الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي و البحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

présentée par

M^{elle} ELMAHDI Hadjer

&

M^{elle} KADI Nadjia

pour l'obtention du diplôme de master en Électronique spécialité Réseaux et Télécommunications

Thème

Transformation du signal vocal par la méthode PSOLA

Proposé par : **Promoteur** : **M. YKHLEF Fayçal**

Co-promotrice : **M^{elle}. CHENTIR Amina**

Année Universitaire 2011-2012

Remerciements

*Nous tenons avant tout de remercier le bon **DIEU** qui nous a donné la volonté et le courage pour la réalisation de ce travail.*

*Nous remercions vivement « **M. YKHLEF Fayçal**» notre promoteur pour la précieuse assistance, sa disponibilité et son soutien qu'il nous accordé tout au long de ce projet.*

*Nous remercions également notre Co-promotrice « **M^{elle}. CHENTIR Amina »** de l'Université de SAAD DAHLEB de Blida, pour ses compétences, son ouverture d'esprit et sa grande disponibilité.*

Nous remercions la direction du Centre de Développement des Technologies Avancées, C.D.T.A et en particulier les dirigeants de la Division Architecture des Systèmes et Multimédias de nous avoir accueilli et d'avoir mis à notre disposition les conditions favorables pour la réalisation de ce présent travail.

Nous remercions aussi tous les membres de jury, qui ont accepté d'examiner et de juger notre modeste travail.

Sans omettre bien sûr de remercier profondément tous ceux qui ont contribué de près ou de loin à la réalisation du présent travail et surtout nos meilleurs amis, particulièrement : Abeladhim, Soumia, Assia, Salim, Rym, Amina, Hadjer et Manal.

Dédicace

Je dédie se modeste travail à la mémoire de mon père qui a été tout pour moi.

A ma très chère mère, qu'il me reste dans cette vie.

A mes sœurs: Meriem, Mouni, Ibtissem et Hadjer.

A mes Frères: Khaled, Abdelrahmane et Abdou.

Hadjer

Dédicace

Je dédie ce travail à mon cher père qui, par ses précieux conseils et son soutien a su me guider vers le droit chemin et vers la voie de la réussite.

À ma très chère mère qui a sacrifié son noble existence pour bâtir la mienne, et qui est pour moi le symbole du courage et du sacrifice.

À mes très chères sœurs Khaoula, Sarah et Khadidja.

À mon cher frère Abdellah.

À tous ceux que j'aime.

Nadjia

ملخص:

الهدف الرئيسي من هذه المذكرة هو عمل دراسة نظرية و تطبيقية حول تقنية TD-PSOLA من أجل تغيير التردد الرئيسي للصوت. وذلك بالاعتماد على الإشارة الوسطية لتعليم التردد الرئيسي.

الأصوات المستعملة هي عبارة عن جمل أخذت من قاعدة البيانات CMU ARCTIC لمتحدثين أمريكيين.

التقييمات الحسابية و السمعية تظهر صحة وجودة نظام TD-PSOLA لتغيير الصوت.

كلمات المفاتيح: إشارة صوتية, التقسيم, التعليم (وضع علامات), TD-PSOLA .

Résumé :

L'objectif principal de ce mémoire est l'étude et simulation de la méthode TD-PSOLA « *Time Domain Pitch Synchronous Overlap-Add*» pour la modification de la fréquence fondamentale de la voix. Les sons utilisés dans notre travail sont des phrases extraites de la base de données CMU ARCTIC *« Carnegie Mellon University ARCTIC »* prononcées en Anglais Américain. Le marquage du fondamental est basé sur le signal moyenneur.

Les évaluations subjectives et objectives montrent que la transformation de la voix par la TD- PSOLA est correctement effectuée.

Mots clés : Signal vocal, Classification, TD- PSOLA, Marquage.

Abstract :

The main objective of this memory is the study and simulation of TD-PSOLA method « *Time Domain Pitch Synchronous Overlap-Add*» for pitch shifting of speech. The sounds used in our work are sentences extracted from the CMU ARCTIC database *« Carnegie Mellon University ARCTIC »*. They are pronounced in American English. Pitch marking uses the mean based signal. Subjective and objective evaluations show that voice transformation by using TD-PSOLA is correctly carried out.

Keywords: Speech, Classification, TD-PSOLA, Pitch marking.

Listes des acronymes et abréviations

AA : Anglais Américain

API : Alphabet Phonétique International

BD : Base de Données

CMU ARCTIC : Carnegie Mellon University ARCTIC

dB : Décibal

DCT : Discreet Cosine Transform

DRT : Diagnostic Rhyme Test

DYPSA : Dynamic Programming Projected Phase-Slope Algorithm

E : Energie

F_0 : Fréquence fondamentale

$\mathbf{F_0'}$: Fréquence fondamentale de synthèse

FA : False Alarm

FBSOLA : Filter Bank Synchronous Overlap and Addition

FD-PSOLA : Frequency-Domain Pitch Synchronous Overlap-Add

Fe : Fréquence d'échantillonnage

F_i : Formant

FT : Fonction de Transfert

GCIs : Glottal Closure Instants

IDCT : Inverse Discreet Cosine Transform

IR : Identification Rate

LP : Linear Predection

LPC : Linear Predictive Coding

LSEE-MTFTM : Least-Squares Error Estimation from the Modified Short-Time Fourier

MBS : Mean Based Signal

MOS : Mean Opinion Score

MR : Miss Rate

MRT: Modified Rhyme Test

PAOLA : Peak-Alignment Overlap and Add

PSOLA : Pitch Synchronous Overlap-Add

QMF : Quadrature Mirror Filter

SOLAFS : Synchronized Overlap-Add

T_0 : Période fondamentale

TDHS : Time Domain Harmonic Scaling

TD-PSOLA : Time Domain Pitch Synchronous Overlap-Add

Te : Période d'échantillonnage

TF : Transformée de Fourier

TFCT : Transformée de Fourier à Court Terme

TFCTI : Transformée de Fourier à Court Terme Inverse

ξ : Identification Error

σ : Identification Accuracy

Table des matières

Chapitre 3

Etude théorique de la méthode PSOLA

Chapitre 4

Résultats et simulations

Liste des figures

Liste des tableaux

Introduction générale

La parole, principal vecteur d'information dans la société humaine, correspond à une variation de la pression de l'air causée par le système articulatoire.

Sa particularité tient du rôle que joue le cerveau dans sa production et sa compréhension par l'emploi automatique de diverses fonctions.

Le traitement de la parole fait l'objet de recherches dans les laboratoires des grands opérateurs de télécommunications depuis leurs premières années d'existence, et dont les travaux se sont intensifiés avec l'apparition du traitement automatique du signal.

Ce vaste domaine est classiquement découpé en quatre grandes catégories.

- Prétraitement.
- Codage.
- Synthèse.
- Reconnaissance de la parole.

La transformation de la voix est une opération qui consiste à modifier les enregistrements audio d'une voix afin d'en changer l'identité perçue. Les modifications sont en général des changements des paramètres prosodiques (fréquence fondamentale, la durée et l'énergie) et l'enveloppe spectrale. Cette technologie est utilisée dans plusieurs domaines qui incluent l'analyse, la synthèse et le prétraitement de la voix.

Plusieurs méthodes de traitement existent dans la littérature. La méthode PSOLA « *Pitch Synchronous Overlap-Add* » est l'une des méthodes les plus utilisées dans la littérature.

L'objectif principal de ce projet de fin d'étude est d'étudier et implémenter la méthode TD-PSOLA « *Time-Domain Pitch Synchronous Overlap-Add* » pour la modification de la fréquence fondamentale du signal vocal. Cette dernière inclut en trois étapes fondamentales qui sont (Analyse, Modification et la Synthèse). On va s'intéresser au niveau de notre étude aux catégories : Voisées, non voisées (Pour la classification de son à l'étape d'analyse) et on va exploiter le signal moyenneur dit en anglais « *Mean Based Signal* » comme une solution pratique pour résoudre le problème de marquage qui doit être effectué pour une transformation complète.

Notre mémoire est divisé en quatre chapitres:

Le premier chapitre regroupe des généralités sur le traitement de la parole, en particulier les caractéristiques acoustiques et phonétiques ainsi que des notions sur l'Anglais Américain.

Le deuxième chapitre est consacré à l'étude des différentes techniques de transformation de la voix.

Le troisième chapitre présente l'état de l'art de la méthode PSOLA « *Pitch Synchronous Overlap-Add* », en particulier la TD-PSOLA.

Le quatrième chapitre donne les résultats et les simulations d'implémentation ainsi que l'évaluation de la TD -PSOLA.

Ce présent document sera terminé alors par une conclusion générale.

Chapitre 1 Généralités sur le traitement de la parole

1.1 Introduction

La parole, manifestation sonore du langage, est sans doute le principal moyen de communication entre les humains. L'avènement des télécommunications, puis du traitement numérique de l'information s'est donc naturellement accompagné d'un vaste effort de recherche visant à comprendre les mécanismes de la communication parlée.

L'information portée par le son de la parole peut être analysée de plusieurs façons. On en distingue plusieurs niveaux de description : acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique et pragmatique.

Dans ce premier chapitre, on ne se focalise que sur les niveaux phonétiques et acoustiques. Ceci amènera, dans l'étude acoustique, à définir les paramètres prosodiques, les étapes du prétraitement et les différentes représentations. Quant au niveau phonétique, on aura à comprendre le mécanisme de production de la parole et en tirer les différentes classes phonétiques. On étendra cette étude à l'Anglais Américain.

Dans le dernier point de ce chapitre, on exposera les propriétés spécifiques du signal vocal rendant complexe son analyse. On terminera par une conclusion englobant les différents points étudiés jusque-là.

1.2 Etude acoustique de la parole

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal

en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone. De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques visant à en mettre en évidence ses traits acoustiques [1].

Les sons de la parole peuvent se décomposer en deux principales catégories : les sons voisés et les sons non voisés (on utilise couramment les termes sonores et sourds pour désigner cette opposition).

Un son est dit voisé si sa production s'accompagne d'une vibration des cordes vocales. Ces signaux se caractérisent par leur quasi-périodicité et une grande énergie concentrée en basses-fréquences (essentiellement entre 0 et 2 kHz) et décroissante rapidement en hautes-fréquences (jusqu'à 4 kHz).Les séquences voisées possèdent une fréquence fondamentale (F_0) et des harmoniques (formants, F_i) multiples de celle-ci.

Un son est dit non voisé si sa production ne s'accompagne pas de vibration des cordes vocales. Il est alors produit par le frottement de l'air dans le conduit vocal entre la glotte et les lèvres. L'énergie des sons voisés est plutôt concentrée en hautes-fréquences (au-delà de 4 kHz) [2].

De ces deux grandes catégories, peuvent être distinguées des catégories secondaires des sons de la parole telles que les sons mixtes et les silences.

Un son mixte est par définition une combinaison entre un son voisé et un son non voisé. Il présente une structure périodique (la présence de la F_0 et des formants) à laquelle s'ajoute une structure aléatoire possédant l'allure d'un bruit blanc légèrement corrélé.

Le silence représente des intervalles temporels où le signal utile est absent. En pratique, il s'agit de bruit d'origines diverses. On dit que c'est un son à amplitude faible ou nulle [3].

Dans ce qui suit, on va présenter les propriétés acoustiques principales du signal de parole.

1.2.1 Paramètres prosodiques

Du point de vue acoustique, la prosodie désigne les phénomènes liés à la variation dans le temps des paramètres de hauteur, d'intensité et de durée. C'est-à-

dire chaque son est caractérisé acoustiquement par certain nombre de données, en particulier la vitesse de vibration (fréquence), l'amplitude de la vibration (intensité) et la durée d'émission [4].

a Fréquence fondamentale

La fréquence fondamentale F_0, nommée « pitch », est la fréquence la plus basse dans le signal de parole. Cette fréquence est celle de vibration des cordes vocales. Elle est généralement donnée sur une échelle logarithmique. La plage de variation moyenne d'un son voisé varie d'un locuteur à un autre en fonction de son âge et de son sexe :

- de 80 à 200 Hz chez les hommes,
- de 150 à 350 Hz chez les femmes,
- de 200 à 600 Hz chez les enfants,

Les sons non voisés sont associés à une fréquence nulle [5].

b Energie

L'énergie ou l'intensité d'un son est la caractéristique permettant de le comparer à d'autres de même hauteur, et ceci en terme de puissance (son aigu, son grave). En d'autres termes, elle permet de distinguer un son fort d'un son faible [6].

Pour un signal échantillonné (S_t) t=0…∞ et à support fini T, elle est donnée par :

$$E = \frac{1}{T}\sum_{t=0}^{T} S_t^2 \qquad (1.1)$$

Étant donné sa dynamique, et afin de respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E = 10 \times \log_{10}\left(\frac{1}{T}\sum_{t=0}^{T} S_t^2\right) \quad dB \qquad (1.2)$$

c Durée

La durée d'un signal (durée acoustique) correspond à son temps d'émission. C'est le paramètre le plus difficile à préciser car rien n'indique comment le système de contrôle, de production ou de perception de la parole, mesure le temps.

