Ministère D'enseignement Supérieur Et De La Recherche Scientifique

UniversitéDr.Yahia Farés-MEDEA –

Faculté De La Technologie

Département De Génie Electrique



Projet De Fin D'études

Pour L'obtention D'un Diplôme De Master En Génie Electrique

Option : Systèmes Intelligents Et Communicants

La Détection du Pitch d'un Signal de la Parole dans un Milieu Bruité.

Réalisé par :

- **BOUMEDIENE Souad**
- **BENAISSA Nawel**

Encadré par :

Mr. ESSAID Billel

Année 2016/2017

Dédicaces

Je dédie ce modeste travail plus particulièrement
A ma mère et mon père qui se sont sacrifiés pour que je puisse achever mes études.

A MES CHERS ET ADORABLES SŒURS ;
FADIA,
AMINA,
SARAH .

À MON CHER FRERE, Mohamed.

A mes grands parents ainsi que mon oncle , tous arrachés a notre affection , Qu'Allah vous accueille dans ses illustres Paradis, Amin.

A toute la famille BOUMEDIENE et BELMADI sans exception.

A Notre respectueux promoteur Mrs ELSSAID BILEL
Et a Mon Binôme NAWEL

A la promotion LMD Master SIC

Et tous ceux que j'aime, tous ceux qui m'aiment et tous ceux qui me sont chers.

Souad

REMERCIEMENT

Nous remercions tous d'abord le bon Dieu , le tout puissant de nous avoir armé de force et de courage pour mener à terme ce projet de fin d'études.

Notre profond gratitude s'adresse à

Monsieur « ESSAID BILLEL » encadreur de ce mémoire pour ses conseils fructueux et pour son aide précieuse qui nous a conduits à concrétiser ce travail.

Nos remerciements à tous les enseignants qui ont contribués à notre formation, et ainsi tous les gens de prés et de loin qui ont aidé à l'élaboration de ce sujet et à tous nos professeurs et travailleurs de l'université YAHIA FARES -MEDEA-

Ainsi que Tous les membres de jury qui nous ont fait l'honneur d'examiner notre travail.

Et enfin, que nos chers parents et familles, et bien avant tout, trouvent ici l'expression de nos remerciements les plus sincères et les plus profonds en reconnaissance de leurs sacrifices, aides, soutien et encouragement afin de nous assurer cette formation de Master dans les meilleures conditions.

ملخص :

الغرض من عملنا هو تحقيق خوارزمية تقدير التردد الأساسي والتي هي قادرة على تحديد القمة العظمى من إشارة الكلام بطريقة مثالية ،حتى مع نسبة الإشارة /الضوضاء سلبية. وتجمع الخوارزمية ضغط السعة غير الخطية لتخفيف مكونات الضجيج الضيقة النطاق مع تطبيق مرشح مشط في المجال الطيفي لقدرة التردد log_frequencyالذي يتم استجابة نبضاته لتخفيف مكونات الضوضاء المتغيرة بسلاسة.

كلمات مفتاحية :

إشارة الكلام ـ القمة العظمى ـ ضوضاء ـ تردد أساسي.

Résumé :

Le but de notre travail est de réaliser un algorithme d'estimation de la fréquence fondamental capable d'identifier le pitch d'un signal de la parole de manière fiable, même avec des rapports signal / bruit négatifs. L'algorithme combine la compression d'amplitude non linéaire, afin d'atténuer les composantes de bruit à bande étroite, avec un filtre en forme de peigne appliqué dans le domaine spectral de puissance à fréquence logarithmique, dont la réponse impulsionnelle est choisie pour atténuer des composants de bruit variant en douceur.

Mots clés :

Signal de la parole – pitch – bruit – fréquence fondamentale.

Abstract :

The purpose of our work is to realize a fundamental frequency estimation algorithm that is able to identify the pitch of voiced frames reliably even at negative signal to noise ratios. The algorithm combines non-linear amplitude compression, to attenuate narrow-band noise components, with a comb-filter applied in the log-frequency power spectral domain, whose impulse response is chosen to attenuate smoothly varying noise components.

Key words :

Speech signal – pitch – noise – fundamental frequency.

SOMMAIRE

LISTE DES FIGURES……I

LISTE DES TABLEAUX ……..III

LISTE DES ACRONYMES ……..IV

Introduction générale.....1

Chapitre I : Généralité sur le signal de la parole

1. Introduction :..... 2

2.Définition de La parole : ……..2

3.Définition du signal de la parole :..... 3

4. Les paramètres de la parole :..... 3

4.1 ***Le fondamental*** **:**……..3

4.2 ***Les formants*** **:**……..4

5. Caractéristiques du signal de la parole :……4

5.1.***L'intensité*** **:**…….7

5.2.***La hauteur*** **:**……..7

5.3.***Le timbre*** **:**……8

6. **Description de l'appareil phonatoire** :..... 8

7. **Les types de sons** :..... 9

7.1. ***Sons voisés*** :……9

7.2.***Sons non voisés*** :……10

8. **Voyelles et consonnes** :..... 10

9. **Conclusion** :..... 13

Chapitre II : Détection de fondamental de la parole

1. **Introduction** :..... 13

2. **Détection de fondamentale de la parole** :..... 13

2.1. ***Pitch vs fréquence fondamentale*** :..... 13

2.1.1. *Fréquence fondamentale* :..... 13

2.1.2. *Pitch* : 14
2.2. ***Algorithmes de détection du pitch*** : 14
2.2.1. *Détection du pitch dans le domaine temporel* : 15
2.2.1.1. *Analyse par auto-corrélation* : 15
2.2.1.2. *Détection de F_0 par la méthode de super résolution* : 16
2.2.1.3. *Détection de F0 par filtre inverse* : 17
2.2.1.4. *Analyse spectrale* : 19
2.2.2. *Détection du pitch dans le domaine fréquentiel* : 23
2.2.2.1 *Harmonic Product Spectrum (HPS)* : 23
3. **Conclusion** : 24

Chapitre llI : Résultats de simulation de la fréquence fondamentale

1. **Introduction** : 25
2. **Environnement de programmation** : 25
3. **Présentation de l'algorithme PEFAC (Pitch Estimation Filter with Amplitude Compression)** : 26
3.1. ***Méthode proposé*** : 27
3.2. ***Définition du filtre*** : 27
3.3. ***Compression*** : 28
3.4. ***Estimation de la fréquence fondamentale*** : 28
4. **Extraction du pitch** : 29
5. **Résultats expérimentaux** : 29
6. **Conclusion** : 39
Conclusion générale……..……..……..……40
Bibliographie……..…41

LISTE DES FIGURES

Chapitre I : **Généralité sur le signale de la parole**

Figure I. 1 : illustration des différences entre un phonème voisé et un phonème non voisé 5

Figure I.2 : illustration des différences entre les formants de différentes voyelles 6

Figure I.3: illustration des différences entre les excitations pour un phonème voisé et un phonème non voisé 7

Figure I.4 : L'appareil phonatoire 8

Figure I.5 **:** *Le Larynx* 9

Figure I. 6 **:** *Exemple de son voisé* 10

Figure I. 8 : Les phonèmes de la langue française 11

Figure I.9 : Les phonèmes de la langue anglaise 12

Chapitre II: **Détection fondamentale de la parole**

Figure II.1 : Signal vocal s(n) et son auto-corrélation temporelle pour les 100 premières millisecondes 16

Figure II.5 : Illustration de la méthode : Super Résolution F0 Détermination 18

Figure II.6 : Illustration de la méthode : SIFT 19

Figure II. 7 : Signal vocal 20

Figure II.8 : Allure du spectre logarithmique d'une trame donnée d'un signal vocal 21

Figure II.9 : Cepstre du signal vocal 22

Figure II.10 : Cepstre du signal vocal sans le premier coefficient 22

Figure II.11 : L'algorithme HPS 24

Chapitre III : Résultats de simulation de la fréquence fondamentale

Figure III.1. Filtrage h(q) depuis la formule (4) pour $\gamma = 1.5$ 28

Figure III.2. Schéma bloc de détection du pitch 29

Figure III.3. Le tracé temporel du signal audio «A8k16m.wav» 30

Figure III.4. Le tracé temporel du signal audio «A.wav» sur [2200 2456] 31

Figure III.5. Fenêtre Hamming 31

Figure III.6. Fenêtrage du signal audio «A.wav» 32

Figure III.7. FFt du signal 32

Figure III.8. Domaine logarithmique 33

Figure III.9. Spectre à long terme 34

Figure III.10. Tracé du spectre à long terme 35

Figure III.11. Détection du pitch 36

Figure III.12. Détection du pitch du signal audio « clean.wav » 36

Figure III.13. Le résultat de détection du pitch en utilisant un bruit " rain.wav" sur le signal audio 37

Figure III.14. Le résultat de détection du pitch en utilisant un bruit " sea1.wav" sur le signal audio 37

Figure III.15. Le résultat de détection du pitch du bruit blanc sur le signal audio 38

LISTE DES TABLEAUX

Tableau I.1. *Spectre à long terme de la voix artificielle* 35
Tableau III.2. *Detection de la fréquence de vibrations des cordes vocales chez l'homme et femme* 38
Tableau III.3. *Detection de la fréquence de vibration des cordes vocales avec différents bruits chez l'homme et femme* 38
Tableau III.4. *Detection de la fréquence de vibration des cordes vocales avec changement de variance du bruit blanc chez l'homme et femme* 39

LISTE DES ACRONYMS

DFT: Transformée de Fourier Discrète.

FFT: Transformé de Fourier Rapide.

FIR: Finite Impulse Response.

HPS: Harmonic Product Spectrum.

IFFT: Inverse Fast Fourier Transform.

PEFAC: Pitch Estimation Filter with Amplitude Compression.

PSD: Densité Spectrale de Puissance.

SIFT: Simplified Inverse Filtering Technique.

SNR: Rapport Signal / Bruit.

SRPD: Super Résolution Pitch Détermination.

Introduction générale :

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. Elle est l'une des formes les plus importantes de la communication humaine car elle joue un rôle important dans le domaine du traitement du signal.

Le but de ce travail est de chercher à détecter la fréquence fondamentale d'un signal de la parole dans un milieu bruité.

Une implémentation sur MATLAB est à réaliser afin de détecter le pitch. Ce dernier étant essentiellement une composante de nombreuses applications de traitement de la parole.

Ce mémoire regroupe trois chapitres, dont les contenus sont comme suit :
Dans le premier chapitre, nous allons présenter une définition de la parole, le signal de la parole et ces caractéristiques, accompagnées de la description du système phonatoire, les types du son et les voyelles et consonnes.

Dans le deuxième chapitre nous présenterons les différents algorithmes de détection de la fréquence fondamentale qui reste généralement la plus pratique et la plus importante, pour son rôle primaire en traitement de la parole.

Le dernier chapitre décrit notre conception, avec les résultats de simulation de notre travail obtenus fournis par MATLAB.

CHAPITRE I

Généralités sur le signal de la parole

1. Introduction :

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. [1]

L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine. L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée. [2]

Dans ce chapitre on a ainsi étudié la détection de la fréquence fondamentale qui est l'une des problèmes les plus importants du traitement de la parole. Il est notamment utilisé dans des systèmes de reconnaissance.

2. Définition de La parole :

La parole est le langage articulé humain destiné à communiquer la pensée, et est à distinguer des communications orales diverses, comme les cris, les alertes ou les gémissements. « Articuler la parole » consiste à former des signes audibles, les syllabes, formant les mots qui constituent des symboles. [13]

Elle est l'une des formes les plus importantes de la communication humaine car elle joue un rôle important dans le domaine du traitement du signal. Il existe deux caractéristiques fondamentales de la parole :

Les caractéristiques prosodiques et les caractéristiques articulatoires :

- Une caractéristique prosodique est construite à partir de trois paramètres acoustiques appropries au signal numérique de la parole produite : sa fréquence fondamentale f_0,

son énergie et son spectre. Chaque trait acoustique est lui-même intimement lié à sa grandeur perceptuelle (respectivement) : pitch, intensité et timbre.

- Une caractéristique articulatoire se traduit physiquement, dans la parole, par une variation de la pression de l'air causée et émise par le système articulatoire.

3. Définition du signal de la parole :

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique, et pragmatique. [2]

Le signal de la parole appartient à la classe des signaux acoustiques produits par des vibrations des couches d'air. Les fluctuations de la pression de l'air produisent des variations de ce signal, en fonction du temps, qui peuvent être enregistrées de façon analogique ou digitale. Ceci constitue une représentation élémentaire du signal de la parole [6]. Ce signal est le résultat du passage du flux laryngé (l'air qui la passe par nos poumons à travers nos cordes vocales) à travers le conduit vocal. Le signal de la parole est donc une onde acoustique qui se propage dans un milieu donné (en général l'air) et qui est le résultat de la modulation par le conduit vocal d'une onde d'excitation. [7]

La parole est un signal quasi stationnaire formé de parties stationnaires et de transitions entre ces différentes parties. C'est un signal non déterministe, dans le sens où deux réalisations d'un même mot auront nécessairement deux formes acoustiques différentes, même si elles sont produites par un même locuteur. Il faut distinguer deux types de variations : la variation acoustique et la variation temporelle non linéaire du signal.

4. Les paramètres de la parole :

Il existe plusieurs paramètres de la parole, on peut citer :

4.1 Le fondamental :

La vitesse à laquelle les cordes vocales s'ouvrent et se referment lors du processus de phonation, produit une vibration d'une hauteur variable appelée fréquence du fondamental dont la valeur est étroitement liée à la taille de l'appareil phonatoire de la personne, Cette fréquence est quasi stationnaire pour un signal de type voisé, elle varie de [6]:

- de 80 à 200 Hz pour une voix masculine.
- de 150 à 450 Hz pour une voix féminine.
- de 200 à 600 Hz pour une voix d'enfant.

Deux sons de même intensité et de même hauteur se distinguent par le timbre, qui est déterminé par les harmoniques du fondamental [14]. Un intérêt majeur pour la fréquence du fondamental se trouve dans les applications de la synthèse de parole.

