الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-

جامعة البويرة

Faculté des Sciences et des Sciences Appliquées

كلية العلوم والعلوم التطبيقية

Référence : ...../MM/2021

المرجع:...../م م/ 2021

Mémoire de Master

Présenté au

Département : Génie Électrique
Domaine : Sciences et Technologies
Filière : Télécommunications
Spécialité : Systèmes des Télécommunications

Réalisé par :

Ben Mesli Nour Nihad
Et
Si Ahmed Cylia

Thème

Analyse et caractérisation d'un signal sonore via différentes techniques

Soutenu le: **06/11/2021**

Devant la commission composée de :

ISSAOUNI Salim	M.A.A	Univ. Bouira	Président
REZKI Mohamed	M.C.A	Univ. Bouira	Rapporteur
BENSMAIL Samia	M.C.B	Univ. Bouira	Examinateur

Année Universitaire: 2020-2021

Dédicaces :

Nous tenons à dédier ce travail à nos chers parents qui ont tout fait pour qu'on arrive à ces moments, nos frères et sœurs, mon mari Hichem et ma belle-famille à nos amis Sarah Reguieg, Imane Ragueb et Amina Boualem, Farouk et Latif. Nous le dédions aussi à l'ensemble des professeurs de l'université Akli Mohand Oulhadj.

Remerciements

Ce travail a été effectué au sein du Département des Sciences et sciences appliquées de l'Université de Bouira.

Nous tenons tout d’abord à remercier Dieu le tout puissant de m’avoir donné la volonté et le courage et de pouvoir terminé mon cursus universitaire.

Toute notre gratitude à notre encadreur Dr. Rezki pour son soutien, ses conseils, son assistance, patience et encouragement tout au long de cette période de travail.

Nous tenons à remercier également Dr Ali Mohad pour son aide et sa patience et toute l’équipe pédagogiques.

Un grand merci à nos chères familles pour leurs patiences, leurs aides et surtout leurs encouragements.

On remercie également tous les membres du jury pour l’intérêt qu’ils ont porté à mon travail :

Enfin, nous associons à ces remerciements tous ceux qui ont contribué à réaliser ce travail.

Résumé

Le traitement de signal a pour objet l'élaboration ou l'interprétation des signaux. Son champ d'application se situe donc dans tous les domaines concernés pour la perception, la transmission ou l'extraction des informations.

L'étude d'un signal audio permet d'extraire plusieurs paramètres selon le but d'étude. Dans notre cas, nous souhaitons détecter le type de genre. De ce fait, on cherche à extraire certains paramètres comme la fréquence fondamentale (Pitch), les formants et l'énergie par différentes méthodes dont certains sont temporelles et d'autre fréquentielles. L'analyse et la caractérisation d'un signal audio se base sur l'étude de ses paramètres.

L'étude expérimentale va nous mener à notre objectif qui se focalise sur la reconnaissance du genre.

Mots clés : traitement de signal, genre, pitch, formant, énergie, signal audio.

Abstract

The purpose of signal processing is the development or interpretation of signals. Its scope is therefore in all areas concerned for the perception, transmission or extraction of information.

The study of an audio signal makes it possible to extract several parameters depending on the study goal. In our case, we want to detect the gender type. Therefore, we try to extract certain parameters such as fundamental frequency (Pitch), formants and energy by different methods, some of which are temporal and others frequency. The analysis and characterization of an audio signal is based on the study of its parameters.

The experimental study will lead us to our goal which focuses on gender recognition.

Keywords: signal processing, genre, pitch, formants, energy, audio signal.

ملخص

الغرض من معالجة الإشارات هو تطوير أو تفسير الإشارات. وبالتالي فإن نطاقه يشمل جميع المجالات المعنية بإدراك المعلومات أو نقلها أو استخراجها.

تتيح دراسة الإشارة الصوتية استخراج العديد من المعلمات حسب هدف الدراسة. في حالتنا، نريد اكتشاف نوع الجنس. لذلك نسعى لاستخراج معلمات معينة مثل التردد الأساسي والصيغ والطاقة بطرق مختلفة بعضها مؤقت وبعضها تردد. يعتمد تحليل الإشارة الصوتية وتوصيفها على دراسة معلماتها.

ستقودنا الدراسة التجريبية إلى هدفنا الذي يركز على التعرف على الجنس.

الكلمات المفتاحية: معالجة الإشارات، النوع، التردد الأساسي، التشكيل، الطاقة، الإشارة الصوتية.

Sommaire

Chapitre 2

Chapitre 3

Liste des figures :

Liste des tableaux :

Liste des symboles et abréviations :

ZCR : zeros crossing rate.

AD : amplitude descriptor .

ADSR: attack Decay Sustain release.

RMS: root mean square energy.

FFT: fast fourier transform (transformée de fourier rapide).

IFFT : inverse fast fourier transform.

DCT : La transformation de cosinus discrète.

LPC : coefficient de prédiction linéaire.

SIFT : simplified inverse filter tracking algorithm.

F : fréquence du signal.

Fs: fréquence d'échantillonage.

F0: fréquence fondmentale.

$\bar{E}$: énergie moyenne.

W : la bande passante.

C_l : seuil d'écrêtage.

Log : fonction logarithme.

TIMIT: Texas Instruments/Massachusetts Institute of Technology.

Introduction générale

La parole apparait 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 premièrement en signal électrique. De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut être soumis à un ensemble de traitements statistiques qui visent en évidence les paramètres acoustiques : sa fréquence fondamentale, son énergie et son spectre.

Le signal de parole n'est pas un signal ordinaire. Il est difficile de le modéliser car ses propriétés statistiques varient au cours du temps. Sa redondance lui confère une robustesse à certains types de bruit.

La variabilité interlocuteur est évidente, la hauteur de la voix, l'intonation, l'accent différent selon le sexe…etc.

L'analyse de signal vocal constitue dans tout système de reconnaissance, de synthèse ou compréhension de la parole.

Notre travail se divise en trois chapitres :

- Le premier chapitre présente une étude et analyse de la parole (production de la parole, spectre, harmoniques...).
- Le deuxième chapitre explore les différentes méthodes pour l'extraction des paramètres d'un signal audio qui nous permet de détecter le genre du locuteur.
- Le troisième chapitre présente une étude comparative des méthodes d'extraction en présentant les étapes de chaque méthode.

Enfin, c'est avec une conclusion générale et perspectives qu'on termine notre mémoire.

Chapitre 1

Analyse de la parole

1. Introduction

Parole et voix sont envisagées de façon pluridisciplinaire, sous des aspects d'ordre physique (traitement du signal, acoustique) linguistique (phonétique, phonologie, syntaxe) ou psychologique (perception et cognition)

Le signal de la parole permet la communication entre humains. Il permet de communiquer la pensée par un système de sons articulés émis par les organes de la phonation et qui est produit par deux processus différents qui sont la vibration des cordes vocales et la turbulence créée par l'air au niveau du conduit vocal **[1]**

Lors de la production, c'est le passage de l'air dans le conduit vocal qui va induire la résonance du canal, et générer un groupe de résonances qui sont appelées les formants. **[1]**

2. Le signal parole :

On définit la parole et on explore les caractéristiques et la modélisation du mécanisme de production de la parole.

2.1 Qu'est-ce que c'est la parole ?

La parole est le principal moyen de communication dans toute société humaine. Son abstraction par rapport à un support physique en fait un moyen de communication très simple à utiliser : il est plus facile de parler à quelqu'un que de lui écrire ou de lui faire un schéma. L'ère industrielle a par ailleurs permis de mettre en place des moyens d'enregistrement, et donc de sauvegarder, de se hisser au rang de l'écrit pour la conservation de la connaissance [2]. L'information portée par le signal parole peut être analysée de différentes façons.

2.2 Synthèse articulatoire

Pour produire de la parole, l'être humain met en mouvement ses organes phonatoires (poumons et cordes vocales) et les articulateurs qui modèlent la forme de son conduit vocal (mandibule, langue, lèvres et velum). La génération d'un signal acoustique de parole perceptible nécessite donc une coordination complexe et précise des différents organes, dans l'espace et dans le temps, et implique le recrutement de plusieurs muscles.

2.3 Fonctionnement de l'appareil vocal :

L'ensemble du système vocal se compose des poumons et du conduit trachéobronchique, du larynx, et du conduit vocal, formé par le pharynx et les cavités nasales

et orales [2]. La figure 1 représente l'appareil phonatoire et modèle mécanique de production de la parole.

- ♣ L'ensemble poumons et conduit trachéobronchique se comporte comme un générateur d'air qui alimente le larynx.
- ♣ Le larynx est l'ensemble de cartilages articulés, ligaments, muscles et muqueuses, qui, grâce à son action sur les cordes vocales (muscles élastiques), permet de déterminer la nature du flux d'air qui va exciter le conduit vocal.
- ♣ Le conduit vocal, par des organes articulateurs, imprime au son émis les caractéristiques spécifiques permettant de distinguer les différents phonèmes et ceci en tant que :
  - résonateur de l'onde glottique pour la production des voyelles.
  - générateur de bruit pour la production des consonnes.

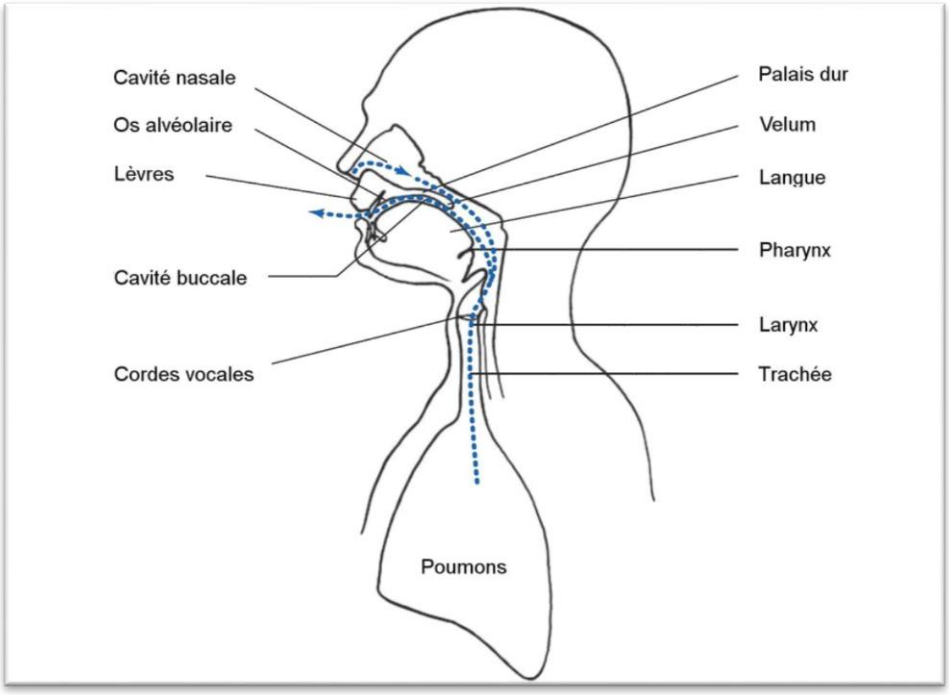


Figure n°1 : Appareil phonatoire.

Un synthétiseur articulatoire, comme d'autres modèles source-filtre de production de la parole, consiste en trois éléments : une source d'excitation glottale, un modèle des propriétés acoustiques du conduit vocal et un modèle des effets du rayonnement de l'air au niveau des deux cavités labiale et nasale. Un modèle des plis vocaux à deux ou plusieurs masses, comme celui montré sur la figure n°2, est souvent utilisé pour fournir le signal d'excitation glottale pour les sons voisés. Dans ce modèle le mouvement des cordes vocales est simulé par les masses couplées par des ressorts [3,4].

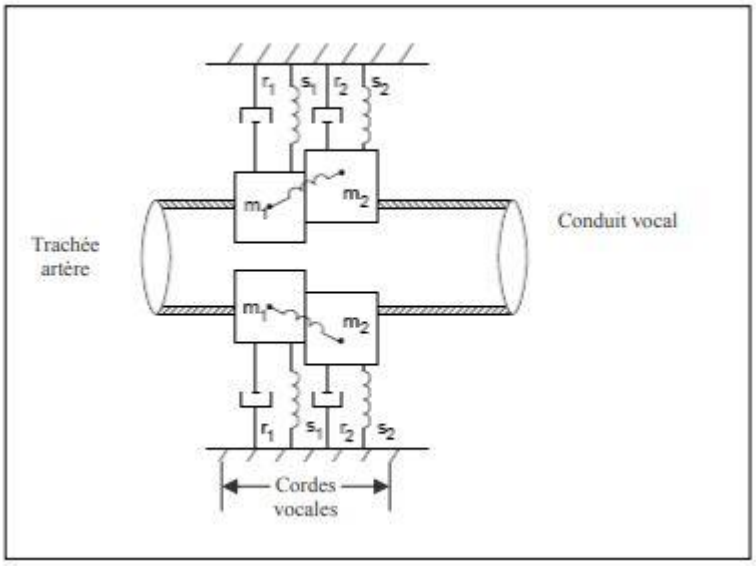


Figure n°2 : Modèle articulatoire des plis vocaux à 2 masses [4]

2.4 Caractéristiques phonétiques :

Les caractéristiques phonétiques sont : les phonèmes, voyelles, consonnes.

2.5 Phonème [2] :

Un phonème est la plus petite unité présenté dans la parole [4]. Le nombre de phonèmes est toujours très limité (normalement inférieur à cinquante) et ça dépend de chaque langue.

2.5.1 Voyelles [2] :

Les voyelles sont des sons voisés qui résultent de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales. Il y a deux types de voyelle : les voyelles orales (i, e, u, ...) qui sont émises sans intervention de la cavité nasale et les voyelles nasales (ã, e~, …) qui font intervenir la cavité nasale. Chaque voyelle se caractérise par les résonances du conduit vocal qu'on appelle "les formants". En général, les trois premiers formants sont suffisants pour caractériser toutes les voyelles.

2.5.2 Consonnes [2] :

Les consonnes sont des sons qui sont produits par une turbulence créée par le passage de l'air dans une constriction du conduit où une source périodique liée à la vibration des cordes vocales s'ajoute à la source de bruit (les consonnes voisées).

Les paramètres acoustiques de la parole : La parole est constituée d'une succession de phonèmes, c'est la plus petite unité phonatoire susceptible de changer un mot en un autre, la langue française comporte 37 phonèmes.

Les traits ou indices acoustiques d'un signal de parole sont son formant, sa fréquence fondamentale (ou Pitch), son énergie et son timbre.