Les indices de durée classique supposent généralement la donnée d'une segmentation, des frontières des unités dont on désire mesurer la durée.

La durée d'une unité est alors mesurée par le nombre de trames qui séparent ses frontières de début et de fin. La plupart des systèmes utilisent une segmentation basée sur le phonème (son élémentaire d'un langage) [7].

Sa durée dépend de :

- la vitesse de débit ;
- ses qualités phonétiques (exemple : plus une voyelle est fermée, plus elle est brève).

1.2.2 Prétraitement du signal vocal

Le prétraitement est une étape importante dans l'analyse d'un signal de parole. Il consiste à le mettre en forme en suivant les étapes : Échantillonnage, filtrage puis fenêtrage.

a Echantillonnage

L'échantillonnage transforme le signal à temps continu x(t) en signal à temps discret x(nTe) défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage Te ; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage Fe.

Afin de garantir la restitution fidèle du signal, le théorème de l'échantillonnage stipule que la fréquence d'échantillonnage Fe doit être supérieure ou égale au double de la fréquence maximale à reproduire [1].

La Figure 1.1 illustre l'échantillonnage d'un signal sinusoïdal.

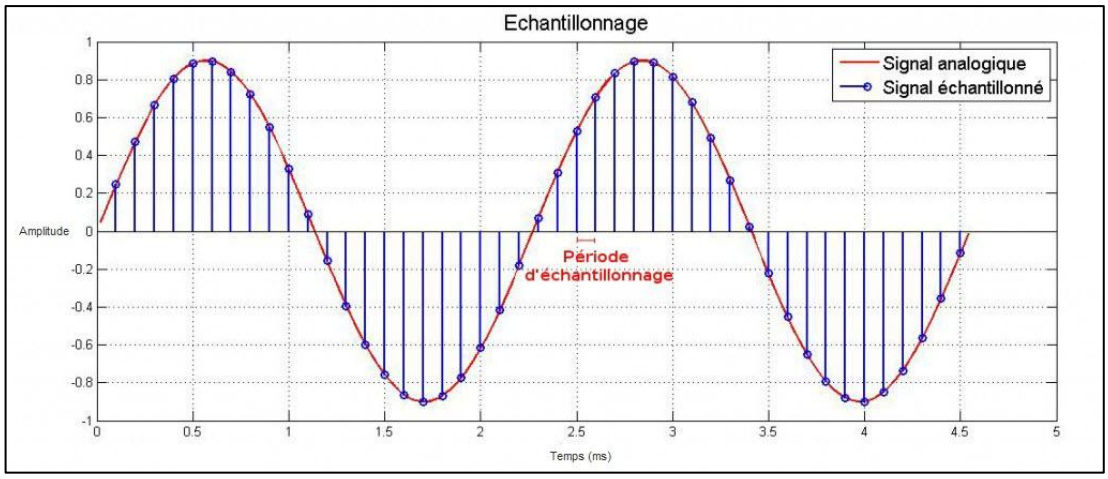


Figure 1.1 : Exemple d'échantillonnage d'un signal sinusoïdal.

b Filtrage

Le filtrage d'un signal est l'élimination de certaines fréquences le composant (autrement dit, la sélection de fréquences particulières) en modifiant certaines parties d'un signal d'entrée dans le domaine temporel et fréquentiel.

Il revient à multiplier le spectre du signal d'origine par celui résultant d'une fonction de transfert (FT) qui agit sur l'amplitude et la phase de ces composantes.

D'après la théorie de Fourier, tout signal réel peut être représenté par une somme de signaux sinusoïdaux (en nombre infini si nécessaire) à des fréquences différentes.

Généralement, on distingue trois types de filtre: filtre passe bas, filtre passe haut, filtre passe bande [8] (Figure 1.2).

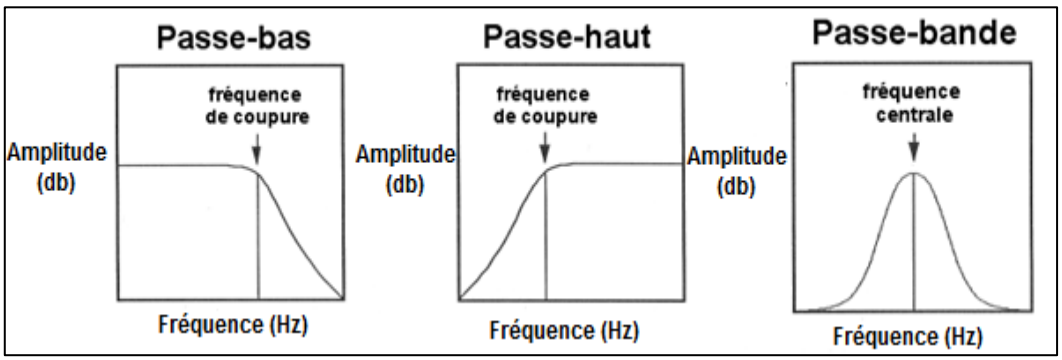


Figure 1.2 : Principaux types de filtres.

La parole ne peut être considérée comme étant un signal périodique stationnaire que sur une durée de quelques millisecondes. Pour ne pas perdre d'information et assurer un meilleur suivi des non-stationnarités, la fenêtre est généralement choisie

glissante où chaque trame couvrant une durée de 20 à 50 ms sur laquelle le signal est quasi-stationnaire.

Plusieurs fenêtres ont été étudiées dans la littérature, on peut citer celles de : Hamming, Hanning, Blackman, Rectangulaire, etc. La plus utilisée est celle de Hamming représentée dans la Figure 1.3 [9].

Elle est définie par :

$$h(n)= \{0.54 - 0.46 \cos (2\pi\frac{n}{N-1})\} \quad \text{si } 0 \leq n \leq N\text{-}1 \qquad (1.3)$$

N étant la taille de la fenêtre en nombres d'échantillons du signal et n l'indice de l'échantillon.

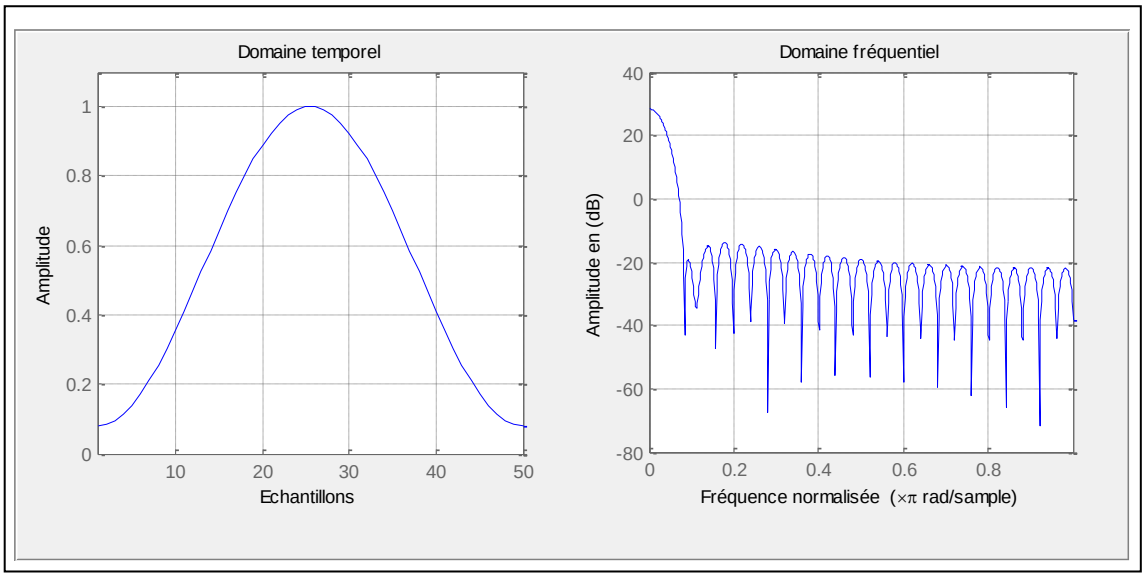


Figure 1.3 : Fenêtre de Hamming dans les domaines temporel et fréquentiel.

1.2.3 Représentation acoustique du signal vocal

Un signal vocal peut être représenté par son audiogramme, son spectre ou son spectrogramme.

a Spectre

Le spectre d'un signal est le résultat de la transformation Fourier de ce signal. A partir d'un spectre on peut savoir la fréquence et l'amplitude des composantes présentes dans le signal analysé. La forme générale des spectres, appelée **enveloppe spectrale**, présente elle-même des pics et des creux qui correspondent aux résonances

et aux anti-résonances du conduit vocal et sont appelés respectivement formants et anti-formants [1].

L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non voisés présentent souvent une accentuation vers les hautes fréquences [10] (Figure 1.4).

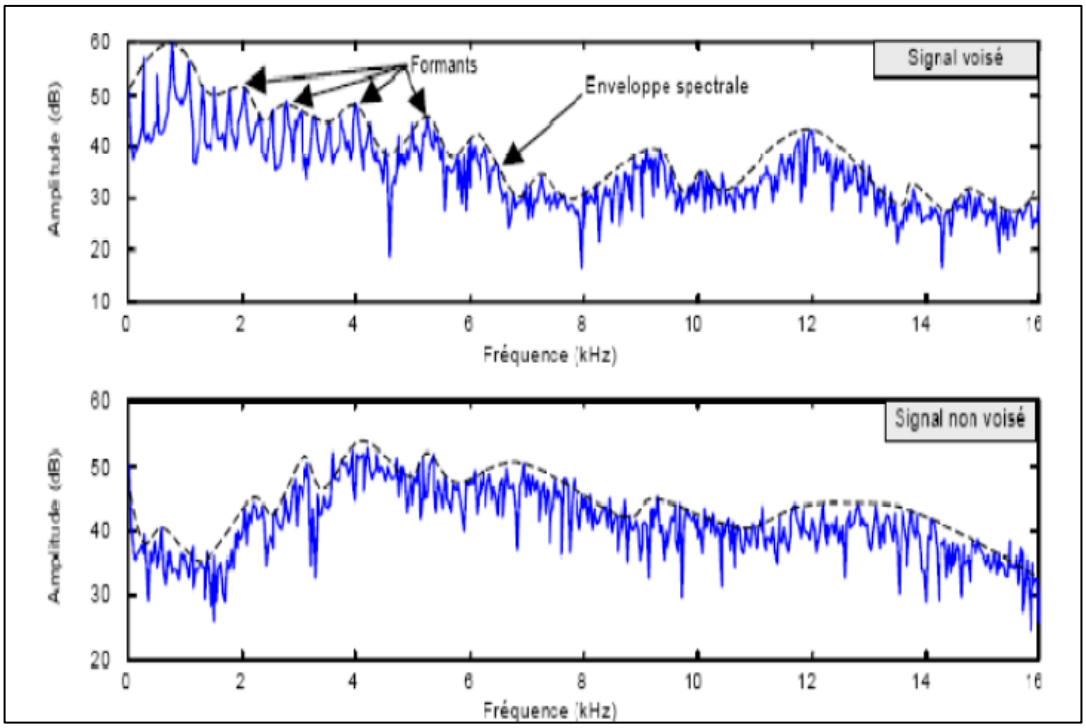


Figure 1.4 : Spectres de sons voisé et non voisé [11].

b Audiogramme

Un audiogramme (Figure 1.5) est une représentation de l'évolution temporelle du signal vocal. Si une limitation du spectre par un filtrage préalable est acceptée, le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable. C'est le rôle du filtre de garde, dont la fréquence de coupure Fc est choisie en fonction de la Fe retenue. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit Fe égale à 8000 Hz [1].