4.2 Les formants :

Le spectre du signal vocal résultant de l'action des sources de sons sur le conduit vocal présente des maximums et des minimums qui correspondent aux résonances et aux anti résonances du conduit vocal, appelés formants et anti-formants. Du point de vue perceptif, seul les trois premiers formants jouent un rôle essentiel pour caractériser le spectre vocal.

On peut caractériser toute voyelle en n'utilisant que ses trois premiers formants. En général la fréquence du premier formant varie de 200 à 900 Hz, celle du second de 500 à 2500 Hz et le troisième se situe entre 1500 et 3500 Hz. Des formants d'ordre supérieur existent même si leur rôle sur le plan perceptif est limité, ils contribuent à caractériser la voix. [14]

5. Caractéristiques du signal de la parole :

La parole est produite lorsque l'air poussé hors des poumons, passe par les cordes vocales et le conduit vocal pour produire un son. Les modes de production des sons diffèrent grandement, ce qui permet une grande variété de phonèmes. On peut diviser ces phonèmes en deux classes : Voisés et non voisés. Les voyelles sont des exemples de sons voisés, alors que les fricatives sont des exemples non voisés. La figureI.1 montre dans le domaine temporel et fréquentiel, une voyelle (/a/) et une fricative (/s/). On remarque que le signal correspondant à une voyelle périodique, ce qui s'explique par le fait que la glotte s'ouvre et se referme à intervalles réguliers. Pour les fricatives, la glotte reste grande ouverte, ce qui explique l'absence de périodicité. Les phonèmes voisés se distinguent dans le domaine spectral par une structure fine harmonique, alors que la structure fine des phonèmes non voisés est stochastique.

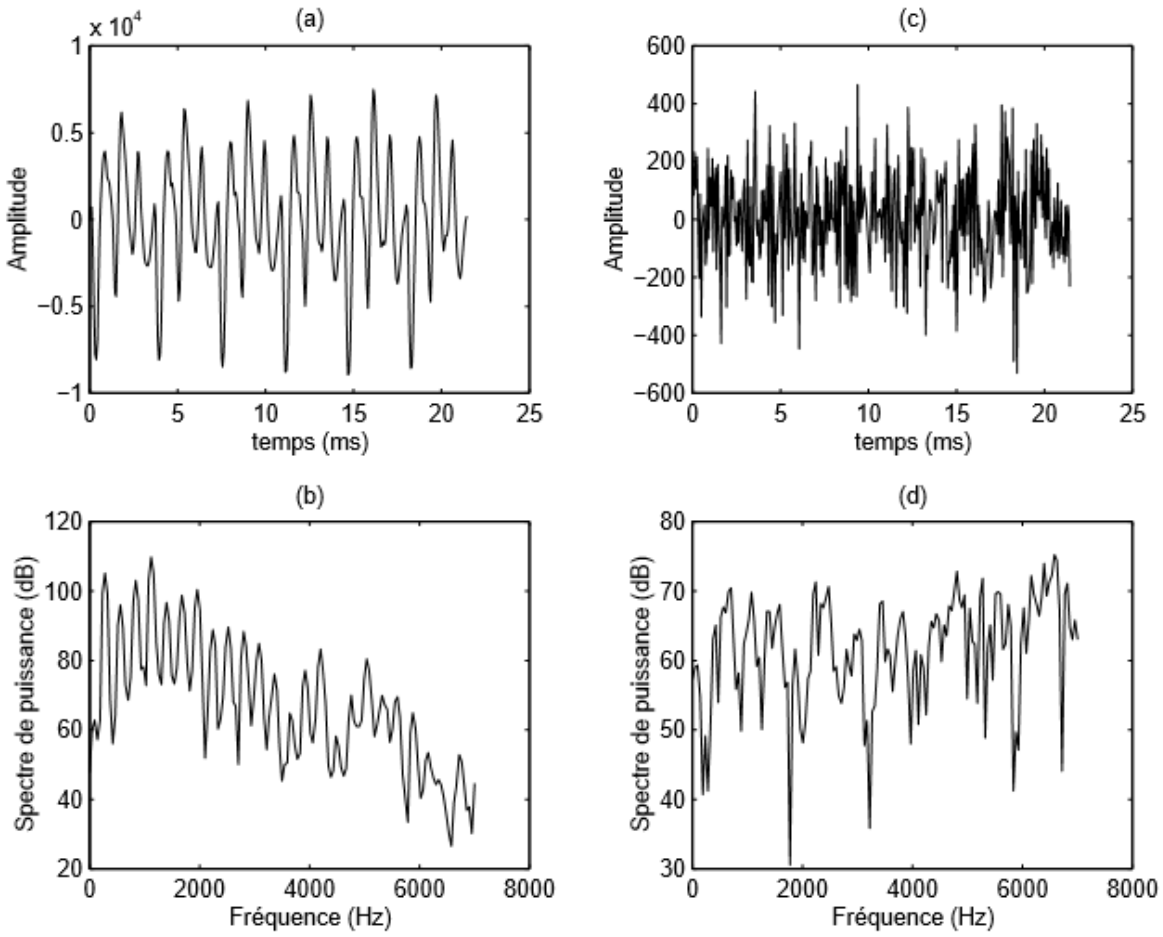


Figure I.1. *Illustration des différences entre un phonème voisé et un phonème non voisé. (a) le phonème voisé /a/ (voyelle) dans le domaine temporel et (b) fréquentiel. (c) le phonème non voisé /s/ (fricative) dans le domaine temporel et (d) fréquentiel.*

En plus d'être différents par leur voisement, les phonèmes se distinguent par la forme générale de leur spectre, soit l'enveloppe spectrale. Les résonances présentes dans l'enveloppe sont appelées « formants » et permettent de discriminer les voyelles entre elles. On peut voir à la figure I.2 la différence entre les formants d'un /a/ et ceux d'un /i/.

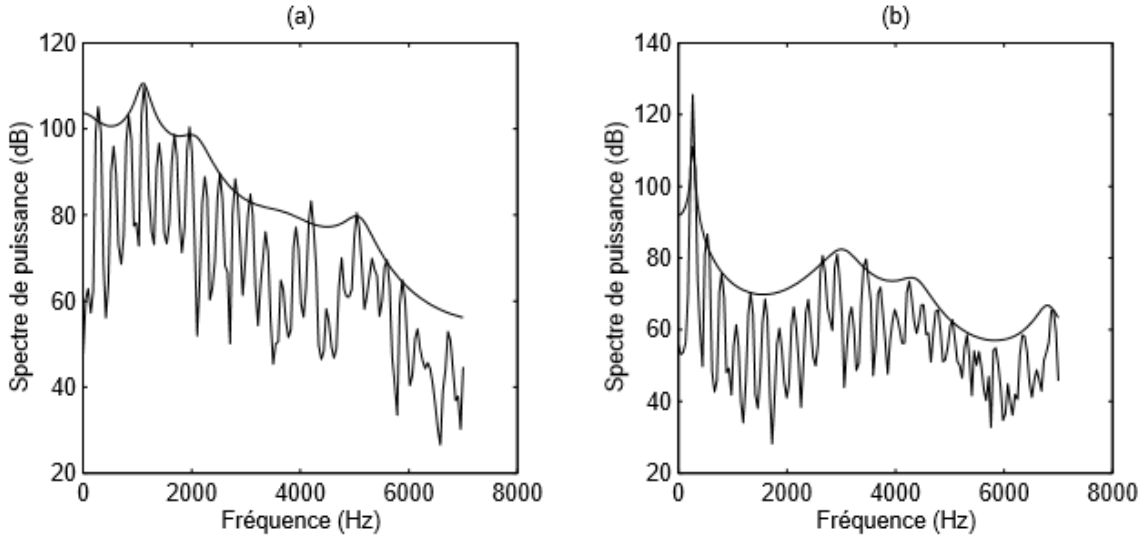


Figure I.2. *Illustration des différences entre les formants de différentes voyelles. (a) le phonème /a/ , et (b) le phonème /i/.*

Le signal de parole peut donc être modélisé simplement comme une source d'excitation (sortie des cordes vocales) filtrée par un filtre résonant (conduit vocal) représentant l'enveloppe spectrale. On connait cette représentation de la parole comme modèle « filtre-excitation ». Afin de modéliser la réponse en fréquence du conduit vocal, on utilise de façon générale un filtre tout-pôles, dont les coefficients évoluent dans le temps. Si le filtre tout-pôles représente bien l'enveloppe spectrale du signal, c'est dire que le spectre de l'excitation doit être généralement plat.

Il ya relativement peu de différences entre les excitations des différents phonèmes. La principale différence se situe entre les phonèmes voisés et non voisés. En effet, tel qu'illustré à la Figure I.3, l'excitation pour les phonèmes voisés (/e/) peut être approximée par un train d'impulsions et a donc un spectre comportant des harmoniques espacées de façon régulière.

L'excitation pour une fricative (/s/), quand à elle, ressemble plutôt à du bruit blanc et son spectre ne comporte pas d'harmoniques.

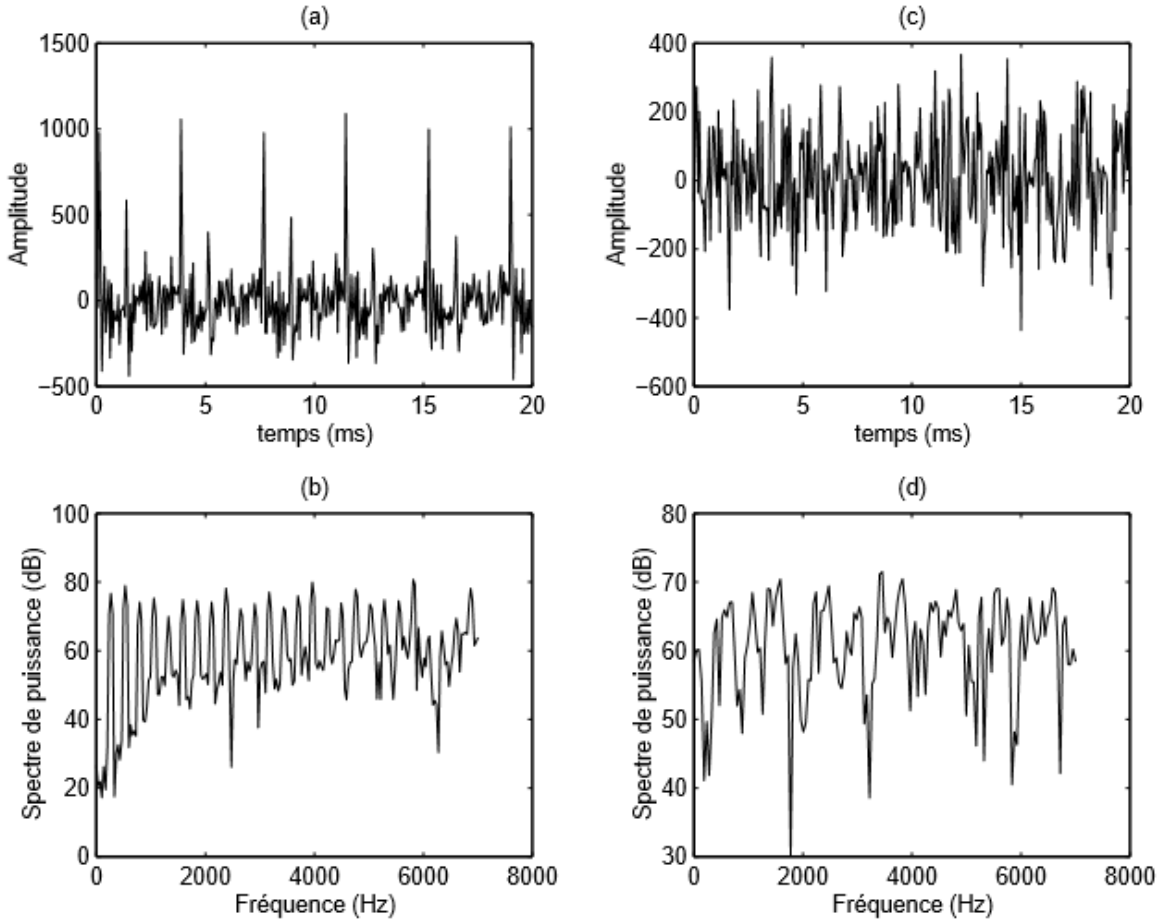


Figure I.3. *Illustration des différences entre les excitations pour un phonème voisé et un phonème non voisé. (a) l'excitation du phonème voisé /e/ (voyelle) dans le domaine temporel et (b) fréquentiel. (c) l'excitation du phonème non voisé /s/ (fricative) dans le domaine temporel et (d) fréquentiel. [3]*

Une parole est définie par trois caractéristiques :

5.1. L'intensité :

L'intensité de la parole c'est son amplitude et elle peut aussi être vue comme son volume sonore mesuré en décibels. Plus l'intensité d'un son est grande, plus son amplitude est grande. De plus, l'intensité du son diminue en fonction de la distance parcourue.

5.2. La hauteur :

La hauteur d'une parole c'est sa position sur l'échelle des graves et des aigus qui est déterminée par sa fréquence, cette échelle divise le spectre audible (20 à 20000Hz) en trois parties: les graves (de 20 à 250 Hz), les médiums (de 250 à 2000Hz) et les aiguës (au delà de 2000 Hz). Plus la fréquence est grande, plus le son est aigu et à l'inverse plus elle est petite plus le son sera grave.

5.3. Le timbre :

Le timbre de la parole est une de ses caractéristiques complexes qui permet de différencier deux paroles (sons) de même hauteur et amplitude. C'est par exemple grâce à ce timbre que l'on est capable de reconnaitre la voix d'un proche dans une foule de voix. C'est aussi le timbre qui explique la différence que l'on entend quand deux mêmes.

D'un point de vue scientifique, le timbre correspond à l'ensemble résultant de l'addition des sons des harmoniques et de l'onde fondamentale. Sa représentation se fait sous la forme d'un spectrogramme ou spectre de fréquences qui admet en abscisse la fréquence des ondes qui composent le son complexe et, en ordonnée, leur intensité à un instant t. [4]

6. Description de l'appareil phonatoire :

L'appareil phonatoire est l'ensemble des organes qui permettent de produire les sons constituant la voix.

