République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire Signal et Communications LSC

Projet de Fin d'Etudes

En vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

Thème :

Détection du Paramètre Prosodique F_0 en Vue de la Reconnaissance Automatique de Locuteur

Réalisé par :

KHELIFI Ahmed Redha

Soutenu Publiquement le 22 Juin 2014 devant le jury composé de :

M. ADNANE	MCA	ENP	Président
M. GUERTI	Professeur	ENP	Rapporteur
B. BOUSSEKSOU	MAA	ENP	Examinateur

REMERCIEMENTS

En premier lieu je remercie Dieu le tout puissant de m'avoir donné le courage et la force pour réaliser ce travail.

Mon profonde gratitude et sincère reconnaissance vont tout d'abord à Mme M. Guerti qui a bien voulu m'encadrer. Je la remercie pour sa disponibilité, son aide, les précieux conseils qu'elle m'a prodigués, ses critiques constructives, ses explications et suggestions pertinentes.

Je remercie Mr. M. ADNANE d'avoir accepté de présider mon jury de PFE. Je remercie également Mr. B. Bousseksou d'examiner mon travail.

Mes remerciements vont également à tous les enseignants de l'Ecole Nationale Polytechnique qui ont contribué à ma formation. Qu'ils trouvent ici l'expression de mon profond respect et ma grande considération.

DEDICACES

A ma très chère mère qui a toujours été là pour moi,
et qui m'a donnée un magnifique modèle de labeur et de persévérance.

ملخص

في إطار عملنا هذا ،إهتمننا بالتعرف الآلي على المتكلم ، إستنادا على المعايير البروزودية مع التركيز على التردد الأساسي.لهذا وضعنا مدونة باللغتين العربية والفرنسية.تم تسجيلهما في المعهد العالي للفنون المسرحية والسمعي البصري من قبل من قبل عدة متحدثين.

من اجل تقسيم هذه المدونة، قمنتا بعملية تحليل بمساعدة البرنامجPRAAT .

طريقتين لكشف التردد الأساسي تم تبيقهما لبرنامجنا SRAL . النتائج المتحصل عليها أعطت نسبة تعرّف DF_0 AC DF_0 Cep وصلت إلى97 %

الكلمات المفتاحية : الكلام، البروزوديا ، كشف التردد الأساسي، طريقة الإرتباط الذاتي، طريقة كبسترال ، التعرف اللآلي على المتكلم

RÉSUMÉ

Dans le cadre de notre travail nous nous sommes intéresses à la Reconnaissance Automatique de Locuteur basée sur les paramètres prosodiques en mettant l'accent sur la fréquence fondamentale (F_0).

Pour cela, nous avons élaboré un corpus composé des mots en Arabe Standard et en Français. Ce dernier a été enregistré à l'ISMAS (Institut Supérieur des Métiers des Arts du Spectacle et de l'Audio Visuel), par des locuteurs : des femmes et des hommes.

Afin de segmenter ce corpus, une analyse a été faite à l'aide du logiciel PRAAT. Deux méthodes de Détection de F_0: l'Autocorrelation DF_0AC (Detecteur de F_0 par la méthode d'AutoCorrelation) et la Cepstrale DF_0Cep (Detecteur de F_0 par la méthode Cepstrale), ont été appliquées à notre Système de Reconnaissance Automatique de Locuteur (SRAL). Les résultats obtenus donnent un Taux de Reconnaissance (TR) qui atteint les 97%.

Mots-clés : Parole, Prosodie, Détection de Fréquence Fondamentale, Méthode Cepstrale, Méthode Autocorrélation, Reconnaissance Automatique Locuteur.

ABSTRACT

In the frame of our work we have been interested in the Automatic Recognition of Speakers based on prosodic parameters with focusing on the fundamental frequency.

For this, we have developed a corpus composed of words in standard Arabic and French. It was recorded in the ISMAS by speakers: women and a men.

To segment the corpus, An analysis was done using the software PRAAT. Two methods to detect the fundamental frequency: the autocorrelation DF_0AC and cepstral DF_0Cep, were applied on our SRAL. The results provide a recognition rate, which reached 97%.

Keywords: Speech, Prosody, Fundamental Frequency detector, Cepstral Method Autocorrelation Method, Automatic Speaker Recognition.

LISTE DES ABRÉVIATIONS

API	**A**lphabet **P**honétique **I**nternational
AR	**A**uto**R**égressif
C-AMDF	**C**lipping **A**verage **M**agnitude **D**iffrence **F**unction
CV	**C**ordes **V**ocales
CT	**C**ourt **T**erme
D	**D**urée
DARD	**DA**ta **R**e**D**uction
DF_0AC	**D**etecteur de **F_0** par **A**uto**C**orrelation
DF_0Cep	**D**etecteur de **F_0** par **C**epstrale
E	**E**nergie
FFT	**F**ast **F**ourier **T**ransform
HPS	**H**armonic **P**rodact **S**pectral
I	**I**ntensité
IFFT	**I**nversed **F**ast **F**ourier **T**ransform
MACC	**M**odified **A**utocorrelation with **C**enter **C**lipping
PPROC	**P**arallel **PR**o**C**essing
RAL	**R**econnaissance **A**utomatique du **L**ocuteur
RAP	**R**econnaissance **A**utomatique de la **P**arole
SRAL	**S**ystème de **R**econnaissance **A**utomatique du **L**ocuteur
TFA	**T**aux de **F**ausse **A**cceptation
TFD	**T**ransformée de **F**ourier **D**iscrète
TFR	**T**aux de **F**aux **R**ejet

LISTE DES FIGURES

Page

Fig.1.1	Appareil phonatoire humain…….…….……	4
Fig.1.2	Mode et lieux d'articulations…….	7
Fig.1.3	Régions principales sur la langue…….…….……	8
Fig.1.4	Trapèze vocalique du Français standard.……...…	12
Fig.1.5	Trapèze vocalique des voyelles anglaises.…….....	18
Fig.2.1	Paramètres prosodiques dans différents domaines de la parole…….	26
Fig.2.2	Dix intonations de base de P. Delattre…….....	31
Fig.3.1	Principe de la Reconnaissance Automatique de la Parole……	38
Fig.3.2	Principe de la Reconnaissance Des Formes…….…….	40
Fig.3.3	Schéma bloc proposé du détecteur de pitch par la C-AMDF.…..…….	45
Fig.3.4	Calcul des coefficients cepstraux par analyse spectrale……	45
Fig.3.5	Schéma bloc proposé du détecteur de pitch par la méthode Cepstrale..	47
Fig. 3.6	Schéma bloc du détecteur de pitch par la MACC…..…….	49
Fig. 3.7	Mise en forme du signal vocal……...	50
Fig. 4.1	Studio d'enregistrement……..……...	56
Fig. 4.2	Microphone électrodynamique……..	57
Fig. 4.3	Station ProTools…….....	57
Fig. 4.4	Interface graphique du DF_0Cep …….....	58
Fig. 4.5	Interface initiale graphique du DF_0AC……...…...……	58
Fig. 4.6	Résultat de la DF_0AC……..…….....	59
Fig. 4.7	Résultat de la DF_0Cep……...…….....	59

LISTE DES TABLEAUX

		Page
Tableau 1.1	Les consonnes et semi-consonnes du Français…….....	9
Tableau 1.2	Description articulatoire des consonnes du Français…….....	10
Tableau 1.3	Description articulatoire des semi consonnes du Français…….....	11
Tableau 1.4	Les voyelles orales du Français…….....……	13
Tableau 1.5	Les voyelles nasales du Français…….....……..	13
Tableau 1.6	Description articulatoire des voyelles du Français.....……..	13
Tableau 1.7	Les phénomènes de coarticulation.....…….....……	16
Tableau 1.8	les consonnes de l'Anglais.....…….....…….	19
Tableau 1.9	Phonétique dialectale du Français québécois.....……..	20
Tableau 3.1	Domaine d'application et produits en Reconnaissance de la parole.	39

TABLE DES MATIÈRES

Introduction Générale 1

Chapitre 1

Notions Fondamentales sur la Parole

1.1. Introduction 3
1.2. Définition de la Phonétique
1.3. Principe de la production d'un son
1.3.1. Eléments constitutifs de l'appareil phonatoire 4
1.3.2. Différents modes phonatoires 5
1.3.3. Les cavités supra-glottiques
1.4. L'Alphabet Phonétique International 6
1.4.1 Production des consonnes Françaises 7
1.4.2. Principes de transcription phonétique 11
1.4.3. Voyelles du Français
1.5. Les signes diacritiques 15
1.6. Phonétique combinatoire
1.7. Transcription phonétique large 17
1.8. Phonétique comparative: l'Anglais 18
1.8.1. Les voyelles
1.8.2. Les consonnes 19
1.9. Phonétique dialectale 20
1.10. Défection pertinentes 21
1.11. Conclusion

Chapitre 2

Etude des paramètres prosodiques

2.1. Introduction 23
2.2. Définition de la prosodie
2.3. But de la prosodie 24
2.4. Paramètres prosodiques 25
2.5. Extraction des paramères prosodiques 27
2.5.1. Durée
2.5.2. Energie 28
2.5.3. Fréquence fondamentale 29
2.6. Autres parapètres prosodiques 50
2.6.1. Intonation
2.6.2. Accent 31
2.6.3. Intensité et débit 33
2.6.2. Les pauses
2.6.2. L'accentuation 34
2.7. Fonctions de la prosodie 34
2.7.1. Structuration de l'énoncé
2.7.2. Focalisation 35
2.7.3. Modalité
2.7.4. Fonctions non linquistiques
2.8. Conclusion 36

Chapitre 3

Les méthodes de détection de F_0 appliquées a la RAL

3.1. Introduction 38
3.2. Reconnaissance Automatique de la Parole
3.3. Reconnaissance Automatique de Locuteur 40
3.4. Complexités de détection de la Fréquence fondamentale 41
3.4.1. Variabilité des paramètres prosodiques et contraintes de production

3.4.2. Les contraintes idiosyncrasiques
3.4.3. Les contraintes inhérentes à la gestion des variations de F0 42
3.4.4. Les phénomènes dits :d'abaissement
3.4.5. Les contraintes interactives: effets microprosodiques 43
3.5. Méthodes de détection de la fréquence fondamentale
3.6. Description des techniques 44
3.6.1. La fonction d'AMDF basée sur le Clippage
3.6.2. La technique Cepstrale 45
3.6.1. La fonction d'Autocorrelation basée sur le Clippage central 48
3.7. Conclusion 51

Chapitre 4

Expériences et résultats

4.1. Introduction 53
4.2. Caractérisation de voix
4.2.1. Enjeux et applications
4.2.2. Études perceptives 54
4.2.3. Principaux paramètres acoustique 55
4.3. Matériels utilisé 56
4.4. Outils utilises 57
4.5. Description de l'interface
4.6. Description de l'application
4.7. Résultats et discussions 59
4.8. Evolution des performances 60
4.9. Conclusion

Conclusions générale et perspectives 62
Références bibliographique 64

Introduction génerale

Du point de vue segmental, la parole peut être vue comme une succession de sons, de segments qui possèdent une structure particulière. Un niveau suprasegmental, celui de la prosodie, intervient à une échelle plus grande et constitue en quelque sorte la mélodie de la parole. Sur le plan physiologique, la production de parole fait intervenir de nombreux muscles et organes qui composent ce phénomène complexe.

La prosodie est complexe puisqu'elle englobe d'une part des phénomènes aussi variés que l'accentuation, l'intonation, les pauses, le rythme, etc., et d'autre part, car elle peut être analysée au niveau phonologique comme au niveau phonétique.

Il n'existe à ce jour ni d'alphabet prosodique international, ni méthode de transcription prosodique universellement admise même si dans l'API on trouve un petit nombre de signes relatifs à des traits prosodiques, comme les symboles relatifs aux accents, aux phénomènes d'allongement, ou encore aux caractéristiques tonales.

La fréquence la plus basse dans le signal de parole est la fréquence Fondamentale (F_0) dénommé « pitch ». Elle représente la fréquence de vibrations 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. La fréquence fondamentale est un facteur prosodique prépondérant contribue également à la caractérisation de l'identité d'une voix et nous nous focaliserons sur celui-ci. De ce fait, l'objectif de notre travail est de procéder par une évaluation qualitative des algorithmes detecteur de F_0. Le choix est porté sur deux techniques, la MACC (Modified Autocorrelation with Center clipping), la CEP (Cepstral Technic). Ces deux techniques ont été appliquées à notre SRAL. Les résultats obtenus donnent un Taux de Reconnaissance (TR) qui atteint les 97%.

Notre projet est organisé en quatre chapitres :

- le premier présent des notions fondamentales sur la parole et la production de sons chez les êtres humains ;
- le deuxième expose la prosodie son but ainsi que ses paramètres;
- dans le troisième nous exposons les deux méthodes appliquées à notre Système de RAL : Autocorrelation et Cepstrale ;
- le dernier est le noyau de notre travail, il décrit notre SRAL ainsi l'évaluation des résultats obtenus.

Chapitre 1:
Notions fondamentale sur la parole

1.1. Introduction

Dans ce chapitre nous commençons en premier lieu par définir la phonétique. Nous expliquerons le principe de base de la production de sons ainsi que l'appareil phonatoire humain, par la suite on définira l'Alphabet Phonétique International et les signes diacritique ainsi la transcription phonétique puis la phonétique comparative et dialectale.

1.2. Définition de la Phonétique

Dans une situation générale de communication, l'un des aspects les plus marquants est sûrement les sons qu'une personne produit ou perçoit. L'étude des sons utilisés dans le langage humain s'appelle la phonétique. En conséquence, les traiteurs des signaux de la parole ont développé une méthode de classification servant bien sûr à classifier mais également à décrire et expliquer la production de sons des langues naturelles. Nous présentons les trois sous branches de la phonétique :

- Auditive : étudie les processus d'audition du langage, la façon dont l'être humain perçoit et reconnaît les sons. En ce qui concerne la perception des messages vocaux, on y étudie ce qui est perçu par l'oreille, or l'oreille juge de façon subjective. En ce sens la phonétique perceptive se distingue de la phonétique acoustique qui elle analyse les sons de manière objective.
- Acoustique : étudie la transmission des sons dans l'air selon ses caractéristiques physiques (fréquence fondamentale, intensité, durée, etc.)
- Articulatoire : la plus ancienne des trois branches de la phonétique, elle étudie la manière dont les sons du langage humain sont produits. La description des articulations se fait à l'aide de trois variables : l'activité du larynx (voisement ou sonorisation), l'endroit où se situe le resserrement maximum de la bouche (point d'articulation) et la façon dont s'effectue l'écoulement de l'air à travers le chenal phonatoire (mode d'articulation) [1].

1.3. Principe de production d'un son

Le processus de production de parole est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique. Il y a une grande quantité d'organes et de muscles qui entrent dans la production de sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain (Figure 1.1) repose sur l'interaction entre trois grandes classes d'organes : les poumons, le larynx, et les cavités supra-glottiques [2].

1.3.1. Eléments constitutifs de l'appareil phonatoire

Les deux premières classes fournissent ce qui est essentiel pour la production de n'importe quel son, qu'il soit musical ou langagier : une source d'air et une source de bruit. La troisième classe renferme les organes qui permettent de modifier le son qui est émis par le travail conjoint des deux premières classes.

Les poumons : La fonction primordiale des poumons est évidemment de permettre au corps de s'oxygéner. Cependant, les poumons fournissent aussi une source d'air qui est utilisée pour produire des sons [3] ;

Lors de la phase d'inspiration, l'action conjointe du diaphragme, qui se contracte et s'abaisse et des muscles intercostaux permet de créer un vide dans les poumons qui est rempli par la pénétration d'air. Lors de l'expiration, le diaphragme se relâche et laisse ainsi s'échapper l'air des poumons qui peut être utilisé pour produire des sons ;

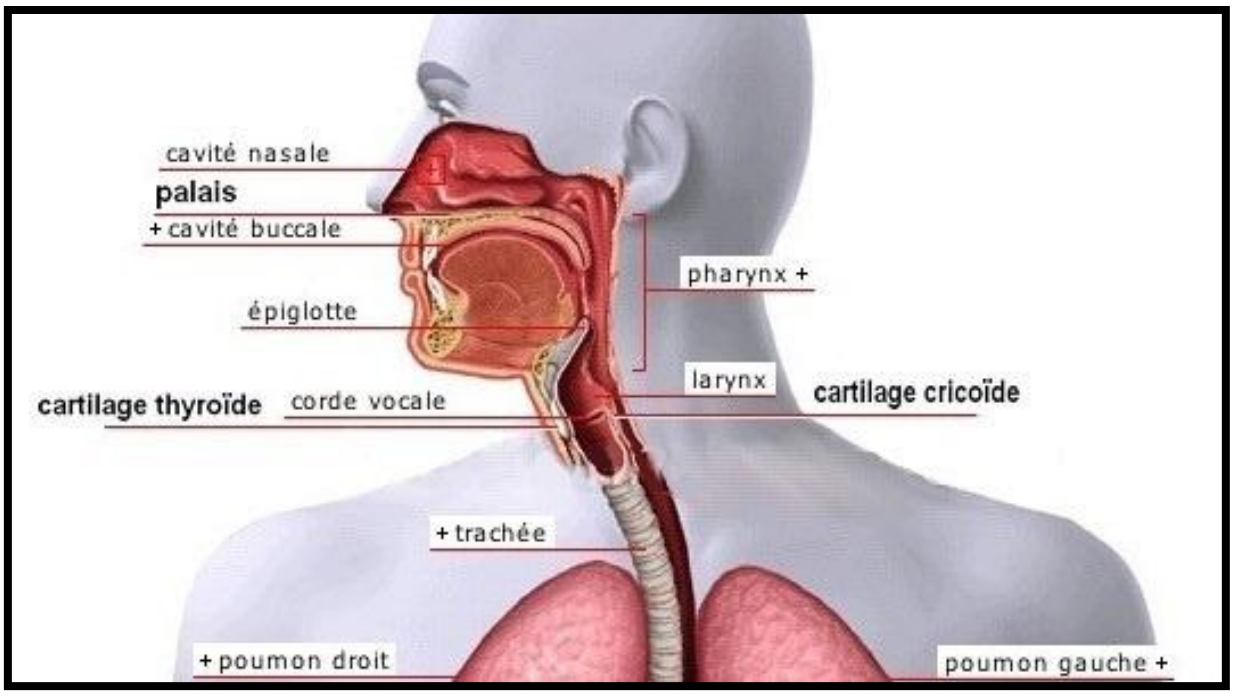


Figure 1.1 : Appareil phonatoire humain

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

Les cordes vocales : sont gardées ouvertes ou fermées par les aryténoïdes. Une voix typique d'un homme résulte de mouvements d'ouverture de 100 à 120 fois par seconde (un

cycle d'ouverture est appelé un Hertz : Hz) alors que celle d'une femme est produite par entre 175 et 250 vibrations des CV par seconde. Ce bruit, qui ressemble à celui d'un ballon qui se dégonfle, sera modifié par les divers organes de la parole qui font partie des cavités supra-glottiques [4].