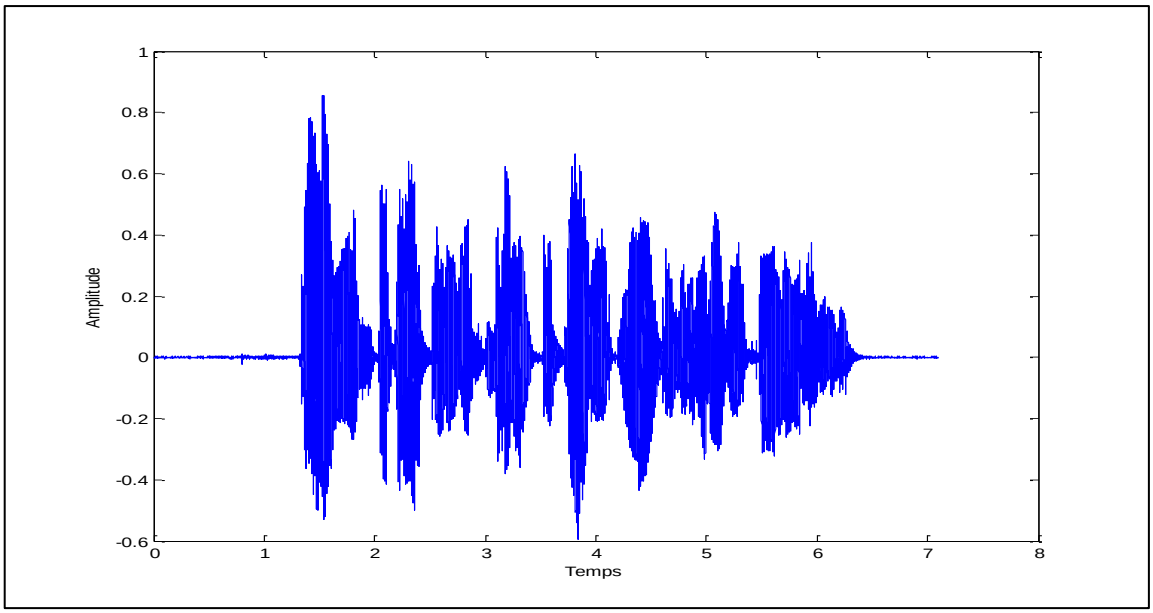
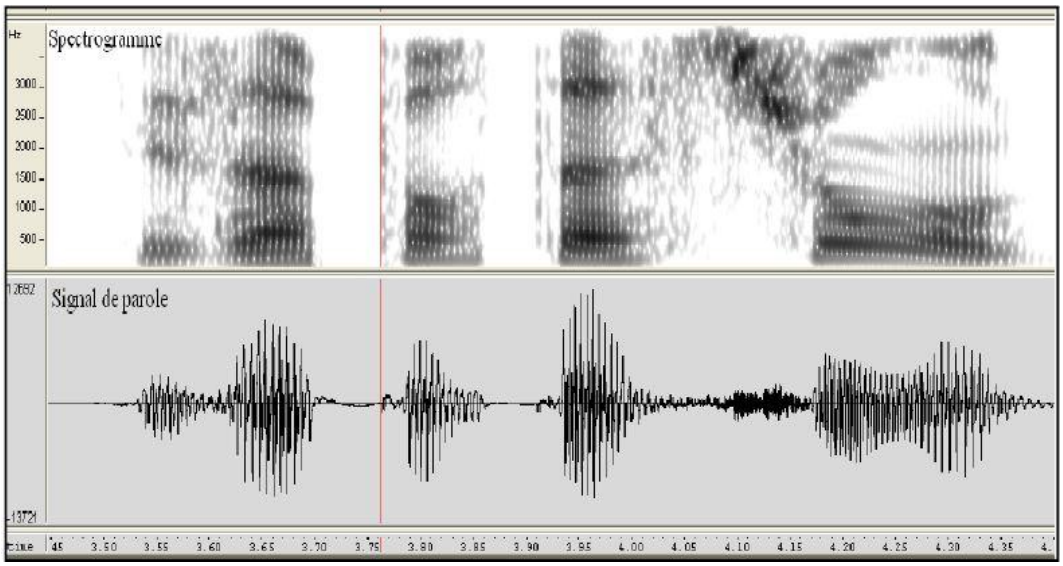


Figure 1.5 : Audiogramme d'un signal vocal.

c Spectrogramme

Le spectrogramme est un diagramme associant à chaque instant t d'un signal, son spectre de fréquence. Dans son format le plus courant, l'axe horizontal représente le temps et l'axe vertical la fréquence. Chaque point à l'intérieur du graphique est doté d'une certaine intensité qui indique l'amplitude (souvent en décibels) d'une fréquence particulière à un temps donné (Figure 1.6).



(a)

(b)

Figure 1.6 : Représentation temporelle et spectrale d'un signal vocal.

(a) Spectrogramme à large bande d'un signal vocal.

(b) Audiogramme du signal vocal.

On parle de spectrogramme à large bande ou à bande étroite selon la largeur de la fenêtre de pondération.

Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms) ; ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées apparaissent sous la forme de bandes horizontales [1].

1.3 Etude phonétique de la parole

Contrairement à l'acoustique, la phonétique ne s'intéresse pas au signal lui-même mais à la façon dont il est produit par le système articulatoire [1].

1.3.1 Production de la parole

C'est une action volontaire et coordonnée d'un certain nombre de muscles du système articulatoire. L'appareil phonatoire (Figure 1.7) se décompose en trois niveaux [12] :

- ***a*** ***Niveau sous glottique*** (diaphragme, poumons, trachée) qui s'apparente à une soufflerie. Il permet de réguler le débit et la pression d'air à l'entrée du système.

- ***b*** ***Niveau glottique*** (larynx avec les cordes vocales) qui intervient dans la production de sons voisés, où il joue le rôle d'un excitateur acoustique.

Le débit d'air qui traverse la glotte est modulé par la vibration des cordes vocales, ce qui génère une onde acoustique qui se propage dans le conduit vocal et qui est rayonnée par les lèvres. Le paramètre qui définit cette onde acoustique est principalement la période fondamentale d'un cycle glottique (T_0).

- ***c*** ***Niveau supraglottique*** (pharynx et cavités buccale et nasale). Il joue le rôle d'un articulateur. Il permet la production des consonnes et des

voyelles. Dans le cas des voyelles, il fait office de résonateur, et permet de sélectionner les bandes de fréquence à renforcer par ajustement des fréquences et largeurs de bande des résonances acoustiques. Dans le cas des consonnes, des sources aéroacoustiques de bruit sont générées selon la position des constrictions.

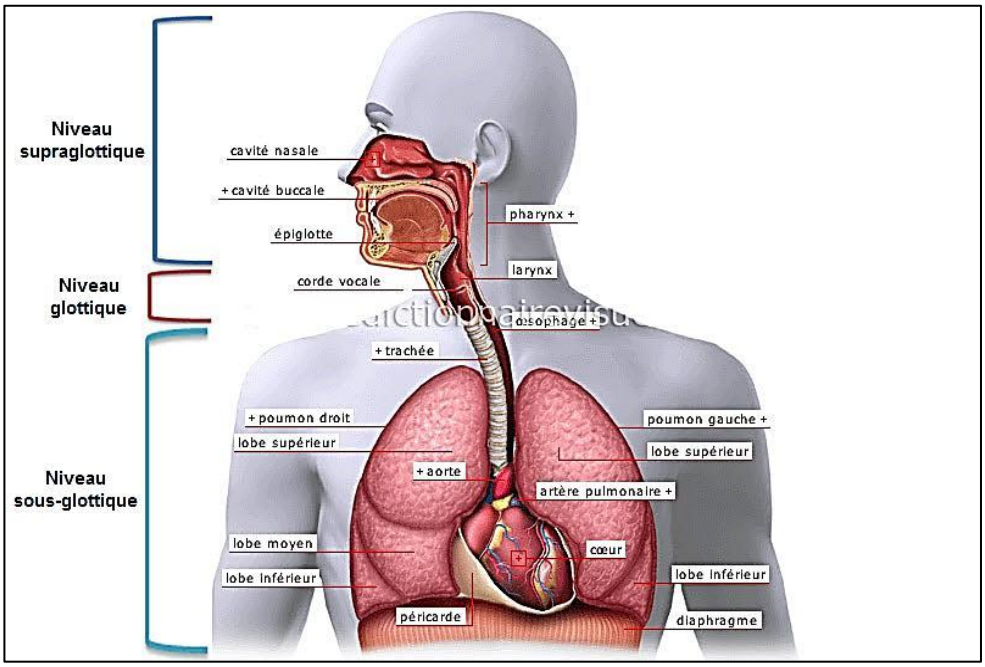


Figure 1.7 : Appareil phonatoire.

1.3.2 Classes phonétiques des sons de la parole

La plupart des langues naturelles sont composées à partir de sons distincts, les phonèmes. Un phonème est la plus petite unité présente dans la parole. Le nombre de phonèmes est toujours très limité et ça dépend de chaque langue. Les phonèmes sont classifiés en classes. Ces dernières permettent de regrouper les sons selon leurs principales caractéristiques qui sont facilement identifiables. A l'intérieur de chaque classe, on retrouve des sons dont les dissimilarités peuvent être faibles. La subdivision des sons en éléments de granularités variables et la division de l'ensemble de ces sons ou phonèmes est à l'origine de la constitution d'alphabets phonétiques qui caractérisent des langues différentes [1].

Les différentes classes phonétiques présentées généralement par les langues sont les voyelles, les consonnes et les semi-voyelles [13].

a Voyelles

Les voyelles qui sont des sons voisés diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal.

Si ce dernier est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle et ceci dépend des facteurs suivants [14] :

- La position de la langue ;
- le degré d'ouverture de la bouche ;
- le mode d'articulation ;
- le lieu d'articulation.

Selon la position de la langue, on distingue les voyelles :

- Antérieures (aigues) : le bout de la langue se déplace vers l'avant de la bouche ;
- postérieures (graves) : le dos de la langue se masse dans l'arrière de la bouche ;
- arrondies : les lèvres sont arrondies et projetées en avant ;
- non arrondies : les lèvres sont écartées ou dans une position neutre.

Selon l'écartement entre l'organe et le lieu d'articulation, on distingue les voyelles :

- Fermées : la langue s'élève et il y a un rétrécissement de la cavité ;
- ouvertes : la langue est en repos ou peu élevée, il y a une ouverture dans la cavité ;
- orales : se prononcent avec le voile du palais relevé, ce qui ferme le passage nasal ;
- nasales : se prononcent avec le voile du palais abaissé, ce qui laisse passer de l'air par la bouche et par le nez.

b Consonnes

Les consonnes sont des phonèmes produits par le passage de l'air dans la gorge et la bouche. Elles évoquent des explosions ou des frottements, produits par le souffle heurtant divers organes. Elles ne peuvent pas constituer des syllabes à elles seules.

Elles les commencent ou les terminent. On les classe en fonction de leur mode d'articulation, leur lieu d'articulation et leur nasalisation [7].

- ***Classification selon le mode d'articulation***

Le mode d'articulation définit le degré de contact, entre les articulateurs, qui existent durant la prononciation d'une consonne.

En fonction de ce mode, on distingue les consonnes [7]:

- voisées / non Voisées : selon la vibration ou non des cordes vocales ;
- nasales/orales : selon que la position de la luette permet ou non l'écoulement de l'air par les fosses nasales ;
- fricatives/occlusives : selon certaines positions des organes d'articulation :

o Les consonnes fricatives (constrictives) résultent de l'écoulement de l'air dans une constriction étroite située en point du conduit vocal, elles peuvent être voisées ou non [4]. Leur énergie est concentrée dans les hautes fréquences [15].

Les fricatives non voisées ont généralement la forme d'un bruit, leur énergie est assez importante.

Les fricatives voisées ressemblent à celles non voisées correspondantes, mais avec, en plus, une vibration des cordes vocales, bien visible sur un spectrogramme [4].

o Les consonnes plosives (occlusives) sont produites par une occlusion momentanée du conduit vocal en un point donné, suivie par une ouverture brusque, elles peuvent être voisées ou non.

Les plosives non voisées sont repérables par un silence court dans le signal (occlusion ou tenue de la plosive), suivi d'une barre verticale ou barre d'explosion (relâchement de la plosive), d'un court délai d'établissement du voisement (pendant lequel un bruit d'aspiration est parfois visible sur le spectrogramme) [4].

Les mouvements formantiques des plosives voisées correspondent à ceux des plosives non voisées ayant le même lieu d'articulation.

La barre de voisement est visible pendant la phase de tenue sur le spectrogramme et sur l'audiogramme. Elles sont caractérisées par une courte durée par rapport à celle des plosives non voisées, leur délai d'établissement du voisement est nettement plus court [4].

- ***Classification selon le lieu d'articulation***

Le lieu d'articulation d'une consonne définit l'endroit où elle se produit. Il peut être :

- bilabial ;
- glottal ;
- labiodental ;
- vélaire;
- palato-alvéolaire;
- dental;
- …etc.

- ***Classification selon la nasalisation***

La nasalisation est caractérisée par un abaissement du voile de palais en faisant intervenir les cavités nasales.

c Semi voyelles

Les semi voyelles sont des phonèmes intermédiaires entre les voyelles et les consonnes. Quand on les prononce, on entend le timbre d'une voyelle auquel s'ajoute un frottement d'une consonne. Leur fréquence d'emploi est liée à la vitesse du débit de la parole. Plus celui-ci est rapide, plus il y aura de semi-voyelles [14].

1.3.3 Notions fondamentales sur l'Anglais Américain

La recherche en traitement de la parole dans une langue donnée doit nécessairement être procédée par l'étude de sa composante phonétique. Cette étude permet de cerner les principales caractéristiques relatives aux différents phonèmes et l'ensemble des paramètres acoustiques [6].

a Classes phonétiques de l'Anglais Américain

L'Anglais Américain (AA) se distingue de l'Anglais Britannique par la prononciation et le vocabulaire mais aussi par l'orthographe et certaines règles de grammaire. L'AA possède 39 phonèmes : Vingt phonèmes consonantiques, quatre semi-voyelles, douze voyelles et trois diphtongues [16], [17] (voir Annexe A).

- ***Consonnes de l'AA***

Les consonnes de l'AA peuvent être classées en plosives, fricatives, nasales, chuchotés (*whispers*) et affriquées (Figure 1.8) [16].

Selon le mode d'articulation, on distingue [18] :

- Les plosives (occlusives) : [b], [d], [g], [p], [t], [k].
- Les fricatives : [v], [ð], [z], [ʒ], [f], [θ], [s], [ʃ].
- Les nasales : [m], [n], [ŋ].
- Le chuchoté : [h].