3.1 Le voisement :

Il est produit par la vibration des cordes et le rapprochement des cartilages aryténoïdes du larynx **[1].** Il caractérise les voyelles et certaines consonnes voisées contrairement aux consonnes sourdes qui sont caractérisées par le non voisement. Ce dernier est dû à l'écartement des cordes vocales et l'ouverture totale de la glotte. L'air s'échappe sans être mis en oscillation par les cordes vocales.

3.2 Le Pitch (fréquence fondamentale) :

La quasi-périodicité du son voisé dans le domaine temporel devient un pic centré sur une fréquence (l'inverse de cette période) appelée fréquence fondamentale (Pitch). Cette information du pitch est utilisée par plusieurs applications comme l'identification du locuteur, la synthèse de la parole, le codage, etc.

La fréquence fondamentale (Pitch) est personnelle pour chaque locuteur et varie selon l'âge et le sexe. Elle est de :

- 50 à 200 HZ pour la voix masculine,
- 180 à 450 HZ pour la voix féminine
- 200 à 600 HZ pour la voix d'enfant **[5]**

3.3 Le formant :

Un formant est une fréquence résonnante du système acoustique. Le formant se caractérise par la présence de maxima spectraux, c'est-à-dire des zones où les harmoniques sont intenses. Il est utilisé généralement dans la phonétique ou l'acoustique pour décrire les vibrations des tractus vocaux ou des instruments musicaux.

- La fréquence du premier formant peut varier de 200 Hz à 800 Hz, celle du second formant de 900 Hz à 2400 Hz. Il existe des formants d'ordres supérieurs pouvant aller jusqu'à 5 KHz ; l'ensemble des formants contribue en particulier à caractériser le « timbre » de la voix.

3.4 L'énergie :

L'énergie d'un son est liée à la pression de l'air en amont du larynx et caractérise son intensité.

3.5 Le timbre :

Le timbre est la caractéristique d'un son permettant de le différencier d'un autre son.

4. Propriétés statistiques du signal parole :

Le signal de parole est une réalisation particulière d'un processus aléatoire non stationnaire c'est-à-dire que ces propriétés statiques changent au cours du temps. Nous faisant l'hypothèse de quasi-stationnarité sur des périodes allants de 10 à 40 ms [6].

4.1 Détection du voisement :

L'étude du signal de parole a montré que l'analyse spectrale constitue une indication importante pour la détection de voisement, les sons voisés contrairement aux sons non voisé présentant d'avantage d'énergie vers les basses fréquences, ainsi que le calcul de nombre d'échantillons successifs signe opposés constitue un deuxième indicateur de voisement ou non voisement du signal parole. Pour le taux de passage par zéro pour un segment de signal de « N » échantillons, la formule est représentée comme suit [7,8] :

$$ZCR=\sum|sgn[x(n)]-sgn[x(n-1)]| \quad (1)$$

L'énergie pour le même segment de signal est [7] :

$$\bar{E}= \sum_{n=0}^{N-1} X_m^2\,(n) \quad (2)$$

4.2 Spectrogramme :

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

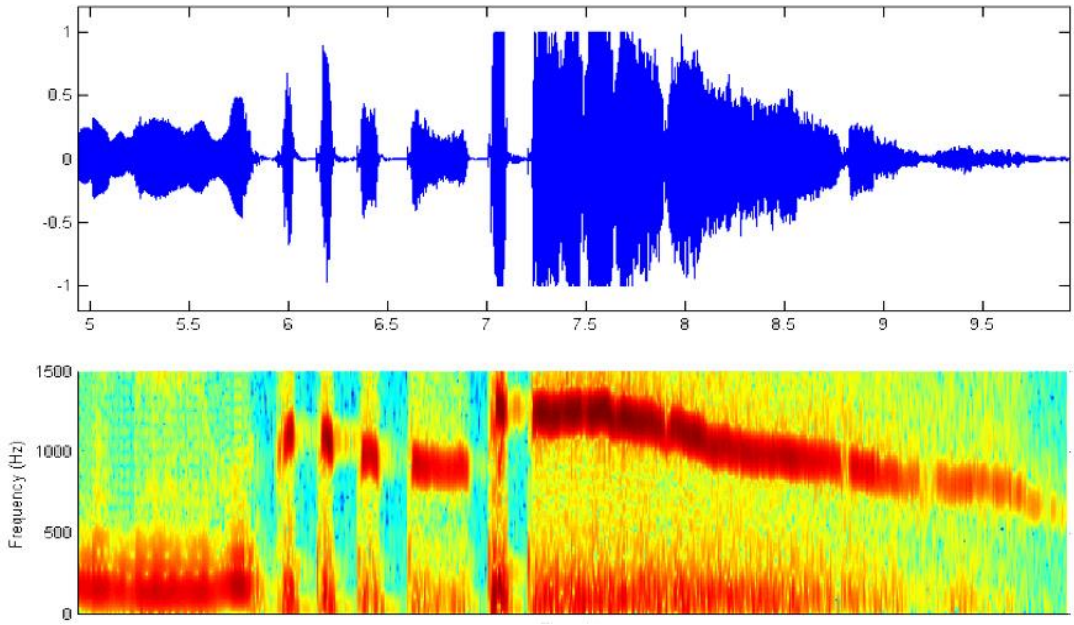


Figure n°3 : spectrogramme d'un signal audio

Le spectrogramme est un outil de visualisation utilisant la technique de la transformée de Fourier. Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné.

4.3 Analyse et extraction des paramètres d'un signal audio :

L'extraction des paramètres joue un rôle très important dans l'analyse d'un signal audio. Les paramètres audio peuvent être sous catégorisées en 4 domaines : domaine temporel, domaine fréquentiel, domaine temps-fréquence commun et paramètres approfondis [9].

Dans ce travail nous ne sommes intéressés que par le domaine temporel et le domaine fréquentiel.

4.3.1. Domaine temporel :

Avant de parler sur les paramètres de ce domaine il faut savoir qu'un signal audio est un signal non stationnaire. Ce type de signaux est difficile à étudier, donc on applique la technique de fenêtrage (windowing) pour diviser ce signal en plusieurs signaux généralement de 20ms pour avoir un signal quasi stationnaire. Parmi les plusieurs types de fenêtres on utilise le Hamming et le Hanning window pour éviter la distorsion de signal lors de l'analyse.

4.3.1.1. Zero-Crossing rate :

C'est le taux de changement de signe des successifs échantillons d'un intervalle de temps, autrement dit, c'est le nombre de fois où le signal change de valeur, de positif à négatif et vice versa, divisé par la longueur de la trame [10].

$$Z(n) = (1/(2*L))* \sum \left| sgn[x(m)] - sgn[x(m-1)] \right| \quad (3)$$

Où $sgn[x(n)] = 1$ si $x(n) \geq 0$

$= -1$ si $x(n) < 0]$

Et L est la longueur de la trame.

Le ZCR est égal au double de la fréquence du signal. Par conséquent, nous pouvons dire que ZCR donne des informations indirectes sur la fréquence du signal f_0.

Prédiction linéaire ZCR: un type de ZCR, C'est le rapport entre le ZCR du signal d'origine et le ZCR de l'erreur de prédiction obtenue à partir d'un filtre de prédiction linéaire.

4.3.1.2 Paramètres basé sur l'amplitude :

C'est une analyse très simple. Les caractéristiques sont basées sur l'analyse de l'enveloppe temporelle du signal. Trois types sont mentionnés ci-dessous :

- **Amplitude descripteur (AD)** : il permet de distinguer les sons avec différentes enveloppes de signaux, étant appliqué, par exemple, pour la discrimination des sons d'animaux. Il est basé sur la collecte de l'énergie, de la durée et de la variation de durée des segments de signal en fonction de leur amplitude haute et basse au moyen d'un seuil adaptatif [11].
- **ADSR enveloppe** : attack,Decay,Sustain,release représente les quatre stages qui décrit la forme d'enveloppe d'un signal .

  Attack : c'est le temps pris pour passer de 0 vers l'amplitude max de signal.

  Decay : c'est le temps de passage de l'amplitude max vers un niveau de tenue défini.

  Sustain : détermine le niveau d'amplitude maintenu lorsque la touche est tenue enfoncée.

  Release : détermine le temps de passage de l'amplitude du niveau de tenue défini à une amplitude nulle lorsque la touche est relâchée.
- **Shimmer :** Il calcule les variations cycle à cycle de l'amplitude de la forme d'onde. Il est défini comme la différence absolue moyenne entre les amplitudes de périodes consécutives divisée par l'amplitude moyenne.

  Il est utilisé dans la détection d'activité vocale, la reconnaissance du locuteur, la vérification du locuteur [12], la classification des sons musicaux [13].

4.3.1.3 Paramètres basés sur l'énergie :

Nous mentionnons la fonctionnalité la plus utilisée « rms ».

- The root mean square energy : c'est de calculer la valeur moyenne de tous les échantillons dans chaque trame. Il est plus utilisé pour classifier les parties vocales / non vocales d'un signal.

La partie vocale d'un signal audio a une énergie élevée en raison de sa périodicité et la partie non vocale du discours à une énergie faible. [14]

4.3.2 Domaine spectral:

Un signal audio est représenté en fonction du temps. Pour analyser un signal en fonction de la fréquence, il faut appliquer la transformée de Fourrier (FFT sur Matlab).

L'analyse du domaine fréquentiel est un outil de grande importance dans le traitement du signal audio grâce au grand nombre des algorithmes qui peuvent être utilisés pour extraire des paramètres.

Il y a beaucoup de fonctionnalités, mais nous discuterons de quelques-unes des caractéristiques importantes dans ce domaine.

4.3.2.1 Coefficients de prédiction linéaire :

C'est l'une des techniques d'analyse de la parole la plus puissante et une méthode utile pour coder une parole de qualité à faible débit. L'idée de base de l'analyse prédictive linéaire n'est qu'un échantillon de parole spécifique à l'heure actuelle pouvant être approximé comme une combinaison linéaire d'échantillons de parole passés.

Il supprime la redondance d'un signal et essaie de prédire les valeurs suivantes en combinant linéairement les coefficients précédents connus [9].

4.3.2.2 Band energy ratio :

Divise le signal en différentes bandes de fréquences généralement en utilisant STFT (Short-Time Fourier Transform), puis divise l'énergie des bandes de fréquences inférieures sur l'énergie des bandes de fréquences plus élevées. Il est généralement utilisé pour la discrimination de la parole ou de la musique.

4.3.2.3 Centroïde spectral :

IL est défini comme le centre de gravité du spectre de magnitude du STFT. Le centre de gravité est une mesure de la forme spectrale et des valeurs de centre de gravité plus élevées correspondent à des textures «plus claires» avec des fréquences plus élevées [15]. En d'autres

termes, il nous donne la bande de fréquences où se concentre le plus d'énergie, sa formule mathématique est :

$$s_t = \frac{\sum_{n=1}^{N} n * m_{t\,(n)}}{\sum_{n=1}^{N} m_{t\,(n)}} \quad (4)$$

où mt est l'amplitude de la transformée de Fourier à une trame et t une bande de fréquence n

4.3.2.4 La bande passante (Bandwidth) :

Elle est liée au centre de gravité spectral. On peut dire que c'est la variance par rapport au « spectral centroid », il décrit le timbre perçu, en d'autres termes, c'est la mesure de la gamme de fréquences présente dans le signal.

$$bw_t = \frac{\sum_{n=1}^{N} |n - s_t| * m_{t\,(n)}}{\sum_{n=1}^{N} m_{t\,(n)}} \quad (5)$$

Ou (n-st) est la distance entre la bande de fréquence (frequency bin) et le « spectral centroid » [16].

5. Conclusion :

Nous avons pu voir au cours de ce chapitre, le phénomène de la production de la parole, les différentes sources permettant la génération des sons.

Nous avons aussi remarqué que le signal vocal est très complexe, du fait de sa grande variabilité, ce qui rend toute tentative de le modéliser ou de reconnaître très délicate.

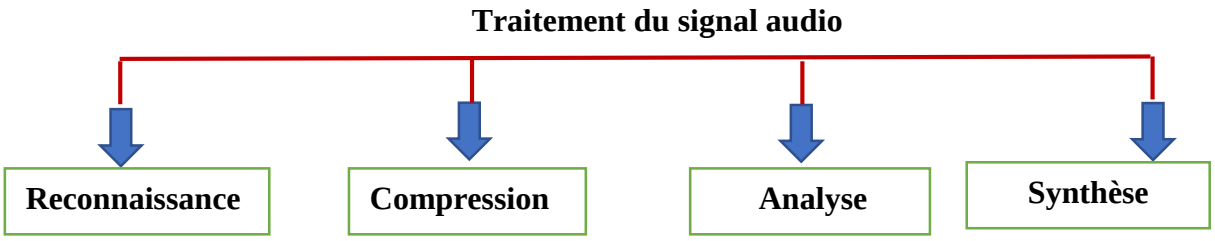
Un signal de parole est une séquence de sons correspondant à une suite d'états de l'appareil phonatoire. Le signal de parole est un processus aléatoire non stationnaire à long terme.

Après avoir vu les différentes notions concernant l'analyse de la parole et de signal audio en général, nous allons découvrir en détails les différentes techniques pouvant être utilisé dans la caractérisation du signal sonore. Ce qui fait l'objet du chapitre qui suit.

Chapitre 2 :

Approche théorique sur la caractérisation de la parole

Approche théorique sur la caractérisation de la parole



1. Introduction

On explore les différents algorithmes et ses classes pour la détection du pitch. Où certain nombre de problèmes se posent, notamment parce que les signaux réels ne sont pas à proprement parlé périodiques, et aussi parce que le paramètre à estimer est variable au cours du temps.

La détection du période du pitch à partir du signal vocal présente une importance considérable dans la reconnaissance, identification du speaker, codage de la parole et sa synthèse. Cependant la détermination de la période du pitch du signal vocal est difficile suite à la complexité du signal vocal qui est considéré comme pour le signal voisé, un signal issu de la sortie d'un système variant dans le temps excité par des trains d'impulsions quasi périodique. Le problème est donc de déterminer la période du signal d'excitation du signal voisé.

Le signal vocal fait partie des signaux non stationnaires. Il s'avère que l'outil de la transformation de Fourier ne résout pas la représentation des signaux non stationnaires. Parmi les méthodes classiques de détection de la période du pitch, la méthode d'auto corrélation qui donne des résultats moins performants dans cette détection, méthode cepstrale, LPC et AMDF …etc.

2. Nécessité de la fréquence fondamentale :

La fréquence fondamentale est parmi les paramètres acoustique d'un signal parole qui sont nécessaire à extraire dans plusieurs domaines comme la synthèse, reconnaissance automatique de la parole, de codage nécessaire aux transmissions (codage LPC), la fréquence fondamentale est par définition l'inverse de la période de vibration des cordes vocales, elle est appelée aussi le pitch.