Depuis deux siècles environ, l'évolution très rapide de la médecine et des techniques d'investigation ont permis de mieux comprendre comment le son était généré avant d'être rayonné dans le milieu extérieur.

Voici tout d'abord en coupe, l'appareil phonatoire humain. [5]

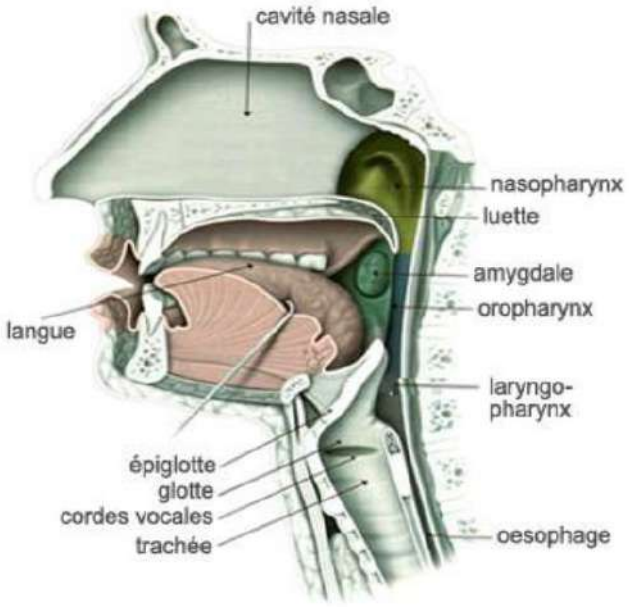


Figure I.4. *L'appareil phonatoire [8]*

Elle est composée principalement de trois éléments qui contribuent l'ensemble à la production de la parole. Ces éléments dont le contrôle et la coordination sont assurés par le système nerveux central, sont :

- L'ensemble poumons+bronches+trachée (l'appareil respiratoire) :

Les poumons permettent en premier lieu d'alimenter les tissus du corps en oxygène par une alternance de cycles d'inspiration et d'expiration. La phonation se produit dans la très grande majorité des cas au cours de l'expiration. Le flux d'air expulsé des poumons se déplace jusqu'à arriver au niveau du larynx où plusieurs cas se présentent.

- Le larynx :

Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. La figure I.5 nous présente une description du larynx. [6]

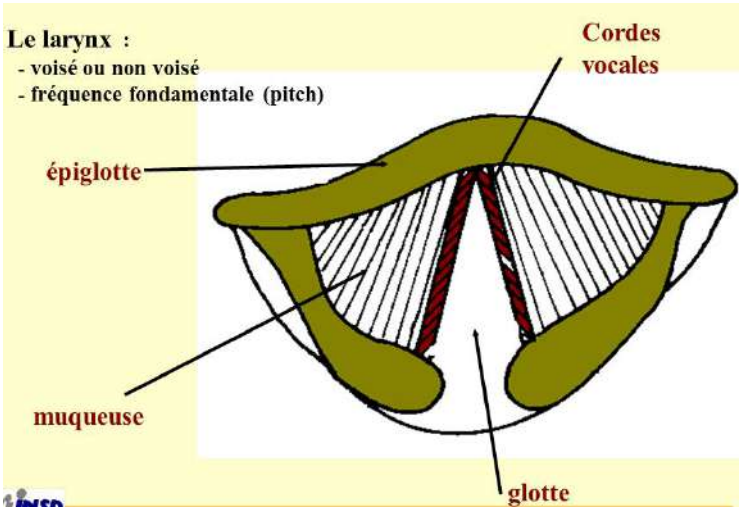


Figure I.5. Le *Larynx. [10]*

7. Les types de sons :

La parole est une succession d'évènements sonores faisant alternativement apparaître deux types de sons, selon que les cordes vocales vibrent (on dit alors que le son est **voisé**) ou pas (le son est alors **non voisé**).

7.1. Sons voisés :

Dans le premier cas, les vibrations des cordes vocales modulent le flux d'air issu des poumons. Cela résulte en une variation de la pression de l'air passant la glotte. Les oscillations des cordes vocales étant sinusoïdales, l'onde sonore produite au dessus de la glotte est également périodique sinusoïdale, en première approximation. Cette onde se propage alors dans le conduit vocal jusqu'aux lèvres.

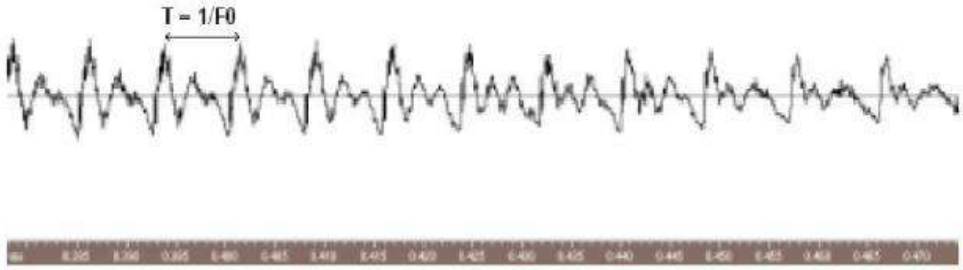


Figure I.6. *Exemple de son voisé.*

7.2. Sons non voisés :

Lorsqu'on produit des sons non voisés, les cordes vocales ne vibrent pas. Le son est alors produit par le frottement de l'air dans le conduit vocal entre la glotte et les lèvres. Si l'air ne rencontre aucun obstacle sur son chemin, le son produit est très faible (son de la respiration).

Par contre, si l'air rencontre une **constriction**, c'est-à-dire un rétrécissement du conduit vocal, sur son passage (par exemple entre les dents et les lèvres pour un 'f'), le son produit est un bruit (son apériodique continu) plus fort. Cela est dû à des propriétés mécaniques de l'écoulement de l'air à travers un tube : dans certaines conditions, le flux d'air qui s'écoule librement et régulièrement avant la constriction (on dit qu'il s'agit d'un écoulement **laminaire)** est transformé en écoulement turbulent plus du tout régulier au passage d'une constriction.

C'est cet écoulement **turbulent**, généré par le frottement de l'air dans la constriction qui provoque le bruit des sons non voisés. [11]

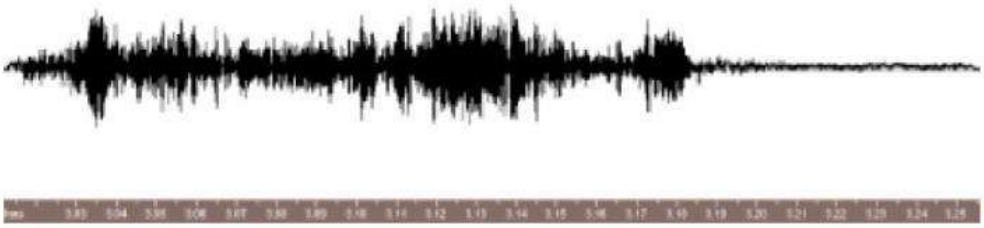


Figure I.7. *Exemple de son non voisé. [12]*

8. Voyelles et consonnes :

Le phonème est la plus petite unité de parole : il s'agit soit d'une voyelle, soit d'une consonne. Le nombre de phonèmes est toujours très limité, normalement inférieur à cinquante.

Par exemple : la langue française comprend 36 phonèmes (figure I.8) et les phonèmes de la langue anglaise sont représentés par la figure I.9.

Les voyelles sont produites lorsque le conduit vocal est ouvert et que le son est voisé.

Elles sont orales ou nasales selon que la cavité nasale n'est pas ou est mise en parallèle à la cavité buccale.

Consonnes			
	[p] paie	[t] taie	[k] quai
	[b] baie	[d] dais	[g] gai
	[m] mais	[n] nez	[ɲ] gagner
	[f] fait	[s] sait	[ʃ] chez
	[v] vais	[z] zéro	[ʒ] geai
	[w] ouais	[ɥ] huer	[j] yéyé
		[l] lait	[R] raie
Voyelles			
	[i] lit	[y] lu	[u] loup
	[e] les	[ø] leu	[o] lot
	[ɛ] lait	[œ] leur	[ɔ] lotte
	[a] là	[ə] le	
	[ɛ̃] lin	[ɑ̃] lent	[ɔ̃] long

Note : Les distinctions vocaliques [e]-[ɛ], [ø]-[œ] et [o]-[ɔ] ne sont pas faites dans tous les contextes et par tous les locuteurs du français. Par contre, certains locuteurs font aussi des distinctions entre patte et pâte, ([a]-[ɑ]) ainsi qu'entre brin et brun ([ɛ̃]-[œ̃]).

Figure I.8. *Les phonèmes de la langue française.*

/p/	**p**in, **p**ope	/i/	b**ea**d, s**ee**, p**ea**
/b/	**b**at, **b**aby	/ɪ/	b**i**d, s**i**t, p**i**ty
/t/	**t**ill, s**t**ripe	/e/	b**a**ke, h**ay**, b**ei**ge
/d/	**d**o, beg**d**ed	/ɛ/	h**ea**d, b**e**d, s**ai**d
/k/	**c**at, **k**ill, **ch**emist, anti**qu**e	/æ/	h**a**t, **a**pple
/g/	**g**oat, **gh**ost, **gu**ard	/a/[10]	p**ar**, c**ar**
/f/	**f**ine, ri**ff**le	/ə/	p**u**rr, w**e**re, ab**o**ve
/v/	**v**ine, ste**ph**en	/ɑ/	p**aw**, h**ea**rt, h**o**t
/θ/	**th**ief, ma**th**s	/ʌ/	f**u**n, c**o**me, d**o**ve
/ð/	**th**at, fa**th**er	/ɔ/	p**aw**, c**augh**t
/s/	**s**it, ri**c**e, **s**cissors	/o/	l**oa**d, g**o**
/z/	buz**z**, disea**s**e, sci**ss**ors	/u/	f**oo**d, n**u**de, wh**o**
/ʃ/	**sh**ip, su**g**ar	/ʊ/	b**oo**k, f**u**ll
/ʒ/	mea**s**ure, colla**g**e	/aj/	b**uy**, h**i**de
/ts/	ri**tz**, bli**tz**	/aw/	h**ou**se, h**ow**
/dz/	a**ds**	/oj/	b**oy**, **oi**l
/tʃ/	**ch**ance, ma**tch**, na**t**ure		
/dʒ/	**j**am, e**dg**e		
/m/	**m**ummy, ra**m**		
/n/	**n**un, **n**ut		
/ŋ/	ba**nk**, sin**g**er, ri**ng**		
/l/	**l**amb, **l**ello		
/ɹ/	**r**an, sti**rr**ing		
/w/	**w**et, **qu**ite,		
/j/	**y**et, **on**ion		
/h/	**h**at, **rh**eat		

Figure I.9. *Les phonèmes de la langue anglaise.*

Dans le cas des voyelles, le niveau supra glottique (pharynx et cavités buccale et nasale) fait office de résonateur, et permet de sélectionner les bandes de fréquence à renforcer par ajustement des fréquences et largeurs de bande des résonances acoustiques.

Les voyelles sont classées selon : la nasalité, l'ouverture du conduit vocal, la position de la construction du conduit vocal et l'arrondissement des lèvres. Les voyelles se différencient principalement les unes des autres par leur lieu d'articulation, leur aperture, et leur nasalisation. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles antérieures, les voyelles moyennes, et les voyelles postérieures, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles fermées et ouvertes. Les voyelles nasales [˜ɛ,ã,˜oe] diffèrent des voyelles orales [i,e,ɛ,a,o,y,u,oe,ø] en ceci que le voile du palais est abaissé pour leur prononciation, ce qui met en parallèle les cavités nasales et buccale.

Dans le cas des consonnes, des sources aéro-acoustiques de bruit sont générées selon la position des constructions. Les consonnes sont produites lorsqu'un rétrécissement apparaît dans l'appareil phonatoire. Elles sont fricatives si le rétrécissement est partiel ([f], [v]) ou occlusives si le rétrécissement est total, provoquant une augmentation de la pression et un relâchement brutal de celle-ci lors de l'ouverture ([p], [t]).

Les consonnes sont classées selon : le voisement, le mode d'articulation (occlusif, nasal, fricatif) et le lieu d'articulation (labiale, dentale, palatale). On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. [5]

8. Conclusion :

Dans le domaine du traitement du signal, la production de la parole concerne surtout la production de sons voisés, qui a un rôle essentiel dans le processus d'ouverture et de fermeture de la glotte. Au cours de ce chapitre, nous avons définit la parole, le signal de la parole et ces caractéristiques, et nous avons décrit le système phonatoire et les types du son, les voyelles et les consonnes, mais celui qui nous intéresse concerne la vibration des cordes vocales, ce qui détermine la nature des sons produits. Le premier propos de notre projet est de chercher à détecter la fréquence fondamentale.

CHAPITRE II

Détection de fondamental de la parole

1. Introduction :

La fréquence la plus basse dans le signal de parole est la fréquence Fondamentale (F_0) dénommé « pitch ». Elle représente la fréquence de vibration des cordes vocales et caractérise les segments Voisés de la parole à l'intérieur desquels elle évolue lentement dans le temps. La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction de son âge et de son sexe. Elle s'étend approximativement de 80 à 200 Hz chez les hommes, de 150 à 450 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [6].

Au cours des trente dernières années, un certain nombre d'algorithmes d'estimation de la F_0 ont été développés et rapportés. Ceci soulève la question évidente, pourquoi de nouveaux travaux sont toujours entrains d'être menés dans ce domaine? Ainsi, aucun des nombreux algorithmes rapportés ne s'est avéré entièrement satisfaisant. Par conséquent, les chercheurs continuent à essayer d'obtenir des techniques améliorés pour l'estimation de la F_0. L'évaluation des algorithmes de détection de la F_0 est une opération importante pour les applications relatives au traitement de la parole [7].

2. Détection de fondamentale de la parole :

2.1. *Pitch vs fréquence fondamentale :*

La fréquence fondamentale F_0 et le pitch ont une relation physiologique bien connue. Le pitch de la voix humaine est l'un des attributs acoustiques les plus facilement et rapidement maitrises. Il joue un rôle central dans, à la fois, la production et la perception de la parole. Le pitch est la fréquence fondamentale perçue par l'oreille. Cette fréquence donne des informations d'intonation de la phrase orale et aussi beaucoup d'informations concernant le locuteur.