Dans la langue AA, il n'existe qu'un seul phonème de chuchotement, le [h]. Il est considéré comme une fricative non voisée.

- Les affriquées : le phonème affriquée est considéré comme une semi occlusive.

Cette consonne particulière se comporte à la fois comme une occlusive et une fricative. Lors de la prononciation, la langue ne s'écarte pas brusquement du palais comme dans le cas des occlusives, mais plutôt d'une manière douce. L'air libéré sera sous forme de friction.

Dans la langue AA, il existe deux affriquées : [tʃ], [dʒ] [16].

Selon le lieu d'articulation, on distingue [17]**,** [18]:

- Plosives
  - bilabiales : [p] (non voisée) et [b] (voisée),
  - alvéolaires : [t] (non voisée) et [d] (voisée),
  - vélaires : [k] (non voisée) et [g] (voisée).
- Fricatives
  - labiodentales: [f] (non voisée) et [v] (voisée),
  - dentales : [θ] (non voisée) et [ð] (voisée),

- alvéolaires : [s] (non voisée) et [z] (voisée),
- palato-alvéolaires : [ʃ] (non voisée) et [ʒ] (voisée).

o Nasale

- bilabiale : [m] (voisée),
- alvéolaire : [n] (voisée),
- vélaire : [ŋ] (voisée).

o chuchoté glottale : [h] (non voisée).

o Affriquées

- alvéolaires : [tʃ] (non voisée) et [dʒ] (voisée).

- ***Semi voyelles de l'AA***

Selon le mode d'articulation, on distingue deux catégories [16] :

- Les glissées: elles ressemblent aux voyelles. De plus, l'air est légèrement plus entravé pour les glides que pour les voyelles. les lèvres presque fermées pour [w], et la langue touche presque le palais pour [y].
- Les liquides : Elles possèdent des caractéristiques spectrales similaires aux voyelles, elles sont faibles en énergie due au fait que le conduit vocal est plus étroit pendant leur production, en AA, on a [l] et [r].

Selon le lieu d'articulation, on distingue [17], [18] :

o Semi voyelles :

- bilabiales : [w] (voisée),
- palatales : [y] (voisée),
- alvéolaires : [l] (voisée) et [r] (voisée).

- ***Voyelles de l'AA***

Les voyelles de l'AA contiennent trois sous-groupes définis selon la position de la langue étant [18] :

- Avant: [i], [ɪ], [e], [æ], [ɛ],
- Centrée: [ɜ], [ʌ],
- Arrière: [a], [ɔ], [o], [U], [u].

Les voyelles arrière, centrée sont arrondies et les voyelles avant sont non arrondies.

- ***Diphtongues de l'AA***

Elles impliquent un mouvement d'une voyelle initiale vers une autre voyelle finale. La différence entre une diphtongue et deux voyelles individuelles est que la durée de la transition est plus grande que la durée de chaque voyelle. De plus la voyelle initiale est plus longue que la voyelle finale.

L'AA possède trois diphtongues : [aI], [aU] et [ɔI] [17].

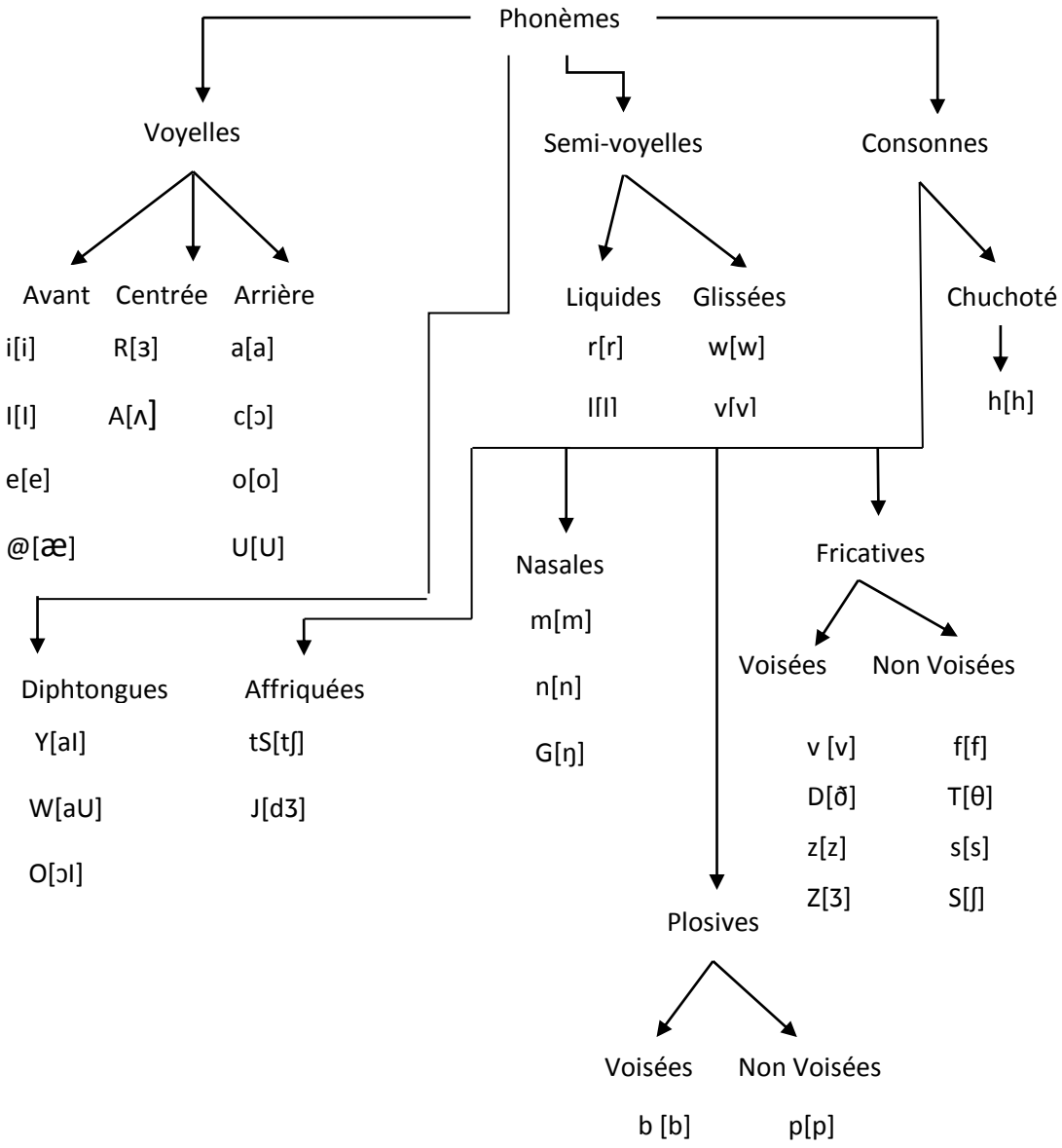


Figure 1.8 : Phonèmes de l'AA [16], [17]. (Les Alphabets Phonétiques sont à gauche, et les phonèmes sont donnés entre crochets).

1.4 Propriétés spécifiques du signal vocal

Le signal de la parole n'est pas un signal ordinaire. Il est le vecteur d'un phénomène complexe : la communication parlée. D'un point de vue mathématique, il est difficile de modéliser le signal de parole, compte tenu de sa variabilité. On va ici tenter de mettre en évidence quelques caractéristiques importantes du signal non stationnaire afin de faire ressortir les problèmes posés lors de son traitement.

1.4.1 Continuité

Le langage oral est une suite continue de sons sans séparation entre les mots. Les silences correspondent en général à des pauses de respiration dont l'occurrence est aléatoire. Il peut très bien y avoir des intervalles de silence au milieu d'un mot et aucun intervalle entre deux successifs. Par conséquent, il est très difficile de déterminer le début et la fin des mots composant la phrase [19].

1.4.2 Redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, en effet, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter la durée du traitement [20].

1.4.3 Variabilité

La parole présente une très grande variabilité dans le domaine temporel et fréquentiel. Ces variations sont dues à [21] :

a Influence de la prise de son

Acoustique de la pièce, milieu ambiant, distance entre le microphone et les lèvres, type de microphone, filtrage et pré-amplification, conversion analogique/numérique.

b Influence du locuteur

- ***Variabilité intra locuteur***

Pour la même personne, le signal varie au cours du temps pendant l'élocution. Ces variations temporelles sont surtout prosodiques [22].

Plusieurs critères peuvent être responsables de ces différences :

- La fatigue ;
- l'état émotionnel du sujet (une émotion telle que la peur affecte le timbre et le rythme de la voix) ;
- Les maladies qui affectent les organes de la voix.

- ***Variabilité interlocuteur***

Due à la différence de l'élocution entre individus. En effet, pour un même mot prononcé par deux individus, on a deux signaux ayant des différences acoustiques importantes [22].

c Influence du contexte

Les mouvements articulatoires peuvent être modifiés de façon à minimiser l'effet à produire pour les réaliser à partir d'une position articulatoire donnée, ou pour anticiper une position à venir. Ces effets sont connus sous le nom de réduction, d'assimilation et de coarticulation.

- La réduction est due au fait que les cibles articulatoires sont moins atteintes dans le parlé rapide ;
- L'assimilation est causée par le recouvrement des mouvements articulatoires et peut aller jusqu'à modifier un des traits phonétiques du phonème prononcé ;
- La coarticulation est définie comme étant le chevauchement et l'interaction des différents articulateurs au cours de la production de segments phonétiques successifs.

Il en résulte que la configuration du conduit vocal à un instant donné peut être mise en relation avec les caractéristiques de différents segments phonétiques [23].

Ces phénomènes sont en grande partie responsables de la complexité du traitement réalisé sur le signal de la parole [24].

1.5 Conclusion

L'étude accomplie dans ce premier chapitre nous a amené à savoir le mécanisme de la communication parlée. Elle nous a aussi permis de comprendre les classes phonétiques de l'Anglais Américain.

Dans le prochain chapitre, on va présenter un état de l'art sur les techniques de transformations prosodiques de la voix.

Chapitre 2 Transformation de la voix

2.1 Introduction

La transformation de la voix est une méthode qui consiste à modifier les paramètres intrinsèques du signal vocal. Elle est souvent utilisée par les développeurs de voix de synthèse pour créer plusieurs voix à partir d'une seule. Pour y parvenir, il faut opérer deux formes de modifications : La modification des paramètres prosodiques et l'enveloppe spectrale.

La première forme consiste à altérer les deux paramètres (la durée et la fréquence fondamentale). La deuxième permet de modifier le timbre de la voix.

Dans ce chapitre on va essayer de donner une vision globale sur les techniques de transformation de la voix, l'évaluation qualitative et les applications de la modification prosodique. On terminera ce chapitre par une conclusion.

2.2 Transformation de la voix

De nombreux chercheurs ont participé depuis les années 1980 à faire évoluer l'état de l'art dans le domaine de la transformation de voix. Parmi eux, on peut citer : T. Dutoit [1], G. Peeters [25], E. Moulines, F. Charpentier et J. Laroche [26], [27].

La transformation de la voix correspond à des besoins courants dans divers domaines de traitement du son et de la parole. La plupart des techniques connues de transformation de la parole visent essentiellement à modifier l'enveloppe spectrale et les paramètres prosodiques de la voix. Dans le cadre de notre étude, on s'intéresse principalement à la modification des paramètres prosodiques. Une brève définition sur la modification de l'enveloppe spectrale est aussi donnée.

2.3 Modification de l'enveloppe spectrale

Plusieurs techniques de modification de l'enveloppe spectrale ont été proposées dans la littérature, parmi lesquelles, on peut citer :

2.3.1 Système à base des réseaux de neurones

En 1999 M. Narendranath et al [28], ont proposé un nouveau système pour la transformation de la voix qui convertit le signal vocal prononcé par un locuteur source à un autre signal qui possède les caractéristiques vocales (timbre de la voix) d'un locuteur cible. Le problème principal revient à transformer les caractéristiques du conduit vocal d'un locuteur à un autre. Les formants sont utilisés pour représenter les caractéristiques du conduit vocal. Le vocodeur de formants est employé pour la synthèse des signaux. Le système se compose d'une phase d'analyse des formants, suivie d'une phase d'apprentissage dans laquelle la transformation des formants est effectuée par une technique à base des réseaux de neurones. Les formants transformés ainsi que le contour de la F_0 modifié (en fonction de la F_0 moyenne du locuteur cible) sont utilisés pour synthétiser la parole avec les caractéristiques du conduit vocal souhaitées.

2.3.2 Méthode à base de la transformée en ondelettes

En 2009, F.Ykhlef et al [29], ont proposé une nouvelle méthode pour des modifications conjointes de l'enveloppe spectrale et la durée du signal vocal nommée FBSOLA « *Filter Bank Synchronous Overlap and Addition »*. Elle consiste en l'utilisation d'une technique hybride à base de l'algorithme SOLA « *Synchronous Overlap and Addition »* (paragraphe 2.4.1(d)) et une décomposition fréquentielle en M sous-bande par un banc de filtres uniforme.