L'extraction du pitch n'est pas une tâche facile pour les trois raisons suivantes :

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

- Séparation source/conduit vocale.
- La plage de dynamique de la fréquence fondamentale est très grande.

3. Extraction des paramètres :

L'analyse du signal vocal constitue dans tout système de reconnaissance, de synthèse ou de compréhension de la parole, l'étape préliminaire et primordiale.

L'objectif principal d'analyse du signal vocal est d'extraire certains paramètres pertinents tels que : le pitch, les formants et l'énergie. Les méthodes de traitement du signal vocal se sont de plus en plus affinés et sont des plus simples comme l'énergie ou le taux de passage par zéro vers les plus complexes comme le spectre court-terme et le codage prédictif linéaire (LPC), la méthode AMDF pour le calcul de pitch.

On peut classer grossièrement les méthodes d'analyse en deux groupes : méthodes temporelles et méthodes spectrales.

3.1 Extraction de Pitch :

3.1.1 Méthode temporelle :

Il existe différentes méthodes temporelles pour la détection du fondamental, citons : l'autocorrélation, l'autocorrélation modifiée, taux de passage par zéro, LPC et l'analyse mixte SIFT (Simplified Inverse Filter Tracking algorithm) [15, 16].

3.1.1.1 Analyse par autocorrélation :

Cette méthode est basée sur la détection des maximas de la fonction d'autocorrélation d'un signal. Les positions de ces maximas nous informe sur l'existence du fondamental d'un signal. On calcule la fonction d'autocorrélation sur une tranche de N échantillons qui recouvre plusieurs périodes du fondamental. La détection de pitch par autocorrélation reste l'un des détecteurs robustes.

La procédure à suivre pour cette méthode peut être résumée comme suit :

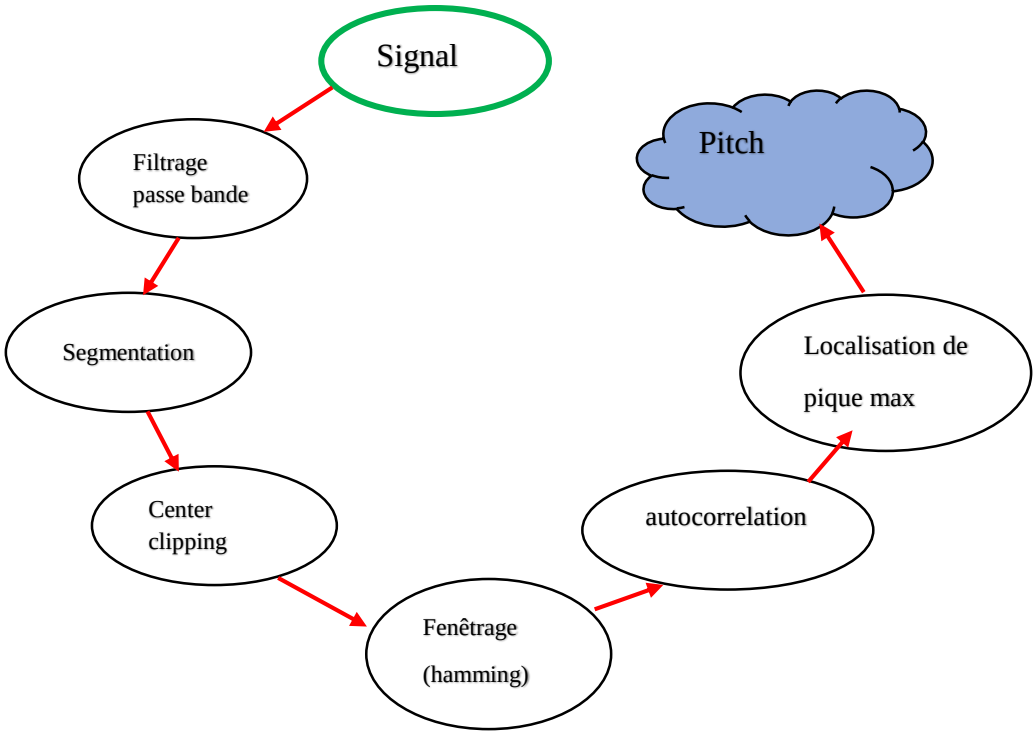


Figure n°4 : Organigramme de l'autocorrélation.

Cet organigramme nous démontre les différentes étapes à suivre pour calculer la valeur de pitch avec la méthode d'autocorrélation

Explication de l'organigramme :

- **Filtrage :** On a utilisé un filtre butterworth d'ordre 4 où la bande passent est de fs/50 à fs/500 car le pitch est inclus entre 50 et 500 Hz pour les hommes, les femmes et les enfants. et on fait toujours la segmentation où la taille d'une trame est de 20 ms donc le nombre d'échantillons est (fs*0.02).
- **Center clipping :** Pour réduire les effets de la structure des formants sur la forme détaillée de l'autocorrélation à court terme et obtenir une forme d'onde utile, on utilise un traitement non linéaire.

  L'une de ces techniques non linéaires est 'center clipping'. La relation entre l'entrée x(n) et y(n) est :

$$y(n) = clc[x(n)] = \begin{cases} (x(n) - c_l) \ , x(n) \geq c_l \\ (x(n) + c_l) \ , x(n) \leq c_l \\ 0 \ si \ non \end{cases} \quad (1)$$

où c_l est le seuil d'écrêtage (clipping).

Généralement c_l est d'environ 30% de l'amplitude maximale du signal. En application, la CL doit être aussi élevée que possible. Pour obtenir le c_l élevé, on peut capturer la valeur

crête du premier 1/3 et du dernier 1/3 du signal et utiliser le moins pour qu'elle soit l'amplitude maximale. Ensuite, on définit les 60-80% de cette amplitude maximale comme c_l.

- ♣ **Fenêtrage de Hamming :** après avoir filtré le signal et afin d'éviter la distorsion on utilise la fenêtre de Hamming.

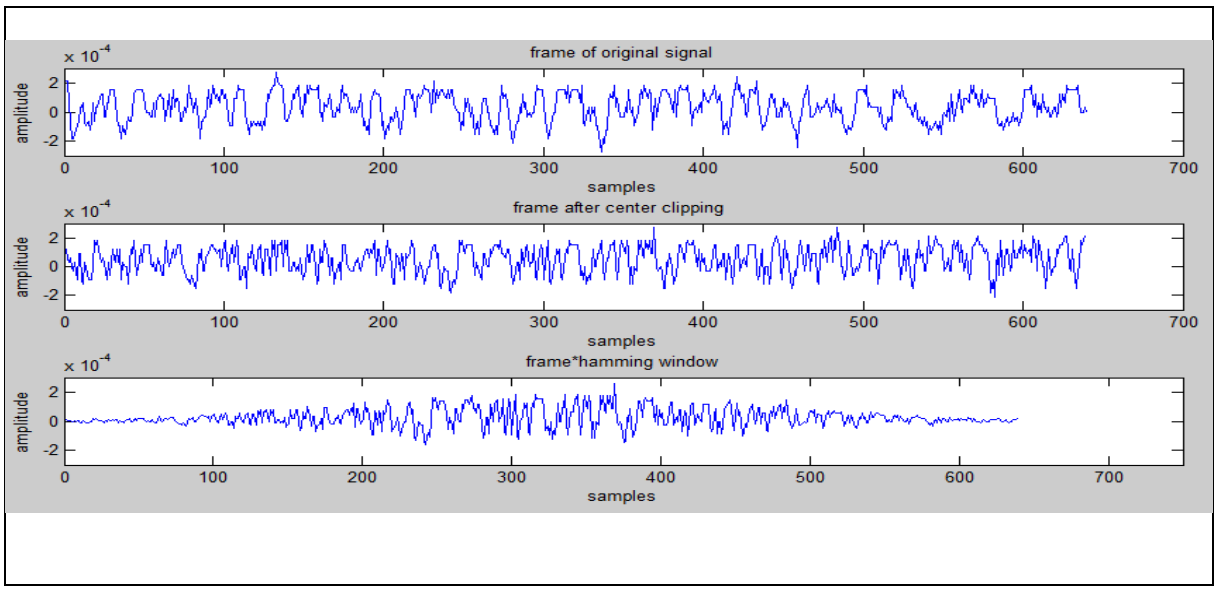


Figure n°5 : Le signal originale d'une trame, signal après l'application de center clipping et signal après application de fenêtre de Hamming.

Après l'application de l'autocorrélation sur le signal on trouve l'emplacement du pic max dans la plage (fs/ 500 à fs/50). L'inverse du décalage (1/lag) est multiplié par la fréquence d'échantillonnage fs et on obtient le pitch.

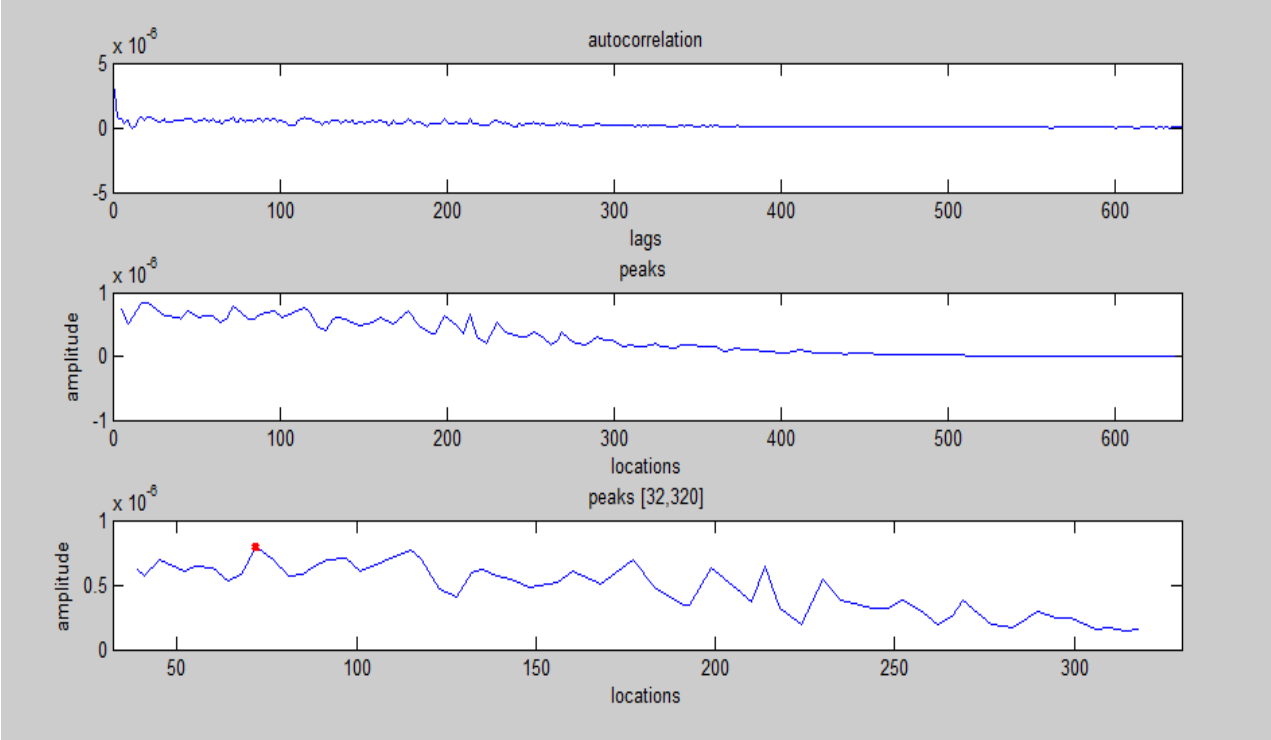


Figure n°6 : 1/2 de signal d'autocorrélation, les piques de signal d'autocorrélation, les piques dans l'intervalle [32,320].

Il y a aussi la méthode AMDF qui est un type d'autocorrélation que nous allons expliquer son fonctionnement d'une manière générale.

3.1.1.2 La méthode AMDF

C'est un autre type d'analyse d'autocorrélation. Au lieu de corréler la parole d'entrée à divers retards (où des multiplications et des sommations sont formées à chaque valeur), un signal de différence est formé entre la parole retardée et l'original, et à chaque valeur de retard l'amplitude absolue est prise. Pour la base de N échantillons, la fonction de différence à court terme AMDF est définie comme :

$$D_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)| \qquad (2)$$

où $x(n)$ sont les échantillons de trame analysée, $x(n+m)$ sont les échantillons décalés dans le temps sur m échantillons et N est la longueur de trame. [17]

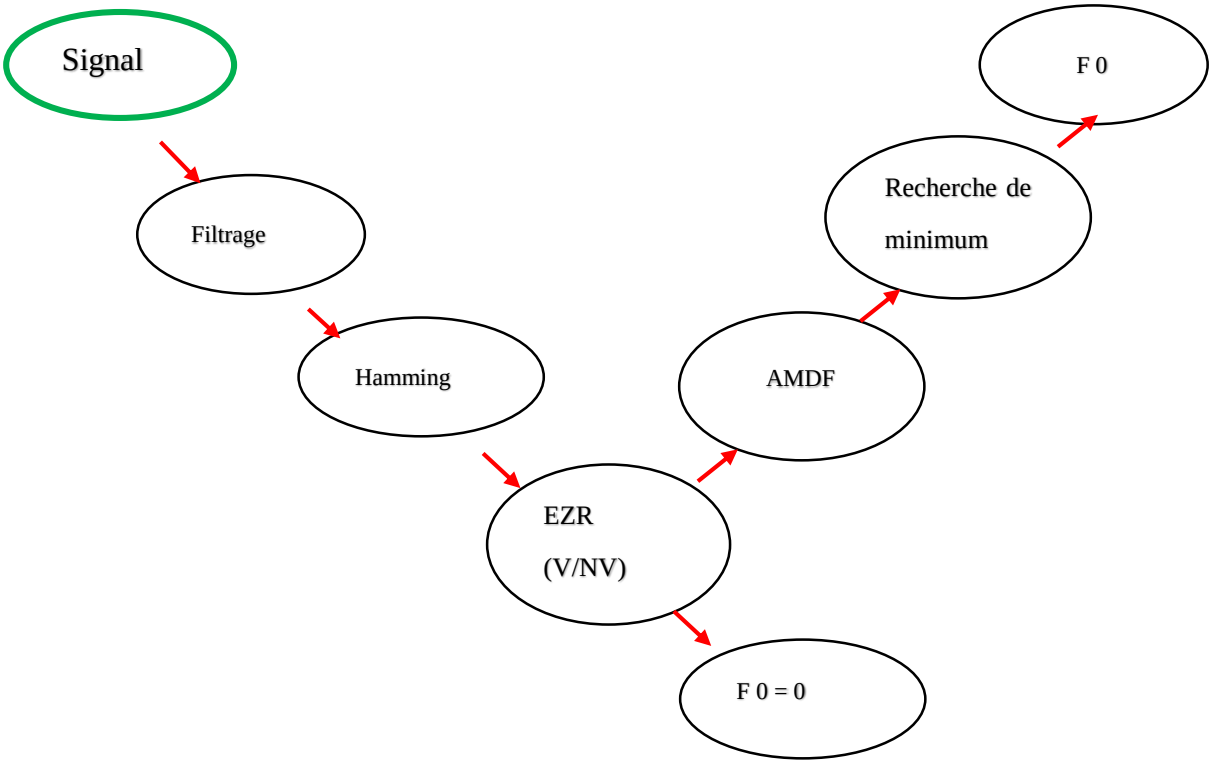


Figure n°7 : bloc diagramme de la méthode AMDF.