2.1.1. *Fréquence fondamentale* :

La fréquence fondamentale F_0 est la fréquence de vibration des cordes vocales. Dans le domaine temporel, c'est la période d'un signal voisé à un instant donné. Cette fréquence fondamentale nous donne un indice sur le pitch du signal vocal. F_0 représente un indicateur émotionnel et est considérée comme étant la quantité à estimer par un algorithme de détection du pitch. Pour le signal de parole, sa fréquence fondamentale n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, déterminée par la tension des muscles qui contrôlent celles-ci. La plage de variation moyenne de cette fréquence varie d'un locuteur à un autre en fonction de son âge et de son sexe. Elle s'étend approximativement de

80 à 200 Hz chez les hommes, de 150 à 450 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [17]. La détection de F_0 joue un rôle essentiel dans le domaine de traitement de la parole et doit être, si possible, pour les applications modernes, calculée en temps réel.

2.1.2. *Pitch :*

C'est grâce au pitch que nous pouvons séparer les différentes phrases d'un message, ou encore que nous pouvons discerner plusieurs parties dans une même phrase. Il est aussi l'indice majeur de l'intonation; donc il joue un rôle extrêmement important dans la segmentation des informations contenues dans un message. Le pitch, ou plutôt les variations du pitch, nous donnent également des informations sur l'émotivité d'un locuteur. C'est grâce à lui que l'on distingue une affirmation d'une interrogation ou d'un ordre. Il est également l'indice de la mélodie de la voix, c'est l'un des paramètres sur lequel nous pouvons agir dans certains cas pour donner plus de naturel aux voix pathologiques. Dans certaines langues étrangères (le chinois par exemple), sa variation code l'information directement. Cette variation est imposée par la langue et parfois peut différencier deux mots qui se prononcent de la même façon à la seule différence presque la variation du pitch est différente. Enfin en musique, on peut jouer sur ce paramètre pour corriger les imperfections vocales ou les fausses notes de certains chanteurs. La détection du pitch rencontre certains problèmes importants, dont nous citons :

- la variation de la fréquence fondamentale F_0 dans le temps (où le suivi de pitch en temps réel est encore difficile mais nécessaire).
- l'apparition d'harmoniques qui faussent la détection.
- la difficulté de réaliser la décision voisée/non voisée dans les contours de pitch.

2.2. *Algorithmes de détection du pitch* :

Le problème de la détermination de F_0 a été approché par le développement d'un algorithme précis de détermination de pitch.

Au fil des années, le processus de suivi du pitch a occupé un large champ d'applications et a préoccupé de nombreux chercheurs à creuser dans le domaine de détection du pitch, que ce soit dans le domaine temporel, dans le domaine fréquentiel , ou encore dans le domaine des ondelettes (multi résolution). Un point très important à noter est que les algorithmes de détection du pitch (ADPs ou PDAs en anglais pour Pitch Détermination Algorithmes), ont conçu une série de techniques sophistiquées pour la décision de voisement, mais la plupart

d'entre eux font appel à un post-traitement. Récemment, un nouveau défi pour les ADPs consiste à éliminer tout type de techniques de post-traitement afin de respecter le temps réel.

2.2.1. *Détection du pitch dans le domaine temporel* :

Les techniques d'estimation du pitch, dans le domaine temporel, opèrent directement à partir du signal.

2.2.1.1. *Analyse par auto-corrélation* :

L'auto-corrélation temporelle est l'une des méthodes les plus connues de détection du pitch dans un signal de parole. Son principal but, dans la plupart des systèmes, est d'aplatir spectralement le signal de manière à éliminer les effets de l'appareil vocal. Bien qu'un grand nombre de méthodes différentes ont été proposées pour la détection du pitch comme la super résolution, filtre inverse et l'analyse spectral. L'auto-corrélation reste depuis toujours l'une des plus robustes et fiables. L'analyse par auto-corrélation est une méthode basée sur la détection des maximas locaux de la fonction d'auto-corrélation d'un signal vocal. (Formule 1)

$$R(n) = \sum_{i=0}^{N-n-1} s(i)s(n+i) \qquad \textbf{(II.1)}$$

Où :

s(n) est le signal vocal échantillonné.

N est le nombre d'échantillons.

En cherchant le maximum de cette fonction, nous pouvons en déduire facilement la fréquence fondamentale F_0

$$F_0 = F_e / n_{max} \qquad \textbf{(II.2)}$$

Où :

Fe est la fréquence d'échantillonnage.

n_{max} = *argmax n(R(n))* (argmax est l'ensemble des points en lesquels une expression atteint sa valeur maximale).

Cependant, l'inconvénient majeur de l'analyse par auto-corrélation est l'apparition d'harmoniques et de sous-harmoniques de la fréquence F_0 (Figure II.1).

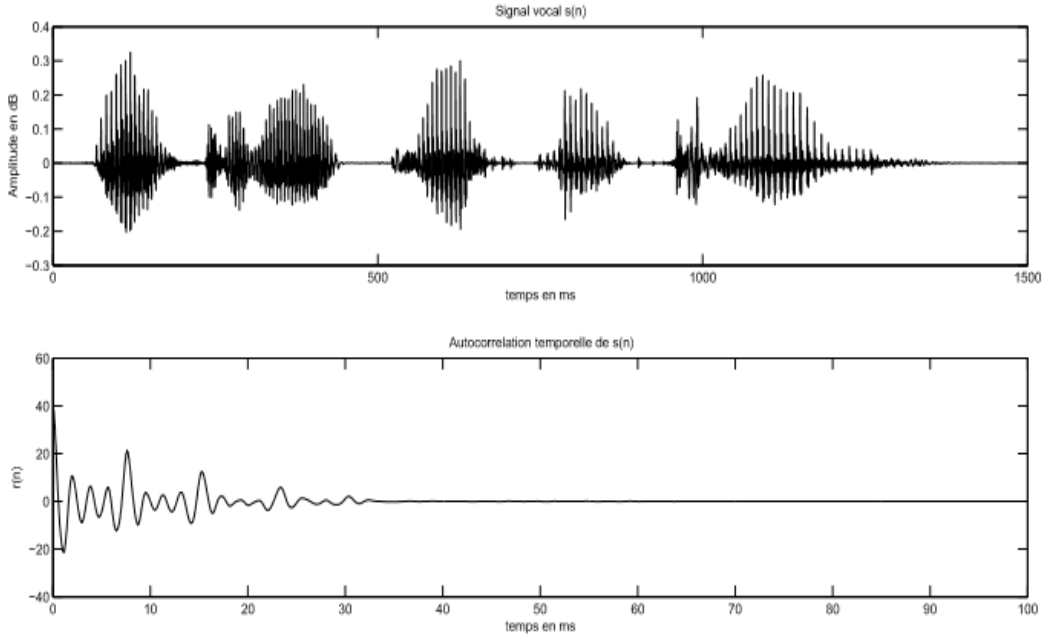


Figure II.1. *Signal vocal s(n) et son auto-corrélation temporelle pour les 100 premières millisecondes.*

2.2.1.2. *Détection de F_0 par la méthode de super résolution :*

La méthode SRPD (Super Résolution Pitch Détermination, en anglais), présentée par Yoav Medan, s'appuie sur l'analyse par auto-corrélation temporelle mais cherche l'existence d'un multiple ou un sous multiple de F_0, au lieu de F_0, afin de corriger le point faible de la méthode précédente.

Pour chaque instant t, on définit deux signaux $x_\tau(t, t_0)$ et $y_\tau(t, t_0)$ comme suit :

$$\begin{cases} x_\tau(t, t0) = s(t)\, \omega\tau(t - t0) \\ y_\tau(t, t0) = s(t + \tau)\, \omega\tau(t - t0) \end{cases} \quad \textbf{(II.3)}$$

Où *s(t)* désigne le signal de parole, et $\omega\tau(t)$ est une fenêtre rectangulaire de longueur τ secondes donnée par :

$$\begin{cases} \omega\tau(t) = 1; & 0 < t < \tau \\ \omega\tau(t) = 0; & sinon \end{cases} \quad \textbf{(II.4)}$$

La méthode suppose que la trame de parole, commence à t = t0 et se compose exactement de deux périodes de pitch de durée τ = T0 ; $x\tau(t, t0)$ la première période et $y\tau(t, t0)$ la seconde période, avec T0 la période du pitch (en secondes) liée à l'instant

t = t0. L'hypothèse est faite de manière à ce que la similitude entre ces deux périodes de pitch successives (modulées en amplitude) soit le plus possible élevée. Ainsi, le modèle de similitude est donné par l'équation suivante :

$$xt0(t,t0) = a(t0)yt0(t,t0) + e(t,t0) \quad \textbf{(II.5)}$$

Avec a(t0) est le facteur de modulation d'amplitude positive (gain), ce qui reflète le changement dans le volume d'impulsion glottale. Le terme d'erreur e(t,t0) reflète les différences entre les deux périodes. Pour optimiser la similitude entre les deux segments $\mathrm{xt0(t,t0)}$et $\mathrm{yt0(t,t0)}$ et réduire au minimum le carré normalisé d'erreur, le problème d'optimisation est donné par la formule suivante :

$$T0 = argmin\ \tau, a(t0)\{\frac{\int_{t0}^{t0+\tau}(x_\tau(t,t0)-a(t_0)y_\tau(t,t0))^2 dt}{\int_{t0}^{t0+\tau}(x_\tau(t,t_0)dt}\} \quad \textbf{(II.6)}$$

Où :

argmin : est l'ensemble des points en lesquels une expression atteint sa valeur minimale. L'optimisation de T0 peut donc être considérée comme une maximisation du rapport signal-sur-bruit (le bruit choisi est un bruit gaussien). Ainsi la fréquence fondamentale F0 est déduite comme suit :

$$F0 = \frac{1}{T0} \quad \textbf{(II.7)}$$

Le calcul de F0 sera réitéré sur la portion [0, T0] jusqu'à ce qu'on ne trouve plus un nouveau T0 multiple du précédent (voir figure 2).

2.2.1.3. Détection de F0 par filtre inverse :

La méthode SIFT (Simplified Inverse Filtering Technique), tente d'éliminer les perturbations dues au troisième niveau de l'appareil phonatoire. En effet, la méthode suppose ici que le signal vocal dont on veut déterminer la fréquence fondamentale, passe dans un filtre (à savoir le niveau supra glottique) avant d'être émis. On cherche donc à filtré le signal reçu par un filtre inverse avant de déterminer F0 en maximisant l'auto-corrélation temporelle. La méthode s'intéresse aussi au voisement du signal. En effet, si le signal n'est pas voisé, la F0 déterminée n'a aucune réalité physique. La classification voisé/non voisée s'effectue par seuillage sur la F0 déterminée.

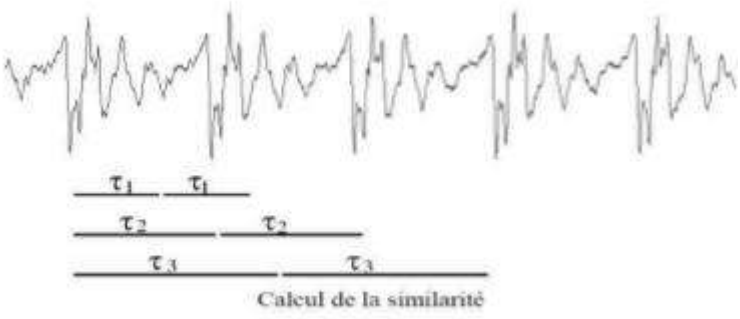


Figure II.2. *Illustration de la méthode : Super Résolution F0 Détermination.*

L'avantage principal de l'algorithme SIFT est qu'il est composé d'un nombre relativement restreint d'opérations arithmétiques élémentaires, L'inconvénient de cette méthode est que le filtrage inverse peut supprimer des informations importantes de la source glottique en ne laissant sortir que du bruit blanc. Un autre problème lié à la méthode SIFT peut se présenter dans le cas d'une F0 élevée comme par exemple 180 Hz : en effet l'auto-corrélation pourra faire apparaitre, dans ce cas, deux pics importants l'un à 180 Hz et l'autre à 90 Hz et fournir par conséquent une fausse détection à 90 Hz. Selon la figureII.3, la méthode SIFT suit les étapes suivantes : le signal vocal est, dans un premier temps, soumis à un filtre passe-bas à l'aide d'un filtre FIR (Finite Impulse Response) avec une fréquence de coupure Fc suivi d'un sous-échantillonnage. Une prédiction linéaire (PL) est effectuée sur le signal résultant en utilisant une longueur de trame de 32 ms et un décalage de trame de 12 ms. La prédiction du signal résiduel est obtenue par un filtrage inverse du signal pour donner un signal spectralement aplati. Dans ce signal résiduel, la recherche des pics se limite à une plage de valeurs, qui correspond à une F0, entre 40 Hz à 500 Hz. Ces pics doivent dépasser une certaine valeur de seuil pour être pris en compte. Dans le cas contraire, si les deux trames précédentes sont voisées, la valeur maximale courante est comparée à un deuxième seuil inférieur au premier. Si ce seuil est effectivement dépassé, la trame est également classée voisée, sinon, elle est classée non voisée. Quand une trame unique non voisée est détectée entre deux trames voisées, la décision sur cette trame sera changée de non voisée à voisée après un calcul de la moyenne des pics des deux trames voisées adjacentes. Pour augmenter la résolution de la période du pitch estimée, l'auto-corrélation de chaque trame est interpolée de manière parabolique dans le voisinage du pic principal et de ses deux valeurs adjacentes.