Le banc de filtres utilisé est un banc pseudo QMF « *Quadrature Mirror Filter »* modulés en cosinus. La méthode exposée exploite la synchronisation offerte par l'algorithme SOLA afin d'effectuer des modifications simultanées de l'enveloppe spectrale et de la durée phonémique. La modification de l'enveloppe spectrale consiste en des amplifications fréquentielles en seize sous bandes équidistantes.

L'application visée dans le cadre de ce travail est l'aide au diagnostic des troubles auditives des personnes malentendantes.

2.4 Modification des paramètres prosodiques

2.4.1 Modification de la durée

De nombreuses techniques ont été proposées dans la littérature pour la modification de la durée du signal vocal, on peut citer :

a Vocodeur de phase

Cette méthode a été introduite par Flanagan et Golden en 1966 et implémentée numériquement par Potonoff [30], [31]. Elle utilise le modèle d'analyse/ synthèse par la Transformée de Fourier à Court Terme (TFCT) dans le but de modifier la durée d'un signal. La méthode consiste à modifier la représentation temps –fréquence du signal original de tel sorte que le signal de synthèse résultant (par l'usage de la transformée de Fourier à court terme inverse (TFCTI)) ait la durée désirée, tout en préservant les caractéristiques spectrales du signal original.

Le vocodeur de phase parvient à modifier la durée d'un signal en posant un intervalle de synthèse différent de celui d'analyse selon le taux de modification de la durée désirée.

b Méthode TDHS

En 1979, Malah [32] a introduit une méthode appelée TDHS « *Time Domain Harmonic Scaling* » qui est une technique de modification de la durée faisant usage de la notion de la F_0. L'algorithme commence par la détermination de la variation de la F_0 du signal d'entrée. Cette variation est utilisée pour segmenter le signal d'entrée en segments qui ont une longueur d'une période de F_0. Évidemment, le signal d'entrée se doit d'être périodique ou quasi périodique pour que cette étape donne de bons résultats. Les segments obtenus sont ensuite chevauchés et additionnés de façon à accomplir la modification désirée.

c LSEE-MTFTM

En 1984 Griffin et Lim [33] ont proposé un algorithme nommé LSEE-MSTFTM « *Least-Squares Error Estimation from the Modified Short-Time Fourier Transform Magnitude* » qui est principalement basé sur l'analyse et la synthèse par l'usage de la TFCT. Cet algorithme permet de minimiser l'erreur quadratique entre la TFCT du signal synthétisé et celle du signal modifié en utilisant seulement le spectre d'amplitude du signal original.

L'algorithme est employé afin de construire itérativement un signal temporel dilaté (ou compressé) à partir de l'information du module de la TFCT modifiée tout en préservant les caractéristiques spectrales du signal original. Les itérations successives consistent à reconstruire une information de phase cohérente avec le module du signal d'entrée de tel sorte que le signal à synthétiser possède une TFCT la plus proche que possible de la TFCT désirée.

d SOLA

En 1985, Roucos et Wilgus [34] ont proposé une méthode de modification de la durée appliquée à la parole qu'ils nomment SOLA. Son formalisme est exposé dans un contexte temps-fréquence. Il est issu de la méthode de Griffin et Lim [35] dont son principe est de produire un signal temporel à partir du spectre d'amplitude à court terme dilaté en reconstruisant de manière itérative le spectre de phase à court terme qui lui soit cohérent.

Le choix de l'estimation initiale du signal est important pour la vitesse de convergence de l'algorithme. Roucos et Wilgus suggèrent d'utiliser comme estimation initiale, le signal original retardé d'une durée fixe. Cette estimation initiale est explicitée dans le domaine temporel et ne nécessite pas d'itération supplémentaire.

Cette méthode est donc beaucoup plus efficace en termes de puissance de calculs, et semble donner des résultats au moins aussi bons que ceux de Griffin et Lim pour la parole [36].

Il existe plusieurs variantes de la méthode SOLA. Elles différent sur le choix des critères de synchronisation. On peut citer :

- SAOLA « *Synchronized and Adaptive Overlap-Add* » [36].
- PAOLA « *Peak-Alignment Overlap and Add* » [36].
- SOLAFS « *Synchronized Overlap-Add with Fixed Synthesis* » [37].

2.4.2 Modification de la F_0

Dans la littérature, plusieurs techniques sont proposées ces dernières années pour la modification de la F_0. On peut citer :

a Algorithme à base de la DCT

En 2003, R. Muralishankar et al. [38], ont proposés un nouvel algorithme pour la modification de F_0. Le signal résiduel est obtenu à partir des trames synchrones à la F_0 par une opération de filtrage inverse. Le signal résiduel est modifié après le calcul de la transformée en cosinus discrète (dite en anglais Discreet Cosine Transform (DCT)).

Selon le facteur de modification de la F_0 voulu, certains coefficients de la DCT vont être ajoutés ou éliminés. Le signal reconstitué est obtenu par le calcul de la DCT inverse (IDCT, dite en anglais Inverse Discreet Cosine Transform). L'inconvénient de cette technique est que l'enveloppe spectrale du signal synthétisé sera modifiée. Des opérations de correction spectrales sont ainsi nécessaires [38].

b Utilisation du vocodeur de phase pour la modification de F_0

Le vocodeur de phase est habituellement présenté comme une solution de haute qualité pour la modification de la durée des signaux [30], [31], [35]. La modification de F_0 étant généralement mise en œuvre comme une combinaison de la (dilatation/compression) de signal et la transformation de taux d'échantillonnage [39].

En se basant sur ce principe, J. Laroche et M. Dolson [39] ont proposé une nouvelle technique pour la manipulation de la F_0 en utilisant le vocodeur de phase, chose qui n'était pas possible par les techniques précédentes. La technique proposée procède directement dans le domaine fréquentiel. Elle est basée sur la détection des pics suivis d'une étape de décalage qui tient en compte d'un chevauchement de 50%.

c Modification de la F_0 par la transformée en ondelettes continue

En 2006, K. Kumar et J. Jain [40] ont proposé une nouvelle technique de modification de la F_0 à base de la transformée en ondelettes continue. Le principe de base consiste en extraction de l'échelle fréquentielle de la F_0 des zones voisées à partir des coefficients complexes de cette transformée. La matrice de phase résultante est altérée en préservant le module de la transformée en ondelettes et cela en fonction de la modification désirée. Les fréquences fondamentales originales et modifiées des trames étudiées sont mesurées par la fonction d'autocorrélation pour la comparaison des résultats obtenus.

d FD-PSOLA

La méthode FD-PSOLA « *Frequency-Domain Pitch Synchronous Overlap-Add* » a été proposée par E. Moulines et F. Charpentier [26] en 1990. Cette méthode est employée seulement pour les modifications de F_0. Elle diffère de sa variante dans le domaine temporel « TD-PSOLA, *Time- Domain PSOLA* » dans la définition des formes d'onde à la synthèse. Les modifications de la F_0 sont effectuées dans le domaine spectral en utilisant la notion des signaux à court terme. La FD-PSOLA est basé sur la reproduction de reéchantillonnage des harmoniques [27]. Plus de détails sur la méthode PSOLA vont être donnés dans le chapitre suivant.

2.4.3 Modification simultanées de la F_0 et de la durée

a TD-PSOLA

La méthode TD-PSOLA « *Time-Domain Pitch Synchronous Overlap-Add* » a été proposée par E. Moulines et F. Charpentier [26] en 1990. Contrairement à sa variante dans le domaine fréquentiel (FD-PSOLA), la TD-PSOLA permet la modification simultanée de F_0 et la durée de la parole. Ces modifications sont nécessaires pour produire un signal de parole compatible avec les consignes prosodiques souhaitées.

On définit d'abord des « marques d'analyse » synchrones à la F_0 pour les parties voisées, positionnées sur la forme d'onde à chaque période [26]. Les modifications d'échelles sont alors effectuées de la façon suivante:

- ***Modification de la durée***

Pour modifier la durée du signal sans en altérer la F_0, on va simplement dupliquer (étirement temporel) ou éliminer (compression temporelle) des périodes de la forme d'onde, en fonction du taux de modification désiré. On est donc conduit à définir des marques de synthèse également synchrones du fondamental, associées aux marques d'analyse.

- ***Modification de la F_0***

Si l'on est capable de positionner dans le signal les marques d'analyse, on conçoit que diminuer/augmenter l'intervalle de temps séparant deux marques d'analyse consécutives va permettre d'augmenter /diminuer la F_0, sans que les formants soient modifiés [41].

b Modèle Source/Filtre

Il est connu de décomposer les signaux de parole selon un modèle dit « Source/Filtre » (Figure 2.1). Dans ce modèle, le signal de parole x(n) est considéré comme une excitation glottique e(n) transformée par un filtre h(n) représentant la variation du conduit vocal. L'excitation est obtenue par un filtrage inverse du signal de parole.

L'excitation peut être un bruit gaussien blanc pour des sons non voisés ou un train d'impulsion pour les sons voisés [42].

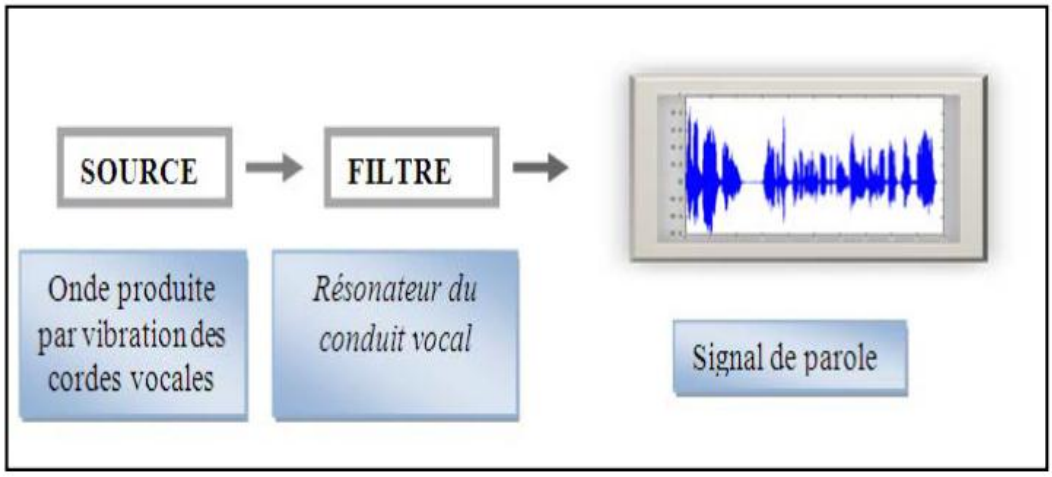


Figure 2. 1 : Modèle Source-Filtre de la production de la parole [42].

La modification de la durée et la hauteur d'un signal de parole par le modèle Source/Filtre est introduite en deux étapes, soit l'analyse et la synthèse :

L'étape d'analyse consiste à décomposer le signal x(n) selon un modèle Source/Filtre (En utilisant la technique LPC « *Linear Predictive Coding* » [42], [43]) et d'extraire les paramètres de ce modèle tels que la période fondamentale du train d'impulsions et les résonances du filtre variant dans le temps.

A la synthèse, pour avoir une modification de l'échelle temporelle (durée), il suffit d'appliquer la fonction de modification du temps "*D*" (dite en anglais T*ime-Warping Function*) définie par [44], [45]:

$$D = \frac{1}{T}\int_0^{nT} \alpha(\tau)\, d\tau \qquad (2.1)$$

Où "T" représente la période d'échantillonnage et α(τ) est le taux de dilatation /compression.

Pour la modification de la F_0, on applique la formule suivante :

$$P'(n) = P(n) / \alpha \qquad (2.2)$$

Où "P (n)" représente la F_0, " P'(n)" la nouvelle F_0 et α le facteur de modification. Cette nouvelle valeur va être imposée comme une grandeur d'entrée dans la paramètrisation de la source d'excitation.

c Modèle sinusoïdal

Les modèles sinusoïdaux sont fondés sur l'hypothèse que le signal de parole peut être représenté sous forme d'une somme d'ondes sinusoïdales. Elles opèrent dans le domaine fréquentiel et se fondent généralement sur la TFCT. Dans [46], le signal de parole est modélisé comme une somme de composantes sinusoïdales :

$$x_k(n) = \sum_{l=1}^{L} A_l^k(n)\cos(\varphi_l^k(n))) \qquad (2.3)$$

Où $A_l^k(n)$ et $\varphi_l^k(n)$ représentent respectivement l'amplitude et la phase de la $l^{ème}$ composante sinusoïdale de la $k^{ème}$ trame. L est le nombre de composantes fréquentielles.

La modification de l'échelle temporelle peut être réalisée en utilisant le modèle sinusoïdal de l'équation suivante :

$$x_{kmod}(n) = \sum_{l=1}^{L} A_l^k\ (n/\alpha)\cos(\varphi'^k_l(n/\alpha)\alpha) \qquad (2.4)$$

Où "α" est le facteur de modification de la durée désirée, A l'amplitude et φ' représente la phase à temps variable non enveloppé (en anglais the Unwrapped Time-Varying Phase) de la voix sinusoïdale [47].

La modification de F_0 est accomplie par l'interpolation fréquentielle des spectres en utilisant le même modèle sinusoïdal [48].