1.3.2. Différents modes phonatoires

Les CV sont utilisées de façon à satisfaire les besoins de la respiration et de la production de sons langagiers. Voici les divers modes d'utilisation avec une brève explication du fonctionnement des CV [4] :

Écartées : comme lors de la respiration, les CV sont totalement écartées l'une de l'autre et aucun son n'est produit par ces dernières. Si on place sa main sur sa gorge, aucune vibration ne devrait être ressentie (son non voisé ou sourd).

Mode vibratoire : les CV s'écartent et se rapprochent très rapidement de façon à interrompre le flot d'air qui passe entre les deux. Les vibrations ainsi produites vont résulter en des sons dits voisés, ou sonores.

Voix partielle : lorsque quelqu'un chuchote, la partie avant des CV se rapproche vers le centre de la glotte (espace entre les deux cordes vocales) alors que la partie postérieure, qui est attachée aux aryténoïdes, est maintenue éloignée de l'autre corde. L'air s'échappe donc de façon forcée et un bruit de friction est créé au niveau de la glotte (production de voix chuchotée).

Voix murmurée: le murmure résulte de vibrations produites par les CV alors qu'elles sont un peu lâches, conservant ainsi un certain relâchement (production de voix voisée avec un bruit de friction).

1.3.3. Les cavités supra-glottiques

Lorsque le son sort de la glotte, il passe à travers les organes vocaux supérieurs appelés cavités supra-glottiques où il est modifié. Ces cavités servent à faire résonner le son et à lui donner une « couleur » particulière qui permettra de différencier les voyelles entre elles par exemple, ou les consonnes. Cette couleur particulière donnée à chaque son provient essentiellement de la modification de la forme des résonateurs à l'aide des mouvements de la langue et des lèvres entre autres choses [5].

Il est possible de faire une analogie avec un instrument de musique comme une clarinette par exemple. Afin de produire des notes avec une clarinette, il faut d'abord faire vibrer une anche, constituée de deux pièces de roseau (ces pièces peuvent être fabriquées en plastique de nos jours) qui produiront des vibrations dans l'air. L'action de cette anche

couplée à l'action des poumons utilisés comme source d'air est similaire à l'action des CV et des poumons dans l'appareil phonatoire humain. Le son ainsi produit est ensuite dirigé dans une colonne dont la longueur est modifiée à l'aide de clefs.

Suivons le cheminement de l'air à travers les organes phonatoires. L'air émis par les poumons, après avoir traversé la glotte, traverse la cavité buccale.

cavité buccale : lorsque la luette est collée à la cavité pharyngale, le son est complètement dirigé vers la cavité buccale. C'est la cavité la plus importante dans le langage humain. L'utilisation de cette cavité donne lieu à des articulations orales. La forme de cette cavité peut ensuite être modifiée, comme nous le verrons plus bas. Ex.: [t, d] etc.

cavité nasale : lorsque la luette, reliée au palais mou, est décollée de la paroi pharyngale, le son peut passer également dans la cavité nasale, créant ainsi une articulation nasale. Ex.: [n, m] etc.

Note: le critère de nasalité est souvent utilisé pour décrire certains accents, comme celui du sud-ouest des États-Unis par exemple.

cavité labiale : finalement, lorsque certaines articulations sont produites en utilisant les lèvres, on parle de sons labiaux. Ex.: [i, u]

cavité pharyngale : cette cavité ne joue aucun rôle en Français. Dans certaines autres langues toutefois, il peut en être autrement, comme pour l'arabe ou elle est utilisée pour la production de certaines consonnes.

1.4. L'Alphabet Phonétique International (API)

L'Alphabet Phonétique International et un système de transcription phonétique utilisé par les linguistes pour représenter les sons du langage. L'API est composé de lettres empruntées à des alphabets connus (surtout les alphabets latins et grecs), de caractères créés [ʃ] qui correspond au *ch* (chuintante sourde) du Français comme dans '*chichi'* et de signes diacritiques [~] pour indiquer la nasalité par exemple.

Le but de l'API est de fournir un répertoire de signes correspondant aux principaux phonèmes réalisés dans les principales langues du monde. Le principe sous-jacent de l'API est : « un seul signe pour un seul son, un seul son pour un seul signe ». Ainsi le signe [ð] transcrit le son que l'on trouve, à la fois à l'initiale du mot anglais *then* « alors », et à l'intérieur du mot espagnol *cada* « chaque ». [6]

On retrouve cette transcription entre crochets, ainsi [...], quand on veut représenter le maximum de nuances phoniques, même celles qui n'ont pas de fonction linguistique ; on

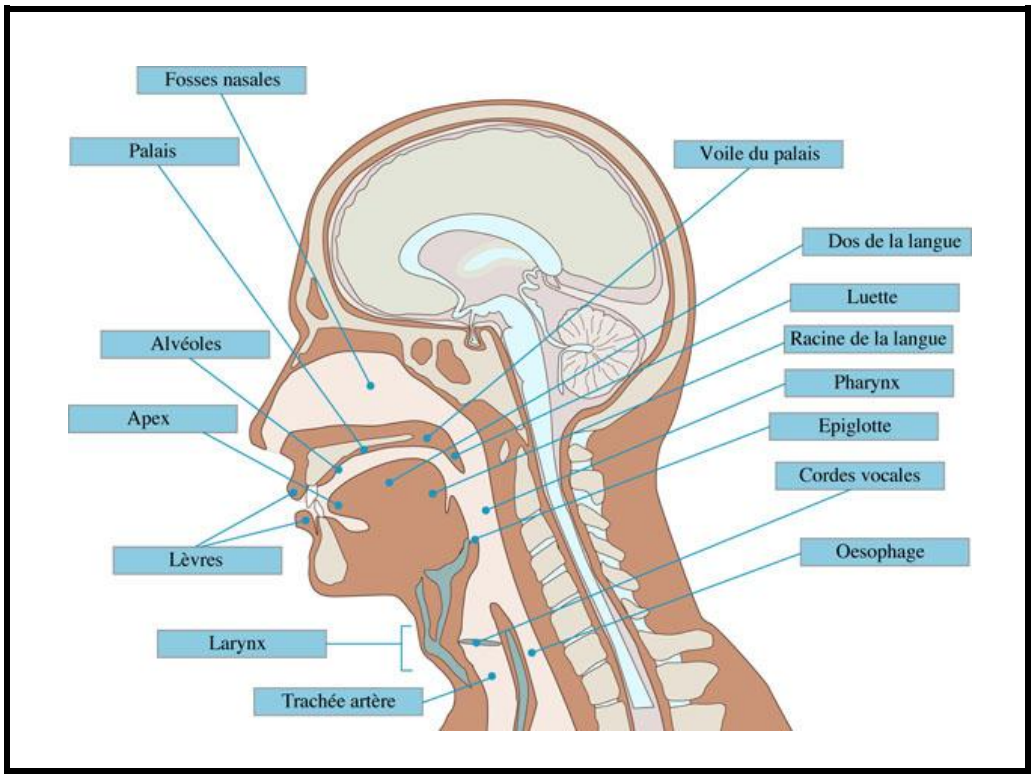
utilisera les barres obliques, /.../, si l'on désire ne représenter que les traits phoniques significatifs au niveau linguistique. Ainsi la consonne initiale du mot Français *rail* sera-t-elle notée /r/ dans une transcription phonologique, mais suivant la prononciation du locuteur, elle sera notée phonétiquement [r], [R] ou [ʁ].

1.4.1. Production des consonnes Françaises

Les consonnes Françaises, comme celles de la plupart des langues naturelles, sont produites en utilisant majoritairement les organes de la cavité buccale et les lèvres. Ces articulations sont décrites essentiellement par deux critères : le mode et le lieu d'articulation.

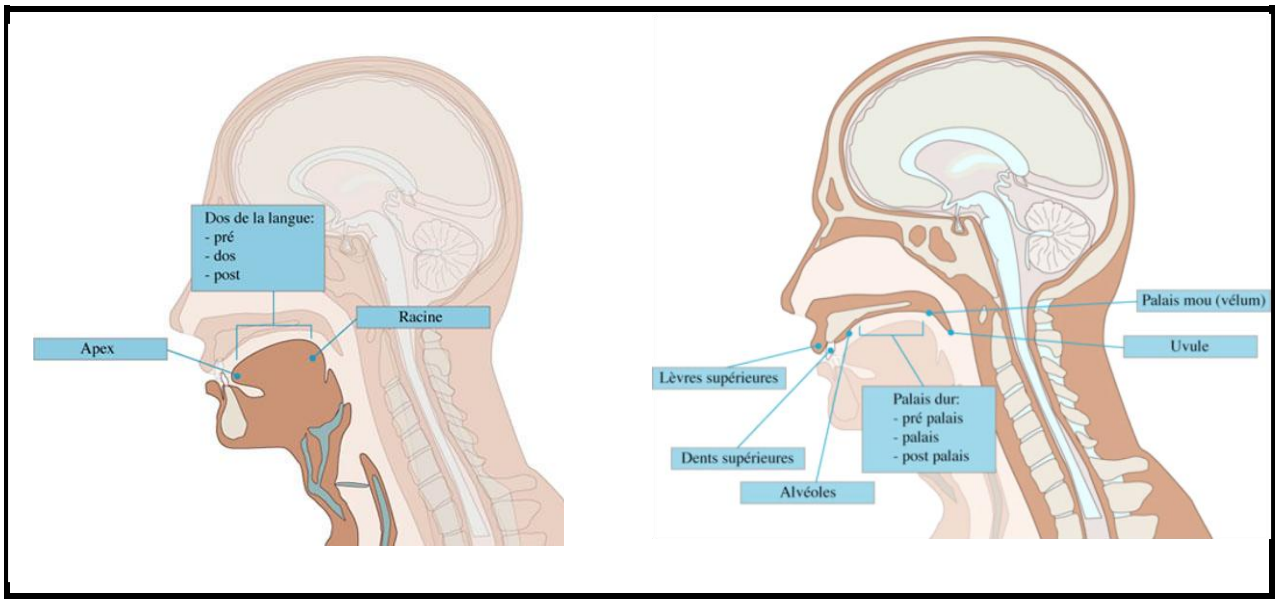
Les modes et lieux d'articulation sont définis d'après les organes articulatoires utilisés dans leur production. Ces organes sont identifiés dans la figure 1.2 et repris avec plus de précision dans la figure 1.3

Ajoutons que trois régions principales sont identifiées sur la langue: l'apex, le dos (le prédos, le dos et le post dos) et la racine figure 1.3



Figue 1.2: Modes et lieux d'articulations

Le lieu d'articulation se défini comme l'endroit de rétrécissement maximal, ce qui veut dire l'endroit où les organes de la cavité buccale sont les plus proches (ou se touchent).



Figue 1.3: Régions principales sur la langue

Le [p] français que l'on retrouve au début du mot pierre par exemple est une articulation labiale car elle est produite avec la lèvre supérieure. Par opposition, le [k] du mot qui est décrit comme étant vélaire, étant produit avec le voile du palais.

À ces lieux d'articulation doit se rajouter un articulateur, qui correspond à l'un des sept organes de la partie inférieure de la cavité buccale. Les articulateurs possibles:

- lèvres inférieures
- dents
- langue: apex
- langue: pré dos
- langue: dos
- langue: post dos
- langue: racine

Par exemple, nous décrirons le [p] comme étant bilabial, car il est produit avec les lèvres inférieure et supérieure respectivement comme articulateur et lieu d'articulation. Le [k] du mot qui décrit auparavant sera décrit comme étant dorso-vélaire, car il est produit avec le dos de la langue et le vélum.

Il existe cependant une autre différence importante dans la production des consonnes en Français. Essayez de produire un [d] (comme dans le mot [dut]) plusieurs fois. Comparez-le maintenant avec un [z] (comme dans le mot [zut]) que vous maintenez pendant quelques secondes. Vous devriez être capable de repérer la principale différence

entre ces deux sons. Le [p] ressemble à une explosion alors que le [s] est un son continu et que l'on peut maintenir pendant plusieurs secondes. Cette différence dans la façon de traiter l'air qui est expulsé pendant la production s'appelle le mode articulatoire. Ce mode est l'un de trois types différents en Français : occlusif, constrictif, nasal.

- **occlusives**: articulations qui comportent une fermeture totale et momentanée du canal vocal.
- **constrictives**: articulations qui comportent un barrage partiel lors de leur réalisation.
  - **fricatives** (prononcées avec le dos de la langue abaissé)
  - **latérales** (produites avec le dos de la langue relevé et les bords abaissés). Il n'y a qu'une seule consonne latérale en Français : le [l].
  - **vibrantes** (articulées avec la langue ou la luette et produisant de lentes vibrations). Il n'y a qu'une seule consonne vibrante en Français : le [R].
- **nasale**: articulations produites avec l'aide de la cavité nasale.

En conséquence, la description de consonnes se fait à l'aide des quatre traits articulatoires présentés ci-dessus. Les exemples suivants illustrent l'utilisation de ces descripteurs :

[p] : occlusive, bilabiale, sourde, orale

[d] : occlusive, apico-dentale, sonore, orale

[f] : fricative, labio-dentale, sourde, orale

[m] : nasale, bilabiale, sonore

Tableau 1.1 : Les consonnes et semi-consonnes du Français

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q G		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	ʙ			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Tableau 1.2 : Description articulatoire des consonnes du Français

[p]	Occlusive, bi-labiale, sourde, orale	pie, pot
[b]	Occlusive, bi-labiale, sonore, orale	bien, beau
[t]	Occlusive, apico-alvéodentale, sourde, orale	type, tôt
[d]	Occlusive, apico-alvéodentale, sonore, orale	disque, doux
[k]	Occlusive, dorso-vélaire, sourde, orale	qui, cou
[g]	Occlusive, dorso-vélaire,sonore, orale	gare, goût
[f]	Constrictive (fricative), labio-dentale, sourde, orale	phare, fou
[v]	Constrictive (fricative), labio-dentale, sonore, orale	vis, vous
[s]	Constrictive (fricative), apico-alvéolaire, sourde, orale	cil, sou
[z]	Constrictive (fricative), apico-alvéolaire, sonore, orale	zouave, zéro
[ʃ]	Constrictive (fricative), prédorso-prépalatale, sourde, orale	chic, chou
[ʒ]	Constrictive (fricative), prédorso-prépalatale, sonore, orale	Gilles, joue
[ʁ]	Constrictive (vibrante), dorso-uvulaire, sonore, orale	riz
[l]	Constrictive (sagittale), apico-alvéolaire, sonore, orale	lire, loup
[m]	Occlusive, bi-labiale, sonore, nasale	mie, mou, ma
[n]	Occlusive, apico-dentale, sonore, nasale	nez, nous
[ɲ]	Occlusive, dorso-palatale, sonore, nasale	agneau, seigneur
[ŋ]	Occlusive, dorso-vélaire, sonore, nasale	camping, trekking

Le Français comporte également trois semi-consonnes ou semi-voyelles (ou glides): [j- ɥ -w]

Ces articulations se distinguent des voyelles équivalentes [i, u, y] par leur brièveté. De plus, ces voyelles sont toujours placées devant ou derrière une seconde voyelle, comme dans les mots suivants:

[i-j]: père/Pierre, fil/fille, qu'il/quille, fer/fière, scie/siège, pied, aïl, caille, etc.

Description : orale, fricative, médio-dorso-médio-palatale (dorso-palatale)

[y- ɥ]: nu/nuage, hutte/huit, fut/fuite, cul/cuite, truite, etc.

Description : orale, fricative, antérieure, arrondie

[u-w]: cou/couette, où/oui, pou/poire, fou/foire, etc.

Description : orale, fricative, postérieure, arrondie

Tableau 1.3 : Description articulatoire des semi consonnes du Français

[j]	constrictive, dorso-palatale, sonore, orale, non arrondie	pied, bien
[ɥ]	constrictive, dorso-palatale, sonore, orale, arrondie	tuer, lui
[w]	constrictive, dorso-vélaire, sonore, orale,	bois, Louis

1.4.2. Principes de transcription phonétique

Les symboles utilisés pour transcrire les consonnes du Français sont relativement semblables à ceux que nous utilisons lorsque nous l'écrivons avec un alphabet orthographique régulier. Cependant, vous aurez certainement noté les symboles qui sont utilisés pour les articulations prépalatales. Ces symboles sont ceux qui sont proposés par un l'Association phonétique internationale. Cette association a été créée dans le but d'uniformiser les diverses transcriptions phonétiques proposées à travers le monde [5].

Cette association a donc proposé un alphabet qui repose sur le principe qui veut qu'à chaque symbole corresponde un seul son. Cela permet de transcrire n'importe quelle langue avec le même jeu de symboles et surtout, de pouvoir lire une transcription phonétique d'une langue que l'on ne parle pas avec une précision relative. Les symboles utilisés dans le tableau des consonnes et des voyelles tels qu'ils sont présentés ci-dessus utilisent ces symboles de l'API.

1.4.3. Voyelles du Français

La production des voyelles est relativement différente de celle des consonnes. Contrairement aux consonnes, la production de voyelles ne nécessite pas la production de bruit de friction ou d'une petite explosion. De façon générale, les voyelles sont produites avec un écoulement nettement plus libre de l'air à travers l'appareil phonatoire.

Les voyelles du Français sont habituellement représentées par une figure géométrique qui contient l'information pertinente à leur classification: le trapèze vocalique.

Ce trapèze, représenté sur deux dimensions, contient deux axes dont chacun renferme un type de donnée:

L'axe vertical du trapèze vocalique indique l'aperture (degré d'ouverture de la bouche) qui se défini comme le degré d'ouverture de la bouche lors de sa réalisation. Par exemple, essayes de prononcer le son [i] de façon isolée comme dans le mot « riz » et le [a] du mot «

rat ». Vous devriez noter que la bouche doit être beaucoup plus ouverte pour la production du [a]. C'est donc l'aperture qui permet de distinguer entre ces deux voyelles.