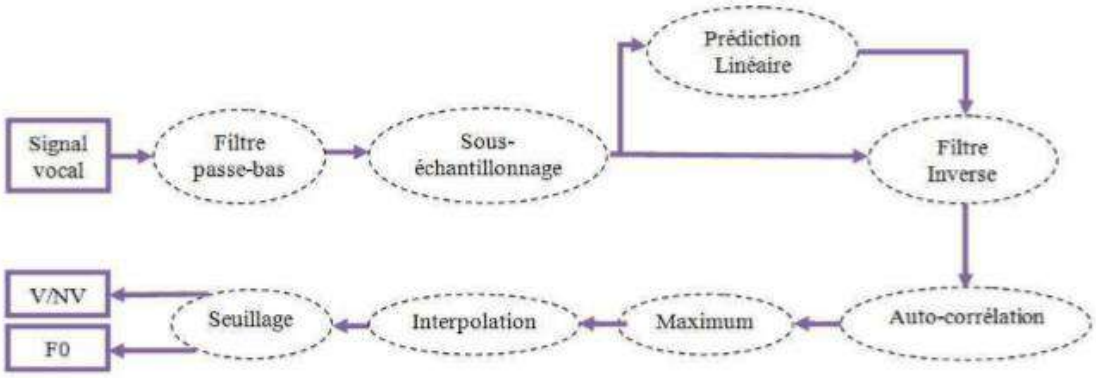


Figure II.3. *Illustration de la méthode : SIFT.*

2.2.1.4. Analyse spectrale :

Dans sa forme la plus basique, le système de production de la parole voisée ne se compose que de la source vocale et du conduit vocal. L'analyse spectrale suppose que le signal source *s(n)*, qui représente les bouffées de l'air périodiques reçues via les cordes vocales, est produit par un signal excitateur *e(n)* qui traverse un système linéaire passif, qui est le niveau supra glottique, de réponse impulsionnelle *h(n)*. L'algorithme CPD (Cepstrum Pitch Détermination) est basé sur l'analyse spectrale.

$$s(n) = e(n) \otimes h(n) \quad \textbf{(II.8)}$$

Où :

$\otimes$ représente l'opération de convolution.

Afin de connaitre l'évolution du spectre de la parole, l'utilisation de la Transformée de Fourier Rapide (FFT : Fast Fourier Transform) est envisageable :

$$F(k) = \sum_{n=0}^{N-1} s(n)\exp(-j\,\frac{2\pi}{N}\,kn) \quad \textbf{(II.9)}$$

Où : *N* : représente la longueur de la fenêtre d'analyse

s(n) : le signal d'entrée.

F(k) : le k^{ème} coefficient spectral complexe.

Pour utiliser la transformée de Fourier, il faut un signal périodique, et comme le signal de parole est purement non stationnaire, l'utilisation d'un fenêtrage est nécessaire. Parmi les fenêtres existantes, nous citons comme exemple la fenêtre rectangulaire, triangulaire ou encore celle de Hamming.

La transformée de Fourier du signal vocal s(n), $FmL_{,}$ sur une période d'échantillonnage T, sera représentée comme suit :

$$F_{ml,\tilde{\varphi}}(\omega) = \sum_{n=-\infty}^{\infty} s(n)\omega(n-ml)exp(-j\,\tilde{\varphi}n) \qquad \textbf{(II.10)}$$

où : ω : est la fenêtre d'analyse appliquée au signal *s(n).*

L : est le décalage temporel (en nombre d'échantillons) entre chaque trame analysée m.

Afin de réaliser une analyse spectrale, le signal vocal (Figure II.4) peut être séparé des contributions de l'excitation et du conduit vocal selon les étapes du traitement homomorphique suivantes :

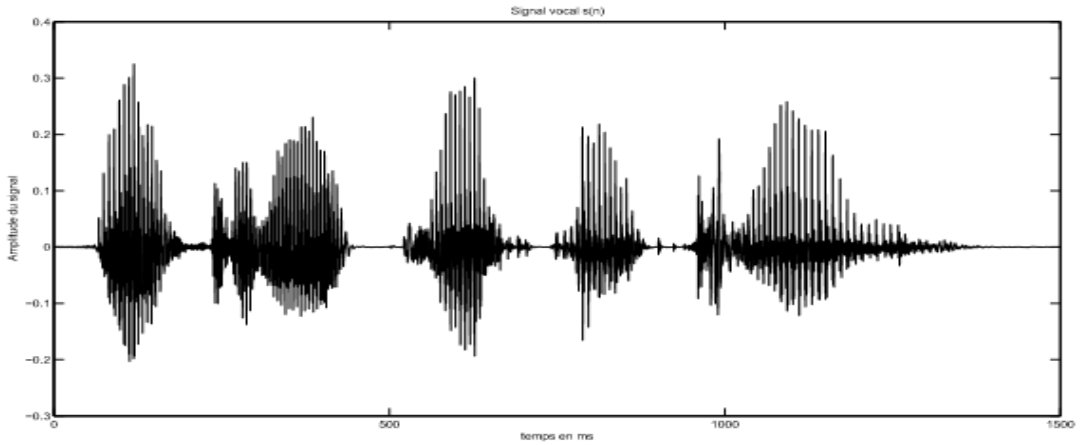


Figure II.4. *Signal vocal.*

a. Transformée de Fourier Rapide (FFT) :

La transformée de Fourier, qui est l'un des outils fondamentaux du traitement dsignal, permet le passage de la convolution (formule 8) vers une multiplication (formule 11) pour obtenir les signaux spectraux.

$$S(f) = E(f).H(f) \qquad \textbf{(II.11)}$$

Où :

E(f) : est le spectre d'amplitude de l'excitation.

H(f): est le spectre d'amplitude du conduit vocal.

S(f) : est le spectre d'amplitude du signal de parole alors égal le produit de *E(f)* par *H(f).*

b. Logarithme :

Une nouvelle fonction a été proposée dans laquelle les contributions de la source vocale et du conduit vocal sont à peu prés indépendantes ou facilement identifiables et séparables. Le logarithme de la transformée de Fourier du spectre d'amplitude permet de séparer les effets du conduit vocal et du signal excitatif. La raison de cette séparation est que le logarithme d'un produit est égal à la somme des logarithmes des multiplicandes. Le spectre logarithmique du signal vocal est donné par la formule II.12, et représenté par la Figure II.8.

$$log[S(f)] = log[E(f)] + log[H(f)] \quad \textbf{(II.12)}$$

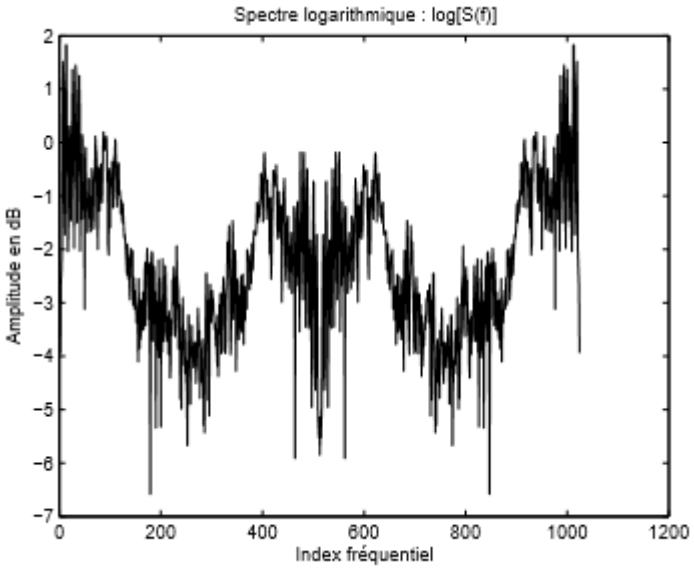


Figure II.8. *Allure du spectre logarithmique d'une trame donnée d'un signal vocal.*

c. Transformée de Fourier Inverse :

Par transformation inverse (IFFT Inverse Fast Fourier Transform), nous obtenons le cepstre, qui est une forme duale temporelle du spectre logarithmique.

$$TF^{-1}(\log[S(f)]) = TF^{-1}(\log[E(f)]) + TF^{-1}(\log[H(f)]) \quad \textbf{(II.13)}$$

Après cette transformation, la période du pitch peut être calculée à partir du signal cepstral (figure II.9), par la détermination de l'index du pic principal du cepstre (hors premiers coefficients).

Le cepstre du signal vocal représente à la fois le cepstre de l'excitation et le cepstre du conduit vocal. La séparation se fait par une simple élimination des premiers coefficients du signal cepstral. En éliminant seulement le premier coefficient, les pics sont plus visibles, comme le montre la Figure II.10.

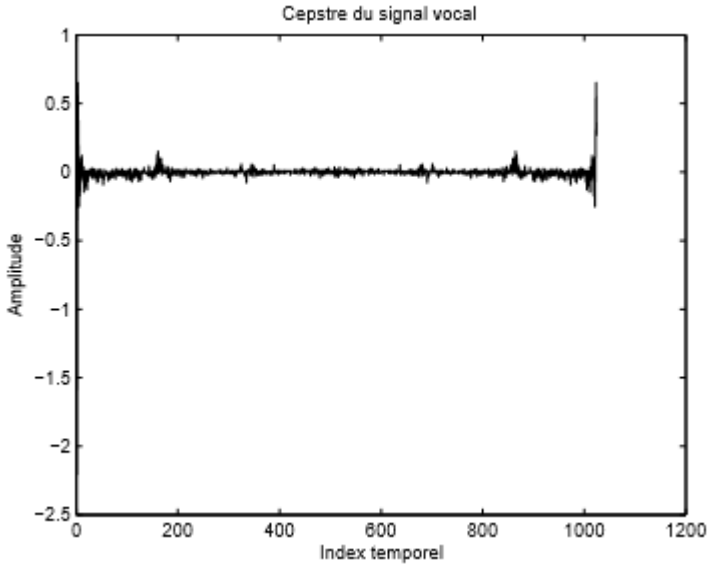


Figure II.9.*Cepstre du signal vocal*.

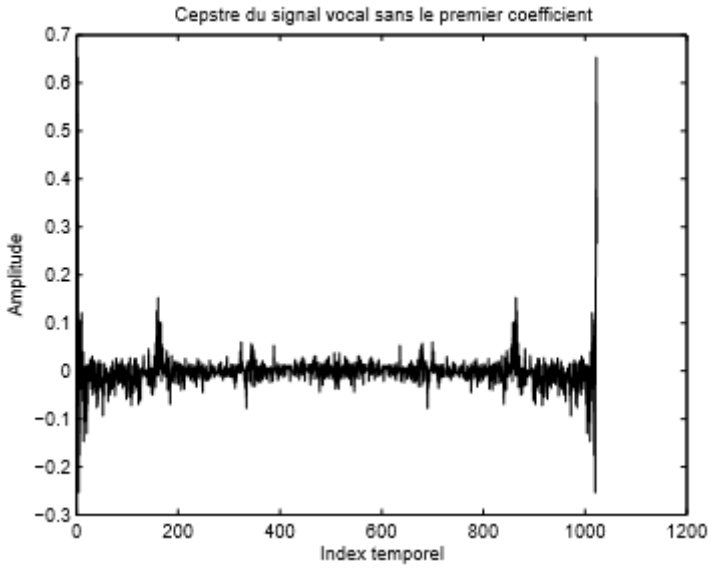


Figure II.10. *Cepstre du signal vocal sans le premier coefficient.*

2.2.2. Détection du pitch dans le domaine fréquentiel :

Dans le domaine fréquentiel, les détecteurs de pitch supposent que si le signal est périodique dans le domaine temporel, le spectre de fréquence du même signal comportera une série d'impulsions relatives à la fréquence fondamentale et à ses harmoniques.

En dépit de son coût de calcul, la transformée de Fourier a rendu le traitement du signal dans le domaine fréquentiel une réalité pratique. Car cette transformation révèle souvent les caractéristiques d'un signal qui sont presque impossibles à détecter autrement.

Presque toutes les méthodes d'estimation du pitch qui opèrent dans le domaine fréquentiel, présentent des pics répétés dans le spectre. Gareth Middleton a exploité ce fait dans son article sur la correction du pitch, dans lequel il cumule des spectres de fréquence compressés dans un processus appelé : Harmonic Product Spectrum (HPS).

2.2.2.1 Harmonic Product Spectrum (HPS) :

L'analyse par HPS utilise le sous-échantillonnage, pour extraire de nouveaux spectres, qu'elle multipliera par le spectre d'origine. Le but de cette analyse est de comprimer le spectre, car l'effet cumulatif de cette multiplication peut changer l'emplacement du pic de fréquence afin de placer correctement la F0.

Pour chaque trame stationnaire du signal vocal *s(n)*, le logarithme de sa densité spectrale de puissance est calculé le long de l'axe des fréquences sur des facteurs entiers. La valeur logarithmique de l'HPS est obtenue par l'addition du spectre logarithmique original et de ses versions compressées (décimées) :

$$HPS(f) = argmax_f \sum_{k=1}^{R} log|S(kf)|^2 \qquad \textbf{(II.14)}$$

Où:

R : représente le nombre total des spectres impliqués dans le calcul.

S(kf) est la k^{ème} transformée de Fourier Discrète de *s(n)*.

Si le signal d'entrée est une note de musique, son spectre se compose d'une série de pics, correspondant à la fréquence fondamentale et de ses composantes harmoniques. C'est pourquoi, lorsque l'algorithme additionne un certain nombre de spectres logarithmiques

compressés, les sommets harmoniques vont s'additionner pour exhiber au final, un proéminent relatif à la fréquence fondamentale (figure II.11). [5]

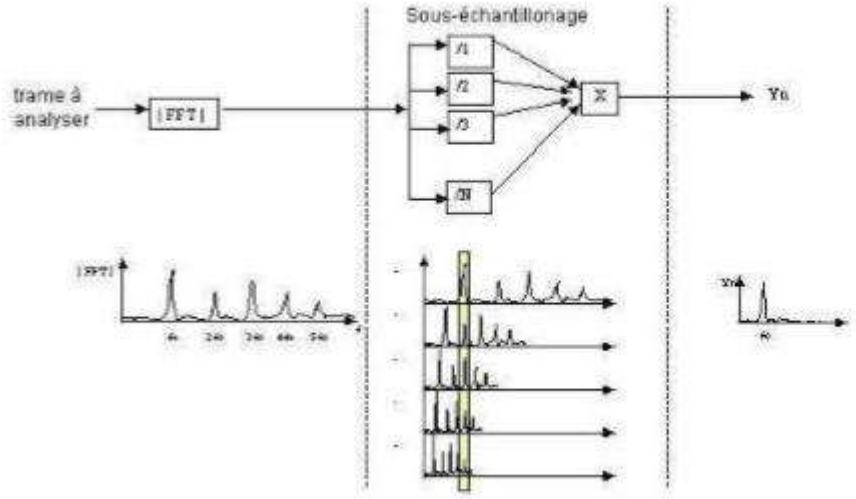


Figure II.11. *L'algorithme HPS.*

3. Conclusion :

Dans ce chapitre nous avons présenté les différents algorithmes de détection du pitch (fréquence fondamentale), qui est l'une des problèmes les plus importants du traitement de la parole. Il est notamment utilisé dans des systèmes de reconnaissance.