2.5 Evaluation des techniques de modifications prosodiques

En général, l'évaluation des techniques de modifications prosodiques est effectuée par des méthodes subjectives ou objectives. D'une part, les méthodes objectives mesurent la qualité sonore par des analyses mathématiques [49]. D'autre part les méthodes subjectives mesurent l'intelligibilité de la parole ou la qualité sonore perçue. Elles sont accomplies par des tests d'écoute. Pour cette raison, il y a une grande demande pour l'évaluation de la qualité sonore dans les applications pratiques. En dépit de l'importance de cette demande, très peu de méthodes sont disponibles [50]. On peut citer :

- DRT « *Diagnostic Rhyme Test*»,
- MRT «*Modified Rhyme Test* »,
- MOS « *Mean Opinion Score* ».

La méthode DRT permet d'évaluer la transparence du message reçu à travers une mesure du degré de dégradation des caractéristiques élémentaires des consonnes lorsque celles-ci se trouvent au début de mots. Une version plus générale du test DRT permet de tester tout aussi bien les voyelles que les consonnes et ce quelle que soit leur position dans un mot [51].

La méthode MRT est une sorte de prolongement du DRT. Le test MRT s'effectue pour les appréhensions de première et dernière consonne. Le test se compose de 50 séries. Chaque série est composée de 6 mots différents contenant une seule syllabe. L'ensemble de séries est prononcé à la fois et l'auditeur marque le mot qu'il entend sur une feuille. La première moitié des mots est utilisée pour l'évaluation des

consonnes initiales. La seconde moitié est utilisée pour l'évaluation des consonnes finales [52].

La méthode la plus utilisée est le MOS. Elle consiste à écouter les sons afin d'évaluer leur qualité selon une échelle de 5 points (excellent, bon, passable, mauvais, et très mauvais) (Tableau 2.1) [53].

Qualité de voix	Echelle (point)
Excellente	5
Bonne	4
Passable	3
Mauvaise	2
très mauvaise	1

Tableau 2. 1 : Test MOS.

2.6 Applications de la modification des paramètres parodiques

Les applications de la modification de la voix sont nombreuses et visent les différents services de télécommunication et multimédia. On peut citer :

2.6.1 Aide aux malentendants

Le problème posé par les malentendants provient essentiellement du caractère spécifique et dégradé de leur perception auditive. La nature précise du déficit auditif est difficile à établir et peut couvrir les aspects fréquentiel, énergétique mais aussi temporel de la perception. Dans le dernier cas, le patient n'a pas le temps d'intégrer les informations phonétiques contenues dans le signal de parole.

Étant donnée la variété des pathologies de l'audition, les modifications de la voix s'étendent d'un simple filtrage à un recodage complet, et peuvent donc nécessiter un certain réapprentissage de la compréhension et l'intelligibilité de la parole. Ces transformations modifient la perception des sons car elles affectent les caractéristiques spectrales et temporelles des sons [54], [55], [56].

2.6.2 Synthèse par échantillonnage

Plusieurs synthétiseurs modernes utilisent une bibliothèque limitée d'échantillons sonores. L'espace mémoire limité dont disposent ces appareils pour y stocker ces échantillons est à l'origine de ce problème. Une solution possible permettant de recréer un éventail plus large de sonorités est de modifier les échantillons en tonalité (couramment nommée le pitch) ou en durée, en temps réel.

2.6.3 Compression de données

La modification de durée a été appliquée au codage de la parole à bas débit. La méthode consiste à réduire la durée du signal à l'encodeur et de restaurer la durée originale au décodeur. Il a cependant été observé que la réduction de débit possible était relativement limitée en utilisant cette technique, dû aux artéfacts introduits lors de la compression et de l'expansion du signal traité.

2.6.4 Voix sur IP

Les systèmes de voix sur IP (Voice over IP) sont particulièrement sensibles au délai introduit par le réseau de communication. De façon à garantir une sortie audio continue au récepteur d'un flux de voix sur IP, la modification de durée est appliquée au signal reçu en fonction du délai induit par le réseau IP.

2.6.5 Livres audio

La lecture auditive permet de consulter des ouvrages littéraires par l'écoute. L'augmentation de la vitesse de diction de l'ouvrage, tout en préservant l'intelligibilité, permet à l'auditeur de consulter des sections d'un ouvrage en survol. La lecture rapide de messages téléphoniques est un exemple d'application similaire.

2.6.6 Assistance d'interface graphique

Les systèmes d'opérations d'ordinateurs personnels permettent aux usagers ayant une accessibilité réduite au contenu visuel d'utiliser une interface graphique à travers une interface audio. Les divers éléments composant une interface graphique

(telle l'étiquette d'un bouton ou le texte d'une page WEB) sont dictés à l'utilisateur à un rythme contrôlé par l'utilisateur.

L'utilisation d'une vitesse d'élocution variable, adaptée aux besoins individuels des utilisateurs, augmente la convivialité d'une telle interface.

2.6.7 Apprentissage d'une langue étrangère

Dans cette situation, il est pratique de réduire artificiellement le débit de parole d'un locuteur s'exprimant dans la langue étrangère enseignée. Ceci facilite l'apprentissage des étudiants, qui peuvent mieux cerner le dialogue.

2.6.8 Postsynchronisation audio-vidéo

Dans le domaine de la production audio-vidéo, la bande sonore d'un ouvrage peut-être créée indépendamment du contenu vidéo. La modification de durée du signal permet de synchroniser a posteriori le contenu audio avec le contenu vidéo.

2.6.9 Mixage audio et composition musicale

Dans le domaine de la musique électronique, les musiciens utilisent la modification de durée de façon à synchroniser deux morceaux de musique de tempos différents, dans le but d'effectuer une transition plus transparente d'un morceau à l'autre. De plus, l'effet de modification de durée d'un signal ouvre la porte à une variété d'effets audio originaux qui peuvent enrichir une pièce musicale [35].

2.7 Conclusion

Divers méthodes sont proposées dans la littérature pour la modification des paramètres intrinsèques de la voix (prosodie et enveloppe spectrale). Au cours de ce deuxième chapitre de mémoire, on a décrit les méthodes de transformation les plus connues et les plus utilisées dans le domaine de traitement numérique de la parole. La méthode TD-PSOLA, pour la modification de la prosodie, semble être une solution efficace de modification en terme de qualité et complexité algorithmique. Elle va être utilisée dans le prochain chapitre pour la modification de la fréquence fondamentale.

Chapitre 3 Etude théorique de la méthode PSOLA

3.1 Introduction

Les techniques de modification de la voix s'avèrent très utiles dans de nombreuses applications de traitement de la parole. Elles permettent de procéder à des modifications prosodiques (modification de la hauteur et la durée) souvent nécessaires pour conférer une intonation acceptable au signal de parole synthétique.

Ce chapitre commence par une description de la méthode PSOLA dans un cadre général. Par la suite, une étude approfondie de la méthode TD-PSOLA « Time Domain *Pitch Synchronous OverLap-Add* » pour la modification de la fréquence fondamentale est effectuée. Une brève conclusion terminera ce chapitre.

3.2 Etat de l'art de la méthode PSOLA

Depuis 20 ans, de nombreuses méthodes de modification du signal reposant sur le principe de superposition/addition temporelle ont été proposées. Parmi les plus importantes, citons les méthodes : TDHS, LSEE-MSTFTM, SOLA, SOLAFS (voir chapitre 2). Les modifications du signal permises par ces méthodes sont essentiellement des modifications de l'axe temporel du signal (compression/dilatation temporelle du signal). Pour cela, ces méthodes procèdent de manière tantôt proportionnelle tantôt synchrone à la période fondamentale, tantôt à l'analyse tantôt à la synthèse.

La méthode PSOLA « *Pitch Synchronous Overlap-Add* », se distingue de ces méthodes par une synchronie à la période fondamentale tant à l'analyse qu'à la synthèse (Tableau 3.1). Ceci permet, à l'inverse des méthodes précédentes, un contrôle à la fois du déroulement de l'axe temporel et de la hauteur du signal [25].

	Analyse	Synthèse
TDHS	Proportionnalité	Proportionnalité
LSEE-MSTFTM	-	reconstruction itérative de la phase
SOLA	-	synchronie (par auto-corrélation) avec le signal déjà synthétisé
SOLAFS	synchronie (par auto-corrélation) avec le signal analytique.	-
PSOLA	Synchronie	Synchronie

Tableau 3. 1 : Critères de synchronisation des méthodes basées sur la « superposition/ addition » [25].

La méthode de superposition/addition synchrone à la période fondamentale, PSOLA, [25] repose sur une décomposition d'un signal en une série de formes d'onde élémentaires. Ces formes d'onde élémentaires sont obtenues par une fenêtre exactement centrée sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition/addition « *Overlap-Add* » de ces formes d'onde élémentaires. La modification de la distance relative entre deux formes d'onde, ainsi que la modification du nombre de formes d'onde, permet de modifier la hauteur et l'axe temporel du signal.

Il existe deux catégories de la méthode PSOLA

- La méthode TD-PSOLA, ($\mu = 2$) « *Time Domain- Pitch Synchronous Overlap-Add* ».
- La méthode FD-PSOLA, ($\mu = 4$) « *Frequency Domain- Pitch Synchronous Overlap-Add* ».

La décomposition du signal en formes d'onde élémentaires est effectuée par multiplication du signal s(n) par une fenêtre de pondération h(n) centrée en des temps m_i appelés « marques d'analyse». Ces marques d'analyse sont positionnées de manière synchrone à la période fondamentale locale du signal. Soient $s_i(n)$ la $i^{ème}$ forme d'onde élémentaire, et m_i la $i^{ème}$ marque d'analyse.

$$s_i(n) = h_i(n - m_i).s(n) \tag{3.1}$$

En notant $h(n)$ La fonction de pondération de longueur normalisée à l'unité, la fenêtre servant au découpage du signal en formes d'onde élémentaires s'exprime :

$$h_i(n) = h(\frac{n}{\mu T(m_i)}) \quad (3.2)$$

Dans lequel $T(m_i)$ désigne la période fondamentale autour du temps m_i [25].

Dans la méthode TD-PSOLA, chaque forme d'onde élémentaire renferme deux périodes fondamentales. Du fait de la largeur fréquentielle de la fenêtre de découpage, le spectre de chaque forme d'onde élémentaire $s_i(n)$ de TD-PSOLA peut être considéré comme une approximation de l'enveloppe spectrale (enveloppe spectrale convoluée par la réponse fréquentielle de la fenêtre d'analyse). Cette approximation provoque cependant un étalement des formants (lissage des résonances étroites), même si celui-ci est difficilement perceptible [26]. La qualité de cette approximation dépend cependant fortement du positionnement de la fenêtre par rapport au signal ainsi que de la nature du signal (dans le cas d'un modèle source/filtre, l'approximation dépend de la durée effective de la réponse impulsionnelle du filtre, du facteur d'atténuation, par rapport à la période fondamentale). Cette interprétation des formes d'onde élémentaires de TD-PSOLA en termes d'enveloppe spectrale est illustrée à la (Figure 3.1), où le spectre d'amplitude d'une forme d'onde élémentaire TD-PSOLA est comparé à la réponse fréquentielle du filtre LP « *linear-predection* » du signal [25].

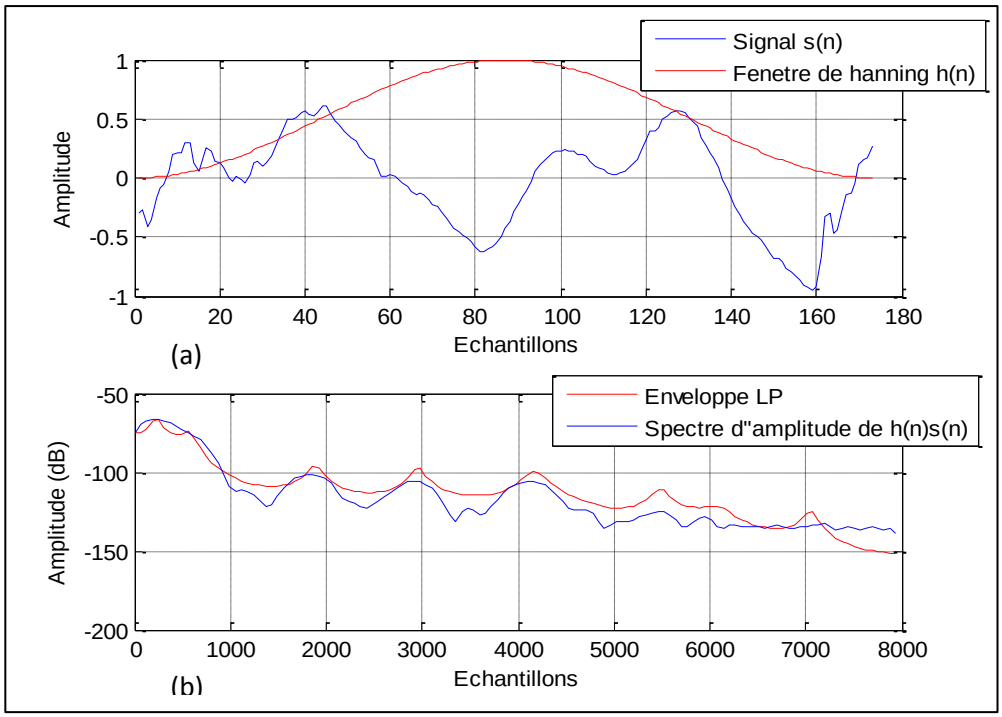


Figure 3. 1 : Méthode TD-PSOLA, (a) signal et fenêtrage, (b) le spectre d'amplitude et l'estimation d'enveloppe spectrale.