Après segmentation, le signal est prétraité pour supprimer les effets des variations d'intensité et du bruit de fond par filtrage passe-bas. Ensuite, la fonction de différence d'amplitude moyenne est calculée sur le segment de parole avec des décalages allant de 32 à 320 échantillons pour fs=16000 Hz. La période de pitch est identifiée comme la valeur du décalage auquel l'AMDF minimum se produit. En plus de l'estimation de pitch, le rapport entre les valeurs maximales et minimales d'AMDF (MAX/MIN) est obtenu. Cette mesure avec l'énergie de trame est utilisée pour prendre une décision si la trame est voisée ou non voisée. [18]

3.1.2 Méthodes spectrales :

3.1.2.1 Analyse par la méthode cepstrale :

La méthode de « cepstre » est une technique fréquentielle pour déterminer la fréquence du signal, L'idée fondamentale derrière le cepstre est que le signal périodique peut être considéré comme la convolution d'un train d'impulsions par un filtre amorti. Dans le domaine des fréquences, les spectres sont multipliés mais en prenant le « log » du résultat, on obtient la somme des résultats indépendants. De cette façon, une convolution dans l'espace des temps

correspond à une addition dans le domaine du cepstre: Si les deux spectres ont des caractéristiques différentes, il devient possible de les séparer.

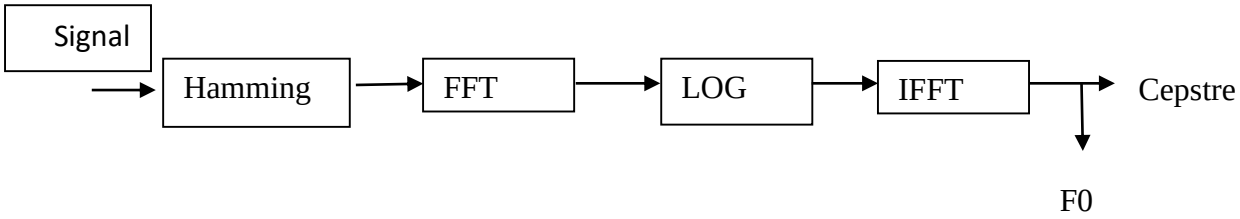


Figure n°8 : Bloc organigramme de la méthode cepstrale.

L'organigramme ci-dessus nous démontre les étapes de la méthode cepstrale en général.

Nous avons travaillé avec ce bloc néanmoins nous avons rajouté quelques étapes comme la fenêtre liftering pour améliorer nos résultats, comme elles sont montrées dans la figure suivante :

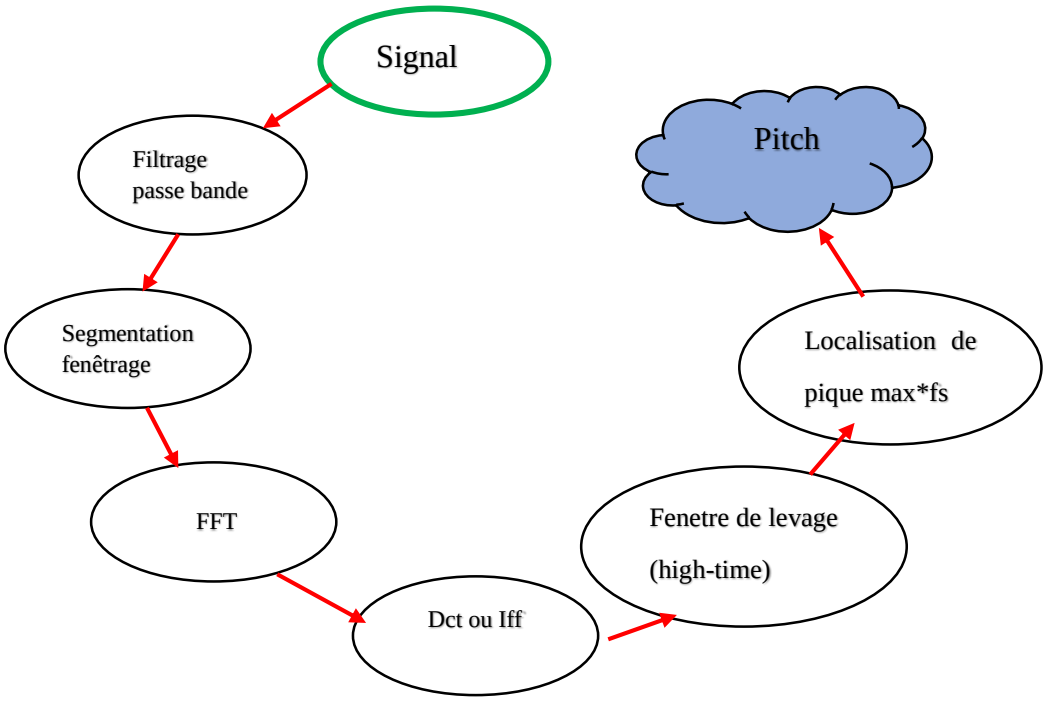


Figure n°9 **:** Organigramme de la méthode de cepstre.

Cet organigramme nous démontre les différentes étapes à suivre pour calculer la valeur de pitch avec la méthode cepstrale.

Explication du schéma :

- **Filtrage :** On a utilisé un filtre passe bande (butterworth) d'ordre 8 et sa bande passante est de 50 à 450Hz. La fréquence minimale est 50Hz et la fmax est 500Hz.

Cette étape nous permet d'enlever ou atténuer les harmoniques afin que le pique correspond à la fraquence fondamentale (Pitch) .

- ♣ **Segmentation :** Le signal parole est par nature non stationnaire. On procède à la segmentation de ce signal, dans la plage de 20 à 40 ms, pour le rendre stationnaire. Ces plages sont appelées trames. Elles sont divisées en N échantillons et séparées par une fréquence f telle que f est inférieure à N.
- ♣ **Fenêtrage :** Pour observer un signal dans un temps fini, nous le multiplions par une fonction de fenêtre [19]. La fenêtre de Hamming est utilisée pour éviter la distorsion.
- **Transformée de Fourier Rapide (FFT) :** FFT un algorithme efficace capable d'effectuer une transformation de Fourier, pour convertir chaque trame de N échantillons du domaine temporel en domaine fréquentielle.
- **Transformée de Fourier inverse (IFFT) :** L'IFFT résulte le "cepstre", l'estimation de pitch se base sur la recherche des maximums du cepstre.

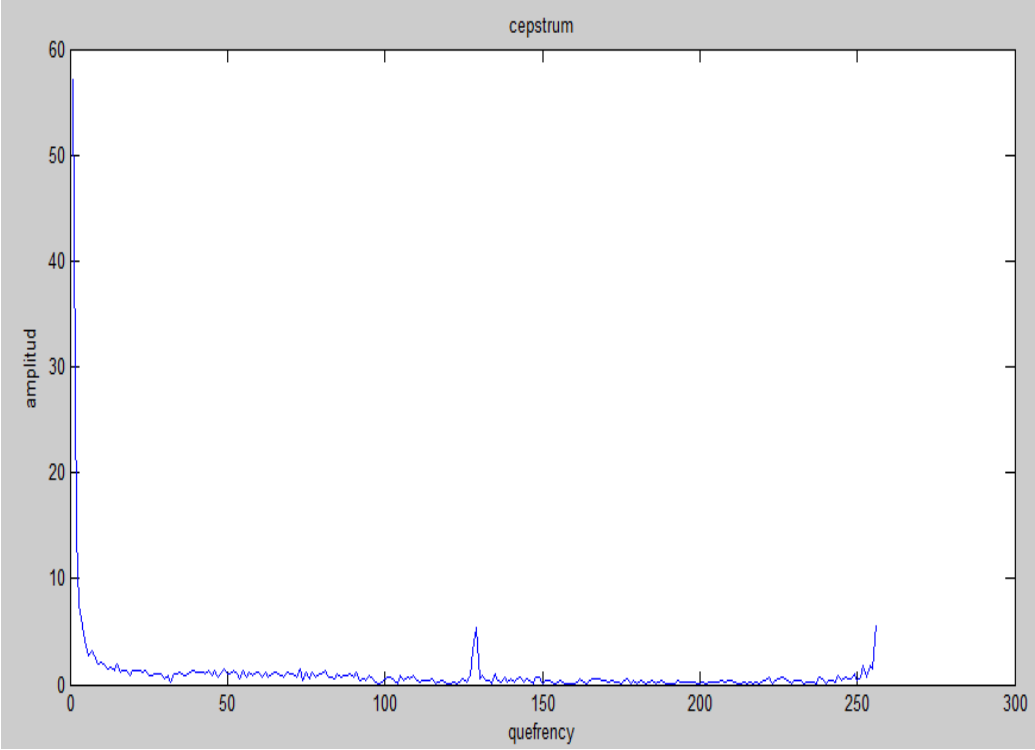


Figure n°10 : Représentation du cepstre d'une trame.

Cette figure nous montre le cepstre d'une trame après l'application de l'IFFT. Pour l'estimation de pitch on doit chercher les maximums du cepstre.

- **Fenêtre de levage (Liftering) :** L'opération de levage est similaire à l'opération de filtrage dans le domaine fréquentiel. Il existe deux types de levage effectués :

- Low-time liftering : est effectuée pour extraire les caractéristiques de conduit vocal dans le domaine de la fréquence.

- High-time liftering : est effectué pour obtenir les caractéristiques d'excitation, la moitié de la longueur du cepstre est prise en compte pour le levage utilisant cette fenêtre. :

$$w(n) = \begin{cases} 1, & l_c < n < N/2 \\ 0, & si\ non \end{cases} \quad (3)$$

où N est la longueur de cepstre et l_c est la longueur coupée de cette fenêtre, généralement Lc est pris entre 15 ou 20.

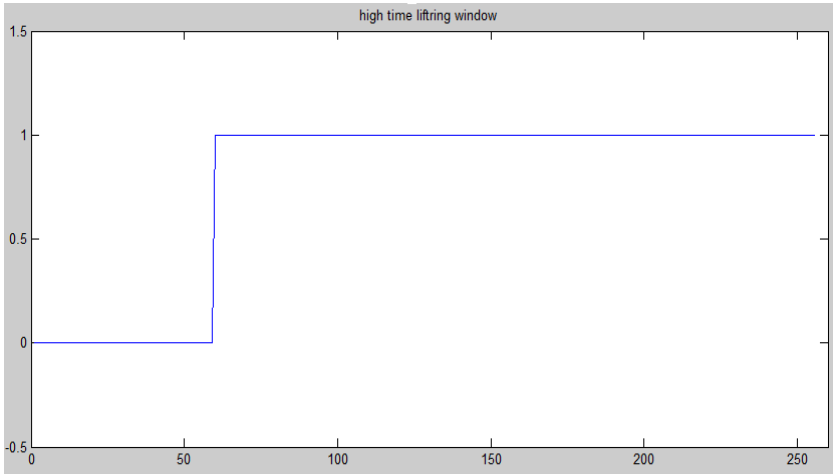


Figure n°11: High Time Liftering Window.

Cette figure représente l'étape de liftering window qui est similaire à l'opération de filtrage dans le domaine fréquentiel.

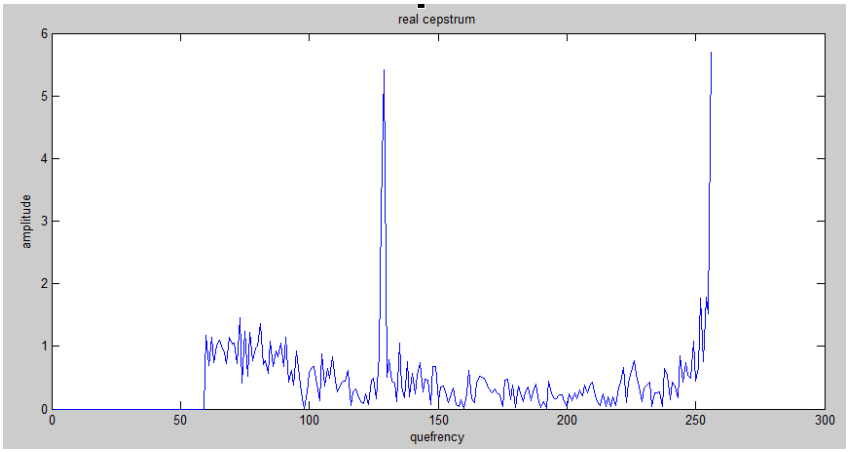


Figure n°12: Cepstre d'une trame après fenêtrage (High time liftering).

Cette figure représente le cepstre d'une trame après l'application de la fenêtre (high time liftering) appliqué sur la moitié du cepstre.

3.1.3 Estimation de pitch :

La période de pitch est l'instant de temps correspondant au plus grand pic dans le cepstre soulevé à temps élevé (high time liftered spectrum). L'inverse de l'intervalle de pitch multiplié par la fréquence d'échantillonnage donne la fréquence de pitch. [20 ; 21].

$$Pitch = fs \times 1/2max \qquad (4)$$

- Remarque : on prend le pique max × 2 car la FFT est symétrique.