Les voyelles comme le [y] de « rue » sont produites avec une projection des lèvres vers l'avant, un peu comme lorsqu'on siffle. Nous leur attribuons le caractère arrondie (ou labialisées).

Finalement, nous ne pouvons toujours pas distinguer entre les voyelles comme le « a » de « pâte » du « a » de « pente ». La production de la deuxième requiert l'ouverture du canal nasal de façon à ajouter une résonance toute particulière à cette voyelle. Le passage vers la cavité nasale est ouvert lorsque le voile du palais (la luette) descend et s'écarte de la paroi pharyngale. L'opposition entre les voyelles nasales et orales est particulièrement utile en Français comme dans quelques autres langues comme le portugais et le polonais. L'anglais, par contre, n'utilise pas tellement cette opposition (les détails seront probablement vus en phonologie). Pour en revenir à nos exemples, la première voyelle sera qualifiée d'orale, alors que la deuxième sera décrite comme étant nasale.

Il existe finalement une voyelle centrale qui n'est pas arrondie et qui est totalement neutre. Il s'agit du « e » appelé chva. Cette voyelle apparaît dans les mots: "le", "serin"' etc. Au total, le Français standardisé possède 16 voyelles, tel que présenté dans la fig 1.4

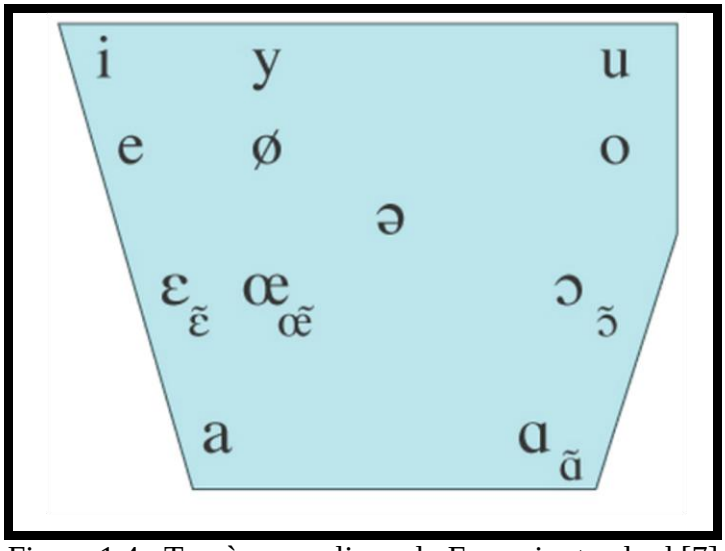


Figure 1.4 : Trapèze vocalique du Français standard [7]

Il est aussi possible de représenter les voyelles orales du français sous forme de tableau en utilisant les descripteurs qui servent à caractériser leur production.

Tableau 1.4 : Les voyelles orales du Français

	Orales				
	antérieures		centrales	postérieures	
	non arrondies	arrondies	non arrondies	non arrondies	arrondies
fermées	i	y			u
mi-fermées	e	ø			o
moyenne			ə		
mi-ouvertes	ɛ	œ			ɔ
ouvertes	a			ɑ	

Et voici le système des voyelles nasales du Français:

Tableau 1.5 : Les voyelles nasales du Français

	Nasales				
	antérieures		centrales	postérieures	
	non arrondies	arrondies	non arrondies	non arrondies	arrondies
fermées					
mi-fermées					
moyenne					
mi-ouvertes	ɛ̃	œ̃			ɔ̃
ouvertes				ɑ̃	

Tableau 1.6 Description articulatoire des voyelles du Français

[i]	non arrondie, antérieure, fermée, orale	île
[e]	non arrondie, antérieure, mi-fermée, orale	été
[ɛ]	non arrondie, antérieure, mi-ouverte, orale	aile
[a]	non arrondie, antérieure, ouverte, orale	patte
[y]	arrondie, antérieure, fermée, orale	puce
[Ø]	arrondie, antérieure, mi-fermée, orale	h**eureu**x

[œ]	arrondie, antérieure, mi-ouverte, orale	peur
[ə]	centrale, moyenne	p**e**tit
[u]	arrondie, postérieure, fermée, orale	boule
[o]	arrondie, postérieure, mi-fermée, orale	beau
[ɔ]	arrondie, postérieure, mi-ouverte, orale	bol
[α]	non arrondie, ouverte, postérieure	pâte
[ɛ̃]	non arrondie, antérieure, mi-ouverte, nasale	pain
[ɑ̃]	non arrondie, postérieure, ouverte, nasale	paon
[œ̃]	arrondie, antérieure, mi-ouverte, nasale	l**un**di
[ɔ̃]	arrondie, postérieure, mi-ouverte, nasale	pont

Contrairement aux voyelles, les consonnes sont produites lorsque le passage de l'air venant des poumons est partiellement ou totalement obstrué. Autrement dit, les consonnes correspondent à des mouvements rapides de constriction des organes articulateurs, donc souvent à des sons peu stables, qui évoluent dans le temps. Pour les fricatives, une constriction forte du conduit vocal provoque un bruit de friction. Les CV peuvent entrer en vibration en même temps que le bruit de friction, la fricative est alors voisée (ou sonore), ou laisser passer l'air sans émettre de son, la fricative est alors non voisée (ou sourde). Les plosives sont des occlusions complètes du conduit vocal, suivies d'un relâchement. Jointe à la vibration des cordes vocales, la plosive est voisée, sinon elle est sourde. Si la dérivation nasale est ouverte pendant la fermeture de la bouche, une nasale est produite. Les semi-voyelles sont des consonnes voisées, mouvements rapides qui passent par la position articulatoire d'une voyelle brève. Enfin, les liquides résultent d'une excitation voisée et de rapides mouvements articulatoires, principalement de la langue [5].

1.5. Les signes diacritiques

Un diacritique est un élément ajouté à une lettre d'un alphabet pour en modifier la valeur. Cet élément peut être souscrit (en indice), surscrit (en chef, en exposant) à cette lettre, à sa droite ou encore à sa gauche. Les accents, le tréma ainsi que la cédille sont des diacritiques en Français. Parmi les plus utiles, mentionnons les deux points [:] utilisés pour allonger une articulation, la nasalisation indiquée par le tilde espagnol telle qu'elle est utilisée pour les quatre voyelles nasales du Français et le dévoisement indiqué par un cercle souscrit: [u̥]

1.6. Phonétique combinatoire

Si vous essayez de transcrire la prononciation de quelqu'un qui parle spontanément, à vitesse normale, vous remarquerez que certains sons peuvent changer considérablement. Par exemple, la suite de mots « porte blanche » peut être produits de la façon suivante :

- [portblanch] si l'on articule avec soin
- [pordblanch] si l'on articule plus rapidement (et normalement).

À quoi peut-on attribuer ces différences de prononciation? Il faut comprendre que la langue est une suite de sons enchaînés avec relativement peu d'interruptions. Dans la production de ces sons, il est plus facile et normal d'essayer de ne pas produire les sons de façon isolée, mais plutôt d'anticiper la prochaine articulation. Cette anticipation, qui change légèrement la qualité des sons, sera permise à condition que la compréhension du message ne soit pas compromise. Ce phénomène est appelé l'économie des changements linguistiques, qui oppose deux types de forces antagonistes :

- L'inertie des organes phonateurs (force articulatoire)
- La nécessité de maintenir les sons distincts pour communiquer (discrimination)

La phonétique combinatoire, définie comme l'étude de l'interaction des sons les uns sur les autres, permet de décrire ces variations.

Il est à noter que l'étude de l'évolution des langues a également mis à jour de nombreux cas de changements phonétiques similaires à ceux qui se retrouvent dans le discours spontané (assimilation, dissimilation, etc.).

La transcription phonétique peut rendre compte de ces différences subtiles de prononciation, lors d'une transcription étroite (qui contient un plus grand nombre de détails) plutôt qu'une transcription large (qui contient peu de détails).

Les phénomènes de la coarticulation sont illustrés à l'aide des exemples tableau 1.7

Tableau 1.7 : Les phénomènes de coarticulation

PHÉNOMÈNES DE COARTICULATION

	PHÉNOMÈNE	EXEMPLES EN SYNCHRONIE OU DIACHRONIE
Phénomènes d'assimilation	**Assimilation** (sons en contact)	*Synchronie :* • « absent » [absɑ̃] > [apsɑ̃] (*régressive*) • « asthme » [asm] > [asm̥] (*progressive*) • « pendant » [pɑ̃dɑ̃] > [pɑ̃nɑ̃] (*double*) • « robe sale » [ʀɔbsal] > [ʀɔpsal] (*régressive*) • « frappe bien » [fʀapbjɛ̃] > [fʀabbjɛ̃] *Diachronie :* • lat. « campum » [kampum] > « champ » [ʃɑ̃] • espagnol chilien : « obscuro » [ɔbskuro] > [ɔxkuro]
	Dilation (sons éloignés)	*Synchronie :* • « définition » [definisjɔ̃] > [defenisjɔ̃] (*progressive*) • « surtout » [syRtu] > [suRtu] (*régressive*) • « disséminer » [disemine] > [disimine] (*double*) • « donner » [dɔne] > [dɔne] (*régressive*)
Phénomènes de différenciation	**Différenciation** (sons en contact)	*Synchronie :* • « dehors » [dəɔʀ] > [deɔʀ] *Diachronie :* • « mei » [mei] > « moi » [mwa]
	Dissimilation (sons éloignés)	*Synchronie :* • « venimeux » [vənimø] > [vlimø] • « réel » [ʀeel] > [ʀejɛl] (insertion d'un [j]) *Diachronie :* • lat. « augustus » [au-] > [o] > [u] « août »
Changement de l'ordre des sons	**Interversion** (sons en contact)	*Synchronie :* • « aéroport » [aeʀopɔʀ] > [aʀeopɔʀ] *Diachronie :* • lat. *formaticum* > fr. « fromage » [fʀɔmaʒ]
	Métathèse (sons éloignés)	*Synchronie :* • « séchoir » [seʃwaʀ] > [ʃeswaʀ] *Diachronie :* • lat. *miraculum* > espagnol : [milagro]
Insertion d'un son	**Épenthèse**	*Synchronie :* • « arc-boutant » [aʀkbutɑ̃] > [aʀkəbutɑ̃] • « ours polaire » [uʀsəpɔlɛʀ] > [uʀsəpɔlɛʀ] • « il va à l'école » [il va a lekɔl] > [il va ta lekɔl] • « moi aussi » [mwa osi] > [mwa zosi]
Effacement d'un son	**Syncope**	*Synchronie :* • « tu veux une chose » [tyvøynʃoz] > [tyvønʃoz] • « plus » [ply] > [py]

Note : le symbole « > » indique « se prononce »

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

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

différenciation : changement phonétique qui a pour but d'accentuer ou de créer une différence entre deux sons contigus.

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

interversion : lorsque deux sons contigus changent de place dans la chaîne parlée.

métathèse : lorsque deux sons non contigus changent de place dans la chaîne parlée.

1.7. Transcription phonétique large et étroite

Il est possible, lors d'une transcription phonétique, de mettre un nombre plus ou moins grand d'information. Ces informations supplémentaires sont généralement superposées, ajoutées aux articulations (sons) principales telles qu'elles ont été présentées dans les lignes précédentes. Le phonéticien peut, lorsque qu'il le juge nécessaire, ajouter le plus de détails possible de façon à faire une analyse plus précise des extraits décrits.

Par exemple, en Français canadien, il est possible de transcrire le mot "petite" dans la phrase "J'ai deux petites sœurs." soit de façon *large* (exemple a), soit de façon *étroite* (exemple b) :

- [p ə t i t]
- [p ə t_s ɪ t]

Vous remarquerez dans l'exemple (b) ci-dessus que les phénomènes d'affrication et d'ouverture de la voyelle haute [i] sont transcrits par des symboles différents, indiquant par le fait même avec plus de précision qu'il s'agit d'une prononciation typique d'un locuteur du Français canadien.

Nous avons vu dans ce cours un certain nombre de diacritiques qui nous permettent de réaliser des transcriptions étroites pour le Français.

1.8. Phonétique comparative: l'Anglais

Il peut être utile de faire une comparaison avec un autre système phonétique, comme celui de l'Anglais, dans le but d'obtenir une meilleure idée des variations phonétiques que nous pouvons rencontrer dans des langues différentes [8].

Les lignes qui suivent feront une comparaison très brève avec le système de l'Anglais.

1.8.1. Les voyelles

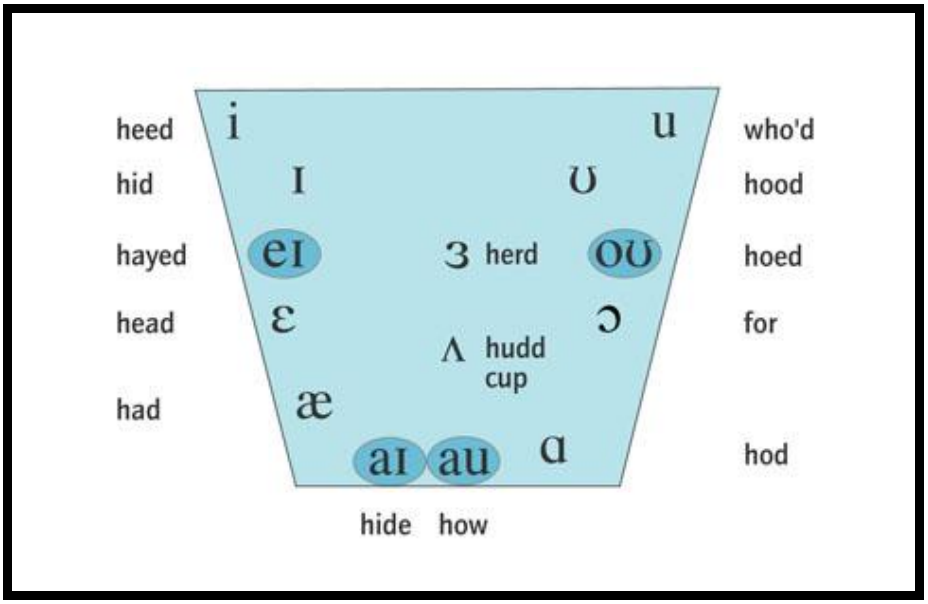


Fig 1.5 : Trapèze vocalique des voyelles anglaises

- Diphtongues (voyelles encerclées dans la figure ci-dessus): Les voyelles anglaises se distinguent de celles du Français notamment par le grand nombre de diphtongues (voyelles à double timbre), alors que le Français standardisé ne contient que des monophtongues (voyelles à timbre simple).

[eI] grey, great, sleigh, engage, gauge, etc.

[au] house, plow

[aI] my, tide, thigh, buy, etc.

- Les voyelles antérieures arrondies : le système des voyelles anglaises n'a pas de classe d'antérieures arrondies comme le Français.
- L'Anglais n'a pas de voyelles nasales alors que le Français standardisé en a quatre.
- L'existence d'un grand nombre de voyelles qui tendent vers un schwa lorsqu'en position inaccentuée.

1.8.2. Les consonnes

Les consonnes de l'Anglais sont sensiblement les mêmes que celles du Français. Tab1.8. L'Anglais possède également deux articulations interdentales totalement inconnues du Français. On articule les interdentales en rapprochant la pointe de la langue vers les dents supérieures tout en dépassant celles-ci pour que la langue soit visible de l'extérieur. ex.: three / tooth / these / bathe

Tableau1.8 : les consonnes de l'Anglais

Mode articulatoire	Lieu d'articulation													
	Bilabiales		Labiodentales		Interdentales		Alvéolaires		Palatales		Vélaires		Glottales	
Occlusives	p	b					t	d			k	g	ʔ	
Fricatives			f	v	ɵ	ð	s	z	ʃ	ʒ			h	
affriquées									tʃ	dʒ				
nasales		m						n				ŋ		
liquides (latérales)								l						
liquides (rétroflexes)								r						
semi-consonnes		w								j				

Note : Les colonnes en blanc et en gris indiquent les symboles sourds et sonores respectivement.

Plusieurs autres langues du monde utilisent ces deux articulations, comme l'Arabe, le Danois, le Grec, etc.

Apico-dentales sont subitement des occlusives alvéolaires en Anglais Il y a une fricative sourde (h)

À noter la consonne glottale [h] qui est présente en anglais mais également en Français comme dans les exemples où est présent un affaiblissement des consonnes fricatives dorso-palatales (comme dans "**ch**ien" et "**j**eu"). Encore une fois cependant, il ne s'agit pas d'un phonème propre au Français mais bien d'un allophone sourd de nos consonnes dorso-palatales.

La présence de consonnes affriquées Certaines langues, et c'est le cas de l'Anglais, utilisent des consonnes complexes. Ces consonnes sont: affriquées: consonnes formées d'une occlusive et d'une fricative. En anglais par exemple, les affriquées sont utilisées dans les mots:

[tʃ] church, lunch, chip, ditch, etc;

[dʒ] judge, jouchrnal, germ, ect;

Note : il est à remarquer que le phénomène d'affrication (sans statut phonologique, nous parlons uniquement du phénomène d'affrication et non de consonnes affriquées) (en Français québécois: [ts, dz]) est également présent en Français canadien.

1.9. Phonétique dialectale

Observez le tableau Tableau 1.9 de Français québécois:

Le "r" en Français canadien est un cas particulièrement intéressant de variation dialectale et sociolectale. Il faut d'abord savoir qu'il y a plusieurs variantes de "r" en

Français, et que l'on en retrouve au moins 5 principales au Canada. De façon à expliquer la distribution de toutes ces variantes, il est nécessaire de considérer l'interaction de plusieurs facteurs: la géographie, les caractéristiques sociales et stylistiques des locuteurs et l'histoire.

