont utilisés conjointement à un modèle de langage afin de prendre des décisions quant-à la suite de mots contenue dans la phrase.

Le rôle du modèle acoustique est de calculer la probabilité qu'un événement linguistique (phonème, mot, ...) ait généré une séquence de vecteurs de paramètres extraits d'un signal de parole [27].

Quelques caractéristiques importantes des modèles acoustiques doivent être prises en compte. D'un point de vue utilisabilité, les modèles acoustiques doivent être robustes puisque les conditions acoustiques de la tâche de reconnaissance sont souvent différentes des conditions d'entraînement. En effet, le signal de parole possède de nombreuses variabilités qui ont pour conséquence d'augmenter la disparité entre la réalisation acoustique et le contenu linguistique. D'un point de vue pratique, les modèles acoustiques doivent être efficaces. Pour que leur utilisation soit acceptable, il est nécessaire qu'ils respectent certaines contraintes temporelles et donc proposer des temps de réponse relativement courts [27].

Les paramètres d'un modèle acoustique sont estimés à partir d'un corpus d'entraînement. Ce corpus d'entraînement est généralement transcrit manuellement. Cela permet d'identifier les segments de parole correspondant à chaque événement linguistique [27].

Actuellement, on distingue deux types de modèles acoustiques couramment utilisés : les modèles de Markov Cachés (Hidden Markov Model - HMM) utilisant des mixtures de gaussiennes (Gaussian Mixture Models - GMM), et les modèles hybrides HMM utilisant des réseaux de neurones (Artificial Neural Network - ANN) [27].

IV.2.2.1 Modèle de Markov caché :

Les modèles de Markov d'ordre 1 sont des automates probabilistes à états finis qui se basent sur l'hypothèse que "le futur ne dépend que de l'état présent". L'état du modèle au temps t ne dépend donc que de l'état du modèle au temps t − 1 : P (qt |q1, q2, ..., qt−1) ∼= P (qt |qt−1) où qt est l'état du système au temps t. À chaque étape temporelle, le modèle évolue suivant la fonction de transition et passe potentiellement dans un nouvel état ; l'évolution du système n'est connue qu'à travers des statistiques [28].

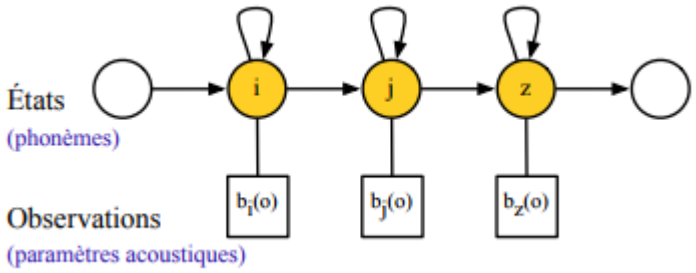


Fig IV.1 exemple de HMM utilisé pour modéliser les phonèmes [28].

IV.2.3 Les réseaux de neurones artificiels :

Les réseaux de neurones ou modèles neuromimétiques sont constitués de cellules élémentaires, appelées neurones, fortement connectées entre elles, Ces neurones émettent en sortie une fonction non linéaire de la somme pondérée de leurs entrées. Une des formes les plus répandues de réseau de neurones est le perceptron multicouche. Un perceptron est un réseau sans contre-réaction, ce qui signifie que les sorties des neurones de la couche i forment les entrées des neurones de la couche i+1. La figure IV.2 montre un perceptron à trois couches dont une cachée, permettant de reconnaître N symboles (phonèmes ou autres) [27].