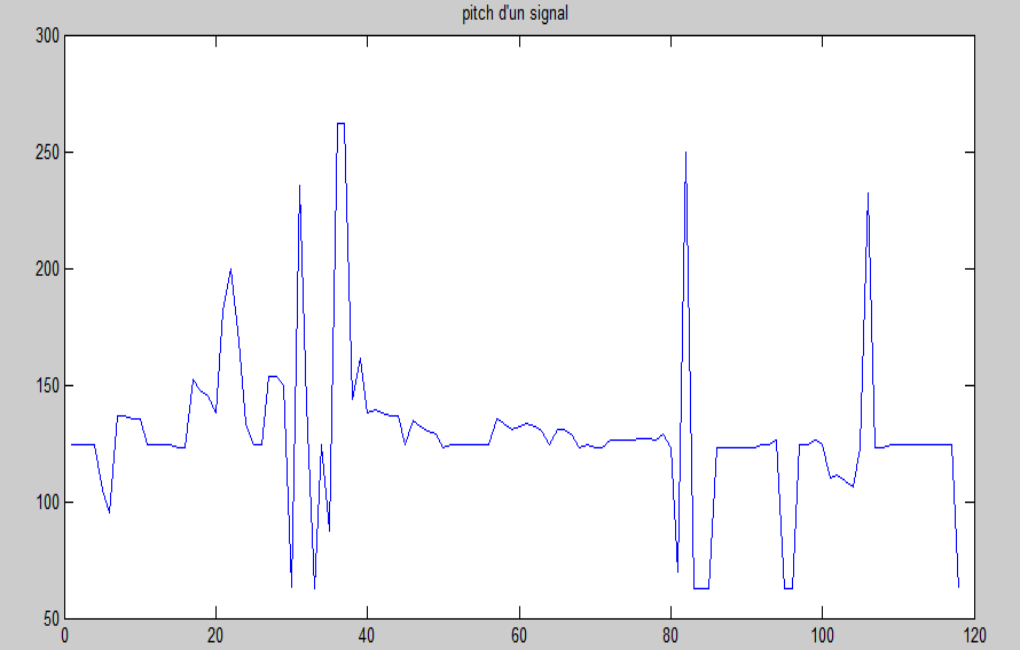


Figure n°13 : Représentation du Pitch d'un signal audio.

Après avoir fait les différentes étapes de la méthode cepstrale, on obtient une représentation de notre pitch comme il est montré dans la figure ci-dessus.

Ces valeurs varient approximativement entre 50 et 270 Hz.

3.1.4 Problèmes et limitations :

Les méthodes cepstrales présentent cependant un Certain nombre de problèmes :

- Il est nécessaire d'appliquer au signal une fenêtre de pondération, ce qui dans le cas de fondamentaux de fréquence basse (faible nombre de périodes dans la fenêtre) atténue fortement les pics cepstraux.
- Si le signal possède peu d'harmoniques, son cepstre ne présente plus de pic à n= T0 (cas limite d'une sinusoïde).

3.2 Extraction de formant :

Il existe plusieurs méthodes pour calculer les formants, dans notre étude on s'est basé sur l'algorithme LPC.

3.2.1 L'analyse par prédiction linéaire (Linear Predictive Coefficients) LPC :

Se fonde sur la corrélation entre les échantillons successifs du signal vocal et fait l'hypothèse du modèle acoustique linéaire.

En introduisant une excitation e(n) de variance σ^2 à l'entrée d'un modèle d'ordre p, l''échantillon du signal à l'instant n s(n) s'écrit comme une combinaison linéaire des p précédents échantillons :

$$s(n) = \left(\sum_{i=1}^{p} a_i s(n-i)\right) + e(n). \quad (5)$$

La fonction de transfert associé à ce filtre linéaire de prédiction est donnée par :

$$H(z) = \frac{S(z)}{E(z)} = \frac{\sigma^2}{1 - \sum_{i=1}^{p} a_i z^{-i}} \quad (6)$$

Les coefficients de prédiction ce modèle, sont calculés en minimisant l'erreur quadratique moyenne induite par le modèle. Le choix de l'ordre de prédiction résulte d'un compromis entre le temps, la quantité de données et la qualité d'analyse.

L'analyse LPC permet de représenter l'enveloppe spectrale du signal à partir des coefficients de prédiction. La réponse fréquentielle du filtre linéaire de prédiction reflète les pics du spectre du signal de parole, ce qui rend cette analyse très utilisée pour la détermination des formants. Les coefficients de prédiction ont été rarement utilisés directement comme paramètres acoustiques. Ils sont plutôt transformés en de plus robustes et moins corrélés paramètres, comme les LPCC et les coef7/*ficients PLP.

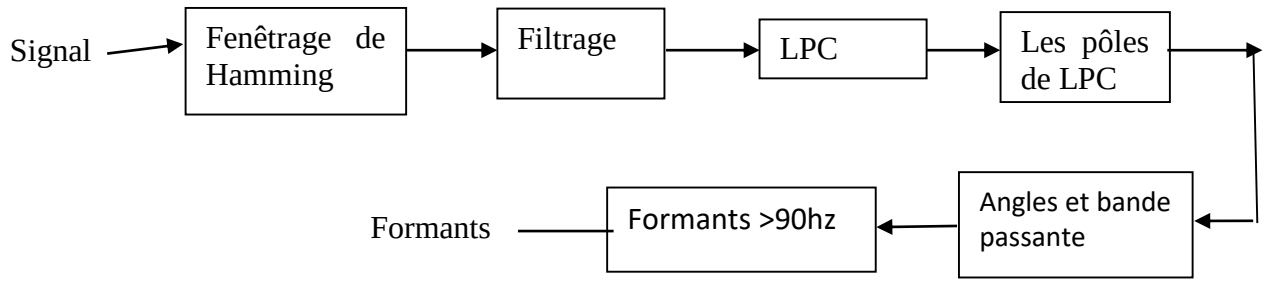


Figure n°14 : Bloc diagramme de la méthode LPC. [22 ;23 ;24]

La méthode utilisée est résumée comme suit :

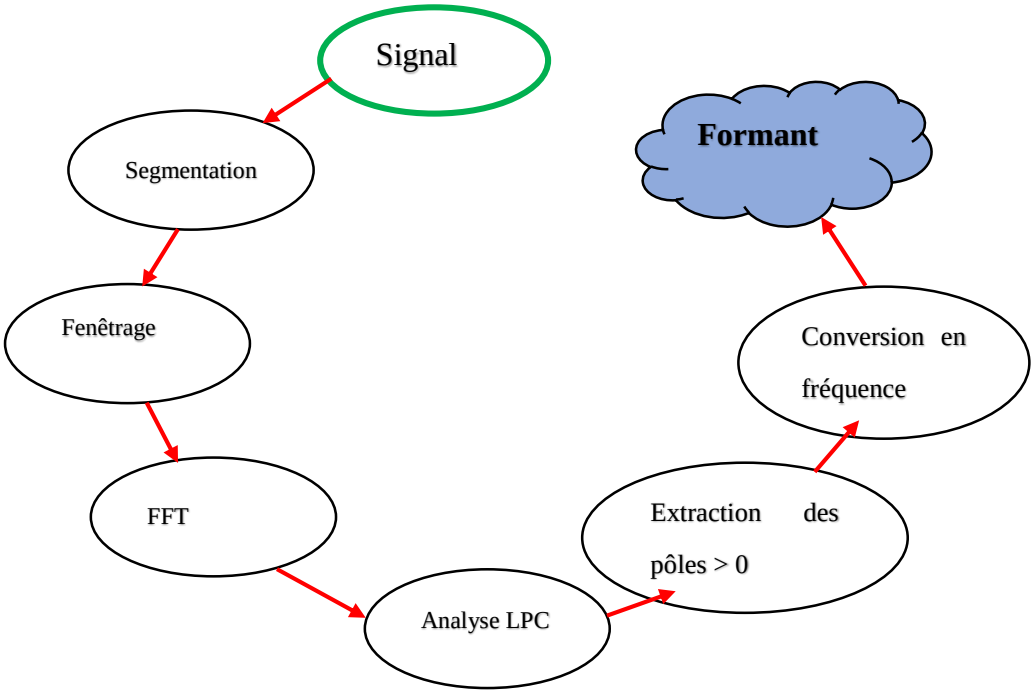


Figure n°15 : Organigramme de la méthode LPC.

Cet organigramme nous démontre les différentes étapes à suivre pour calculer la valeur de du formant avec la méthode LPC.

Explication du schéma :

Après la segmentation du signal, on a utilisé le fenêtrage par la fenêtre de Hamming.

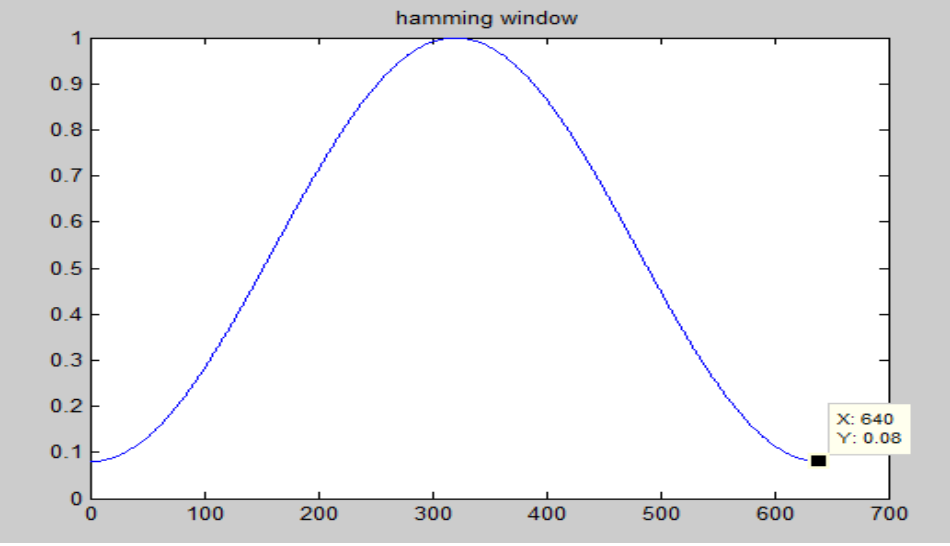


Figure n°16 : Fenêtre de Hamming.

Cette figure est une représentation graphique de la fenêtre de Harmming qui par principe divise notre signal en trame afin d'éviter la distorsion.

La fenêtre a le même nombre d'échantillon que la Trame.

- On calcule la valeur absolue de la FFT.

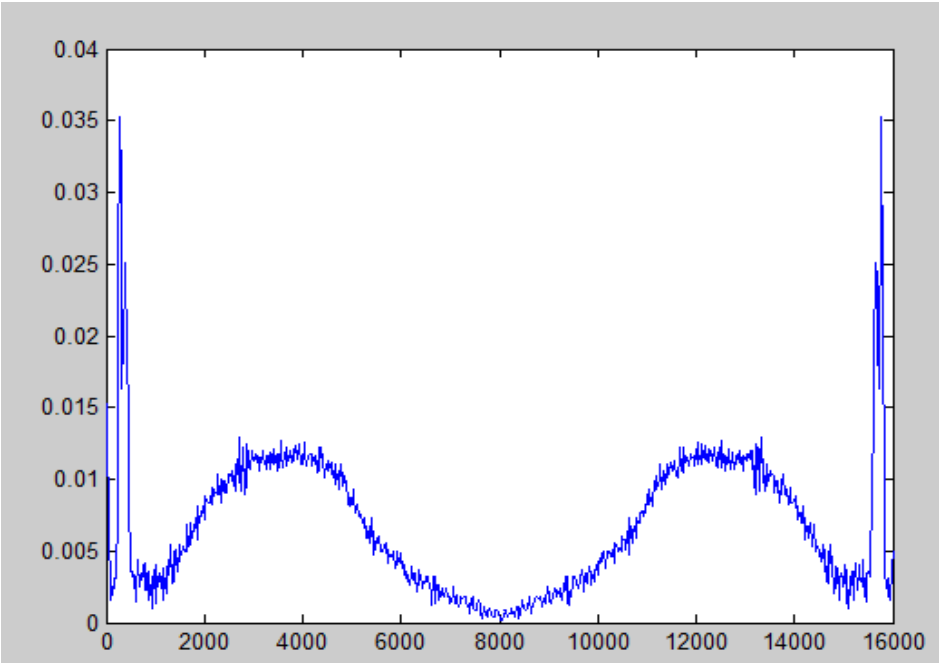


Figure n°17 : Transformée de Fourier rapide (FFT).

Cette figure est une représentation graphique de note signal après l'application de la transformée de Fourier rapide avec une fréquence d'échantillonnage qui est égale à 16 000 Hz.

- Ensuite, on applique le filtre LPC d'ordre q [q= 2+ (fs/1000)].[23 ; 24]

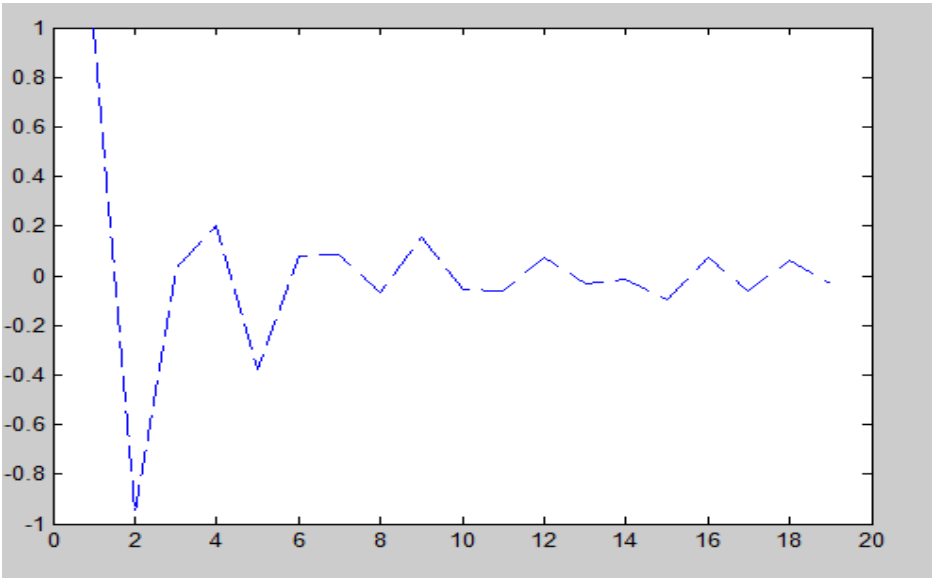


Figure n°18 : Coefficient de prédiction linéaire.

Cette figure nous présente les coefficients de prédiction linéaire d'ordre q.

Après avoir extrait les polynômes de LPC on cherche ses solutions (pôles) (les solutions sont des nombres complexes). On choisit les solutions qui ont des angles positifs, par la suite on va les convertir en fréquence suivant l'équation [25]:

$$f = \sin^{-1}\left(\frac{imag(r)}{real(r)} * \frac{fs}{2*\pi}\right) . \quad (7)$$

- Classifier en ordre descendant puis prendre les quatre premières fréquences basses.