CHAPITRE III

Résultats d'estimation

1. Introduction :

Dans ce chapitre, nous nous présentons les résultats de simulation pour l'estimation de la fréquence fondamentale de la parole. Dans une première partie, nous commençons à présenter l'environnement de programmation et d'expliquer le principe de l'algorithme « PEFAC » (Pitch Estimation Filter with Amplitude Compression). Une deuxième partie sert à estimer la fréquence fondamentale d'un signal audio.

2. Environnement de programmation :

Nous avons implémenté nos algorithmes à l'aide de l'environnement de programmation MATLAB.

Ce dernier ayant les propriétés suivants :

- ✓ MATLAB : est un environnement puissant, complet et facile à utiliser destiné au calcul scientifique. Il apporte aux ingénieurs, chercheurs et à tout scientifique un système interactif intégrant calcul numérique et visualisation. C'est un environnement performant, ouvert et programmable qui permet de remarquables gains de productivité et de créativité. Comme il est ouvert et extensible pour le calcul, Dispose de plusieurs centaines (voire milliers, selon les versions et les modules optionnels autour du noyau Matlab) de fonctions mathématiques, scientifiques et techniques. L'approche matricielle de MATLAB permet de traiter les données sans aucune limitation de taille et de réaliser des calculs numériques et symboliques de façon fiable et rapide. Grâce aux fonctions graphiques de MATLAB, il devient très facile de modifier interactivement les différents paramètres des graphiques pour les adapter selon nos souhaits.
- ✓ MATLAB : possède son propre langage, intuitif et naturel qui permet des gains de temps de CPU spectaculaires par rapport à des langages comme le C, le Turbo Pascal et le Fortran.
- ✓ MATLAB : permet le travail interactif soit en mode commande, soit en mode programmation, tout en ayant toujours la possibilité de faire des visualisations graphiques. Considéré comme un des meilleurs langages de programmations (C ou Fortran), MATLAB possède les particularités suivantes par rapport à ces langages :

- la programmation facile.
- la continuité parmi les valeurs entières, réelles et complexes.

- la gamme étendue des nombres et leurs précisions.
- la bibliothèque mathématique très compréhensive.
- l'outil graphique qui inclut les fonctions d'interface graphique et les utilitaires.
- la possibilité de liaison avec les autres langages classiques de programmations (C ou Fortran). [15]

3. Présentation de l'algorithme PEFAC (Pitch Estimation Filter with Amplitude Compression) :

PEFAC, un algorithme fondamental d'estimation de fréquence capable d'identifier le pitch des sources sonores de manière fiable, même avec des rapports signal / bruit négatifs. L'algorithme combine la compression d'amplitude non linéaire, afin d'atténuer les composantes de bruit à bande étroite, avec un filtre en forme de peigne de Dirac appliqué dans le domaine spectral de puissance, dont la réponse impulsionnelle est choisie pour atténuer des composants de bruit variant en douceur.

Les estimateurs de pitch peuvent être largement divisés en trois groupes selon qu'ils fonctionnent dans le domaine temporel, fréquentiel ou du temps-fréquence. En règle générale, la première catégorie trouve des pics dans la fonction d'auto corrélation, la seconde recherche des pics harmoniques dans le spectre de puissance tandis que le troisième effectue une analyse dans le domaine temporel sur les sorties d'un banc de filtres passe-bande.

Dans les situations où le bruit acoustique est élevé ou lorsque la distance entre le microphone et le locuteur est grand, le rapport signal / bruit (SNR) d'un signal vocal acquis peut être très faible. Dans de telles circonstances, la performance des algorithmes d'estimation de pitch se dégrade [1], et de nombreuses méthodes deviennent inutilisables en dessous de 0 dB SNR. Au cours des dernières années, un certain nombre d'algorithmes robustes au bruit ont été conçus, mais une estimation de fréquence fondamentale fiable aux SNR négatives demeure un problème difficile.

On met en évidence le spectre dans le domaine log-frequency avec un train d'impulsions de Dirac, et sélectionnant le sommet le plus élevé. L'algorithme utilisé dans notre mémoire, de même, estime la fréquence fondamentale de chaque trame en convertissant sa densité spectrale de puissance dans le domaine log-frequency avec un filtre qui résume l'énergie des harmoniques de pitch tout en rejetant un bruit additif qui a un spectre de puissance en douceur

variable. La compression d'amplitude est appliquée avant de filtrer pour atténuer les composants du bruit à bande étroite.

3.1. Méthode proposé :

Pour une source parfaitement périodique à la fréquence f_0, le modèle du signal au temps t dans le domaine de densité spectrale de puissance est :

$$\boldsymbol{Y_t\ (f) = \sum_{k=1}^{k} a_{k,t}\delta(f - kf_0) + \quad N_t(\mathrm{f})} \qquad \textbf{(III.1)}$$

Où :

$N_t\ (f)$: représente la densité spectrale de puissance du bruit indésirable, $a_{k,t}$ la puissance de la $k^{ième}$ harmonique. Dans l'espace Log-fréquence, le modèle du signal peut être exprimé comme suit:

$$\boldsymbol{Y_t\ (\mathrm{q}) = \sum_{k=1}^{k} a_{k,t}\delta(q - logk - logf_0) + \quad N_t(q)} \qquad \textbf{(III.2)}$$

Où : $q = log\ f$.

Dans ce domaine, l'espacement des harmoniques est indépendant de f_0 et son énergie peut donc être combinée par convolution $Y_t\ (q)$ avec un filtre à train d'impulsion de Dirac :

$$\boldsymbol{h(q) = \sum_{k=1}^{k} \delta(q - logk)} \qquad \textbf{(III.3)}$$

La convolution $Y_t(q)*h(q)$ inclura un pic à $q_0 = log\ f_0$ et des pics additionnels correspondant à des multiples rationnels de f_0.

3.2. Définition du filtre :

En pratique, la largeur de chaque pic harmonique sera élargie en raison de la fenêtre d'analyse et du taux de changement de f_0. Nous utilisons donc un filtre à pic élargis ayant la réponse impulsionnelle suivant :

$$\boldsymbol{h(q) = \beta - log(\gamma - cos(2\pi eq))} \qquad \textbf{(III.4)}$$

Pour $log\ (0,5) < q < log\ (K + 0,5)$ et $h\ (q) = 0$ sinon. $\boldsymbol{\gamma}$ est un paramètre d'algorithme qui contrôle la largeur du pic alors que β est choisi de telle sorte que $\int h(q)dq = 0$.

Le nombre de pics, K, est restreint à 10 afin de réduire la réponse de $Yt\ (q) * h\ (q)$ à des valeurs de q correspondant à des sous-harmoniques de f_0. Figue III.1 Montre $h\ (q)$ pour $\boldsymbol{\gamma}$=1,5.

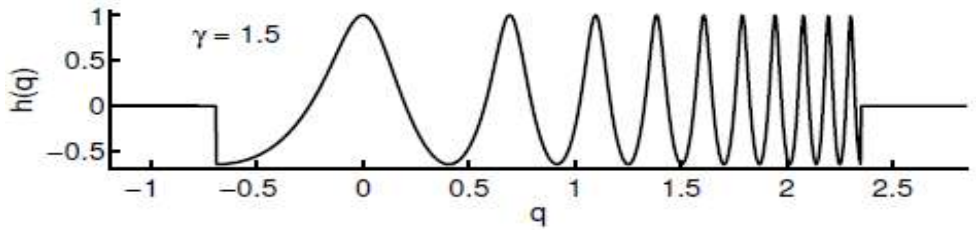


Figure III.1. *Filtrage h(q) depuis la formule (4) pour* $\gamma = 1.5$.

Puisque *h(q)* est choisie pour avoir une moyenne nulle, le terme de bruit blanc, *Nt(q)*, sera supprimé par le filtre. De plus, comme les pics de *h(q)* sont tous approximativement symétrique avec une moyenne nulle, *Nt(q)* sera supprimé si *dNt(q)/dz* est approximativement constante sur la plage *Log (k - 0,5) <q <log (k + 0,5)* pour chaque *0 <k ≤ K*. Dans la pratique, cela signifie que toute puissance de bruit sera fortement atténuée par le filtre.

3.3. Compression :

Bien que, comme nous l'avons vu, le bruit avec un spectre sera supprimé par le filtre *h(q)*, certaines sources de bruit contiennent des composantes à bande étroite d'amplitude élevée qui peuvent dominer la sortie du filtre. Pour éviter cela, Nous appliquons la compression au spectre dans chaque intervalle de temps avant de la convoluer avec *h(q)* en

$$\boldsymbol{Y'_t\ (q) = Y_t\ (q)^{a_t(q)}} \qquad \textbf{(III.5)}$$

Où :

t : est l'indice de temps. Pour déterminer l'exposant de compression $a_t(q)$, nous calculons d'abord le spectre lissé $Y''_t\ (q)$ par filtrage passe-bas de $Y_t\ (q)$ dans le domaine temporel et le domaine log-fréquence. En l'absence de bruit, nous pouvons écrire $Y''_t\ (q) \approx L\ (q)$, spectre moyen à long terme de la parole. Par conséquent, on Normalise $Y''_t\ (q)$ et $Y_t\ (q)$ à la puissance de $L\ (q)$ et on fixe l'exposant de compression à :

$$\boldsymbol{a_t\ (q) = \frac{logL(q)}{logY_t\ (q)}} \qquad \textbf{(III.6)}$$

3.4. Estimation de la fréquence fondamentale :

L'algorithme PEFAC (Pitch Estimation Filter with Amplitude Compression) comprend donc les étapes suivantes :

- Transformation de signal d'entrée en domaine temps-fréquence en utilisant la transformée de Fourier à courte terme (STFT), *Yt (f).*
- Interpoler la densité spectrale de puissance (PSD) de chaque trame sur une échelle logarithmique de fréquence, *Yt (q).*
- Trouver $\boldsymbol{a_t}(\boldsymbol{q})$ et calculer la densité spectrale de puissance (PSD) compressée.
- Convolution du PSD compressée *Y 't (q),* avec le filtre d'analyse *h (q),* et sélectionner le pic le plus élevé dans la plage fréquentielle.

4. Extraction du pitch :

Le schéma bloc suivant (figure III.2), montre les différents étapes aboutiront à la détection du pitch :

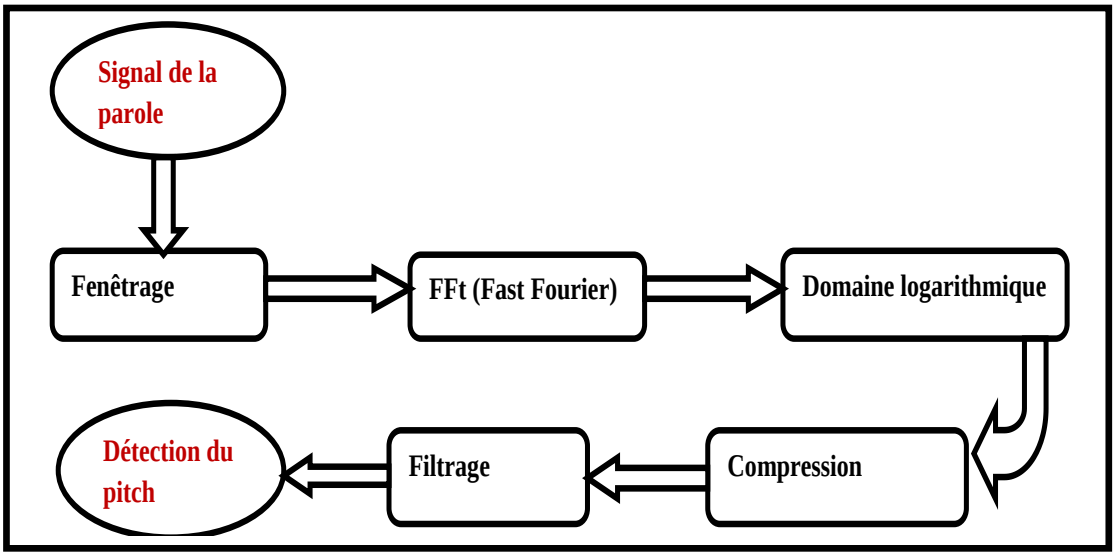


Figure III.2. *Schéma bloc de détection du pitch.*

5. Résultats expérimentaux :

Un fichier audio contenant le phonème de la voyelle A, a été enregistré sous le nom «A8k16m.wav». L'enregistrement est sous format 16 bits, à échantillonnage 8 KHz.