Tableau 1.9 : Phonétique dialectale du Français québécois

MOT	TRANSCRIPTION	MOT	TRANSCRIPTION
petite	[pətsɪt]	père	[paᵉʀ]
fête	[faᵉt]	dieu	[dʒjø]
ruche	[ʀʏʃ]	mur	[mʏʀ̥]
tarte	[taʀt]	beau	[bo]
excuse	[ɛkskyz]	tien	[tʃjɛ̃]
additionner	[adzisjɔne]	cadre	[kadʀ̥]
jaune	[ʒoᵘn]	mot	[mo]
entourer	[ɑ̃tuʀe]	muret	[mʏʀɛ]
toute	[tʊt]	roue	[ʀu]
bidule	[bidzʏl]	adoucir	[adusiʀ]
chien	[ʃjɛ̃]	char	[hɑʀ]
curé	[tʃyʀe]	culottes	[tʃylɔt]
pépin	[pepɛ̃]	bébé	[bəbəᵉ]
tien	[kjɛ̃]	diable	[dʒɑb]
le bon dieu	[lə bɔ̃ jø]	prendre	[prɑ̃dr̥]
cinq	[sɑ̃ᵋk]	encore	[ɑ̃koᵘʀ̥]
mettre	[mɛtʀ̥]	plus	[plʏs]
musique	[mʏzɪk]	âge	[aᵒʒ]
plus	[ply]	crêpe	[kʀaᵉp]
pâte	[paᵒt]	père	[peⁱʀ]
neige	[nɛ ːʒ]	pige	[pi ːʒ]
pipe	[pɪp]	mère	[mɛ ːʀ]

L'information détaillée sur ce phénomène est affichée sur le site du Centre interdisciplinaire de recherche sur les activités langagières (CIRAL) de l'université Laval à Québec.

1.10. Définitions pertinentes

Diphtongue : articulation vocalique complexe qui comporte une variation de lieu d'articulation ou de mode articulatoire en cours d'émission. Les diphtongues impliquent un relâchement des organes articulatoires et un changement de timbre vocalique.

Monophtongue : voyelle qui est articulée de façon constante, sans changement de timbre en cours d'émission

Diacritique : un signe diacritique est un signe graphique (point, accent, symbole quelconque) qui est ajouté à une lettre de l'alphabet pour en changer la valeur. L'A.P.I. utilise de nombreux signes diacritiques. Ils peuvent être suscrits, antéposés, postposés, voire superposés. Chacun altère la valeur du son sur lequel il porte.

Affriquée : une consonne affriquée combine successivement une occlusion et une constriction

1.11. Conclusion

Ce chapitre a permis dans sa première partie d'introduire certains concepts de base de la phonétique et le principe de production de son. En deuxième partie on a expliqué l'API et la TOP avec donnant des exemples illustratives

Chapitre 2 :
Etude des paramètres prosodiques

2.1. Introduction

Dans ce chapitre nous présentons la prosodie qui est définie de façon diverse, qui confirme la complexité du phénomène et la multiplicité de ses facettes. Même en se restreignant au domaine de la linguistique. En suite nous donnons sa fonction et son but principal ainsi que ses principaux paramètres sur le plan articulatoire, perceptif et acoustique.

2.2. Définition de la prosodie

Dans la littérature linguistique, les définitions du terme prosodie recouvre plusieurs faits dont le domaine d'application s'étend au-delà du phonème : syllabe, accent, rythme, ton, intonation, pause, débit, etc. Dans d'autres contextes, le terme est aussi défini par référence à la poésie comme étant l'ensemble des règles qui régissent la composition des vers, et en musique, le terme concerne l'étude des règles de concordance des accents d'un texte et de la musique qui l'accompagne. En résumé, redéfinir la prosodie avant chaque étude est devenu une sorte de compromis entre les chercheurs qui abordent les études prosodiques sous des angles différents.

Le mot prosodie est originaire du grec Προσωδια, qui signifie ″accent, quantité, dans la prononciation″. Ces derniers l'utilisaient pour faire référence aux traits du discours, plus précisément aux tons ou accent mélodique. La mélodie de la prosodie est restée dans l'oubli jusqu'à la fin des années 1940, lorsque Firth [9] utilisa ce terme à nouveau pour décrire l'approche qu'il préconisait pour l'analyse linguistique.

La première expérience sonore prosodique est l'écoute d'une langue étrangère, dont nous ne maîtrisons aucun aspect. Ainsi, les impressions qui nous parviennent sont transmises par le chant, la force, ou le timbre de la voix, qui nous permet d'identifier un ami par exemple, même si nous ne comprenons pas ses propos. La prosodie est donc liée à l'impression musicale que fournit un locuteur lorsqu'il parle.

Dans ce travail, nous avons retenu la définition de la prosodie, proposée par Di Cristo, puisqu'elle nous semble englober les différents domaines impliqués dans l'étude du phénomène :[10], « La prosodie est une branche de la linguistique consacrée à la description et à la représentation formelle des éléments de l'expression orale tels que les accents, les tons, et l'intonation, dont la manifestation concrète, dans la production de la parole, est associée aux variations de la fréquence fondamentale (F_0), de la durée (D) et de l'intensité (I) (paramètres prosodiques physiques) ». Ces variations étant perçues par

l'auditeur comme des changements de hauteur (ou de mélodie), de longueur et de sonie (paramètres prosodiques subjectifs). Les signaux prosodiques véhiculés par ces paramètres sont polysémiques et transmettent à la fois des informations para-linguistiques et des informations linguistiques déterminantes pour la compréhension des énoncés et leur interprétation pragmatique dans le flux du discours.

A travers les différentes composantes de la prosodie, nous retiendrons qu'elles ne se définissent pas uniquement à partir des caractéristiques physiques du signal, c'est-à-dire l'acoustique. La prosodie traite des éléments de l'expression orale qui se manifestent physiquement par des variations de F_0, de durée et d'intensité. Ces éléments de l'expression orale transmettent notamment des informations sur le sens d'un énoncé.

Sur l'exemple suivant, à l'écrit, une marque syntaxique (?) permet la distinction entre les deux phrases et il n'y a pas d'ambiguïté sur leur sens :

- Le train arrive à midi.
- Le train arrive à midi ?

À l'oral, la situation est différente et il est nécessaire pour se faire comprendre de transposer la marque interrogative de l'écrit dans le message oral. On peut noter que la nature des sons, les phonèmes ne changent pas dans les deux exemples. C'est un autre procédé qui va permettre de modifier le sens de la phrase et c'est l'intonation qui est utilisée. Le sens de la phrase dépend donc de l'intonation et plus généralement de la prosodie. On voit sur cet exemple que la prosodie est un procédé non univoque. Autrement dit, un même énoncé peut être prononcé avec des prosodies différentes. Ces différences de prosodie influente sur le sens de l'énoncé, c'est pourquoi la prosodie paraît essentielle à la compréhension et au naturel de la parole.

2.3. But de la prosodie

Améliorer les performances d'un système indépendant du locuteur. Le rôle de la prosodie dans le système de la langue est plus ou moins contesté. La difficulté de prêter une fonction propre à la prosodie vient de la notion de continuité et donc de la difficulté de dégager des unités intonatives.

La conception de la prosodie qui a prévalu pendant longtemps dans la théorie linguistique est celle qui la considère comme un phénomène parallèle à la parole et extralinguistique. Parmi toutes les fonctions de la prosodie qui ont pu mettre à jour, la fonction syntaxique semble être la plus débattue actuellement.

Deux points sont essentiels :

Malgré les divergences théoriques et méthodologiques, il semble qu'il y ait un consensus sur l'importance et l'intérêt de la prosodie envisagée comme un phénomène autonome possédant ses propres paramètres acoustiques.

Un 2ème consensus semble être établi au sujet de l'existence dans la chaîne parlée d'unités prosodiques, unités pouvant aider à la segmentation automatique de l'énoncé (en structures syntaxiques, en thème, en groupes de sens, en mots ou même en syllabe selon les auteurs).

En d'autres termes, tous les auteurs s'accordent pour prêter à la prosodie une fonction démarcative, quelle que soit par ailleurs, la théorie sur le lien de celle-ci avec les divers niveaux linguistiques.

2.4. Paramètres prosodiques

La prosodie dépend non seulement du niveau de la syntaxe mais aussi de la sémantique. De plus, selon Vannier, la prosodie a la particularité d'être à la fois universelle et spécifique à une langue. Autrement dit, chaque langue possède sa propre prosodie, même si elle partage certaines propriétés avec d'autres langues. Pour une même langue, on note également l'existence d'une diversité intra-locuteur et interlocuteur qui peut par exemple être liée à l'état d'esprit du locuteur ou encore à son origine socioculturelle.

Nous étudierons plus spécifiquement par suite les principaux paramètres de la prosodie que sont la durée, l'intensité et la fréquence fondamentale.

On peut considérer que l'information prosodique se résume essentiellement à l'évolution de F_0 qui, de ce fait, apparaît comme un élément prépondérant de la prosodie [11].

Néanmoins, la prosodie est un des constituants de la parole qui reste accessible à chacun d'entre nous sans connaissance particulière (Figure 2.1).

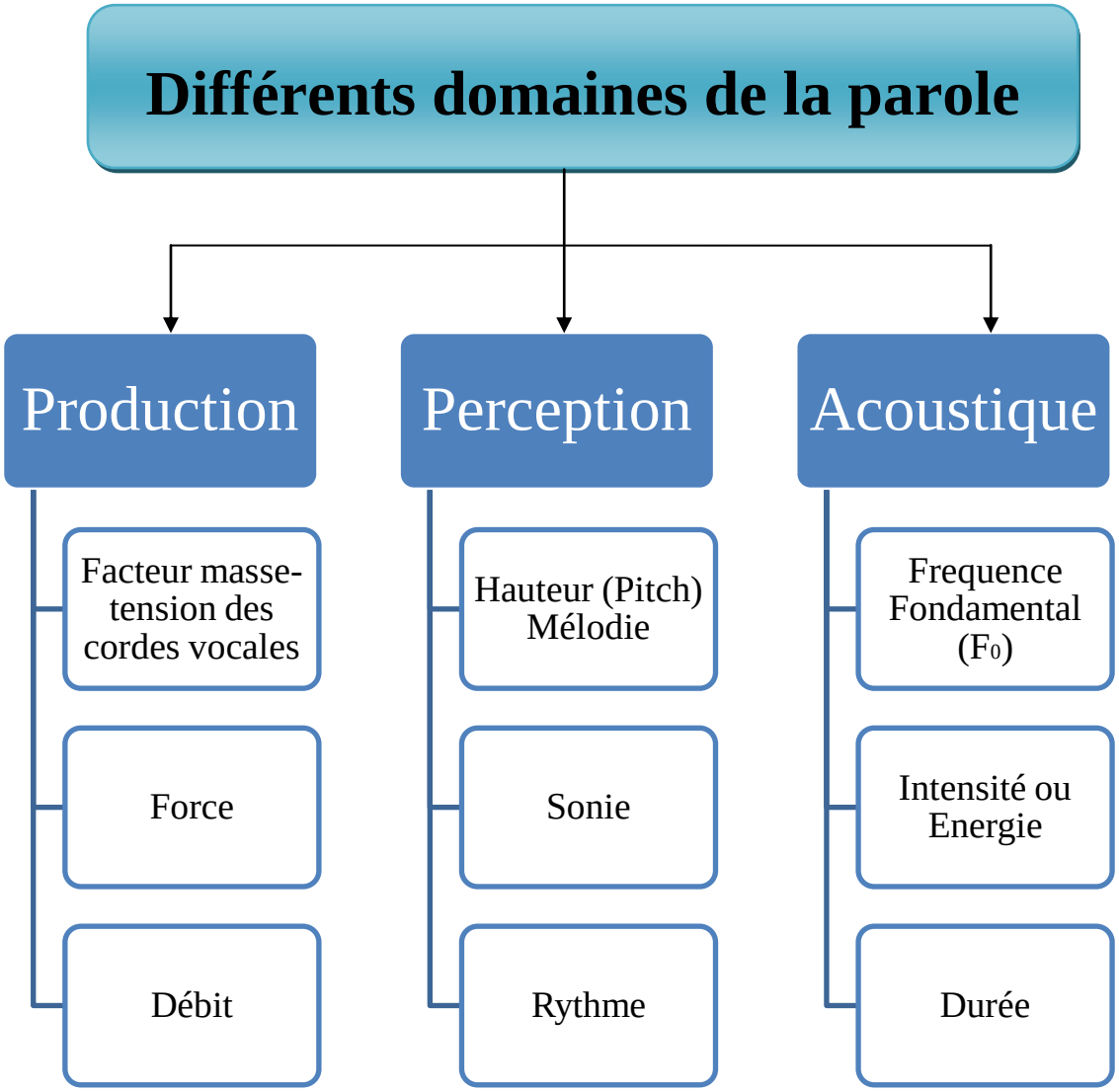


Figure 2.1 : Paramètres prosodiques dans différents domaines de la parole

La substance prosodique repose sur 4 propriétés acoustiques [12] :

• le rythme : cette notion comprend le débit de parole, la longueur et la répartition des pauses, les allongements syllabiques, la durée de divers événement sonores (syllabes, phonèmes), etc. Étant donné son lien avec tous les autres traits acoustiques de la prosodie, il est particulièrement difficile d'isoler les paramètres acoustiques qui puissent résumer cette dimension ;

• l'intonation: elle est souvent présentée comme le paramètre primordial de la prosodie. Elle contient le caractère chantant de la parole. Contrairement au rythme, l'intonation est définie à partir d'un seul paramètre acoustique (F_0). Il est la conséquence de la vibration des cordes vocales et de la pression trans-glottique ;

• le volume sonore **:** il correspond au paramètre physique de l'intensité (I), c'est-à-dire l'énergie contenue dans le signal au cours d'un intervalle de temps donné ;

• le timbre **:** il est spécifique des instruments ou des voix. Il est perçu indépendamment de sa hauteur ou de son intensité. L'évolution du timbre est provoquée par la superposition des composantes harmoniques durant l'émission du son. Une fois encore, cette dimension est difficile à expliciter physiquement puisqu'elle s'appuie sur l'ensemble des valeurs spectrales du signal.

Ainsi, un très grand nombre de définitions de la prosodie ont été proposées, suivant notamment les contextes dans lesquels ce terme est employé. D'un point de vue acoustique, par exemple, nous trouverons une définition comme: « étude de la durée, de F_0 et de l'intensité (ou énergie) du son », ou bien pour des aspects plus linguistiques : « partie de la phonologie qui échappe à l'analyse en phonèmes et traits distinctifs, tels que le ton, l'intonation, la durée et l'accent».

2.5. Extraction des paramètres prosodiques

Les trois paramètres prosodiques classiquement extraits du signal acoustique sont la fréquence fondamentale, la durée et l'intensité (ou l'énergie). La première étape de l'extraction de paramètres prosodiques est généralement une analyse à court terme du signal, en faisant l'hypothèse que sur une fenêtre de faible longueur le signal est quasi-stationnaire, ce qui signifie que les caractéristiques statistiques du signal évoluent peu. En général une fenêtre d'analyse de 30 ms, appelée trame, est utilisée. L'analyse est répétée à intervalles réguliers, typiquement toutes les 10 ms. Compte tenu du caractère suprasegmental de la prosodie, une analyse sur 30 ms ne suffit pas. C'est pourquoi il est nécessaire de calculer d'autres paramètres à partir d'une analyse sur plusieurs trames successives afin de traduire le rythme, l'intonation et l'accentuation.

2.5.1. Durée

Des trois paramètres prosodiques, la durée (D) est le plus difficile à préciser, car elle n'est pas directement associable à un corrélat biologique du système phonatoire.

Avant de mesurer des durées, il faut cerner correctement les entités à mesurer. On distingue les durées des unités phonétiques, des syllabes, des phonèmes ou même la distance entre voyelles et les durées des pauses. Comme les autres paramètres, la durée de l'entité choisie est largement dépendante du locuteur et du débit de parole. Ainsi, aucune

mesure ne peut donner de modèle absolu de la durée. La considération des résultats des observations devra plutôt s'orienter vers un modèle relatif qui pourra s'exprimer en termes d'allongements ou de réductions.

Chaque phonème a une durée intrinsèque et co-intrinsèque. Ces durées sont des caractéristiques des phonèmes. On se rend compte aisément que le phonème [a], pris seul, est plus long que le phonème [b], par exemple.

Les pauses en parole spontanée ne sont pas toutes des silences. On distingue les pauses silencieuses des pauses non silencieuses (qui peuvent être remplies, faux départs, répétitions, ou syllabes allongées). En situation de lecture seule, les pauses qui se traduisent acoustiquement par une absence de signal (les pauses silencieuses) sont considérées.

La durée des différentes unités constitue le phénomène central pour la prosodie. En effet, chaque variation de F_0 ou d'intensité s'établit sur un certain laps de temps.

Etudier l'organisation temporelle de la parole est incontournable. Etudier la durée, c'est observer et modéliser les durées d'unités bien déterminées.

Pour cela, la durée et la nature de ces unités ont fait l'objet de nombreuses études, principalement motivées par la nécessité de la modéliser dans des systèmes de synthèse de la parole.

Dans sa thèse, P.Barbosa [13] consacre tous le chapitre 2 pour classer les différents modèles de prédiction des durées, selon un ordre croissant de taille des unités utilisées. Parmi les unités qui ont servi de base à ces modélisations, on en trouve principalement quatre : le phonème, la syllabe, le pied et le GIPC (Groupe-Inter-Perceptuel-Center).

2.5.2. Energie

L'énergie (E) (ou l'intensité (I)) du signal sonore de la parole est perçue comme la force de la voix. Son niveau est lié au fonctionnement des systèmes respiratoire et phonatoire et à la pression sous glottale : si la pression sous glottale augmente, l'intensité de la voix augmente également et inversement. Cette intensité est relative à l'énergie contenue dans le signal au cours d'un intervalle de temps donné. Ce terme correspond au corrélât acoustique de la pression sous glottique et d'ouverture du conduit vocal.

Sachant que l'énergie est un paramètre couramment utilisé en traitement du signal, c'est le paramètre prosodique le plus facile à calculer. L'énergie à CT d'un signal est échantillonnée sur une fenêtre de longueur T.

2.5.3. Fréquence fondamentale (F_0)

La vibration, qui est en fait l'accolement puis la séparation, des CV portées par le larynx détermine F_0 appelée encore pitch ou F_0. Elle est comprise entre 75 et 150 Hz chez les hommes, 150 et 300 Hz chez les femmes, et est supérieure ou égale à 300 Hz chez les enfants [11]

Autrement dit, c'est l'estimation de la fréquence laryngienne à partir du signal acoustique à un instant donné. Les algorithmes d'extraction de F_0 utilisent une représentation temporelle ou spectrale du signal.

Un algorithme d'extraction de F_0 se décompose en trois phases successives [14] :

• un prétraitement et un changement de représentation ;

• l'extraction du fondamental ;

• un post-traitement visant à corriger les erreurs.

La deuxième phase consiste à extraire F_0 et dépend donc du domaine utilisé. Généralement, cela revient à optimiser une fonction de F_0 (fonction de coût, résultat d'une transformation, corrélation, densité de probabilité).