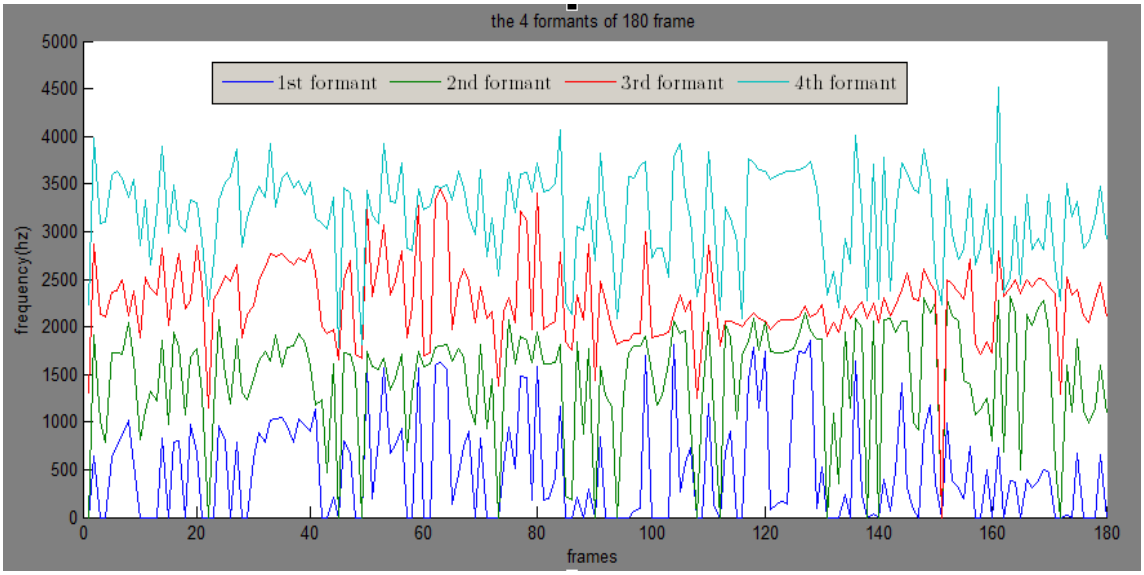


Figure n°19 : les quatre premières fréquences des formants.

Cette figure nous illustre les quatre premiers formats sur un ensemble de trame. Le premier est représenté en bleu et varie entre 50 et 1800 Hz, le second représenté en vert varie entre 100 et 2300 Hz, le troisième en rouge varie entre 1150 et 3400 Hz, et le quatrième en turquoise varie entre 2000 et 4500 Hz.

3.3 Calcul d'énergie :

L'énergie est un paramètre important dans l'analyse d'un signal audio. Nous tirons le résultat de l'énergie à partir du théorème de Perseval.

3.3.1 Energie et théorème de parseval :

La définition à long terme de l'énergie du signal est la suivante [26]:

$$E = \sum_{n=-\infty}^{\infty} |x(n)^2| \quad (8)$$

Dans l'expression ci-dessus, E représente l'énergie du signal x(n). Il y a une très faible ou presque aucune efficacité de cette définition pour les signaux variant dans le temps, tels que la parole. Nous calculons donc l'énergie à court terme en utilisant la formule suivante [27] :

$$E = \sum_{n=-\infty}^{\infty} |x(n)^2/N| \quad (9)$$

Où N est la longueur de fenêtre de Hamming utilisé.

En multipliant le signal par la fonction d'énergie on obtient la valeur d'énergie,

Loi d'énergie [5]:

$$\bar{E} = \sum_{n=0}^{N-1} X_m^2(n) \quad . \quad (10)$$

3.3.2 Spectre d'énergie :

Le spectre d'énergie est représenté dans la figure suivante :

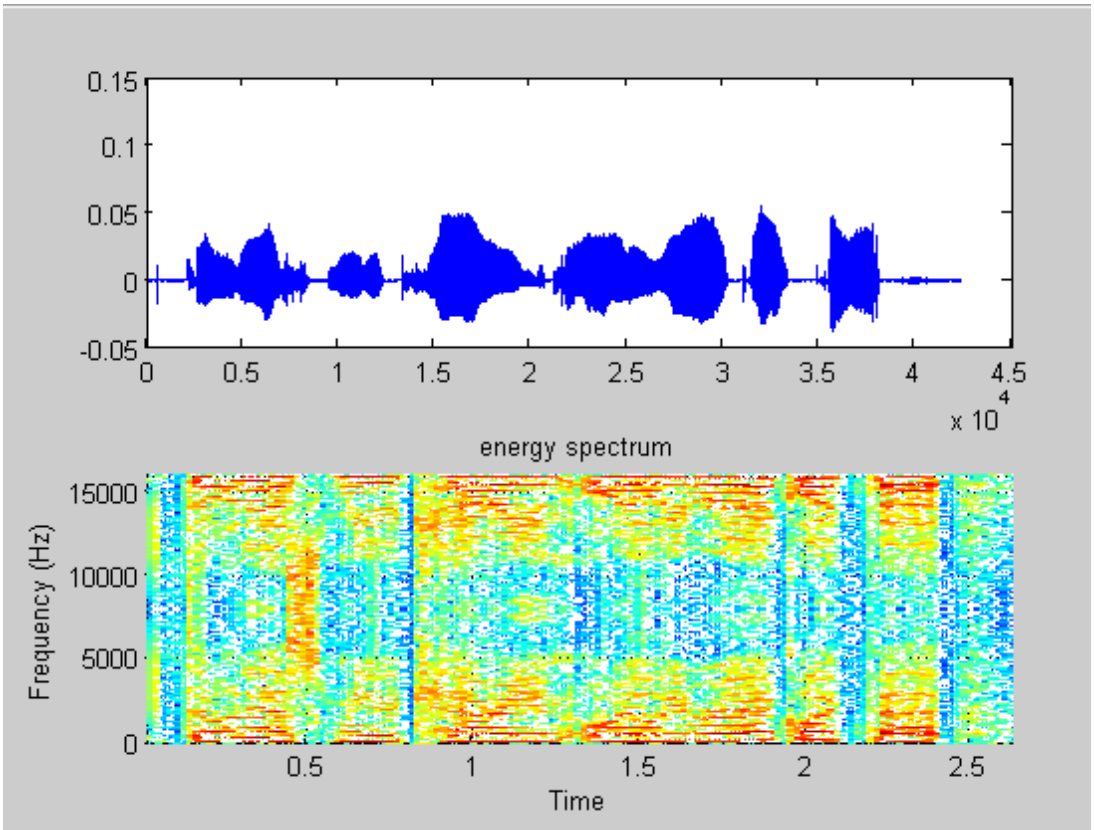


Figure n°20 : Spectre d'énergie.

Cette figure nous présente le spectre d'énergie. La zone en bleu signifie la présence de la parole qui varie selon son énergie.

Calcul du taux de précision :

Afin d'évaluer la précision de nos algorithmes proposés, nous avons utilisé le pourcentage de correction (POC). Le POC est défini comme le rapport entre le nombre de genres correctement détectés et le nombre réel de locuteurs.

$$poc = \frac{N_C}{N_t} \quad (11)$$

Ou N_C est le sexe correctement détecté et NT est le nombre total des locuteurs.

D'après l'expérience la longueur des trames peut aller jusqu'à 0,05s.

4. Conclusion

La fonction cepstre est plus performante parce qu'elle sépare le spectre de conduit vocal avec le signal source, ainsi le pitch est facile à calculer.

L'analyse du signal vocal n'est pas complète tant qu'on n'a pas mesuré l'évolution de la fréquence fondamentale, c'est un paramètre très important pour l'analyse de la parole. L'estimation du pitch est bien sûr liée à la localisation des tranches voisées. Cette tâche est difficile pour des nombreuses raisons telles que la non-stationnarité du signal de parole, l'existence de certaines irrégularités dans l'excitation glottique encore une interaction avec les formants.

Après avoir fait le tour de différentes techniques utilisées pour la caractérisation du signal d'un point de vue théorique, nous passons directement à l'expérimentation ainsi que la validation de la méthode la plus fiable dans le chapitre suivant.

Chapitre 3

Etude comparative des méthodes de détection du genre

1. Introduction :

On procède à une étude comparative des méthodes de détection du genre, calcul de fréquence fondamentale (Pitch), formants et énergie. Dans notre étude, nous avons utilisé la méthode d'autocorrélation et cepstrale pour la détection du Pitch, l'analyse LPC pour les formants et l'énergie avec le théorème de Parseval.

L'étude a été faite sur l'enregistrement de 70 personnes (35 hommes et 35 femmes). Ces personnes prononcent la même phrase.

Dans notre travail, nous nous sommes focalisé sur l'étude des caractéristiques d'un signal audio. Cette étude nous permet de constater le paramètre le plus fiable pour la détection du genre.

Pour effectuer cette expérience, on se réfère à la base des données TIMIT afin d'utiliser quelques enregistrements vocaux.

Le logiciel utilisé est Matlab.

Nous discutons les résultats obtenus afin de caractériser la meilleure méthode selon leur taux de précision.

A propos de la base de données TIMIT :

Cette version du TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1) possède tous les fichiers de formes d'onde formatés avec des en-têtes ms-wav / RIFF, pour rendre le corpus plus accessible à un public plus large.

Le corpus TIMIT de la parole lue est conçu pour fournir des données vocales pour les études acoustiques-phonétiques et pour le développement et l'évaluation de systèmes de reconnaissance automatique de la parole. TIMIT contient des enregistrements à large bande de 630 locuteurs de huit dialectes majeurs de l'anglais américain, chacun lisant dix phrases phonétiquement riches. Le corpus TIMIT comprend des transcriptions orthographiques, phonétiques et verbales alignées dans le temps, ainsi qu'un fichier de forme d'onde vocale 16 bits, 16 kHz pour chaque énoncé. La conception du corpus était un effort conjoint du Massachusetts Institute of Technology (MIT), de SRI International (SRI) et de Texas Instruments, Inc. (TI). Le discours a été enregistré à TI, transcrit au MIT et vérifié et préparé pour la production de CD-ROM par le National Institute of Standards and Technology (NIST).

Les transcriptions du corpus TIMIT ont été vérifiées manuellement. Des sous-ensembles de test et d'apprentissage, équilibrés pour la couverture phonétique et dialectale, sont spécifiés.

Des informations tabulaires consultables par ordinateur sont incluses ainsi que de la documentation écrite.

2. Etude comparative des paramètres :

L'extraction et la sélection de la meilleure représentation paramétrique des signaux acoustiques est une tâche importante dans tout système de reconnaissance vocal, cela affecte considérablement la performance de reconnaissance.

Nous distinguons deux phases dans le processus de détection du genre : apprentissage et test.

L'apprentissage :

Pour les trois paramètres et quatre méthodes, on va prendre 50 enregistrements des échantillons de la base des données TIMIT de forme wav (car cette forme est sans perte), 25 hommes et 25 femmes.

Concernant le paramètre ***pitch*** on va étudier et comparer deux méthodes, l'une est temporelle et l'autre est fréquentielle.

Après la détermination des pitchs en va prendre la moyenne de la moyenne des hommes et la moyenne des femmes ce qui donne le seuil. Ce dernier est enregistré dans la base de données.

Pour le deuxième paramètre, les ***formants***, après l'extraction des 4 premiers formants de chaque échantillon, on obtiendra 200 formants au total. Ensuite, on calcule la moyenne de ième formant des deux sexes et l'enregistrer dans la base des données.

La méthode du paramètre ***énergie*** va être un peu différente car on va inverser les échantillons de test et d'apprentissage dans les deux phases pour évaluer la fiabilité de cette méthode. Puis on enregistre la valeur d'énergie moyenne des deux genres.

Test : La partie test contient 20 exemples où les enregistrements sont différents de ceux d'apprentissage.

Théoriquement il y a des gammes pour détecter le sexe du locuteur, mais dans cette étude on va les extraire pratiquement et ceci est réalisé dans la phase précédente.

Pour cette phase, 10 échantillons sont pris. Chaque pitch est comparé à la valeur enregistrée dans la base. S'il est supérieur à cette valeur, le locuteur est une femme, si non donc c'est un homme.

Formants :

On extrait des formants pour chaque échantillon des 10 enregistrements sachant que la règle utilisée est basée sur les résultats de la partie apprentissage, on compare les valeurs des formants avec les valeurs dans la base des données s'ils sont supérieurs aux valeurs de la base donc le locuteur est une femme si non c'est un homme.

Energie :

On a mentionné précédemment qu'on va faire deux méthodes. Utilisant le même algorithme, on calcule l'énergie des 5 enregistrements d'échantillon en changeant ces échantillons d'une méthode à l'autre, et on compare toujours à la valeur du seuil, et là qu'on va savoir la règle à appliquer selon les résultats de la partie d'apprentissage.

La phrase prononcée est la même pour tous les exemples « don't ask me to carry an oily rag like that ». Sa durée est de 3 à 4s.

2.1 Fréquence fondamentale (pitch) par la méthode cepstrale :

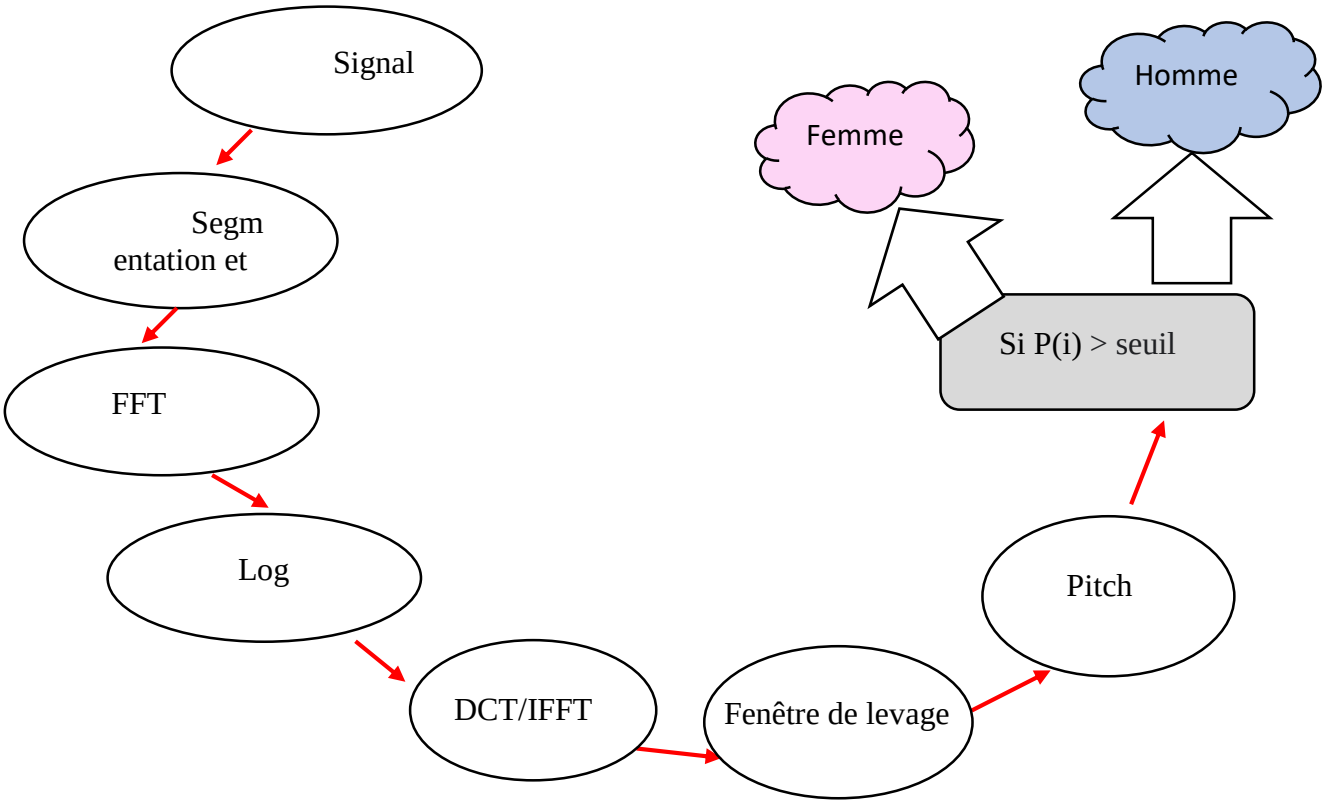


Figure n°21 : étapes de calcul de Pitch.