Dans la méthode FD-PSOLA, chaque forme d'onde élémentaire renferme quatre périodes fondamentales.

Du fait de la durée de la fenêtre de découpage, le spectre de chaque forme d'onde élémentaire $s_i(n)$ présente une structure fine dans laquelle les harmoniques sont résolues.

Du fait de la résolution des harmoniques le FD-PSOLA ne permet pas d'interprétation directe en termes d'approximation de l'enveloppe spectrale, ni en termes de décomposition source/filtre. De ce fait, une modification de la hauteur du signal nécessite une correction des spectres des formes d'onde élémentaires .Cette correction introduit un certain nombre de nouveaux problèmes tels que la transposition des caractéristiques harmoniques/non-harmoniques, la transposition des détails fins du spectre, des relations de phase dans de nouvelles bandes de fréquence, ou l'apparition de trous dans le spectre. Des solutions ont cependant été proposées pour ces problèmes [26]. Un autre problème inhérent au fenêtrage large (μ=4) de FD-PSOLA est l'apparition d'une certaine réverbération dans le signal, (Tableau 3.2) [25].

	Avantages	Inconvénients
TD-PSOLA	coût de calcul faible, signal peu réverbérant	limité à un certain type de signal (atténuation rapide du filtre), nécessité d'un marquage à la fois synchrone à la période fondamentale et contraint par l'énergie locale
FD-PSOLA	ne nécessite pas de formes d'onde localisées temporellement	signal plus réverbérant nécessité d'une correction du spectre lors d'une modification de hauteur

Tableau 3. 2 : Avantage et Inconvénients des méthodes PSOLA [25].

3.3 Principe de la méthode TD-PSOLA

La méthode TD- PSOLA permet d'opérer des modifications prosodiques sur le signal de parole telles que la dilatation/compression de durée ou le changement de la F_0 tout en conservant une bonne qualité sonore. Les transformations prosodiques de la voix ne sont accompagnées d'aucune modification du timbre [57].

La méthode TD-PSOLA peut être réalisée en trois étapes fondamentales :

- Analyse,
- Modification,
- Et synthèse.

3.3.1 Analyse

a Extraction des zones voisées

L'extraction des zones voisées a une grande importance pour la modification de la F_0 (voir chapitre 4). De plus si la classification voisées /non voisées est erronée, la qualité du signal synthétique est fortement dégradée. Les artéfacts les plus audibles se produisent dans les zones voisées ou mixtes. Ils sont essentiellement dus à une mauvaise représentation de l'évolution des paramètres de voisement [58].

Dans les zones voisées la F_0 est bien définie, tandis que dans les zones non voisées elle n'est pas définie (elle est généralement assignée à une fréquence nulle). Pour cette raison les zones non voisées seront recopiées directement à la sortie (signal reconstitué).

b Découpage du signal en formes d'onde élémentaires

Le signal s(n) est découpé en formes d'onde élémentaires par une fenêtre h(n) exactement centrée sur les périodes fondamentales m_i (Equations 3.1, 3.2). Les marques d'analyse m_i déterminent le centre des fenêtres de découpage. Chaque fenêtre est définie sur une longueur égale à deux fois les périodes fondamentales locales (du signal original) [25]. La fenêtre utilisée est de type Hanning (voir Annexe B).

c Positionnement des marques d'analyse

Les marques d'analyse représentent des éléments importants pour réaliser la modification de F_0. Elles sont placées sur le signal d'analyse. Dans les zones voisées, le marquage s'effectue suivant le respect de deux contraintes :

- le centrage sur les maximas d'énergie (pour ne pas détériorer le signal après fenêtrage),
- le respect d'une distance entre deux fenêtres proche de la période fondamentale.

Le marquage du fondamentale n'est effectué que sur les zones voisées.

En tenant compte de ces deux contraintes, les marques d'analyse peuvent être positionnées sur trois instants différents du signal :

- les instants de fermeture de la glotte « *Glottal Closure Instants* » (GCIs) : correspondent aux pics globaux du signal résiduel (Ce signal est obtenu par une analyse LPC « *Linear Predictive Coding »* inverse*)* (Figure 3.2) [59].
- les instants qui correspondent aux pics globaux obtenus sur le signal vocal (Figure 3.3) [60].
- les instants qui correspondent aux vallées globales obtenues sur le signal vocal (Figure 3.4) [60].

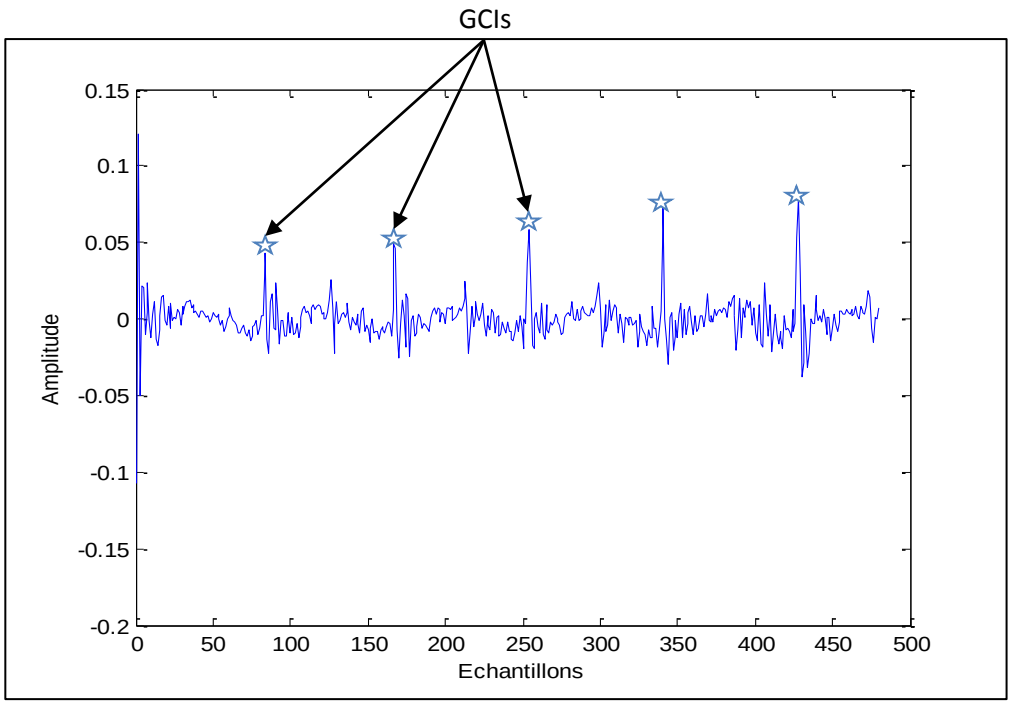


Figure 3. 2 : Positionnement des marques d'analyse qui correspondent aux GCIs du signal résiduel.

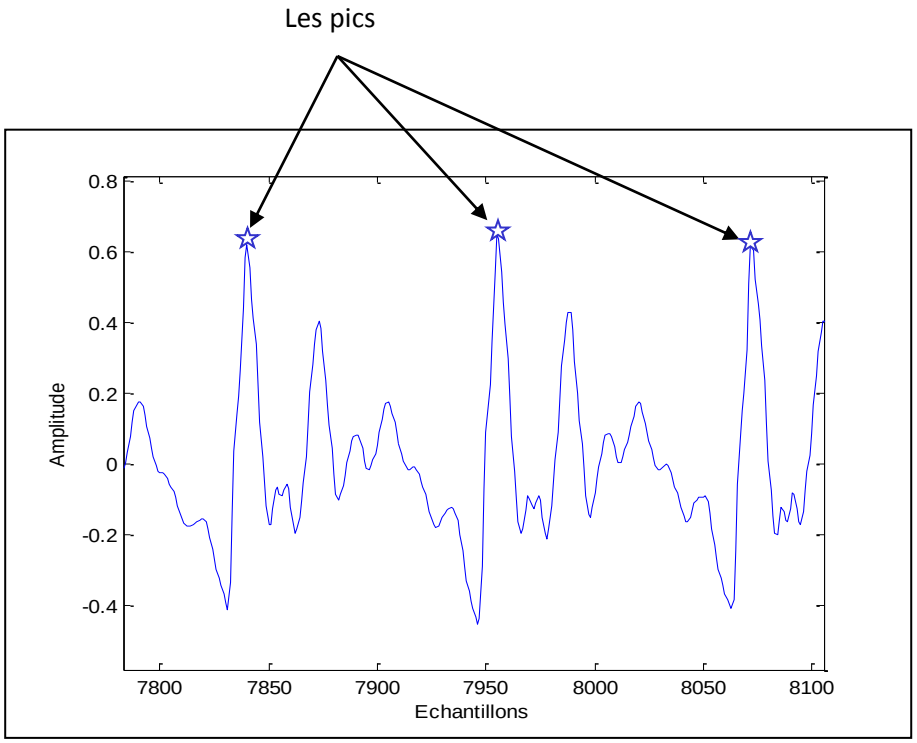


Figure 3. 3 : Positionnement des marques d'analyse qui correspondent aux instants des pics globaux.

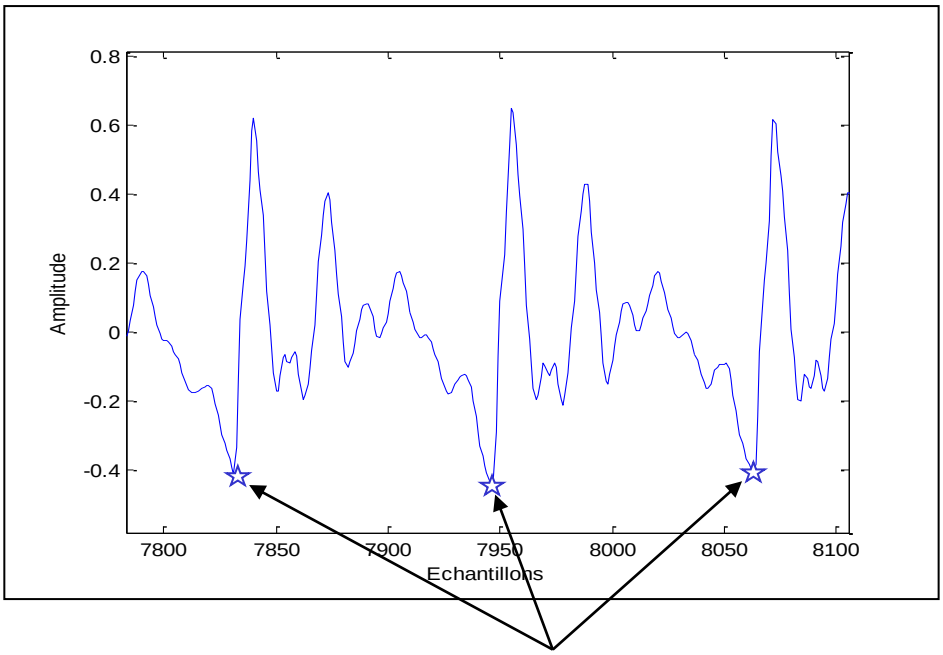


Figure 3. 4 : Positionnement des marques d'analyse qui correspondent aux instants des vallées globales.

Une solution pratique qui repose sur l'extraction des instants des pics globaux va être exposée dans le chapitre prochain.

3.3.2 Modification

Le principe général de modification des paramètres prosodiques à base de la TD-PSOLA consiste à extraire les marques de synthèse à partir des marques d'analyse. Ces marques de synthèse vont être utilisées par la suite pour la construction du signal modifié.

Dans le cadre de notre travail on va se focaliser seulement sur la modification de F_0.

$$F_0' = \alpha \times F_0 \tag{3.3}$$

Où

F_0' : La fréquence fondamentale de synthèse,

F_0 : La fréquence fondamentale,

α : Le facteur de modification varie entre 0.1 et 2.

Le but de l'étape de modification est de trouver un ensemble de marques synthétiques en fonction du facteur de modification voulu de telle sorte que la durée totale des zones voisées ne sera pas modifiée.

Dans la littérature plusieurs approches ont été proposées pour l'extraction des marques de synthèse. On peut citer :

- Les travaux de G. Peeters [25], qui a introduit la notion des marques de correspondances. Il s'agit des instants virtuels de liaison entre l'étape d'analyse et celle de synthèse. Il s'est intéressé à la modification simultanée des deux paramètres prosodiques,
- Les travaux de E. Moulines et J. Laroche [27] : Ils ont introduit la notion des périodes fractionnelles en tenant compte du facteur de modification voulu. Ils se sont intéressés aussi à la modification simultanée des deux paramètres prosodiques.

Dans notre travail on s'est inspiré des deux approches proposées afin de créer une nouvelle approche pour modifier la F_0 (sans changer la durée totale du signal).