Le script suivant permet de lire et de tracer le signal audio :

```
[s,fs]=wavread('A8k16m.wav'); % un segment de la voyelle "A"
(256 échantillons)
s=s*1000;
% plot waveform
t=(0:length(s)-1)/fs;% times of sampling instants
plot(t,s);
legend('Waveform');
```

```
xlabel('Time(s)');
ylabel('Amplitude');
```

La figure suivante donne le tracé temporel du signal de test :

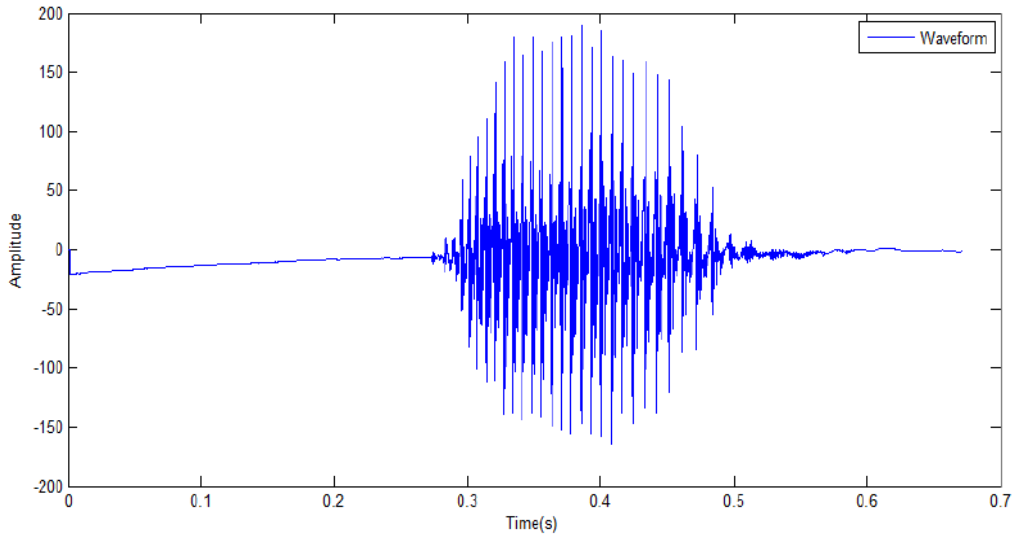


Figure III.3. *Le tracé temporel du signal audio «A8k16m.wav».*

Dans la figure suivant La zone utile entière sans silence, est située entre les échantillons n°2200 et n°2456. Soit 256 échantillons.

Le script suivant permet de lire et de tracer le signal audio :

```
[s,fs]=wavread('A8k16m.wav',[2200 2456]); % un segment de la
%voyelle "A" (256 échantillons)
s=s*1000;
% plot waveform
t=(0:length(s)-1)/fs;% times of sampling instants
subplot(4,1,1);
plot(t,s);
legend('Waveform');
xlabel('Time(s)');
ylabel('Amplitude');
```

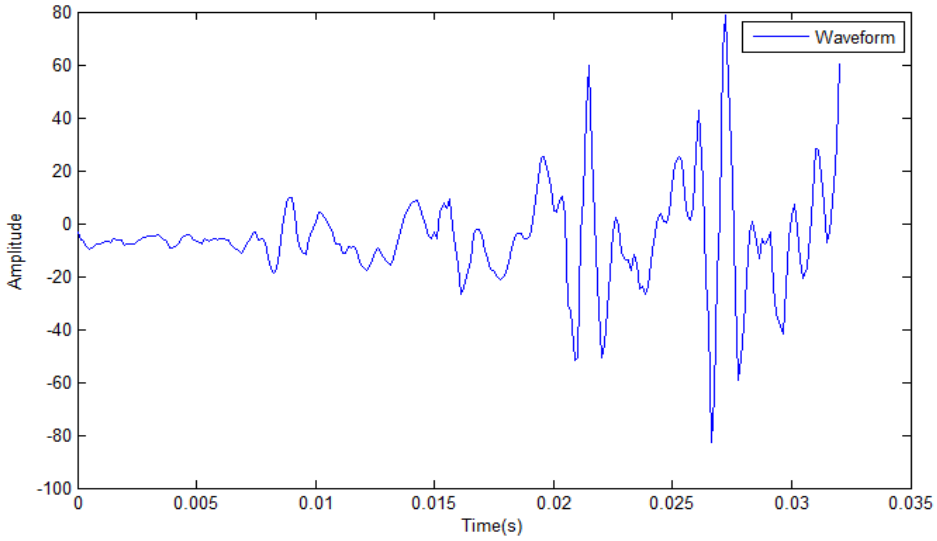


Figure III.4. *Le tracé temporel du signal audio «A.wav» sur [2200 2456].*

Fenêtrage Hamming :

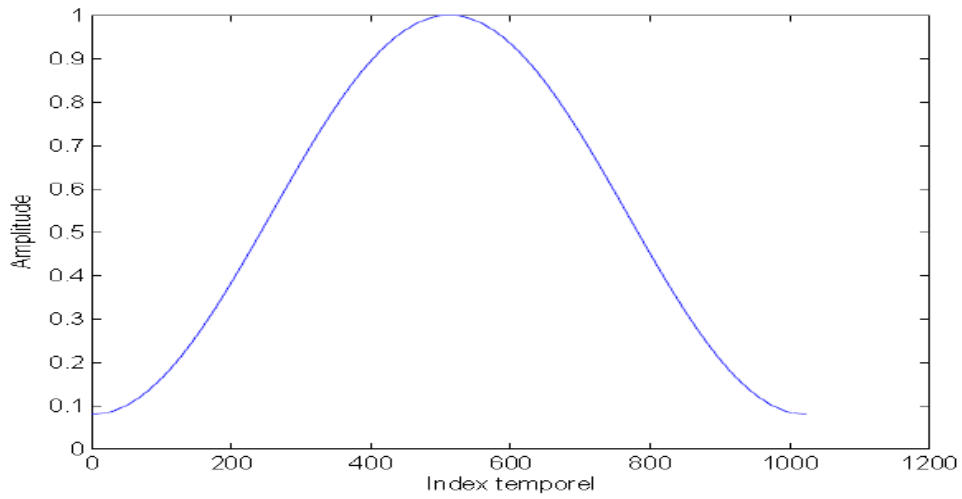


Figure III.5. *Fenêtre Hamming.*

Un prétraitement du segment consiste à son fenêtrage Hamming, figure.III.5. Le traitement quant à lui, vise à aboutir aux filtrages passe-bas du segment.

L'idée principale de l'algorithme est de pouvoir manipuler le spectre logarithmique d'excitation (au lieu du spectre brut du signal vocal) qui est obtenu avant FFT par la mise à zéro des premiers coefficients du signal cepstral, en exécutant le script suivant :

```
Yf=(s.*hamming(length(s)));

plot(t,Yf);
```

Le résultat de ce script est représenté par la **(**Figure III.6).

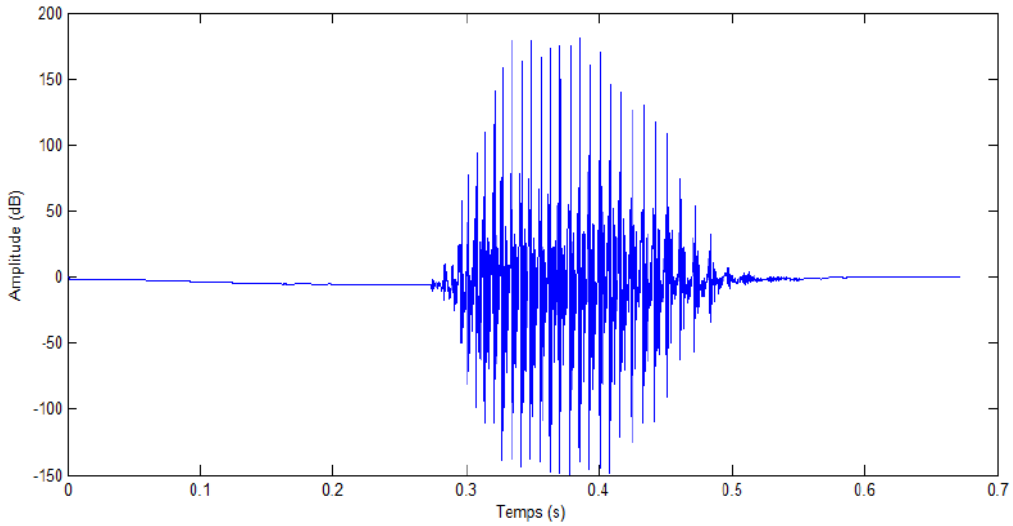


Figure III.6. Fenêtrage *du signal audio «A.wav».*

Transformée de Fourier discrète :

Après cette mise en forme du signal (commune à la plupart des méthodes d'analyse de la parole), une transformée de Fourier discrète DFT en particulier FFT (Transformé de Fourier Rapide) est appliquée pour passer dans le domaine fréquentiel.

Dans ce contexte, on passe par le calcul du module du spectre et son logarithme, qui sont retrouvés en exécutant le script suivant :

```
Yf=fft(s.*hamming(length(s)));
f=linspace(0,fs/2,length(Yf)/2+1);
plot(f,10*log10(abs(Yf(1:length(f)))));
```

Le résultat de ce script est représenté par la **(**Figure III.7).

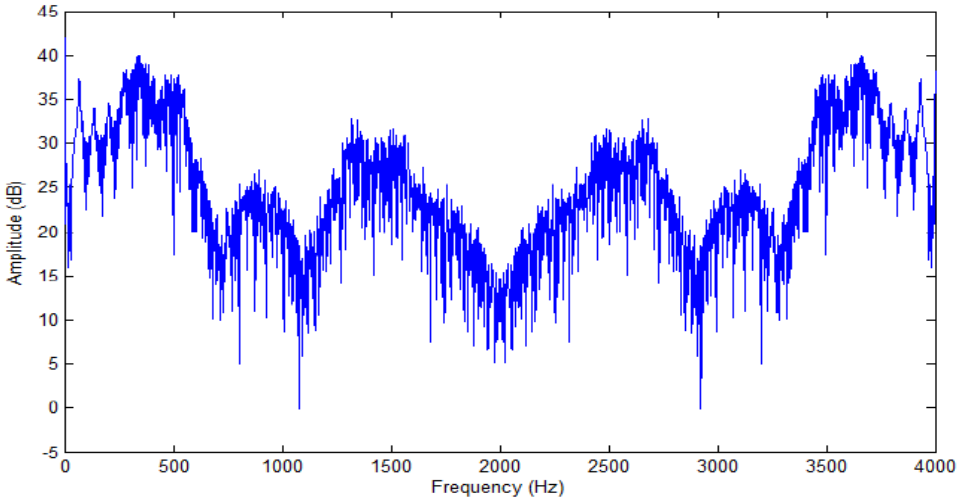


Figure III.7. *FFT du signal.*

Echelle logarithmique :

Une échelle logarithmique aura plutôt tendance à « rétrécir » l'amplitude de la courbe que nous étudions.

Le script suivant décrit cette tâche :

```
l=log10([f(2) f(end)]);
flog=linspace(l(1),l(2),length(Yf));
f=10.^flog;
O=Yf'*trans;
subplot(4,1,3);
O1=10*log10(abs(O(1:length(f))));
plot(f,O1);
```

Le résultat de ce script est représenté sur la figure suivante :

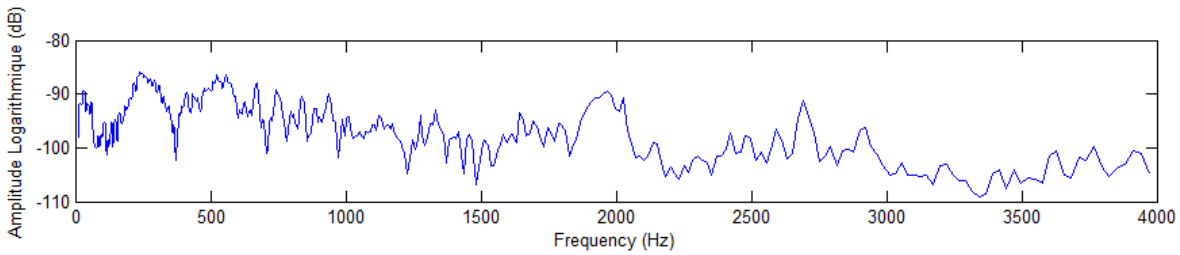


Figure III.8. *Domaine logarithmique.*

Spectre moyen à long terme :

Le spectre moyen à long terme filtré en tiers d'octave de la voix artificielle est représenté à la Figure 9 et au Tableau 1, normalisé pour un niveau de pression acoustique à large bande de −4,7 dBPa.

Les valeurs du spectre à long terme de la voix artificielle au point de référence bouche sont données par la formule:

$$S(f) = -376,44 + 465,439(log_{10}f)^1 - 157,745(log_{10}f)^2 + 16,7124(log_{10}f)^3 \quad \textbf{(III.7)}$$

où $S(f)$ est la densité spectrale en dB par rapport à une intensité sonore de 1 p W/m² par Hz à la fréquence f. La bande de fréquences de travail est de 100 Hz à 8 KHz.

Les valeurs de $S(f)$ aux fréquences définies par l'ISO correspondant à des tiers d'octave sont données dans la 4e colonne du Tableau 1. Les tolérances sont indiquées dans la 5e

colonne du même tableau. Les tolérances au-dessous de 200 Hz ne s'appliquent qu'à la voix d'homme artificielle.

Le niveau de pression acoustique totale du spectre défini par la formule (III.7) est de −4,7 dBPa. Cependant, ce spectre est valable également pour les niveaux compris entre −19,7 et +10,3 dBPa. Cela signifie que le premier terme de la formule (III.7) peut être compris entre –391,44 et –361,44.

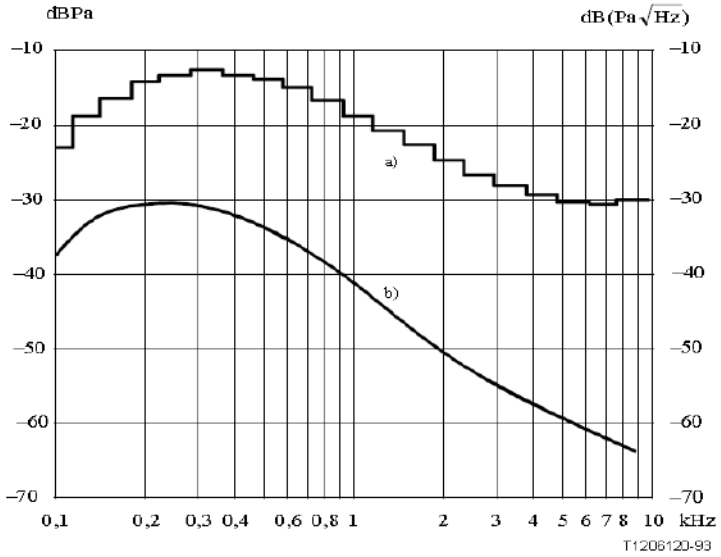


Figure III.9. *Spectre à long terme*

a) spectre au 1/3 d'octave [Tableau 1, colonne (3)].
b) densité du spectre [Tableau 1, colonne (3)-(2)].