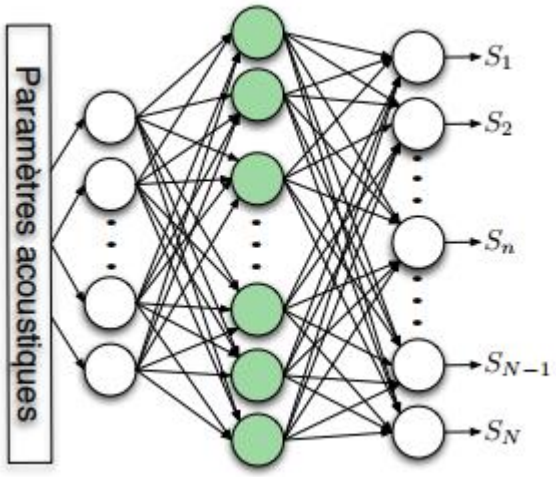


Fig IV.2 Architecture d'un perceptron à trois couches dont une cachée (en gris) permettant de reconnaître N symboles [27].

IV.2.4 Modèle de langage :

Le langage est la faculté de mettre en œuvre un système de symboles linguistiques (qui constituent la langue) permettant la communication et l'expression de la pensée. Cette faculté peut être mise en œuvre, notamment, par des moyens vocaux (parole), graphiques (écriture) et/ou gestuels (langue des signes). Les modèles de langages ont pour objectif de décrire un langage. Deux types de modèles sont principalement utilisés. Les premiers sont à base de grammaires formelles mises au point par des experts en linguistique. Les autres sont des modèles stochastiques qui utilisent un corpus pour estimer des probabilités d'une suite de mots d'un langage de manière automatique. La génération d'un ensemble de règles décrivant un langage est un processus long et difficile, c'est pourquoi les modèles probabilistes sont privilégiés dans les systèmes de reconnaissance automatique de la parole, Les modèles de langage stochastiques les plus utilisés sont les modèles N-grammes permettant d'estimer calculer la probabilité d'une suite de mots [27].

IV.2.5 Modèle de prononciation :

Le lexique d'un système de reconnaissance vocale précise une ou plusieurs prononciations pour chaque mot. Pour le français, les prononciations multiples sont en partie dues aux événements de liaison ou de réduction, dans le cadre desquels un locuteur peut prononcer ou pas un certain phonème dans un certain contexte. Les accents et les dialectes peuvent aussi générer diverses variantes de prononciations.

La liaison implique la prononciation d'un phonème de liaison entre deux mots. Pour donner un exemple : les mots "les oiseaux" se prononcent séparément "l e" et "w a z o", alors qu'ensemble ils se prononcent "l e z w a z o". Modèle acoustique La réduction implique l'omission d'un phonème a priori présent dans la prononciation standard d'un mot, comme dans le cas de "ce" qui se prononce normalement "c ə", mais qui peut également être prononcé simplement "c" dans le cas d'une prononciation rapide [28].

IV.3. L'apprentissage :

IV.3.1 L'apprentissage mono locuteur :

Pour la reconnaissance mono-locuteur généralement on distingue deux types d'apprentissage [29]:

IV.3.1.1 Apprentissage simple :

Pour un locuteur entraîné capable de garder le même rythme d'élocution, et dans des conditions idéales d'enregistrement, tel que l'absence de bruits, un procédé simple d'apprentissage consiste à utiliser chaque mot prononcé durant la session d'entraînement comme mot de référence [29].

IV.3.1.2 Apprentissage robuste :

Pour des mots qui ne sont pas acoustiquement voisins, la méthode précédente n'est pas suffisante pour espérer obtenir des résultats de reconnaissance consistants, alors une amélioration du procédé est suggérée. Dans ce cas le locuteur est invité à prononcer chaque mot du vocabulaire plusieurs fois et le dictionnaire de référence est conçu à partir de ces énoncés. C'est le cas par exemple de l'apprentissage appelé apprentissage robuste [29].

IV.3.2 L'apprentissage multi-locuteurs :

Le regroupement ou le "clustering " en anglais consiste à regrouper L occurrences prisent à partir de plusieurs locuteurs, en N classes ou cluster, Une bonne méthode de regroupement permet de garantir une grande similarité à l'intérieur de la même classe et une faible similarité avec les autres classes.