La phase de post-traitement a pour but de diminuer les erreurs qui sont de plusieurs types :

• les erreurs de voisement, lorsqu'une valeur de F_0 a été trouvée sur une zone non voisée, ou lorsque aucune n'a été trouvée sur une zone voisée ;

• les erreurs grossières (« gross-errors » en anglais), F_0 correspond à une harmonique ou une sous-harmonique. Ce type d'erreur peut facilement être corrigé en tenant compte du voisinage ou en effectuant un lissage ;

• les erreurs fines, la valeur trouvée est située à plus ou moins 10 % de la valeur réelle.

Le calcul de F_0 se fait donc sur les sons voisés qui ont un caractère pseudo périodique. Cela concerne principalement les voyelles, mais aussi quelques consonnes. Il existe plusieurs algorithmes pour l'estimation de F_0 et qui ne donnent pas toujours des résultats identiques. Ainsi, de nombreuses recherches ont été menées dans le domaine de l'extraction de F_0 des signaux de parole. On peut citer l'ouvrage de référence de Hess [15], où un grand nombre d'algorithmes est détaillé.

Les plus performants d'entre eux, sont cependant incapables de fournir des valeurs toujours correctes de F_0 dans toutes les circonstances (sons, bruits, locuteurs, etc.). Les principaux problèmes rencontrés sont :

- les sauts d'octaves : l'analyseur fournit une valeur de F_0 qui ne correspond pas au premier harmonique. Cela peut arriver pour un spectre dont le deuxième harmonique correspond au premier formant ou dans le cas d'une insuffisance passagère de l'amplitude du fondamental ;
- les non-détections : il existe une fréquence « théorique » que l'algorithme n'a pas détectée. Ceci arrive très souvent dans des portions peu énergétiques et / ou bruitées du signal de parole ;
- la finesse du détecteur : les valeurs proposées sont éloignées faiblement des valeurs théoriques ;
- la décision de voisement : cette décision, bien que difficile à prendre dans certaines situations (faible énergie, parole bruitée, …) serait cependant fort utile, au-delà du bon fonctionnement du détecteur, à des fins de segmentation du continuum sonore ;

On recense deux grandes catégories d'algorithmes de décision ; Ceux qui opèrent dans le domaine temporel comme la technique d'AMDF (Average Magnitude Difference Function). Et ceux qui travaillent dans le domaine spectral: les valeurs de F_0 sont calculées à partir des maxima des spectres d'amplitude.

2.6. Autres paramètres prosodiques

Dans ce paragraphe nous allons site d'autres paramètres prosodiques

2.6.1. Intonation

Dans plusieurs études, le terme intonation est employé indifféremment de celui de la prosodie. Ainsi, Di Cristo suggère de compléter la définition précédente par : « la prosodie est une structure grammaticale possédant une organisation qui lui est propre. ». 'Pour la notion d'intonation, il emprunte la définition : «[le terme intonation] fait référence à l'usage qui est fait des traits phonétiques suprasegmentaux pour véhiculer, au niveau post-lexical ou de la phrase, des signifiés pragmatiques d'une façon linguistiquement structurée. »[10].

L'article de Delattre établit une classification des différentes intonations possibles dans un énoncé. Pour établir les principaux types d'intonation du français, Delattre a utilisé des extraits de conversations, de pièces de théâtre et de conférences. Le résultat de cette analyse met en évidence l'existence de dix types d'intonation de base figure 2.2.

Avec ce modèle d'intonation, il existe quatre niveaux d'intonation : basse, moyenne, haute et aiguë. Cette modélisation met en jeu les trois modalités suivantes : interrogation, exclamation, affirmation. En particulier, Delattre montre qu'en faisant des substitutions entre les intonations de base dans une phrase de même contenu, on obtient des changements de sens. Cela montre notamment le rôle important de l'intonation pour la compréhension du message oral.

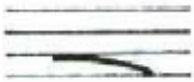
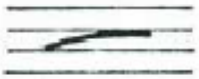
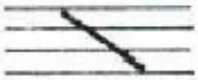
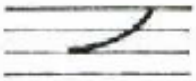
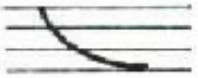
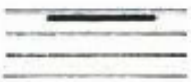
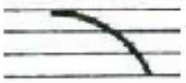
		représentations graphiques
déclaratives	finalité	2-1 
	continuation majeure	2-4 
	continuation mineure	2-3 
	implication	2-4_ 
	commandement	4-1 
interrogatives	question	2-4+ 
	interrogation	4-1 
parenthétiques	parenthèse	4-1 
	écho	4-4 
	exclamation	4-1 

Figure 2.2 : Les dix intonations de base de Delattre

Le mot ton désigne le ou les niveaux de hauteurs observés dans une syllabe donnée. Le ton coïncide donc avec la partie de la courbe mélodique qui se rattache à une seule syllabe ». L'intonation d'un énoncé se présente comme une succession de tons. On distingue quatre niveaux de hauteur : haute, basse, infra-bas et suraiguë.

2.6.2. Accent

En Français, l'unité porteuse de l'accent est la syllabe et l'accentuation, comme l'écrit, elle peut être définie de la manière suivante : « Une syllabe est dite accentuée quand elle ressort sur son entourage par sa force particulière, par un contraste d'intensité subjective »

On distingue généralement trois classes de langues suivant leur comportement par rapport à l'accentuation :

- les langues à accent libre (cas de l'Anglais) : on ne peut pas déterminer sa place à l'avance et il y a autant de possibilités de placement de l'accent que de syllabes (dans le cas où l'unité accentuable est la syllabe) ;
- les langues à accent déterminé, ou « fixe » (cas du français) : l'application des lois de placement de l'accent suppose le décompte préalable des syllabes. En règle générale, on détermine la place de l'accent en partant de la fin d'un mot ;
- les langues à tons : les variations tonales y sont utilisées à des fins sémantiques pour distinguer plusieurs significations linguistiques.

Deux grandes catégories d'accents sont distinguables en français : l'accent final (primaire) et l'accent non-final (secondaire) associé aux fonctions linguistiques (focalisation) ou paralinguistiques (expressivité). Sur le plan lexical, l'accent français n'est pas distinctif ni pour les mots ni pour les morphèmes.

Plusieurs dénominations existent pour parler de l'accent final du Français : accent logique, objectif, tonique, normal ou interne. La plupart des études s'accordent à dire que l'accent primaire est assigné à la dernière syllabe pleine (qui ne contient pas de schwa) du dernier item lexical d'un groupe accentuel. Il se caractérise principalement par un allongement important de la durée.

L'accent secondaire est optionnel. De manière générale, sur le plan acoustique, l'accent secondaire se traduit par une augmentation de F_0 et de l'intensité. Il obéit à des contraintes de nature rythmique, pragmatique et expressive.

Trois types d'accents secondaires peuvent ainsi être distingués :

- rythmique est lié à la réalisation d'un décompte syllabique et permet la mise en relief de la mélodie. Il peut être associé à l'augmentation de l'intensité.
- pragmatique, également qualifié d'accent énonciatif ou d'accent de focalisation, permet de mettre en relief une partie de l'énoncé.
- expressif ou emphatique, exprime l'attitude du locuteur à l'égard de ce qu'il dit.

2.6.3. Intensité et débit

Le rôle de l'intensité est strictement communicatif. L'enjeu pour un locuteur est de montrer s'il désire conserver la parole ou bien la céder à son interlocuteur, de montrer une certaine insistance sur une partie de son discours ou encore de montrer son adhésion plus ou moins forte à l'énoncé.

Même si le débit est généralement exprimé en unités de parole par unité de temps, par exemple en nombre de syllabes par seconde pour le cas du français, augmenter ou diminuer le débit d'une phrase de manière globale ne produit pas un effet naturel. En effet, le débit est influencé par plusieurs facteurs qui sont notamment les pauses, l'allongement ou le raccourcissement des segments, l'ajout de sons. Bien que le dernier point puisse sembler inattendu, il correspond notamment à un phénomène que l'on pourrait qualifier « d'hyper-articulation » et qui provoque un ralentissement du débit de la parole. Les variations de durée peuvent par exemple traduire l'hésitation du locuteur, une certaine incertitude ou encore une émotion. Concernant le débit du français, il se situe entre 4 et 7 syllabes par seconde.

2.6.4. Les pauses

Elles correspondent à des silences ente les mots, des groupes de mots ou des phrases. Elles constituent des éléments très importants en prosodie même si au plan acoustique elles correspondent à des phases de non signal. Les pauses sont souvent les seuls éléments fiables qui permettent de segmenter un énoncé en phrase ou syntagmes donc en ‘intonèmes’ (unités intonatives). Souvent on va donc chercher les pauses avant même d’étudier les variations des autres paramètres.

2.6.5. L'accentuation

Elle se matérialise selon les langues par la variation d'un des paramètres acoustiques (intensité, hauteur, timbre ou durée). En français ce sont les accents d'intensité ou de durée qui sont privilégiés. Les accents jouent un rôle essentiel puisque ce sont eux qui donnent le rythme à la parole. On trouve l'accent

fixe / libre : Dans les langues à accent fixe la place de l'accent est toujours la même. Dans les langues à accent libre la place de l'accent n'est pas prévisible, l'accent peut alors avoir une valeur distinctive pour opposer par exemple 2 catégories syntaxiques.

syntaxique / lexical : L'accent syntaxique porte sur un syntagme. En Français l'accent syntaxique et l'accent fixe se confondent. L'accent lexical porte sur tout mot lexicalement plein (en morphologie on parle de lexèmes). Ces accents ont une fonction démarcative.

emphatique / interne : L'accent emphatique a une fonction expressive et une fonction d'insistance (mise en relief de certains termes). Il est placé sur la 1ère syllabe du mot ou du groupe qui doit être mis en relief. On dit aussi qu'il a une fonction contrastive ou culmination car il met en relief certaines syllabes. L'accent interne, par opposition, se place toujours sur la dernière syllabe et a une fonction démarcative (il se confond avec l'accent fixe ou l'accent syntaxique). Il signale les fins de groupes ou de phrases.

2.7. Fonctions de la prosodie

Dans ce paragraphe, nous allons simplement lister les fonctions principales de la prosodie.

2.7.1. Structuration de l'énoncé

Dans les langues à accent libre, la prosodie permet la distinction entre homonymes.

En Anglais par exemple, elle permet de distinguer un nom d'un verbe (en gras, position de l'accent lexical) :

- **seg**ment (un segment)
- seg**ment** (segmenter)

Dans les langues à accent fixe, la position de l'accent lexical est la même pour tous les mots. En Français, l'accent lexical intervient sur la dernière syllabe. Il permet de distinguer les frontières des mots (fonction démarcative).

De manière générale, la prosodie permet de déterminer la structure d'un énoncé.

Par exemple, certaines ambiguïtés syntaxiques peuvent apparaître :

« La petite brise la glace »

Cette phrase peut prendre deux sens différents selon que le verbe est « briser » ou « glacer ». Ici, la prosodie nous aide à déterminer quel est le sens adapté à la situation.

Par exemple, une pause du locuteur après « La petite » permet de focaliser l'attention sur le sujet et de le délimiter. Ainsi, cela permet de lever l'ambiguïté syntaxique de la phrase. Dans ce cas, on peut aisément identifier que le verbe est « briser » et non « glacer ». L'expression « une tasse de thé russe » est un autre exemple d'ambiguïté pouvant être levée grâce à l'intonation. Est-ce la tasse qui est russe ou bien le thé ? On voit alors clairement apparaître la fonction de structuration que possède la prosodie.

2.7.2. Focalisation

La focalisation est un moyen d'insister sur certains mots. Dans l'exemple suivant, l'accentuation apporte un sens particulier à la phrase :

- Je vais terminer ;
- Je vais terminer ;
- Je vais terminer.

Ici, le premier cas montre une insistance sur le sujet pour montrer qu'il est important que ce soit moi qui termine. La suivante traduit le fait que cela sera terminé mais qu'il faut encore du temps. Enfin, la dernière assure que la tâche sera bel et bien terminée, c'est une certitude. Ces différentes accentuations apportent donc des nuances de sens qui traduisent une volonté du locuteur par rapport au message qu'il souhaite transmettre.

2.7.3. Modalité

La prosodie, et plus particulièrement la mélodie, est liée au mode de la phrase. On peut distinguer quatre modes : affirmatif, interrogatif, impératif et exclamatif. Elle permet, par exemple, d'identifier une question sans qu'il y ait besoin d'indices syntaxiques tels qu'une inversion sujet/verbe. Ainsi, la phrase _ tu vas bien _ peut être produite selon les modes interrogatif, affirmatif ou encore exclamatif.

2.7.4. Fonctions non linguistiques

La prosodie peut apporter des informations sur l'état psychologique du locuteur : calme, énervé, triste, etc. Aussi, elle varie très largement suivant la provenance

géographique, le niveau social et permet d'identifier un individu en tant que membre d'un groupe social ou culturel. L'exemple le plus frappant est sûrement l'accent régional comme celui du sud ou du nord/ est ou l'ouest.

De plus, la prosodie permet de véhiculer l'attitude du locuteur envers l'interlocuteur et celle vis-à-vis de l'énoncé (adhésion plus ou moins forte). La manière de parler à un enfant peut facilement être différenciée de celle utilisée pour parler à un adulte. Le type d'intervention orale fait également appel à des prosodies différentes que ce soit pour une narratrice de contes populaires ou pour le discours d'un homme politique. La prosodie est liée à la stratégie de communication du locuteur.

2.8 Conclusion

Dans ce chapitre, nous avons décrire la prosodie et ses paramètres, sur les différents plans ainsi son but principale on donnant plus de détails sur le plans acoustique avec les paramètres énergie, durée et fréquence fondamentale suivi par sa fonction et les caractéristiques de la voix

Chapitre 3 :
Méthodes de détection de la F0 appliquées à la RAL

3.1. Introduction

Dans cette partie, nous allons définir brièvement la reconnaissance automatique (RAP) de la parole et la reconnaissance automatique de locuteur (RAL), leurs utilisations dans différents domaines. En suit les complexités pour détecté F_0 ainsi cité quelque méthodes existantes et on finira par une description de ces techniques

3.2. Reconnaissance Automatique de la Parole

L'utilisation de la parole comme mode de communication entre un homme et une machine a été largement étudiée au cours des dernières décennies. Nous nous intéressons dans cette partie à la RAP, c'est-à-dire à l'ensemble des techniques permettant de communiquer oralement avec une machine. La RAP présente un intérêt pratique indéniable, dans certaines conditions d'utilisation tableau 3.1. Des produits commerciaux existent depuis plus de trente ans, d'abord essentiellement pour la reconnaissance de mots isolés et enchaînés puis maintenant pour des phrases prononcées continûment. La plupart sont fondés sur des algorithmes de programmation dynamique et des modèles stochastiques figure 3.1 (sources de Markov). Néanmoins, des problèmes restent à résoudre pour accroître la robustesse de ces systèmes et pour étendre leurs capacités de dialogue. Les recherches menées actuellement portent ainsi sur la reconnaissance de parole bruitée, le traitement d'énoncés incomplets ou incorrects, la définition de procédures de dialogue, etc.

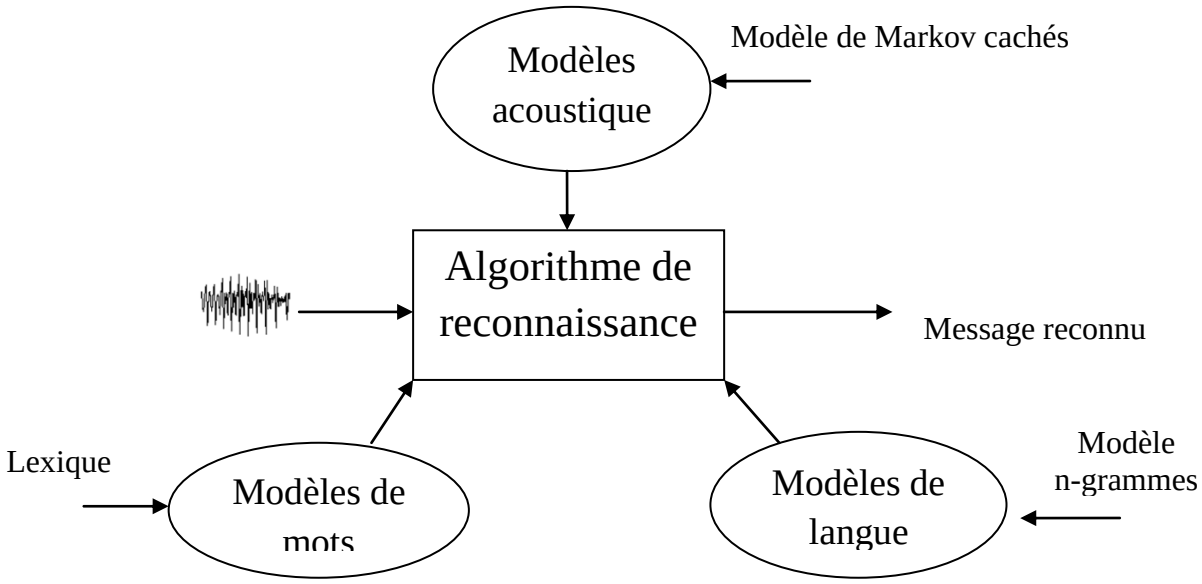


Fig. 3.1 : Principe de la reconnaissance de la parole

Tableau 3.1 : Domaine d'application et produits en Reconnaissance de la parole

Type de fonction	Description	Exemples
Contrôle/commande	Commende vocale d'appareils ou de logiciel : - Chaise roulante - Appareillage - Commandes à un système d'exploitation d'ordinateurs - Numérotation téléphonique Mots isolés, petits vocabulaires	- Différents logiciels ou composants reconnaissant des mots isolés. - Téléphones avec numérotation vocale (Marta, Northern Telecom, Uniden) - Dragon Dictate et Via Voice comportent des commandes vocales au système d'exploitation.
Saisie de données	Entrée à la voix de données dans un ordinateur (remplissage de formulaires, contrôle de qualité, passage d'une commande, etc...),	Plusieurs prototypes dans différents domaines, pas de produits commercialisés
Télématique vocale	Messagerie vocale. Accès a une base de données ou un centre de renseignements, Opérations bancaires Mots isolés, petits ou moyens vocabulaire	Système d'assistance au téléphone d'ATT Voice FONCARD de Sprint
Dictée vocal	Production de lettres ou de documents écrits par dictée Parole continue, grands vocabulaires (plusieurs dizaines de mots), adaptation au locuteur	Natutally Speaking de Dragon Via Voice d'IBM Speech Magic de Philips Voice Xpress de Lerout et Hauspie

3.3. Reconnaissance Automatique du Locuteur

Le terme générique « reconnaissance automatique du locuteur » est utilisé aussi bien pour définir l'identification et la reconnaissance du locuteur. La vérification consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé. L'identification consiste en la reconnaissance d'un locuteur particulier parmi un ensemble fini de locuteurs possibles. Aussi bien la reconnaissance, que l'identification du locuteur se font en calculant un modèle stochastique sur la base de l'expression vocale du locuteur à reconnaître. Une fois calculé, ce modèle est comparé à des modèles préentraînés sur la base de différentes phrases prononcées par le(s) locuteur(s).