Dans ce bloc diagramme, nous montrons les étapes de calcul de pitch pour la détection du genre.

Cette méthode consiste à faire une segmentation et un fenêtrage de Hamming pour les diviser en trames. Puis, nous utilisons la FFT afin d'assurer le passage temporel au fréquentiel. Ensuite, nous appliquons le logarithme et la FFT inverse ainsi la fenêtre Liftering (High Time Liftering). Enfin, nous comparons le résultat obtenu du Pitch au seuil pour pouvoir décider le genre du locuteur.

2.1.1 Résultats expérimentaux :

Les résultats obtenus après l'application de la méthode de cepstre sont détaillés dans le tableau suivant :

Pitch homme test	Pitch femme test
129,1581	157,6874
118,9976	155,2353
130,2967	167,1535
127,7309	165,2699
127,3336	155,5405
120,2450	186,7569
125,7342	178,1511
124,6559	179,3913
119,9973	134,1471

Tableau n°01 : Les valeurs de Pitch pour les deux genres.

Ce tableau résume les résultats obtenus pour vingt enregistrements (dix hommes et dix femmes). Ces valeurs de Pitch varient selon le genre du locuteur. Nous avons un Pitch qui varie entre 118,9976 et 130,2967 Hz pour les hommes, et de 134,1471 à 186,7569 Hz pour les femmes.

Nous disons que ces valeurs sont corrects où acceptables car elles sont dans l'intervalle des fréquences de chaque type.

Nous calculons maintenant la moyenne pitch de chaque genre et le seuil.
Le tableau suivant représente nos résultats :

	Pitch homme apprentissage	Pitch femme apprentissage
Moyenne	124,91	163,9
Seuil	144,41	
Taux de précision	100%	90%
	95%	

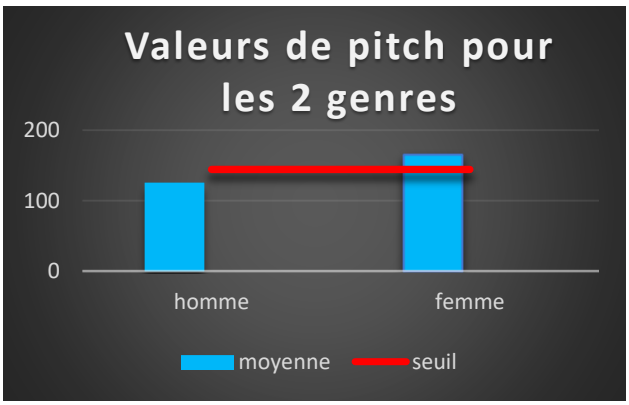


Tableau n°02 : Moyenne et seuil du pitch.

Figure n°22 : Moyenne et seuil du pitch.

Comme il est montré dans le tableau, nous avons un seuil de 144,41Hz. A partir de ce seuil nous pourrons décider le genre par la condition suivante : si P(i)> seuil donc c'est une femme, et si P(i)<seuil c'est un homme.

Nous calculons le POC de cette méthode :

Poc (hommes) = 100% ; Poc(femmes) = 90% , donc le Poc= 95%.

2.2 Formants par la méthode d'analyse LPC :

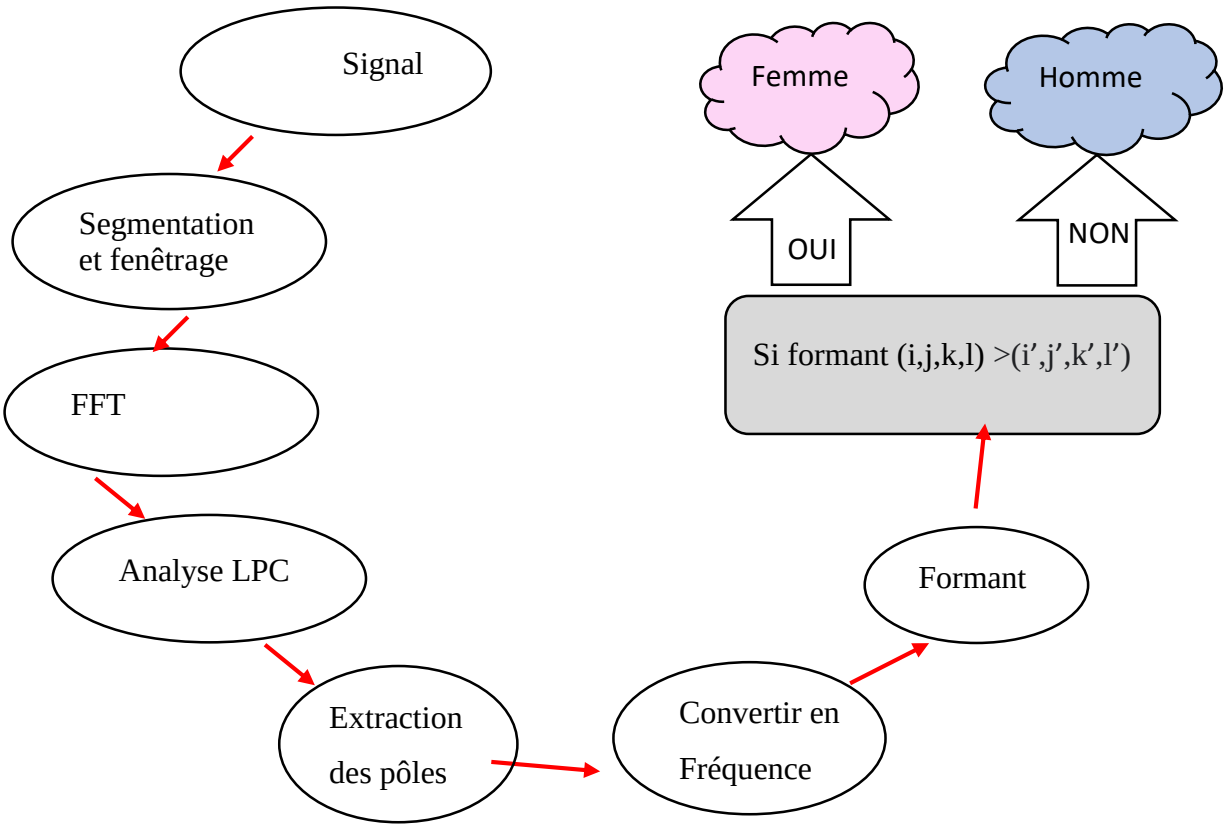


Figure n°23 : étapes de calcul de formants par LPC.

Ce diagramme nous résume les différentes étapes pour calcer les formants par la méthode LPC (coefficients de prédiction linéaire).

D'abord, nous faisons une segmentation et un fenêtrage de Hamming pour découper notre signal en trames. Puis, nous utilisons la transformée de Fourier rapide (FFT) pour passer au domaine fréquentiel. Ensuite, nous appliquons le LPC afin d'extraire les pôles qui seront par la suite convertis en fréquences. Enfin, nous comparons les résultats des formants aux seuils pour qu'on puisse décider le genre du locuteur.

2.2.1 Résultats expérimentaux :

Les résultats sont détaillés dans les tableaux suivant :

Homme

Formant1	Formant2	Formant3	Formant4
418,8178	1308,5393	2227,9015	3083,7500
450,8109	1391,9298	2408,8242	3259,8407
473,9365	1396,0113	2390,0274	3329,7616
413,6115	1304,4074	2331,8007	3248,5106
453,3869	1429,5445	2484,3454	3228,3258
414,9753	1295,8808	2373,7165	3238,8085
348,6697	1263,4976	2233,0010	3068,4683
396,9680	1235,0661	2251,4739	3205,1305
423,6183	1265,9001	2249,1167	3256,9888
423,4059	1363,9047	2436,1827	3218,2673
372,5453	1230,6106	2335,6761	3205,6375
423,6183	1265,9001	2249,1167	3256,9888
388,1712	1287,3198	2371,1117	3236,4615
390,3422	1239,8563	2313,1415	3143,5826
395,8855	1205,0996	2355,5969	3341,7102
385,9887	1276,8319	2332,8306	3157,6693
450,8109	1391,9298	2408,8242	3259,8407
337,9196	1202,3378	2275,5603	3053,1848
322,3459	1223,3029	2311,1359	3074,2076
484,0498	1416,3209	2424,0250	3276,0768
387,8357	1306,2796	2276,0482	3097,8795
355,3364	1208,7886	2233,7889	3193,7647
423,4059	1363,9047	2436,1827	3218,2673
395,5267	1324,9114	2407,8624	3313,1127
342,2564	1210,7814	2170,3509	3219,5990
Moyenne			
402,9696	1296,3543	2331,5057	3207,4334
Seuil			
463,2327	1375,5924	2293,7189	3196,9516
Taux de précision		62,5%	

Tableau n°03 : Représentation des valeurs des 4 premiers formants pour les hommes par la méthode LPC.

Ce tableau nous présente les valeurs obtenues des quatre premiers formants qui concernent les 25 hommes. Le premier formant varie entre 322,3459 et 484,04 Hz, le second varie entre 1202,78 et 1429,54 Hz, le troisième entre 2227,9 et 2484 Hz et le quatrième varie entre 3068 et 3341 Hz.

Nous avons remarqué qu'il y'a certaines valeurs du 3 ème et 4 ème formants qui dépassent le seuil. Cela, influe sur le taux de précision.

Le tableau nous présente aussi, la moyenne de chaque formant ainsi le seuil.

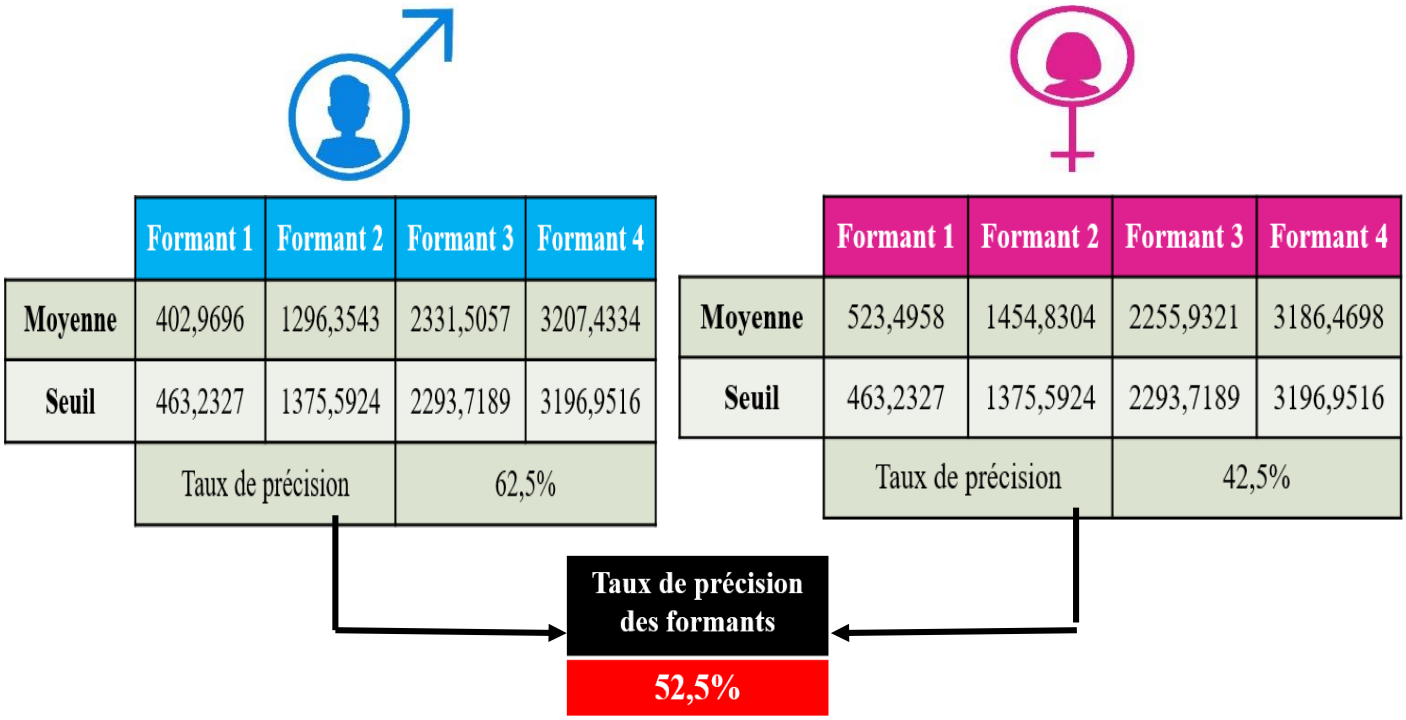
Après tous ces calculs, nous avons trouvé un taux de précision de 62.5%.

Femme

Formant 1	Formant 2	Formant 3	Formant 4
500,3958	1454,5605	2249,7805	3131,7359
701,2892	1558,8257	2278,8787	3313,5177
605,4262	1498,3812	2191,2525	3208,9613
645,7470	1457,1720	2240,3777	3313,4297
534,0602	1589,9182	2417,2208	3173,9852
395,2145	1332,7185	2154,6140	3028,5679
592,6717	1562,4184	2302,4670	3291,0103
621,9634	1513,9597	2308,1218	3254,5228
500,3958	1454,5605	2249,7805	3131,7359
521,3675	1463,5150	2228,1113	3185,1847
414,1502	1395,3058	2181,1336	3213,3978
521,3675	1463,5150	2228,1113	3185,1847
467,9053	1311,8303	2224,8883	3064,8715
554,1643	1507,3887	2250,4922	3226,6849
601,8591	1535,1276	2262,2415	3301,5384
363,7680	1282,0100	2309,2383	3143,5717
532,1454	1587,3344	2277,4809	3177,7213
605,4262	1498,3812	2191,2525	3208,9613
645,7470	1457,1720	2240,3777	3313,4297
534,0602	1589,9182	2417,2208	3173,9852
395,2145	1332,7185	2154,6140	3028,5679
561,8666	1486,1841	2360,4306	3212,0947
414,1502	1395,3058	2181,1336	3213,3978
386,1943	1275,2976	2285,4270	3010,6276
470,8443	1367,2419	2213,6562	3155,0596
Moyenne			
523,4958	1454,8304	2255,9321	3186,4698
Seuil			
463,2327	1375,5924	2293,7189	3196,9516
Taux de précision		42,5%	

Tableau n°04 : représentation des valeurs des 4 premiers formants pour les femmes par la méthode LPC.