Fréquence au centre du 1/3 octave (Hz) (1)	Facteur de correction de la largeur de bande 10 $\log_{10}\Delta f$ (dB) (2)	Niveau de pression acoustique (1/3 octave) (dBPa) (3)	Densité du spectre (dB) (3) – (2)	Tolérance (dB)
100	13,6	–23,1	–36,7	–
125	14,6	–19,2	–33,8	+3, –6[a)]
160	15,6	–16,4	–32,7	+3, –6[a)]
200	16,6	–14,4	–31,7	+3, –6
250	17,6	–13,4	–31,7	±3,0
315	18,6	–13,0	–31,6	±3,0
400	19,6	–13,3	–32,9	±3,0
500	20,6	–14,1	–34,7	±3,0
630	21,6	–15,4	–37,7	±3,0
800	22,6	–17,0	–39,6	±3,0
1000	23,6	–18,9	–42,5	±3,0
1250	24,6	–21,0	–45,6	±3,0
1600	25,6	–23,0	–48,6	±3,0
2000	26,6	–25,1	–51,7	±3,0
2500	27,6	–26,9	–54,5	±3,0
3150	28,6	–28,6	–57,2	±3,0
4000	29,6	–29,8	–59,4	±6,0
5000	30,6	–30,6	–61,2	±6,0
6300	31,6	–30,9	–62,5	±6,0
8000	32,6	–30,5	–63,1	–

[a)] Les tolérances indiquées s'appliquent au spectre à long terme de la voix d'homme et doivent aussi s'appliquer au bruit modelé de la parole. Toutefois, elles ne s'appliquent pas au spectre de la voix de femme, dont le contenu d'énergie dans cette gamme de fréquences est pratiquement négligeable.

Tableau III.1. *Spectre à long terme de la voix artificielle.* [16]

Nous avons utilisé le paramètre « S(f) » pour la compression du signal de domaine logarithmique.

La figure suivante montre le tracé du spectre :

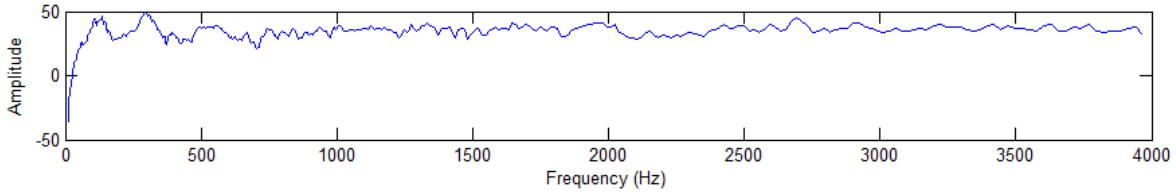


Figure III.10. *Tracé du spectre à long terme.*

Dans cette étape, on montre la convolution du PSD compressé *Y't (q)*, avec le filtre d'analyse *h(q)*, et sélectionne le pic le plus élevé dans la plage possible.

Nous voyons que le bruit de basse fréquence qui masque le fondamental dans (figure 7) est fortement atténué en (figure 10) et qu'un pic clair à 150 Hz est visible en (figure 11) bien

qu'il soit absent dans le spectre d'origine. L'algorithme n'impose aucune contrainte de continuité temporelle aux estimations du pitch.

La figure suivante montre la détection de la fréquence fondamentale :

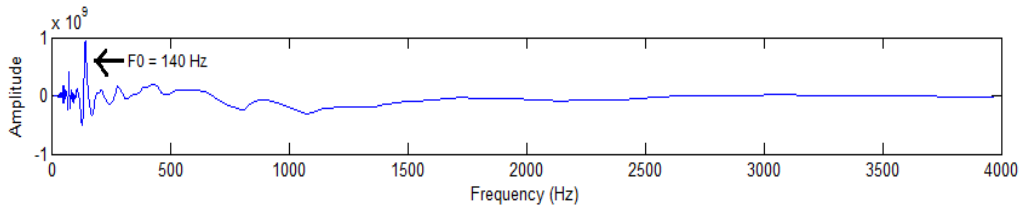


Figure III.11. *Détection du pitch.*

La figure suivante présente la détection de la fréquence fondamentale sans bruit du signal de parole femme : « clean.wav »

Nous voyons qu'un pic clair à 150 Hz est visible en (figure 12).

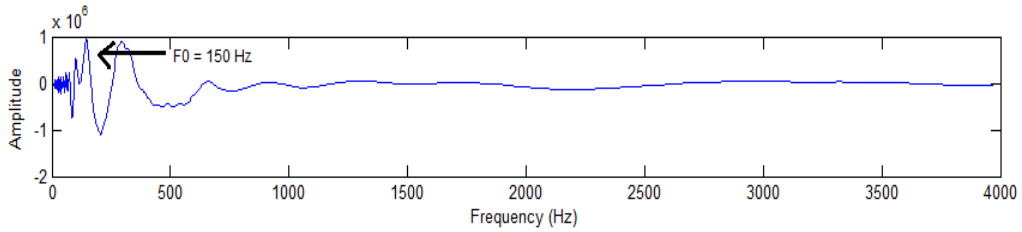


Figure III.12. *Détection du pitch du signal audio « clean.wav ».*

On a appliqué sur le signal de parole un bruit qui a été enregistré sous le nom " rain.wav".

Le script suivant montre l'ajout du bruit au signal audio :

```
[y,fs]=wavread('rain.wav');
[x,fs]=wavread('clean.wav');
s=x(1:size(y,1),:)+y(:,2);
```

La figure III.13 montre le résultat de détection du pitch de notre signal audio bruité :

Nous voyons qu'un pic clair à 275 Hz est visible en (figure 13).

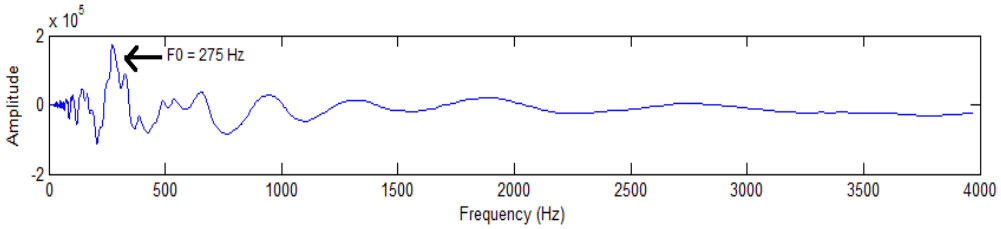


Figure III.13. *Le résultat de détection du pitch en utilisant un bruit " rain.wav" sur le signal audio.*

Un autre bruit a été appliqué sur le signal de parole et enregistré sous le nom " sea1.wav".

Le script suivant montre l'ajout du bruit au signal audio :

```
[y,fs]=wavread('sea1.wav');
[x,fs]=wavread('clean.wav');
s=x(1:size(y,1),:)+y(:,2);
```

La figure III.14 montre le résultat de détection du pitch en utilisant le bruit " sea1.wav" sur notre signal audio :

Nous voyons qu'un pic clair à 285 Hz est visible en (figure 14).

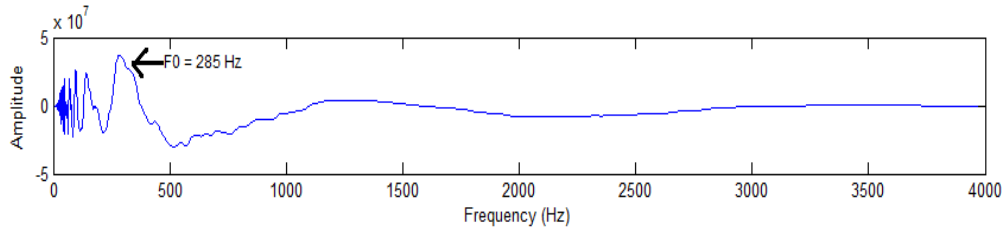


Figure III.14. *Le résultat de détection du pitch en utilisant un bruit " sea 1.wav" sur le signal audio.*

Un troisième bruit « bruit blanc » gaussien qui suit une loi normale de moyenne et variance données a été appliqué sur le signal de parole.

Le script suivant montre l'ajout du bruit au signal audio :

```
[s,fs]=wavread('clean.wav');
bruit=0.05*randn(size(s));
s=s+bruit;
```

La figure III.15 montre le résultat de détection du pitch en utilisant un bruit blanc sur notre signal audio :

Nous voyons qu'un pic clair à 300 Hz est visible en (figure 15).

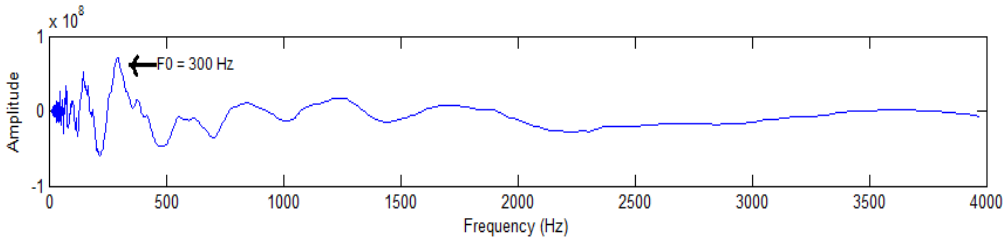


Figure III.15. *Le résultat de détection du pitch du bruit blanc sur le signal audio.*

Détection de La fréquence de vibration des cordes vocales chez l'homme et femme :

Sexe / Signal	Signal A	Signal Clean
Homme	140 Hz	
Femme		150 Hz

Tableau III.2. *Détection de La fréquence de vibration des cordes vocales chez l'homme et femme.*

Fréquence Fondamentale		
Bruit / Sexe	Homme	Femme
Blanc	112 Hz	150 Hz
Pluie	137 Hz	275 Hz
Plage	122 Hz	285 Hz
Voiture	96 Hz	180 Hz

Tableau III.3. *Détection de La fréquence de vibration des cordes vocales avec différents bruits chez l'homme et femme.*

Bruit	Sexe / Variance	Homme	Femme
Blanc	0.04	115 Hz	150 Hz
	0.05	143 Hz	160 Hz
	0.2	148 Hz	183 Hz

Tableau III.4. *Détection de La fréquence de vibration des cordes vocales avec changement de variance du bruit blanc chez l'homme et femme.*

A chaque fois que l'on augmente la variance du bruit blanc gaussien, la fréquence fondamentale augmente. Même résultat pour les autres bruits.

Conclusion :

Nous avons réalisé l'algorithme d'estimation de la fréquence fondamentale et nous avons montré qu'il est capable de donner des estimations de pitch fiables. L'algorithme comprend un étage de compression d'amplitude qui atténue les composantes de bruit à bande étroite avec un filtre d'estimation de pitch qui rejette le bruit à large bande ayant un spectre de puissance lisse.

Conclusion générale :

Dans ce travail nous avons réalisé un algorithme de détection de la fréquence fondamentale d'un signal audio. En premier lieu, nous avons donné des généralités sur le signal de la parole, nous avons défini la parole, le signal de la parole et ces caractéristiques ainsi que les paramètres et types, l'appareil phonatoire et les voyelles et consonnes. Dans le deuxième chapitre nous avons présenté un certain nombre d'algorithmes d'estimation de la fréquence fondamentale, qui est notre objectif dans notre travail. En règle générale, la première catégorie trouve des pics dans la fonction d'autocorrélation, la seconde recherche des pics harmoniques dans le spectre de puissance tandis que le troisième effectue une analyse de domaine temporel sur les sorties d'une banque de filtres passe-bande. Dans de nombreux cas, un algorithme identifie plusieurs candidats de terrain dans chaque intervalle de temps, puis utilise des contraintes de continuité temporelle pour choisir entre eux.

Dans le dernier chapitre, nous avons réalisé l'algorithme d'estimation de la fréquence fondamentale et nous avons montré qu'il est capable de donner des estimations de hauteur fiables.

Enfin, ce travail nous a été bénéfique dans notre formation, il nous a aidé à considérer nos connaissances théoriques qui nous serons utiles dans notre projet de fin d'études, aussi pour notre vie professionnelle.

Bibliographie :

[1] F. Artières, G. Geffriaud, N. Gresillon, M. Mondain, J-P. Piron, S. Romdhane, M. Sicard, M. Sillon, V. Tesseydre, A. Uziel, V. Ventosa, A. Vieu, "Les implants cochleaires", Document réalisé par les membres du GEORRIC, Edité dans la monographie numéro 33 du CCA Groupe, édition en 2002.

[2] Introduction au traitement automatique de la parole.

[3] KLEIJN, W. B PALIWAL, K. K Speech Coding And Synthesis, Elseviar,1995.

[4] Généralités sur les ondes , par Nikhil Banerjee, Ladislas Quazza et Victor Ruelle , 2013.

[5] Détection du fondamental de la parole en temps réel: application aux voix pathologiques , par Fadoua Bahja, 2013.

[6] R. Boite et M. Kunt, Traitement de la Parole, Presses Polytechniques Romandes, 1ère édition 1987.

[7] Calliope, Traitement automatique de la parole, Masson, 1989.

[8] http://outilsrecherche.over-blog.com/pages/Notes_131_Lappareil_Phonatoire_Humain-3083095.html [consulter en 2011].

[10] Jean-Marie Gorce, Traitement de la parole, Télécommunication, service & usage.

[11] http://www:cyberphon.ish-lyon.cnrs.fr/partie3/p3c1:html.

[12] CALLIOPE «La parole et son traitement automatique ». Paris. Masson. 717p.1989.

[13] https://fr.wikipedia.org/wiki/Parole.

[14] J. Koreman, B. Andreeva, et H. Strik, Acoustic parameters versus phonetic features in ASR, In Proceedings of the International Congress of Phonetic Sciences, 1999.

[15] http://mcours.net/cours/pdf/info/Presentation_de_Matlab.pdf.

[16] Recommandation UIT-T P.50, Voix Artificielle.