Pour les systèmes indépendants de locuteur, des algorithmes de classification sont utilisés pour la création du dictionnaire des mots de référence. Généralement on retient pour chaque mot plusieurs représentants de huit à douze [30], ce qui permet d'avoir une bonne image de l'ensemble des prononciations d'un même mot, et constitue un choix raisonnable pour tenir compte de la variabilité interlocuteurs. Le dictionnaire peut être ensuite stocké en mémoire non-volatile dans le

système de reconnaissance. Plusieurs méthodes de classification ont été développées par les chercheurs, tel que les méthodes dites semi-automatique Ces dernières nécessitent la supervision d'un expert qui guide le processus de classification [31].

IV.4 les algorithmes de classification :

On distingue deux approches de classification automatique [29]:

• **les algorithmes hiérarchiques** : ils permettent de créer une décomposition hiérarchique de l'ensemble d'apprentissage selon certains critères, le résultat est un dendogramme (ou arbre hiérarchique) qu'il faudra ensuite interpréter.

• **les algorithmes non hiérarchiques (ou de partitionnement) :** l'ensemble d'apprentissage est décomposer en sous-groupes (classes) selon certain critère, ces classes se distinguent les uns des autres soit par des frontières soit par le centre de gravité qui caractérise chaque classe appelée aussi centroïde.

IV.5 conclusion :

Dans ce chapitre en a donné une vue générale sur les méthodes utilisée dans la reconnaissance automatique de la parole.

La reconnaissance automatique de la parole est un domaine qui a captivé le public ainsi que de nombreux chercheurs. À ses balbutiements, les projections sur ses applications étaient très optimistes, Malheureusement, malgré l'incroyable évolution des ordinateurs et des connaissances, la reconnaissance automatique de la parole n'en demeure pas moins un sujet de recherche toujours actif... et les résultats obtenus sont encore loin de l'idéal qu'on aurait pu en attendre il y a vingt ans.

Conclusion générale

Conclusion générale

Dans ce travail, nous avons décrit les différentes techniques d'analyse et d'extraction d'informations pertinentes du signal de parole en vue de sa reconnaissance.

Nous avons proposé d'analyser des signaux de la base de données NOIZEUS, sous l'environnement Matlab. Cette base de données permet également de traiter un corpus de parole bruitée par un bruit blanc gaussien.

L'analyse dans le domaine temps-fréquence, du signal de parole nous a paru une approche pertinente qui permet d'appliquer efficacement les algorithmes de reconnaissance de la parole.

Les recherches effectuées dans le domaine de la reconnaissance vocale permettent d'envisager un éventail toujours plus large d'applications industrielles (commande de systèmes, robotique,

aéronautique, renseignement militaire, domaine médical, ...) ou grand public (téléphonie et plus généralement télécommunication, aide aux personnes handicapées, ...).

Les perspectives dans ce domaine restent encore très riches et très importantes.

Les études qu'on a faites dans ce mémoire nous ont a permis, en plus de travail effectué, de découvrir de nouvelles voies des recherche que nous envisageons d'explorer dans le futur.

Références bibliographiques :

[1] A. Bendahmane, « Cours de Traitement Automatique de la Parole », Polycopié de l'USTO, Oran, Algérie, 2014.

[2] S. Nefti, « Segmentation automatique de parole en phones », Doctorat, Université de Rennes 1, Rennes, France, 2004.

[3] J.Tubach, « Reconnaissance automatique de la parole », Doctorat, Université Joseph-Fourier - Grenoble I, Grenoble, France, 1970.

[4] N. Sturmel, « Analyse de la qualité vocale appliquée à la parole expressive », Ecole Doctorale, Université Paris-Sud 11, Paris, France, 2011.

[5] L. Buniet, « Traitement automatique de la parole en milieu bruité », Doctorat, Université Henri Poincaré - Nancy 1, France, 1997.