Ce tableau nous présente les valeurs obtenues des quatre premiers formants qui concernent les 25 femmes. Le premier formant varie entre 363,7680 et 701,2892 Hz, le second varie entre 1275,2976 et 1589,9182 Hz, le troisième entre 2154,6140 et 2417,2208Hz et le quatrième varie entre 3131,7359 et 3313,4297 Hz.



	Formant 1	Formant 2	Formant 3	Formant 4
Moyenne	402,9696	1296,3543	2331,5057	3207,4334
Seuil	463,2327	1375,5924	2293,7189	3196,9516
Taux de précision			62,5%	

	Formant 1	Formant 2	Formant 3	Formant 4
Moyenne	523,4958	1454,8304	2255,9321	3186,4698
Seuil	463,2327	1375,5924	2293,7189	3196,9516
Taux de précision			42,5%	

Tableau n°05\06 : Moyenne et seuil des formants pour des deux genres.

Nous avons obtenu ces résultats par la méthode LPC. Nous constatons que certaines valeurs ne dépassent pas le seuil, ce qui engendre une diminution du taux de précision.

Taux de précision pour les femmes égales à 42.5%. Après tous ces calculs, nous avons trouvé un taux de précision des formants égal à 52,5%.

2.3 Energie :

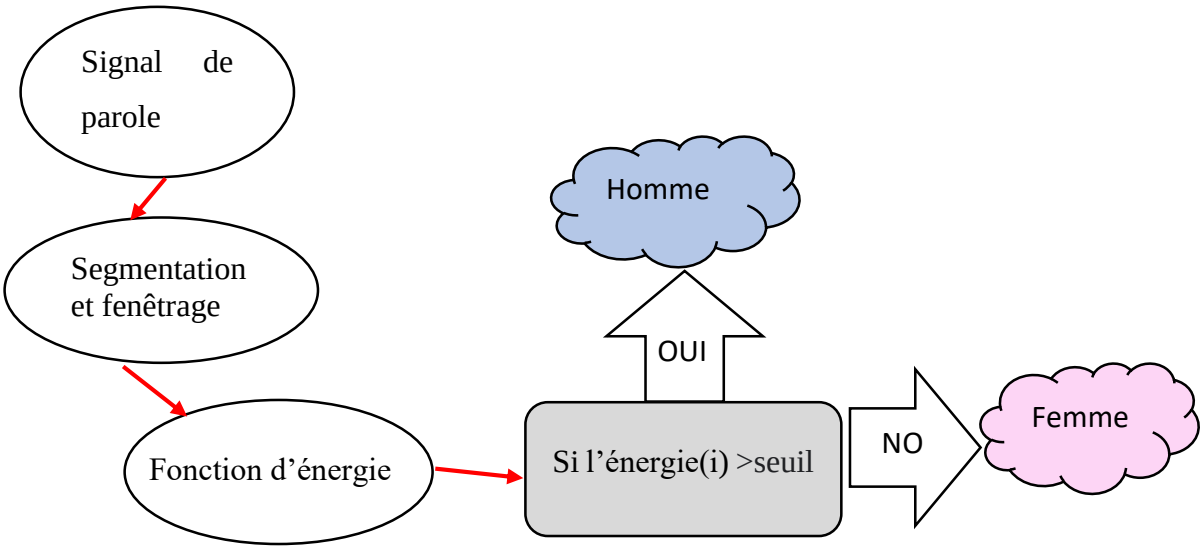


Figure n°24 : Etapes de calcul de l'énergie.

Cette figure nous démontre les étapes à suivre pour calculer l'énergie de notre signal audio. Après avoir fait la segmentation, fenêtrage, on applique la fonction d'énergie. Si le résultat de l'énergie est supérieur au seuil, ça veut dire que le type de notre locuteur est un homme, dans le cas contraire c'est une femme.

4.3.3 Résultats expérimentaux :

Les résultats sont détaillés dans le tableau suivant :

Méthode 1		**Méthode 2**	
Hommes	**Femmes**	**Hommes**	**Femmes**
4,7154	4,8962	22,9688	11,0247
18,8860	4,3369	4,3546	4,5784
1,1678	6,6115	17,7871	2,0495
13,5990	4,2362	4,4668	16,1376
2,3275	6,2297	9,9276	5,7544
9,4551	30,2792	13,3217	5,1283
36,2877	5,9360	18,8860	3,6133
5,5809	2,0047	31,7275	6,6115
4,3546	4,8962	19,9209	4,2362
3,9804	4,5784	8,3384	6,2297
22,9688	11,0247	9,3758	30,2792
4,3546	4,5784	0,7880	16,1436
17,7871	2,0495	3,9804	11,0247
4,4668	16,1376	3,4414	44,0673
9,9276	5,7544	10,4602	74,1900
13,3217	5,1283	32,5694	2,3426
18,8860	3,6133	2,5508	17,4940
31,7275	6,6115	21,1389	1,8995

19,9209	4,2362	3,4950	8,4151
8,3384	6,2297	4,9036	9,6538
9,3758	30,2792	22,9688	2,0495
0,7880	16,1436	17,7871	16,1376
3,9804	11,0247	4,4668	5,7544
3,4414	44,0673	9,9276	5,1283
10,4602	74,1900	13,3217	3,6133

Tableau n°07 : représentation des valeurs de l'énergie pour les 2 genres par 2 méthodes.

.

	Méthode 1(apprentissage-test)	
	Hommes	Femmes
Moyenne	12,6	11,2
Seuil	11,9	
Taux de précision	50%	80%
	65%	

	Méthode 2(test-apprentissage)	
	Hommes	Femmes
Moyenne	12,54	12,51
Seuil	12,50	
Taux de précision	30%	90%
	60%	

Taux de précision d'énergie
62,5%

Tableau n°08/09 : moyenne et seuil de l'énergie des 2 genres.

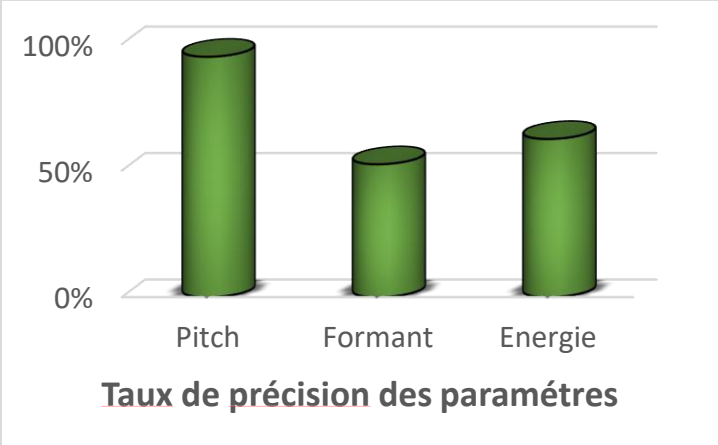
D'après les résultats que nous avons obtenus et l'écoute des enregistrements vocaux, nous avons remarqué que l'énergie dépend de la personne plus que le genre. Il y'a des hommes qui parlent doucement et lentement, ce qui explique leur faible énergie et contrairement pour les femmes.

3.Comparaison des résultats :

Tableau n°10 : Le taux de précision des paramètres pour la détection du genre.

	Pitch	**Formant**	**Energie**
Taux de précision	95%	52,5%	62,5%

Figure n°25 : Taux de précision des paramètres.



.

Dans cette étude le pitch nous a donné les meilleurs résultats par rapport aux autres paramètres.

4. conclusion :

L'étude des paramètres sur un signal audio nous permet de détecter le genre du locuteur. Cette étude nous montre que chaque méthode a une occurrence différente.

La méthode de calculs des formants qui a un taux de précision de 52,5%. Celle de l'énergie a donné un taux de précision de 60,5%. Ce sont des résultats acceptables mais pas assez convaincants ou suffisants.

Dans le cas de Pitch, on a atteint un taux de précision très important qui est égal à 95%. De ce fait, on peut considérer la méthode cepstrale comme étant la meilleure méthode par rapport à sa précision et sa fiabilité pour la détection du genre.

Conclusion générale

Dans notre étude, nous avons parlé du signal sonore précisément du signal de la parole, de ses différents paramètres dans le domaine temporel et du domaine fréquentielle, ainsi que quelques méthodes pour extraire certains de ses paramètres.

Nous nous sommes concentrés sur le pitch, les formants et l'énergie. Pour la détection de Pitch, nous avons choisi les méthodes autocorrélation et cepstrale, et pour les formants la méthode d'analyse par les coefficients de prédiction linéaire LPC a été choisie et nous avons calculé l'énergie.

Pour évaluer la performance de ces paramètres sur l'analyse du signal, nous avons choisi de l'appliquer à la reconnaissance du genre pour voir lequel donne le meilleur taux de précision en utilisant les mêmes enregistrements audio (.wav) (70 entre femmes et hommes).

Comme indiqué dans le troisième chapitre, le taux de précision atteint pour le pitch en utilisant la méthode cepstrale était très élevé et il est égal à 95% : 100% pour les hommes et 90% pour les femmes.

Concernant les formants, nous avons utilisé la méthode d'analyse par LPC et nous avons eu une précision qui est égale à 52.5%.

La reconnaissance par énergie a été calculée selon deux méthodes proposée par notre promoteur et nous a donné deux valeurs de POC (pourcentage de correction) pour la première 65% et 60% pour la deuxième où l'on remarque qu'elles sont proches l'une de l'autre donc on peut dire que notre programme est fiable mais nous donnent encore une faible précision.

D'après notre étude expérimentale, nous constatons que le paramètre de pitch calculé par la méthode cepstrale a donné les meilleures performances par rapport aux autres paramètres et puis vient l'énergie.

Pour obtenir de meilleurs résultats concernant les formants et l'énergie, nous pouvons les combiner.

Ce travail a été fait sans tenir compte du bruit et de la technique VAD (Voice Activity Detection).

Références :

[1] La parole et son traitement automatique. Collection technique et scientifique des télécommunications.

[2] Damien Vincent. Thèse « Analyse et contrôle du signal glottique en synthèse de la parole» l'École Nationale Supérieure des Télécommunications de Bretagne 2007.

[3] Ishizaka, K. and Flanagan, J. L., "Synthesis of voiced sounds from a two-mass model of the vocal chords", Bell systems Technology Journal, 50:1233-1268, (1972).

[4] Flanagan, J. L. and Ishizaka, K., "Computer model to characterise the air volume displaced by the vibrating vocal chords", Journal of the Acoustical Society of America, 63:1559-1565, (1978).

[5] M.Kunt. « Traitement de la parole », presses polytechniques Romandes, EPFL, 1987

[6] Codage et décodage LPC de la parole, avril 2003.

[7] Joseph Picone ,"Fundamentals of speech recognition" institute for signal and information processing". Mississippi State University. 1998.

[8] Bojan Kotnik1, Harald Höge, Zdravko Kacic1 "Evaluation of Pitch Detection Algorithms in Adverse Conditions", University of Maribor, Slovenia, Siemens AG, Corporate Technology, Germany 2006.

[9] Fant, C. G. M., "On the predictability of formant levels and spectum envelopes from formant frequencies", for Roman Jackobson, (1956), pages 109-120.

[10] Garima Sharma Kartikeyan Umapathy Sridhar Krishnan, " Trends in audio signal feature extraction methods", The Department of Electrical and Computer Engineering, Ryerson University, ON M5B 2K3, Canada, Available online 23 September 2019.

[11] Theodoros Giannakopoulos. Aggelos Pikrakis, "Introduction to Audio Analysis " A MATLAB Approach,2014, Pages 59-103.

[12] Mitrovi´c, D.; Zeppelzauer, M.; Breiteneder, C. Discrimination and retrieval of animal sounds. In Proceedings of the 12th International Multi-Media Modelling Conference, Beijing, China, 4–6 January 2006.

[13] Farrús, M.; Hernando, J.; Ejarque, P. Jitter and shimmer measurements for speaker recognition. In Proceedings of the 8th Annual Conference of the International Speech Communication Association (InterSpeech), Antwerp, Belgium, 27–31 August 2007;p779.

[14] K. Jensen,"Pitch independent prototyping of musical sounds",1999 IEEE third workshop on multimedia signal processing (Cat. No. 99TH8451), IEEE (1999), pp. 215-220.

[15] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D.Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal, Electrical Engineering Department School of Engineering, University of Bridgeport.

[16] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.pp.293-30

[17] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," vol.ASSP-22, no. 5, pp. 353–362, Oct. 1974.

[18] L. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," IEEE Transactions on ASSP, vol. 24, pp. 399-417, 1976.

[19] R. Boite, M. Kunt, "Traitement de la parole", Presse polytechnique romandes, 1987.

[20] M. Taba, "Analyse-Synthèse de la parole par vocoder numérique à canaux", Université de Annaba, 1992.

[21] Jorge martinez*, hector perez, enrique escamilla masahisa mabo suzuki 'speaker recognition using mel frequency cepstral coefficients (mfcc)', national polytechnic institute (ipn) ,the university of electro-communications(uec).

[22] M. N. A. Aadit, S. G. Kirtania, and M. T. Mahin, "Pitch and formant estimation of bangla speech signal using autocorrelation, cepstrum and lpc algorithm," in 2016 19th International Conference on Computer and Information Technology (ICCIT), pp. 371–376, IEEE, 2016.

[23] https://vlab.amrita.edu/?sub=3&brch=164&sim=615&cnt=1 , consulté le 12/10/2021.

[24]https://slideplayer.com/slide/4407572/,5fbclid=IwAR1jsFMD4dt09yMsl_CMjsasoHgEBP5sHhy5cVRR_5tb2FO0oECMm-rAYX0 ,consulté le 05/10/2021.

[25] Snell, R. C., Milinazzo, F., 1993. Formant location from LPC analysis data, IEEE Transactions on Speech and Audio Processing 1, pp. 129-134.

[26] Speech Signal Processing, School of Electronic Information, Chapter 3, Wuhan University.

[27] Sumit, K.B., Dekate, S.K.," Text-dependent method for gender identification through.