On classifie également les systèmes de reconnaissance et d'identification du locuteur en deux catégories :

- Indépendant du contenu de la phrase prononcé (text-independent).
- Dépendant du texte et donc effectué sur la base d'un texte imposé (text-dependent)

Les applications potentielles des systèmes de reconnaissance de locuteur incluent le contrôle d'accès à distance de bases de données, les services d'information et de réservation à distance, les services bancaires à distance, etc. La tendance actuelle montre une évolution ers l'exécution de diverses transactions en utilisant les téléphones mobiles.

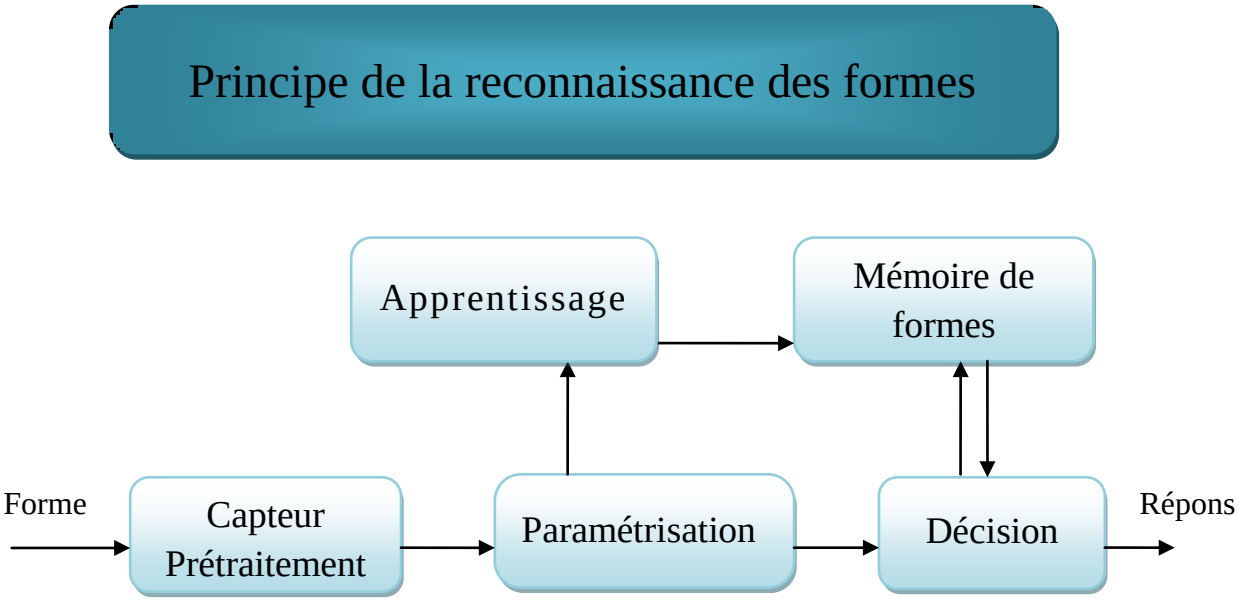


Fig. 3.2 : Principe de la reconnaissance des formes

3.4. Complexités de détection de la Fréquence fondamentale

La complexité d'évaluation du Fondamental est une tâche difficile pour de nombreuses raisons telles que la non-stationnarité du signal vocal, une certaines irrégularités dans l'excitation glottique ou encore une interaction avec le premier formant, la décision du voisement, la distinction entre les segments non voisée et les segments voisée à énergie réduite, la difficulté inhérente en définissant le début et la fin exacts de chaque période de F0 durant les segments de la parole Voisée et dernièrement le doublement de période local [16]. C'est un type d'erreur qui affecte pratiquement toutes les méthodes d'estimation de F0.

3.4.1. Variabilité des paramètres prosodiques et contraintes de production

Nous examinerons principalement le cas de F0, qui est le paramètre prosodique physique auquel les analyses se réfèrent le plus fréquemment. Si nous tentons d'établir, en premier lieu, la nature des relations qui lient les variations d'un paramètre physique comme la fréquence fondamentale (F0) à l'actualisation des entités phonologiques, lexicale et supra lexicale que représentent le ton, l'accent et l'intonation, il s'avère utile, voire indispensable, de connaître la nature et l'importance des contraintes de production qui affectent l'évolution de ce paramètre. Ces contraintes, qui sont associées à plusieurs facteurs, engendrent des *effets universels* (quoique quantitativement variables d'une langue à l'autre) que nous proposons de décrire brièvement.

3.4.2. Les contraintes idiosyncrasiques

Il est clairement établi que les variations de F0 sont en partie déterminées par les caractéristiques physiologiques des locuteurs (notamment par la masse volumique des cordes vocales) et que par conséquent la tessiture tonale peut varier d'un sujet à l'autre, et varie de toute évidence systématiquement entre les voix d'homme, de femme et d'enfant. Cette source de variation non linguistique étant reconnue, il importe de la neutraliser avant toute interprétation linguistique, ce qui peut être réalisé par la mise en œuvre de procédures de normalisation. Ces dernières consistent principalement à convertir, à l'aide de formules appropriées, les valeurs absolues de F0 mesurées en Hertz (Hz) en valeurs relatives (ou en valeurs logarithmiques). Plusieurs échelles sont alors disponibles à cet effet, telles que l'échelle des demi-tons ou l'échelle ERB, qui présentent également l'avantage de

correspondre à des échelles auditives (pour obtenir plus de détails, lancer la recherche: «*auditory scales of pitch perception*» sur le web).

3.4.3. Les contraintes inhérentes à la gestion des variations de F0

Une autre contrainte de production qui affecte également les variations de F0 dans la parole concerne la vitesse avec laquelle celle-ci peut croître ou décroître dans un temps imparti, défini, par exemple, par les limites temporelles qu'impose la prononciation d'une syllabe (en admettant que cette dernière constitue une unité de programmation des variations de F0, comme cela a été observé par Di Cristo, 1978). On peut s'attendre ainsi à ce qu'une importante variation de Fo, activée par une *instruction linguistique* particulière, ne puisse atteindre la *cible acoustique* planifiée à la demande de cette instruction, si la variation est associée à une syllabe particulièrement brève (intrinsèquement brève ou abrégée par suite d'une accélération du débit de parole). L'effet de troncation des variations de F0, auquel nous venons de faire allusion, est attesté par *l'analyse expérimentale*. Le recours à des expériences de perception permet également de constater que le système cognitif «*connaît*» cette contrainte et peut la surmonter en reconstituant subjectivement la partie tronquée et, du même coup, la cible planifiée par le locuteur. À la lumière de ces remarques, la prise en considération des contraintes de production qui affectent les trajectoires de F0 et leurs alignements avec le matériau segmental paraît donc s'imposer comme un préalable à l'interprétation des variations de ce paramètre prosodique, en vue d'établir des liens entre ces variations et les représentations phonologiques qu'elles actualisent.

3.4.4. Les phénomènes dits «d'abaissement» *(actualisation de plusieurs niveaux de contraintes)*

Outre l'influence des contraintes propres à la gestion des variations de F0 dans les dimensions fréquentielle et temporelle, il convient de mentionner les effets qui résultent de l'incidence d'un autre paramètre de production. Il s'agit en l'occurrence des variations de la *pression sous-glottique*. En raison de la baisse régulière du volume d'air pulmonaire dans le cours de l'énoncé, la pression sous-glottique (PS) tend à diminuer aussi de façon graduelle. Etant donné la relation physique qui lie les paramètres F0 et PS, il est attendu que F0 décroisse pareillement de façon progressive (toutefois, cette décroissance graduelle n'est pas due uniquement à l'influence du paramètre PS, mais aussi à des ajustements

laryngés comme le déplacement vertical du larynx: Honda, 2004). Cet effet d'abaissement graduel est couramment appelé: *déclinaison.* Comme il affecte à la fois les minima et les maxima de la courbe de F0 (que l'on peut relier par deux lignes distinctes), il est d'usage de distinguer entre la ligne de déclinaison basse (*Baseline*) et la ligne de déclinaison haute

3.4.5. Les contraintes interactives: effets microprosodiques

Les variations de F0, en particulier, et celles des paramètres prosodiques physiques, en général, sont soumises à des contraintes de production «*interactives*», qui résultent dans ce cas d'une interaction entre ces paramètres et la prononciation du matériau segmental. Les phénomènes engendrés par ce type de contrainte sont appelés *microprosodiques*, car leur empan est très limité et n'excède pas la taille du segment phonémique. Il existe deux sortes de phénomènes microprosodiques, qui sont qualifiés par les termes: *intrinsèques* et *co-intrinsèques*, respectivement.

3.5. Méthodes de détection de la fréquence fondamentale

D'après les travaux de Hess [18], les algorithmes de détection de pitch sont classées en trois groupes principaux : temporelles, spectrales et Hybrides.

Les méthodes temporelles permettent l'estimation de F0 avec des calcules très simples. Elles sont relativement peu couteuses en temps de calcul car elles nécessitent peu d'opérations arithmétiques de multiplications et d'additions [19]. Toute fois, elles manquent de précision. De variétés de techniques temporelles sont décrites dans la littérature. Parmi les techniques de base on peut citer : la Fonction d'Autocorrélation (FAC) et ses versions modifiées [19], la Fonction de différence d'AMDF (Average Magnitude Difference Function) et ses variantes [20], la Fonction de réduction de donnée, DARD (DAta ReDuction method) [21] et la Fonction du calcul parallèle, PPROC (Parallel PRoCessing method) [22].

Les méthodes spectrales sont définies comme étant celle qui permet d'obtenir une F0 en traitant le spectre de la parole directement. Parmi ces techniques, on peut citer : la technique Cepstrale (CEP) [23], le Produit Harmonique Spectral (HPS), et l'inter corrélation avec le Peigne Spectrale (PS) [24].

Les méthodes hybrides, visent à combiner différentes approches pour augmenter les performances globales du système d'extraction. Elles appliquent différents analyseurs simultanément sur le signal et combinent les différents estimateurs [19].

3.6. Description des techniques

Dans la plupart des algorithmes d'extraction de F0, trois phases essentielles durant le traitement s'implique : le prétraitement, le traitement et le pos traitement.

La phase de prétraitement est réservée à la préparation du signal issue d'un microphone. Elle consiste à choisir la durée des trames d'analyse et du recouvrement afin de moins compromettre la condition de stationnarité exigée par les algorithmes de traitement et l'effet de bord lié aux fenêtres de pondération appliquées.

La durée de la trame est généralement choisie entre 20 et 50ms avec un recouvrement de 30 à 50%, pour assurer la présence d'au moins une période du Fondamental [25]. Nous trouvons souvent d'autres techniques permettant d'améliorer la rapidité d'extraction tel que le filtrage, la décimation et les techniques de transformation non linéaire dites Clippage. La phase de traitement est réservée à l'extraction de F0 et dépend donc de l'algorithme utilisé.

La phase de post traitement à pour but de diminuer les erreurs qui peuvent être de plusieurs types. Ces erreurs vont être détaillées au cinquième paragraphe. On présente dans ce paragraphe les techniques choisies pour une éventuelle évaluation des performances.

3.6.1. La fonction d'AMDF basée sur le Clippage

Plusieurs versions d'AMDF existent pour la détection de F0 [21]. On a limité notre étude sur la C-AMDF (Clipping -Average Magnitude Diffrence Function). En premier lieu, le signal vocal est filtré par un filtre passe bas de type Butterworth à une Fréquence de coupure Fc de 900 kHz. Ensuite, segmenté en trames de 30 ms.

L'opération de clippage consiste à appliquer sur les fenêtres à court termes résultantes une transformation non linéaire définit par le clippage central donné par l'équation (3.5). Un seuil de clippage CL doit être calculé pour chaque trame. Dans notre application, le seuil de clippage est choisi égal à 30% de l'amplitude du pic maximal de la trame en cour de traitement. L'AMDF est calculée sur le signal clippé pour chaque trame d'analyse. La valeur de F0 est déterminée avec la localisation de la vallée minimale entre 70 Hz et 600 Hz. En fin, la décision du voisement-silence s'effectue avec le calcul du Taux de Passage par Zéros (TPZ) et de l'énergie (Fig.3.3).

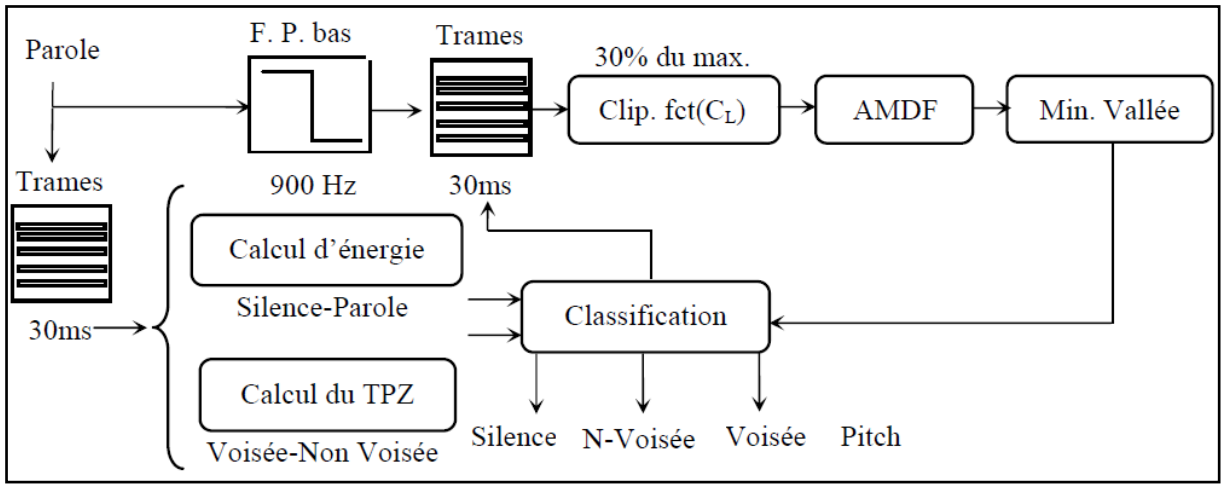


Figure 3.3: Schéma bloc proposé du détecteur de pitch par la C-AMDF

3.6.2. La technique Cepstrale

L'opération « cepstre » est une méthode numérique permettant d'étudier séparément la source et le conduit vocal. Une façon simple de décrire le cepstre est de dire qu'il tend à séparer un composant harmonique fort du reste du cepstre. Il existe deux méthodes de calcul des coefficients cepstraux:

Analyse spectrale.

Analyse paramétrique.

Analyse spectrale

Le signal de parole est modélisé en utilisant le spectre. Il est possible de calculer directement les coefficients cepstraux en suivant le processus décrit à la Figure.3.4 :

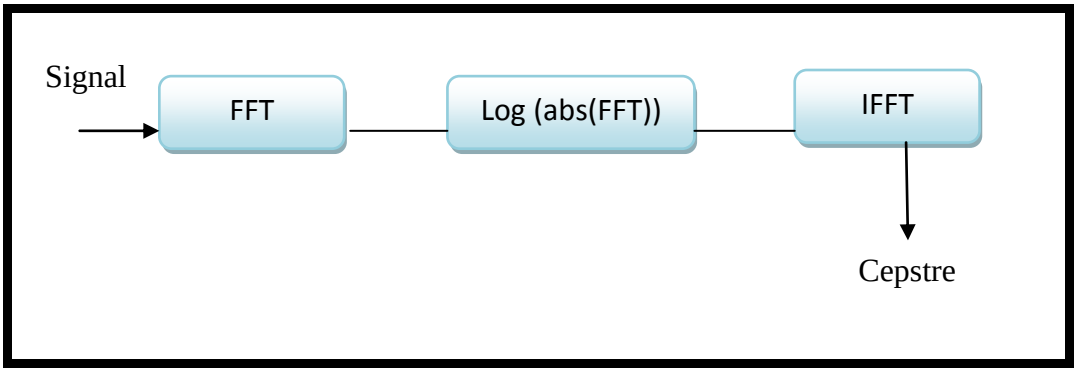


Fig.3.4 : Calcul des coefficients cepstraux par analyse spectrale

Il suffit pour obtenir les coefficients cepstraux de prendre le spectre (c.à.d. transformée de Fourier, puis passer dans le domaine logarithmique et enfin de prendre la fonction inverse de FFT (IFFT).

Le cepstre est basée sur une reconnaissance du mécanisme de la production de la parole. On part de l'hypothèse que la suite constituant le signal vocal est les résultats de la convolution du signal de la source par le filtre correspondant au conduit.

$$s(t) = u(t) * b(t) \tag{3.1}$$

Avec s(t) : signal temporel, u(t) : le signal excitateur, b(t) : la contribution du conduit.

Le but du cepstre est de séparer ces deux contributions par déconvolution. Une transformation de Fourier permet de transformer la convolution en produit :

$$S(f) = U(f).B(f) \tag{3.2}$$

$$Log\ S(f) = LogU(f) + Log\ B(f) \tag{3.3}$$

par transformation inverse, nous obtenons le cepstre, enfin nous aurons une relation dans le domaine temps donnée par :

$$TF^{-1}\big(Log\ S(f)\big) = TF^{-1}(LogU(f)) + TF^{-1}(Log\ B(f)) \tag{3.4}$$

Les résultats du calcul cepstral est une séquence temporelle comme le signal d'entrée lui-même. Si le signal d'entrée possède une période de hauteur fondamental forte, elle apparaît dans le cepstre sous forme de pic, en mesurant la distance entre le temps zéro et le temps pic on trouve la période fondamentale de cette hauteur.

Engendré par un filtre dont il faut trouver les coefficients a(i). Ce modèle de production du signal de parole est appelé AR (autorégressif).

L'estimation de la période de pitch peut être faite sur le Cepstre réel. La Figure 3.5 représente la description du détecteur de pitch par la méthode Cepstrale [16]. Chaque segment de 51.2ms est pondéré par une fenêtre de type Hamming.