[6] V. Barreaud, « Reconnaissance automatique de la parole continue », Doctorat, Université Henri Poincaré - Nancy 1, France, 2004.

[7] I. Zitouni, « Modélisation du langage pour les systèmes de reconnaissance de la parole destinés aux grands vocabulaires », Doctorat, Université Henri Poincaré - Nancy 1, France, 2000.

[8] J. Le Grand, « Amélioration des systèmes de reconnaissance de la parole des personnes âgées », Mémoire de Master 2, Université Stendhal, Grenoble, France, 2012.

[9] B. Lecouteux, « Reconnaissance automatique de la parole guidée par des transcriptions a priori », Doctorat, Université d'Avignon, Avignon, France, 2008.

[10] T. Hézard, « Production de la voix », Université Pierre et Marie Curie - Paris VI, Paris, France, 2013.

[11] R. Ajgou, « Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux », Doctorat en sciences en : Génie électrique, Université Mohamed Khider, Biskra, Algérie, 2016.

[12] M. Didiche, «Modélisation neuro-prédictive pour La classification phonétique», Thèse Doctorat, Université Mohamed Khider, Biskra, Algérie, 2014.

[13] G. Richard, « Eléments de Reconnaissance de la Parole pour PACT », Extraits du polycopié de cours de l'UE SI340, Télécom ParisTech, Paris, France, 2012.

[14] C. Hadri, « La recherche des paramètres de la trace acoustique et son application dans la reconnaissance de la parole », université Badji Mokhtar, Annaba, Algérie, 2008.

[15] H. Carfanton, « Traitement numérique du signal », Note de cours, cours de l'Université Paul Sabatier de Toulouse, France, 2003.

[16] M. Kunt, « Traitement numérique des signaux », Traité d'électricité, 2e édition, 1989.

[17] S. Lasaulce, « De la transformée de Fourier à la transformée en ondelettes », chapitre1, Module Ondelettes du DEA TIS, 2010.

[18] J. Dumas, « L'analyse temps – fréquence », (Document réalisé par: Groupe MVI technologies), limonest Lyon, Version février 2001.

[19] B. Meriane, « Analyse du Signal de Parole par Les Ondelettes », Thèse de de Magister, Université de Batna, Algérie, 2009.

[20] J.P. Ovarlez, « Distributions Temps-Fréquence », Stage ONERA, Châtillon, France, 1997.

[21] F. Mudry, « Cours de Traitement du Signal », École d'ingénieurs du canton de Vaud, Suisse, 2004.

[22] P. Loizou, « NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms », http://ecs.utdallas.edu/loizou/speech/noizeus/

(Consulté le 30.07.2017).

[23] Y. Ben ayad, « Détection de mots clés dans un flux de parole », Interface homme-machine [cs.HC]. Télécome ParisTech, France, 2003.

[24] Fréjus A. A. Laleye, « Contributions à l'étude et à la reconnaissance automatique de la parole en Fongbe », Thèse de Doctorat, Université d'Abomey-Calavi et Université du Littoral Côte d'Opale, France, 2016.

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

[26] V. Lê, « Reconnaissance automatique de la parole pour des langues peu dotées », doctorat, Discipline : Informatique, Université Joseph-Fourier - Grenoble I, Grenoble, France, 2006.

[27] L. Barrault, « Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole », École Doctorale 166 I2S « Mathématiques et Informatique », Université d'Avignon, Avignon, France, 2008.

[28] L. Orosanu, « Reconnaissance de la parole pour l'aide à la communication pour les sourds et malentendants », École doctorale IAEM Lorraine, Université de Lorraine, Metz et Nancy, France, 2015.

[29] Calliope, « la Parole et son Traitement Automatique, Masson, Paris », 1989.

[30] L. Rabiner, S. Levinson, A. Rosenberg, and J. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, pp. 336-349, 1979.

[31] S. Levinson, L. Rabiner, A. Rosenberg, and J. Wilpon, "Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, pp. 134-141, 1979.