Le principe de la procédure de calcul de pitch fondé sur le Cepstre est plutôt simple. On recherche dans le Cepstre un pic dans la région autour de la période du pitch (P). Si le

pic est supérieur à un seuil fixé (P_0), le segment de parole en entrée est probablement Voisé, et la position autour du pic est la zone dans laquelle on peut estimer le pitch. Si le pic n'est pas supérieur au seuil, il est alors probable que le segment de parole en entrée est non Voisé.

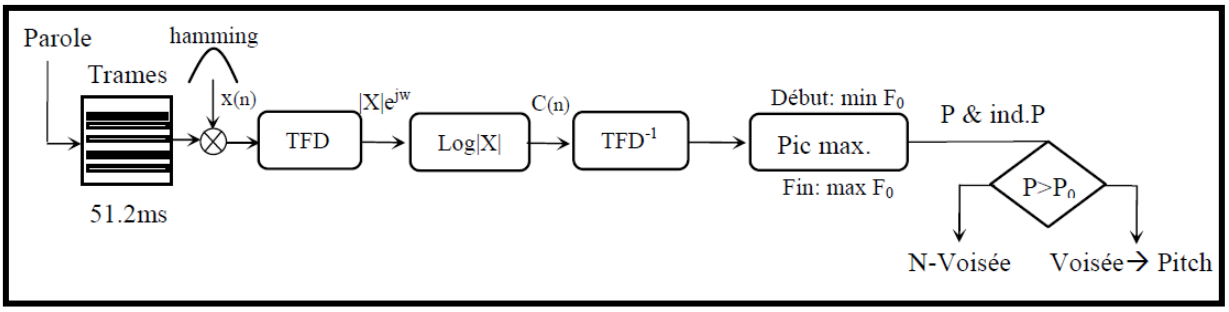


Fig. 3.5: Schéma bloc proposé du détecteur de pitch par la méthode Cepstrale

Analyse Cepstral fournit un moyen pour l'estimation du pitch. Si nous supposons que la séquence de parole voisée est le résultat de la convolution de la séquence d'excitation glottique $e[n]$ avec la réponse impulsionnelle discret du conduit vocal $\theta[n]$ Dans le domaine fréquentiel, la relation de convolution devient une multiplication. Puis, en utilisant la propriété de la fonction log, la multiplication peut être transformée en une addition. $log\ AB\ =\ log\ A\ +\ log\ B$

Enfin, la vraie Cepstre d'un signal $s[n] = e[n] * \theta[n]$ est défini comme

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|S(w)|e^{jwn}\ dw \qquad (3.5)$$

D'où

$$S(w) = \sum_{n=-inf}^{inf} s[n]e^{-jwn} \qquad (3.6)$$

Autrement dit, le Cepstre est une analyse de Fourier du spectre d'amplitude logarithmique d'un signal. Si le spectre en amplitude contient de nombreuses harmoniques régulièrement espacées, donc l'analyse de Fourier du spectre présenté un pic correspondant à l'espacement entre les harmoniques ; c'est à dire la fréquence fondamentale. En effet,

nous traitons le spectre du signal comme un autre signal, puis recherche de périodicité dans le spectre lui-même.

3.6.3. La fonction d'autocorrélation basée sur le clippage central

Elle a été à l'origine proposée par L. Rabiner [16]. L'appellation Anglophone de cette technique est dite MACC (Modified Autocorrelation with Center Clipping) (Fig.3.6).

Le processus commence avec un filtre passe-bas, dont le but est d'atténuer l'influence des fréquences autres que F0. Le filtre coupe à 900 Hz, du fait qu'une valeur de pitch est comprise entre 70 et 600 Hz. La deuxième phase de traitement est la segmentation du signal vocal à des trames de 30 ms pour assurer la stationnarité du signal. La troisième phase est le calcul du seuil de clippage (CL) pour chaque trame d'analyse par la recherche les deux pics maximums dans la première (P1) et la troisième (P3) portion de 10 ms et de prendre le minimum de ces deux valeurs. Ce minimum est multiplié par la suite avec un niveau de clippage k. C'est un paramètre très important qu'il faut l'optimiser avec soin. On prend en général des valeurs variant entre 30% et 80% de l'amplitude de l'échantillon maximal de la trame. La fonction de clippage qui est implémentée ici, est le clippage central avec compression [17] :

$$y(n) = clc[x(n)] = \begin{cases} x(n) - C_L & x(n) \geqslant C_L \\ 0 & |x(n)| < C_L \\ x(n) + C_L & x(n) \leqslant C_L \end{cases} \qquad (3.5)$$

La quatrième phase de traitement est le calcul de la FAC normalisée et la recherche du pic maximum (P) et son indice (ind. P) dans la gamme d'existante de F0 qui nous permettra par la suite de calculer la valeur du Fondamental.

La dernière phase de cet algorithme consiste à choisir un seuil de décision du voisement (V0) en fonction du pic calculé. Si le pic maximum de chaque trame obtenue lors de la phase précédente dépasse le seuil de voisement, la trame est classifiée comme Voisée, sinon elle est classifies Non-Voisée. Dans le cas du silence, la détection se fait grâce à l'énergie à courte terme suivant un seuil bien défini. Si la valeur de l'énergie dans chaque trame ne dépasse pas ce seuil, la trame est considérée comme silence [16].

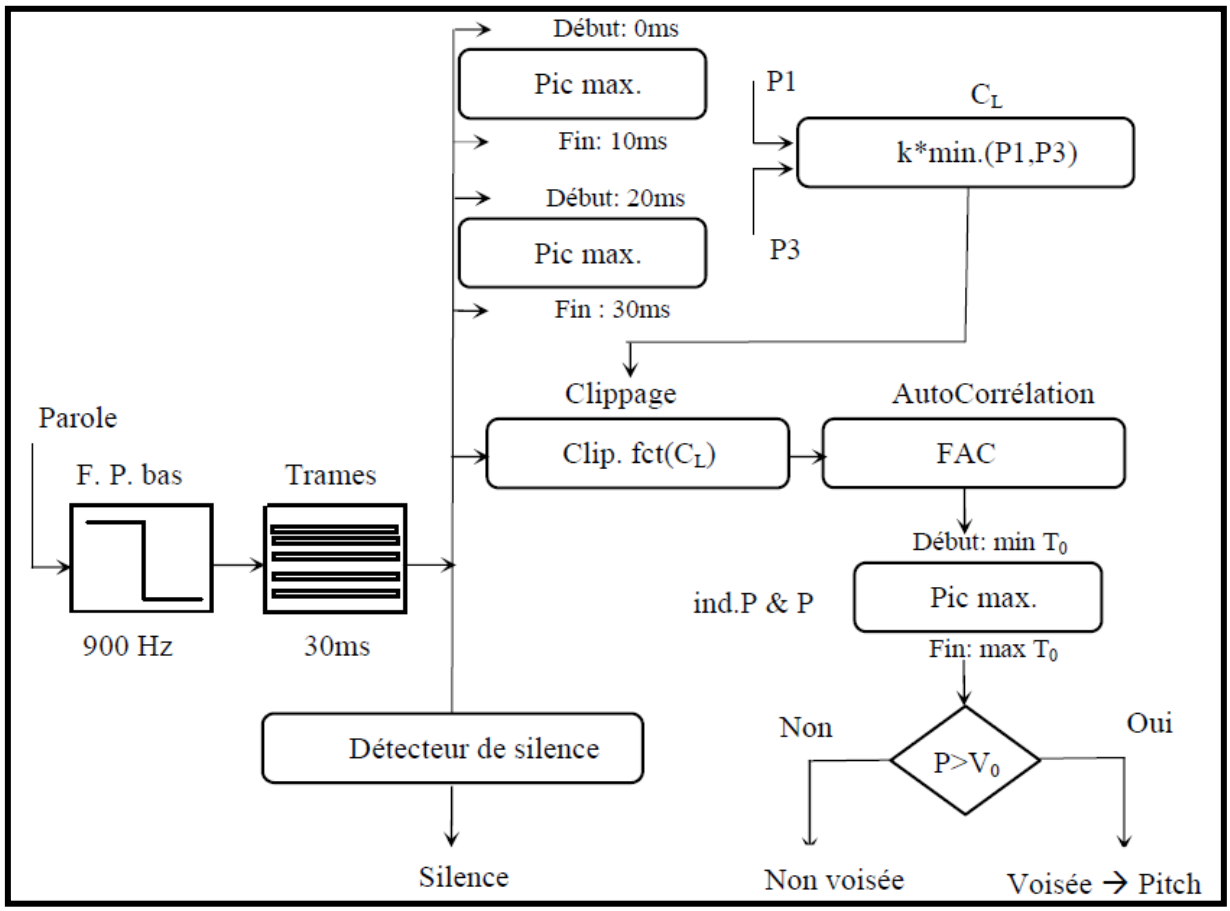


Figure 3.6: Schéma bloc du détecteur de pitch par la MACC

L'autocorrélation statistique d'un processus sinusoïdale aléatoire

$$x[n] = cos\,(w_0 n + \emptyset) \tag{3.6}$$

Donnée par

$$R[m] = E\{x*[n]\,x[n+m]\} = \frac{1}{2}\,cos(w_0 m) \tag{3.7}$$

Qui présente un max pour $m = lT_0$ la période de pitch et ses harmoniques, afin que nous puissions trouver la période de pitch en calculant la valeur la plus élevée de l'auto-corrélation. De même, il peut être démontré que tout processus périodique WSS avec période T_0 a également une auto-corrélation qui présente ses maxima à $m = lT_0$.

Dans la pratique, nous avons besoin d'obtenir une estimation $\hat{R}\,[m]$ seulement de N échantillons. La fonction d'auto-corrélation empirique est donnée par

$$\hat{R}\,[m] \;=\; \frac{1}{N}\sum_{n=0}^{N-1-|m|}(w[n]x[n]w[n\;+\;|m|]x[n\;+|m|]) \qquad (3.8)$$

Ou $w[n]$ est une fenêtre de longueur N, Pour que le processus aléatoire dans l'équation. 1 conduit à une valeur prévue de

$$E\{\hat{R}[m]\} \;=\; (1\;-\frac{|m|}{N})\;\frac{cos(w_0 m)}{2},\; |m|\;<\;N \qquad (3.9)$$

Dont maximale coïncide avec la période de pas pour $m > m0$

Puisque périodes de pitch peuvent être aussi bas que 40 Hz (pour une voix masculine très grave) ou aussi haut que 600 Hz (pour une femme très aigu ou voix de l'enfant), la recherche du maximum est menée dans une région.

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 on Fig.3.7

Acquisition du signal y(n)=y (n.Te)

Pré-accentuation éventuelle par passage du signal dans un filtre de transmission: (1-μ.z-1) ; μ=0.95 cette opération vise à accentuer la partie haute fréquence.

3 Segmentation en tranche dont la durée varie de 27 à 32ms.

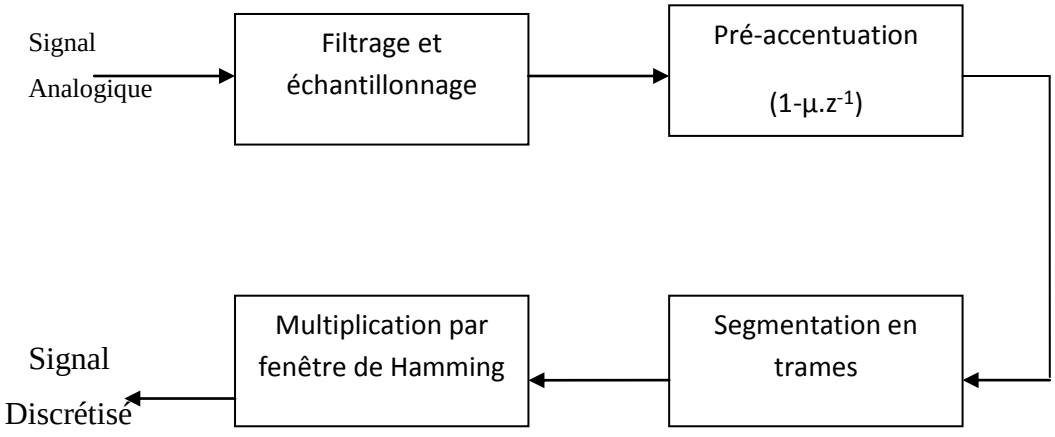


Figure 3.7: Mise en forme du signal vocal.

Application de la pondération consiste à pondérer la tranche par une fenêtre Hamming dont l'expression est la suivante:

$$w(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N}\right) \quad (3.10)$$

Calcul de la fonction d'autocorrélation par la formule suivante

$$r(k) = \sum_{n=1}^{n-1+k} x(n).x(n+k) \quad (3.11)$$

L'autocorrélation travaille directement sur le signal temporel, elle est largement insensible à la variation de phase.

3.7. Conclusion

Ce chapitre nous a permis d'analyser quelque méthode pour la détection de la F0

Chapitre 4 :
Expériences et résultats obtenus

4.1. Introduction

Dans ce chapitre, nous présentons une caractérisation de la voix, les étapes suivies dans l'élaboration du programme de la détection de F_0. Par la suite nous décrivons l'outil utilisé ainsi que les résultats obtenus.

4.2. Caractérisation de la voix

Lors d'une communication téléphonique, il est aisé de reconnaître notre interlocuteur par la seule écoute de sa voix. L'indice qui nous permet cette identification est ce que l'on appelle le timbre de la voix. Ce terme est assez difficile à définir objectivement et la définition donnée par l'American Standards Association emploie même le terme de sensation : « *... that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar* ». Le timbre reposerait donc sur un jugement subjectif de l'auditeur. Ceci montre une influence de l'auditeur sur les critères et la stratégie mise en place pour discriminer deux voix par rapport à leur timbre. Un autre point serait que le timbre est indépendant de la force sonore et du pitch.

Dans la suite, nous allons présenter quelques enjeux et applications issus de la problématique de caractérisation de la voix, puis nous présenterons quelques résultats qui reflètent la difficulté de caractériser l'identité d'une voix. Nous terminerons ce paragraphe par un bilan des principaux paramètres acoustiques utilisés dans ce domaine.

4.2.1. Enjeux et applications

La parole fait intervenir les trois niveaux linguistique, paralinguistique et extralinguistique. Le niveau paralinguistique regroupe les facteurs qui caractérisent le locuteur à un niveau comportemental, par exemple ses émotions. Le niveau extralinguistique est caractérisé par les propriétés physiologiques et physiques de l'appareil phonatoire d'un locuteur. Nous distinguons la dimension socio/psychologique de la dimension physiologique. La première est à rapprocher du niveau paralinguistique et comprend des facteurs tels que l'âge le statut social, le dialecte et la communauté d'appartenance.

On peut noter que les facteurs impliqués sur les plans paralinguistique et extralinguistique peuvent varier non seulement entre deux locuteurs mais également pour un même locuteur. La problématique est alors de déterminer quels paramètres permettent d'expliquer la variabilité interlocuteur de la voix, tout en étant insensibles à la variabilité

intra-locuteur. Ainsi, le lien entre les facteurs qui influencent la perception de la voix et les paramètres acoustiques doit être déterminé. Il définit également les caractéristiques idéales de ces paramètres :

- La représentation de l'information dépendante du locuteur doit être efficace.
- L'acquisition du paramètre doit être facile.
- Le paramètre doit se montrer stable à travers le temps.
- Il doit apparaître naturellement et fréquemment dans la parole.
- Un imitateur ne doit pas pouvoir le reproduire.

De nombreux travaux de recherche existent sur la caractérisation de l'identité d'une voix. Les applications de ses travaux se situent au niveau de la biométrie ou encore de la synthèse de la parole. Dans le premier cas, il s'agit d'identifier une personne par les caractéristiques de sa voix. Les conséquences juridiques de cette application ne sont pas anodines et il convient d'être prudent. Pour ce qui concerne la synthèse de parole, connaître les caractéristiques qui font l'identité d'une voix permettrait d'améliorer le naturel des voix de synthèse mais aussi de diversifier les voix de synthèse en modifiant les paramètres acoustiques importants.

4.2.2. Études perceptives

Pour identifier les caractéristiques propres à la voix d'un locuteur, on peut s'intéresser à la manière dont les auditeurs distinguent les voix les unes des autres. Ainsi, montrent que le fait d'être familier avec un locuteur permet de l'identifier de manière assez fiable. Dans le cas contraire, une accoutumance est nécessaire pour atteindre un score plus élevé que le hasard. Ils montrent également que si un locuteur déguise sa voix, cela provoque une certaine confusion pour tous les auditeurs, même pour ceux qui sont familiers de la voix du locuteur en question.

Concernant les facteurs qui permettent d'identifier une voix, de nombreuses études dont les résultats sont souvent contradictoires existent. Certaines montrent l'importance de paramètres de nature suprasegmentale (débit et mélodie), d'autres présentent le F_0 moyen, la pente spectrale de l'onde de glotte et les trois premiers formants comme des facteurs prépondérants.

4.2.3. Principaux paramètres acoustiques

Les principaux paramètres acoustiques utiles à la caractérisation de la voix sont:

Intensité : il s'agit de l'un des paramètres les plus faciles à obtenir. Pour des signaux non stationnaires comme la parole, l'intensité est définie en fonction du temps par l'equation (4.1)

$$E(t) = \int_{t-T/2}^{t+T/2} s^2(\tau) d\tau \qquad (4.1)$$

T est choisi arbitrairement. Sa valeur est habituellement comprise entre 10 et 30ms. Les variations de l'intensité de la parole sont causées par la variation de la pression sub-glottique et la forme du conduit vocal. Elle est liée à des caractéristiques dépendantes du locuteur.

pitch : La réalisation acoustique du pitch est la fréquence fondamentale, F_0. Les variations temporelles du pitch représentent une caractéristique importante de la parole et des travaux ont montré leurs importances pour la reconnaissance automatique du locuteur.

spectre à court-terme : Il s'agit d'une représentation à trois dimensions de la structure temps-fréquence du signal de parole. Le spectre à court-terme offre une description complète des caractéristiques acoustiques du signal. Cette information semble efficace pour la reconnaissance automatique du locuteur même si elle n'est pas très compacte.

coefficients de prédiction : La prédiction linéaire est une méthode efficace pour représenter les propriétés spectrales du signal de parole. Avec cette méthode, un échantillon est vu comme une combinaison linéaire des p échantillons passés.

fréquence et bande passante des formants : Les fréquences centrales des formants sont définies comme les fréquences de résonance du conduit vocal et sont dépendantes du locuteur. La principale difficulté consiste à estimer de manière efficace les formants à partir du spectre à court-terme.

coarticulation nasale : En parole continue, la forme du conduit vocal à un instant t dépend non seulement du phonème courant mais également des phonèmes voisins. Le phénomène de coarticulation résulte de l'influence du contexte phonémique sur le mouvement des articulateurs. Des travaux montrent que la coarticulation pendant la production de nasales permet de différencier les locuteurs.

corrélation spectrale : Un degré significatif de corrélation existe entre les spectres à court-terme différentes fréquences. Ces corrélations sont obtenues par des moyennes de spectres sur le long terme et elles varient de manière consistante d'un locuteur à un autre.

débit d'élocution et événements temporels : La durée de certains événements dans la parole est différente d'un locuteur à un autre.

4.3. Matériels utilisé

Nous avons procédé à l'enregistrement dans un studio de l'ISMAS à Bordj-El-Kiffan - Alger. Celui-ci est constitué de deux cabines : cabine speaker et cabine technique (figure 4.1).

Les enregistrements ont été effectués sous le format WAV, avec une fréquence d'échantillonnage de 48 kHz.

Nous avons utilisé comme matériel hardware un microphone électrodynamique de marque Beyer-dynamic M69TG (figure 4.2) et software, la station ProTools LE version 8 (figure 4.3), qui est une station audionumérique (en Anglais : DAW, pour **D**igital **A**udio **W**orkstation). Pro Tools est utilisée par une grande partie de l'industrie de la production sonore. On la trouve dans des domaines aussi variés que l'enregistrement et le mixage musical, le post production audio film et télévision, le montage son, la création et l'illustration sonore, la création et la composition musicale, etc.

Figure 4.1 : Studio d'enregistrement

Fig. 4.2 : Microphone électrodynamique

Fig. 4.3 : Station ProTools

4.4. Outils utilises

Le SIAL est construit par deux outils : MATLAB et PRAAT.

- MATLAB: version 8.1.0.604 (R2013b);
- PRAAT : version 5.3.49 est un logiciel libre scientifique gratuit conçu pour la manipulation, le traitement et la synthèse de sons vocaux (phonétique). Il a été conçu à l'institut de sciences phonétiques de l'université d'Amsterdam par Paul Boersma et David Weenink[26].

4.5. Description de l'application

Concernant l'outil MATLAB, nous avons utilisé des fonctions pour chaque méthode répartir en des fichiers.m

Cepstral_pitch.m / Ac_pitch.m : ils contiennent la fonction principale du programme

- charge les fichiers sonores des locuteurs en utilisant la fonction wavread ;
- paramétrisé le signal parole etfaire le calcul : Cepstrale / Autocorrélation.

Callbacks_cepstral_pitch.m / Callbacks_ac_pitch.m : ils font appel à des fonctions pour afficher l'interface graphique les différents boutons et texte.

Notre application consiste a enregistré la voix des locuteurs sur la base de données en suites faire un deuxième enregistrement indépendant du texte d'un des locuteurs pour teste la reconnaissance.

4.6. Description de l'interface

Pour mieux présenter les résultats obtenus, nous avons réalisé des interfaces graphiques DF_0Cep et DF_0AC (Figures 4.4 et 4.5).

- une barre de titre ;
- des boutons pour charger et lire le fichier audio, d'autres boutons pour choisir les paramètres et la méthode de détection et enfin un bouton pour fermer et quitter l'interface.

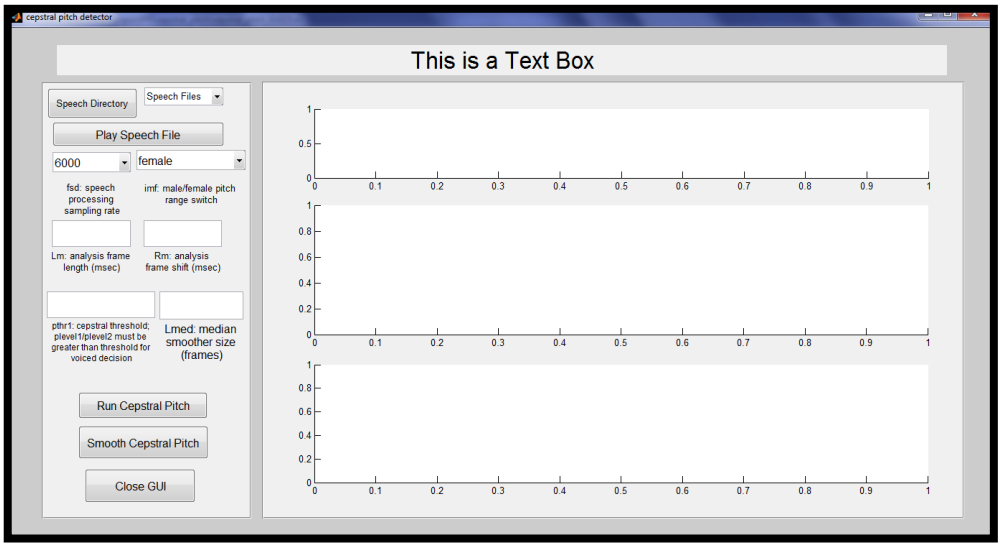


Figure 4.4 : Interface graphique du DF_0Cep

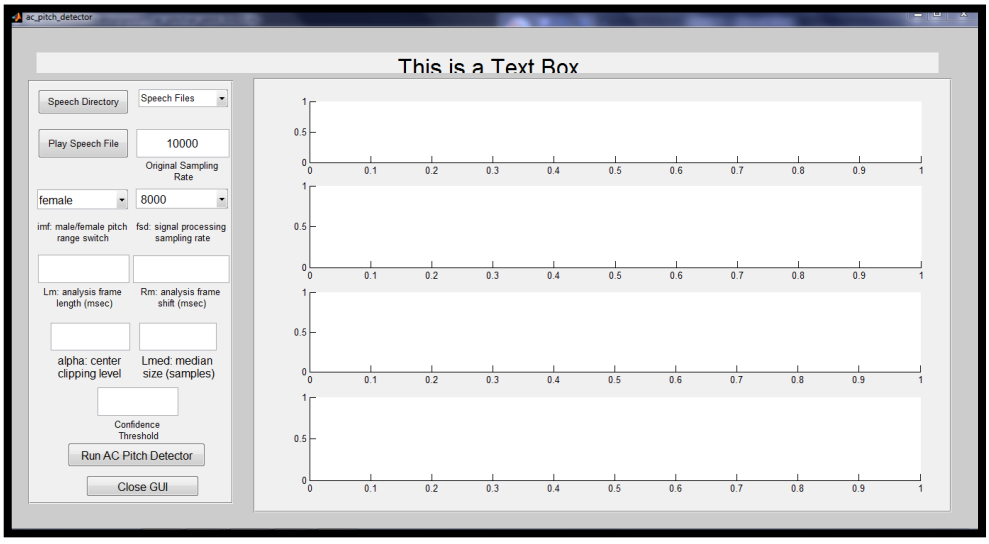


Figure 4.4 : Interface graphique du DF_0AC

Ces programmes mettent respectivement, en œuvre :

- un algorithme de détection de période de pitch en utilisant la méthode d'autocorrélation modifiée avec un centre spécifié seuil d'écrêtage. Il analyse un énoncé de parole désigné sur une base trame par trame ;
- un détecteur de période de pitch sur la base de la détection et le suivi des pics dans des régions de cepstre de parole voisée. La caractéristique principale de ce détecteur de période de pitch est l'utilisation d'un détecteur de crête cepstrale secondaire, pour chaque trame de parole, afin de détecter les erreurs de détection de la période pitch correctes à cause des effets tels que pitch doublement de période

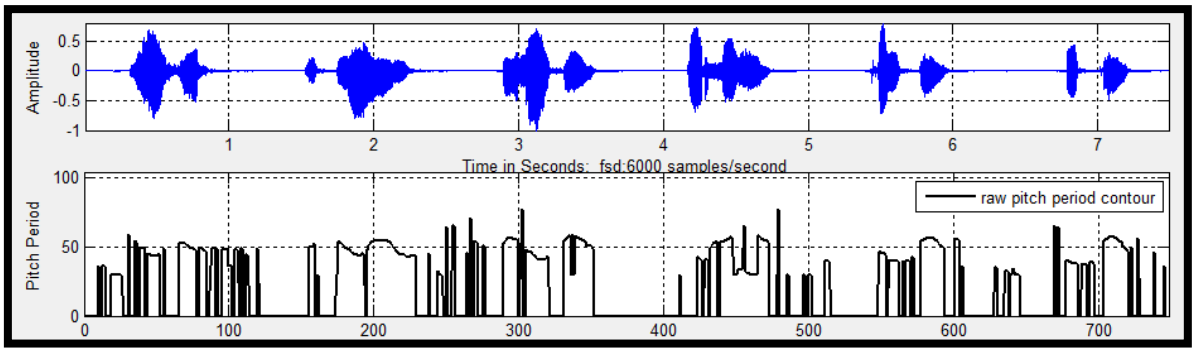


Figure 4.6 : Résultat du DF_0AC

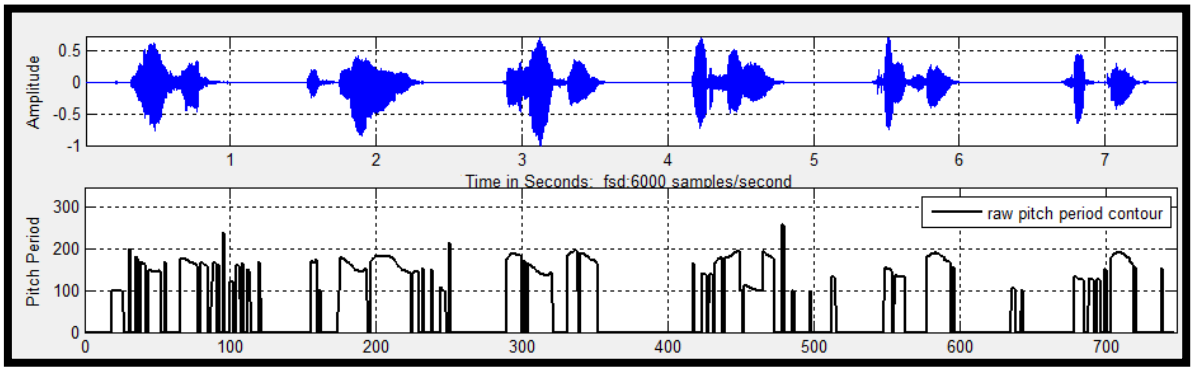


Figure 4.7 : Résultat DF_0Cep Cepstral

4.7. Résultats et discussions

Pour justifier plus l'utilité de la précision de la fréquence fondamentale, nous étudierons la variation de cette dernière comme caractéristique intrinsèque acoustique. Le tableau 4.1 montre les résultats obtenus.

On constate que la méthode d'autocorrélation donne des bons résultats pour la détection du pitch par rapport à la méthode cepstrale. Cette dernière représente des inconvénients sachant qu'elle exige une grande taille mémoire et nécessite un temps de calcul relativement long.

4.8. Evolution des performances

L'identification qui est un procédé permettant de déterminer l'identité d'une personne, ne comprend qu'une étape. Son évaluation se fait donc uniquement par calcul du **T**aux de **F**ausse **A**cceptation (TFA), contrairement à un système de vérification qui nécessite le calcul du **T**aux de **F**aux **R**ejet (TFR).

Fausse Acceptation : Événement ayant lieu lorsqu'un système biométrique accepte une personne alors qu'elle n'est pas dans sa base d'utilisateurs. Cet événement doit être le plus rare possible pour assurer la sécurité d'un système biométrique.

Faux Rejet : Événement ayant lieu lorsqu'un système biométrique refuse une personne alors qu'elle est dans sa base d'utilisateurs. Cet événement est souvent dû à une mauvaise acquisition des données biométriques et est perçu comme un gène par l'utilisateur.

TFA - Taux de fausse acceptation : Indique la probabilité qu'un utilisateur inconnu soit identifié comme étant un utilisateur connu. Ce taux définit la sécurité du système biométrique. Le Taux de fausse acceptation est égale au nombre de fausses acceptations divisé par le nombre de tests imposteur de la base (N_i).

$$TFA(\tau) = \frac{FA(\tau)}{Ni} \tag{4.2}$$

TFR - Taux de faux rejet : Indique la probabilité qu'un utilisateur connu soit rejeté par le système biométrique. Ce taux définit en partie le confort d'utilisation du système biométrique.

$$TFR(\tau) = \frac{FR(\tau)}{Nc} \tag{4.3}$$

Le Taux de faux rejet est égale au nombre de faux rejets divisé par le nombre de tests cible dans la base (N_c)

Le taux d'erreur de décision sont dépendant su seuil de décision fixé dans le module de décision et sont en générale en fonction du seuil.

Tableau 4.1 : Résultats obtenus

	Femmes	Hommes	Taux de reconnaissance
DF_0AC	13	19	96,87%
DF_0Cep	13	19	96,87%

Lors des essaies effectué, nous avons obtenu un taux de reconnaissance de 97%.

4.9. Conclusion

Dans ce chapitre nous avons présenté les expériences et les résultats obtenus à partir des deux méthodes : DF_0AC et DF_0Cep, appliquées à notre SRAL.

Conclusions générale et perspectives

L'analyse du signal vocal n'est pas complète tant qu'on n'a pas mesuré l'évolution de la F_0. Ce paramètre est très important pour la Traitement Automatique de la Parole.

L'estimation du pitch est, bien sûr, liée à la localisation des segments voisés. Cette tâche est difficile à cause de nombreuses raisons, telles que la non-stationnarité du signal de parole, l'existence de certaines irrégularités dans l'excitation glottique et l'interaction avec les formants.

Dans notre travail, nous avons traité le problème de la détection de F_0. Il s'agit d'extraire le pitch, à partir des signaux de paroles enregistrés par deux locuteurs (une femme et un homme). Pour l'analyse paramétrique nous avons utilisé la fonction d'autocorrélation et la technique cepstrale.

Notre travail est divisé en deux parties :

- la première est consacrée à l'enregistrement des fichiers sons dans une station audionumérique « ProTools». Une segmentation a été faite à l'aide du logiciel PRAAT ;
- la seconde partie présente les deux algorithmes de la détection de F_0 (DF_0Cep et DF_0AC) ont été réalisés avec MATLAB.

Les résultats obtenus donnent un Taux de Reconnaissance (TR) qui atteint les 97%. Et indiquent que DF_0AC donne de bons résultats par rapport à la DF_0Cep.

Comme perspectives à ce travail nous proposons :

- la mise au point d'une Base de Données sonores, plus riche à enregistrer par plusieurs locuteurs, dans un milieu ambiant ;
- l'utilisation d'autres détecteurs de F_0 comme la MFCC ainsi d'autres paramètres prosodiques comme la durée et l'énergie.

Références bibliographiques

[1] B. Matmberg, La phonétique, Que -sais-je ? N° 637, Presses Universitaires de France, Paris 1984.

[2] E. Garde, La voix, Que sais-je ? N°627, Presses Universitaires de France, Paris 1970.

[3] Le dictionnaire des médecines naturelles Editions Marabout, Verviers 1980.

[4] M. Kob, Physiologie des lèvres et des cordes vocales. Journée d'étude « Lèvres vibrantes et cordes vocales », ENST, Paris, France, Juillet 2004.

[5] J. Vaissière, Phonétique et Phonologie. Cours Deug 2. Second semestre, Laboratoire de

Phonétique et Phonologie, Paris III, France, 2002.

[6] Dernier consultaion : 23/06/2014 http://alis.isoc.org/langues/api.fr.htm

[7] P. Martin, Éléments de phonétique avec application au Français, Canada: Presses de l'Université Laval. 1996

[8] Dernier consultaion : 23/06/2014

http://faculty.washington.edu/dillon/PhonResources/vowels.html

[9] J.R. Firth, Papers in linguistics. London: Oxford University Press, UK, 1951.

[10] A. Di Cristo, Interpréter la prosodie. Actes des XXIIIèmes Journées d'Etude sur la Parole, Aussois, France, 2000.

[11] Calliope, La parole et son traitement automatique, CNET-ENST, Masson, Paris 1989.

[12] J.M. Blanc, Traitement de la Prosodie par un Réseau Récurrent Temporel. Thèse de

Doctorat en Sciences Cognitives, Université Lumière Lyon II, France, 2004.

[13] P.A. Barbosa, Caractérisation et génération automatique de la structuration rythmique du Français. Thèse de Doctorat en Signal, Image et Parole. Institut National Polytechnique de Grenoble (INPG), France, 1994.

[14] J.L. Rouas, Caractérisation et identification automatique des langues. Thèse de Doctorat en Informatique, Université Toulouse III – Paul Sabatier, France, 2005.

[15] W. Hess, Pitch Determination of Speech Signals - Algorithms and Devices. Springer Verlag, 1983.

[16] L.R Rabiner, A Comparative Performance Study of Several Pitch Detection Algorithms. IEEE Trans.Acoust., Speech, And Signal Processing, Vol. ASSP-24, N°.5, October 1976.

[17] F.Bimbot, I. Magrin-Chagnolleau, L. Mathan, Second-order statistical measures for text independent speaker identification. Speech Communication, Volume 17, Number 1, pp. 177-192(16), Elsevier, 1995.

[18] P. Bagshaw, S. Hiller, et M. Jack, 1993. Enhanced pitch tracking and the processing

[19] J. Dubnowski, R. W. Schafer, L. R. Rabiner, Real-Time Digital Hardware Pitch Detector. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 2-8, 1976.

[20] Yu-Min Zeng at al., Modified AMDF Pitch Detection Algorithm. Proceedings of the Second International Conference on Machine Learning and Cybernetics Wan, 2-5, 2003.

[21] N. J. Miller, Pitch Detection by Data Reduction. IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-23, pp. 72-79, 1975.

[22] A. E. Rosenberg, M. R. Sambur, New Techniques for Automatic Speaker Verification. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, 1975.

[23] R. W. Schafer, L. R. Rabiner, System for Automatic Formant Analysis of Voiced Speech. J. Acoust. Soc. Amer., Vol. 47, pp. 634-648, 1970.

[24] F. Flego, Fundamental Frequency Estimation Techniques for Multi Microphone Speech Input. Phd Dissertation, University of Torento, USA, 2006.

[25] R. Boite, Traitement Automatique de la Parole. Edition Masson, France, 1989.

[26] Dernier consultaion : 23/06/2014 http://www.fon.hum.uva.nl/praat/