Le positionnement des marques d'analyse est effectué sur les instants qui correspondent aux pics globaux obtenus sur la forme d'onde de la voix (Figure 3.3).

La première étape de notre approche consiste à calculer les périodes à court terme des zones voisées (calculer des distances entre deux marques consécutives T). La première marque de synthèse correspond à la première marque d'analyse.

Le positionnement de la deuxième marque de synthèse est obtenu par l'ajout d'une fraction de période à la première marque d'analyse (voir équation 3.5).

Pour achever le positionnement des marques de synthèse restantes, l'organigramme de la Figure 3.5 est utilisé.

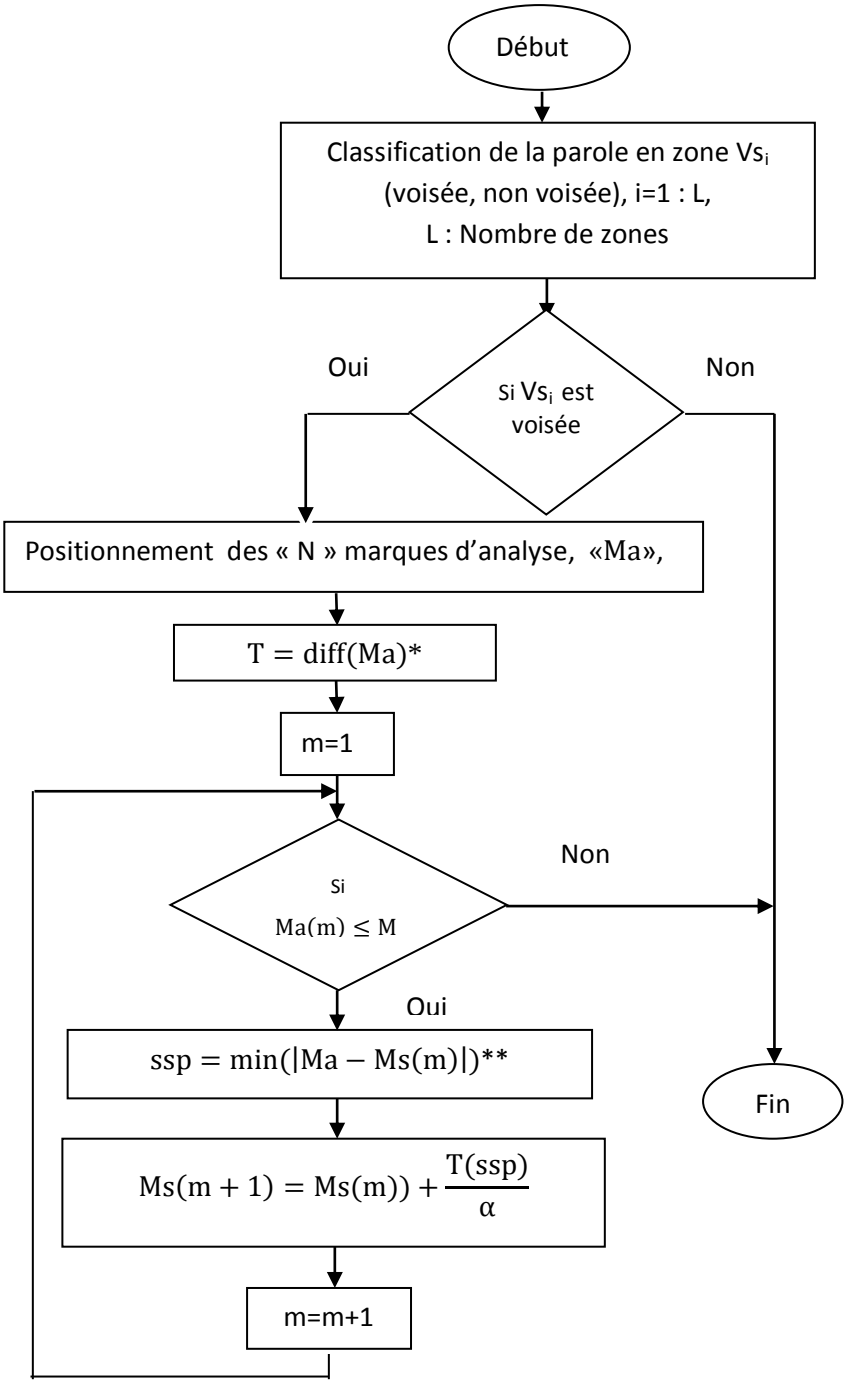


Figure 3. 5 : Organigramme de modification de la fréquence fondamentale.

Avec :

α: Le facteur de modification de la F_0 (varie entre 0.1 et 2).

M : la taille de signal vocal.

*diff est une fonction qui calcule la différence entre deux éléments adjacents d'un vecteur donné :

$$\boldsymbol{diff(Ma) = Ma(n+1) - Ma(n)} \quad (3.4)$$

Pour n allant de 1 jusqu'à N-1

N : le nombre de marques analytiques.

**min : est une fonction qui permet de trouver le point le plus proche de la $m^{ième}$ marque synthétique (Ms) parmi toutes les marques analytiques (Ma).

$$\mathrm{Ms(m)} = (\mathrm{Ma(1)}, \mathrm{Ma(1)} + \frac{\mathrm{T(1)}}{\alpha}) \qquad (3.5)$$

La Figure 3.6 représente une opération d'extraction des formes d'onde élémentaire d'une zone voisée.

La Figure 3.7 représente une opération d'augmentation de la F_0 sans modification de l'axe temporel. Cette augmentation est obtenue par duplication d'une même forme d'onde élémentaire.

La Figure 3.8 représente une opération d'abaissement de la hauteur par l'élimination de certaines formes d'onde élémentaires.

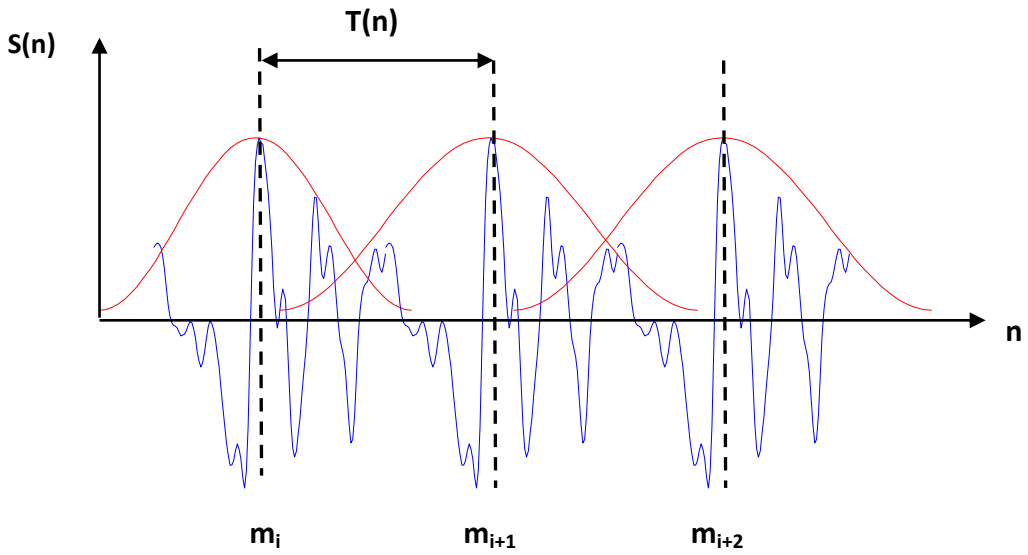


Figure 3. 6 : Positionnement des marques d'analyse sur le signal original.

$\mathbf{m_i}$ correspond aux marques analytiques « Ma ».

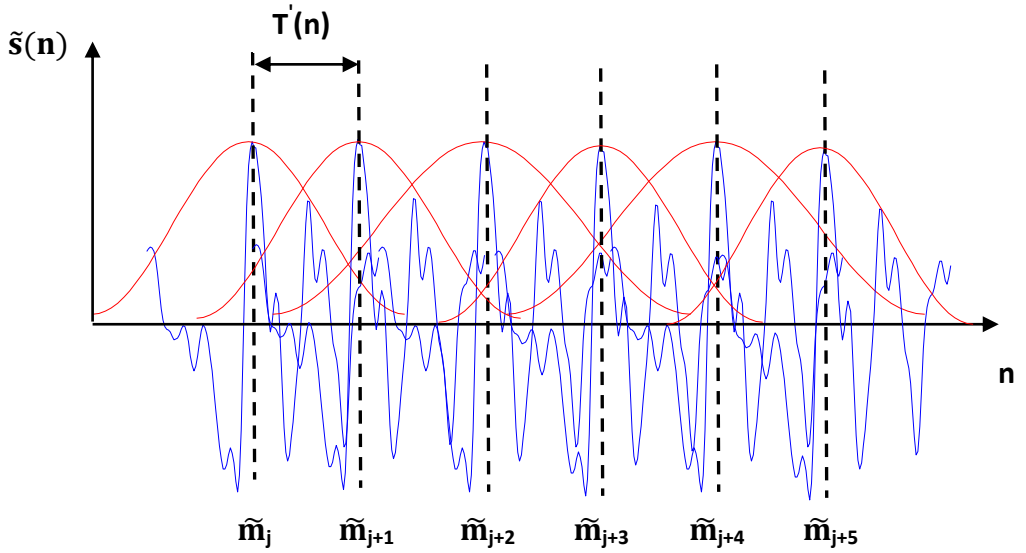


Figure 3. 7 : Duplication des formes d'onde élémentaires.

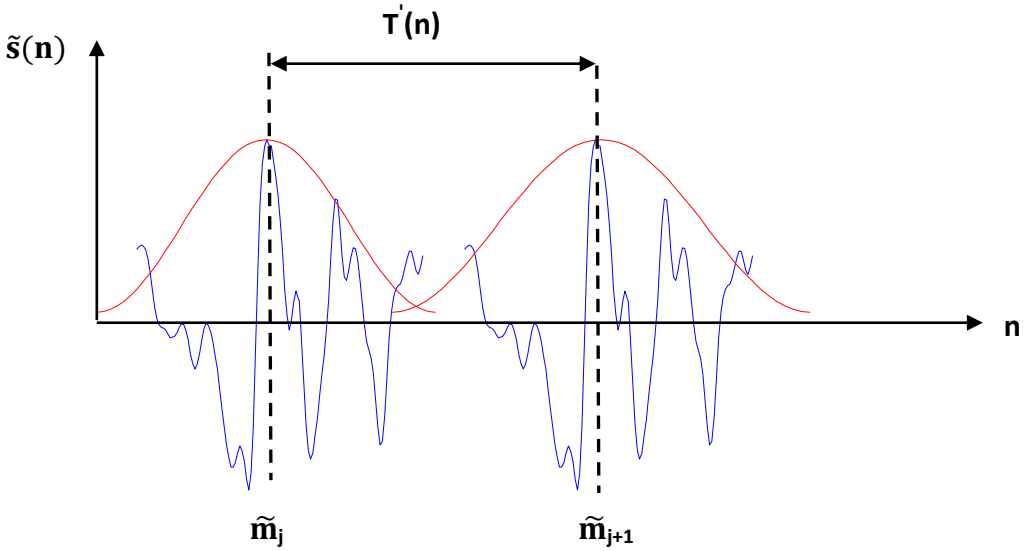


Figure 3. 8 : Elimination des formes d'onde élémentaires.

$\tilde{m}_j$ correspond aux marques synthétiques « Ms ».

3.3.3 Synthèse

Le signal de synthèse est construit par superposition/addition des formes d'onde élémentaires placées dans de nouvelles positions $\tilde{m}_j$ appelées marques de synthèse.

Ces marques de synthèse sont déterminées par les modifications voulues du signal : modification de hauteur [25].

$$\tilde{s}(n) = \sum_j \tilde{s}_j\,(n - \tilde{m}_j) \tag{3.6}$$

Où $\tilde{s}(n)$ représente le signal de synthèse et $\tilde{s}_j(n)$ désigne les formes d'onde élémentaires successives du signal de synthèse.

Il existe d'autres types de superposition/addition qui sont suivis par une procédure de normalisation, de manière à tenir compte du facteur de recouvrement éventuellement variable au cours du temps. Les deux types de normalisation sont :

- ***Méthode d'Allen :*** Il s'agit de la procédure normale de normalisation faisant suite à une méthode de superposition/addition [25], dans lequel $\boldsymbol{s_j(n)}$ désigne les formes d'onde élémentaires successives du signal de synthèse, et $\boldsymbol{f_j(n)}$ la fenêtre de synthèse.

$$\tilde{s}(n) = \frac{\sum_j s_j(n)}{\sum_j f_j(m_j - n)} \tag{3.7}$$

- ***Méthode de Griffin :*** La méthode de Griffin est obtenue par minimisation de l'erreur énergétique entre spectre du signal original et spectre modifié correspondant *à* $\boldsymbol{y_j}$.Cette méthode n'est utile que lorsque des modifications fréquentielles sont appliquées [25].

$$\tilde{s}(n) = \frac{\sum_m f_j(m_j - n) y_j(n)}{\sum_m f_j^2(m_j - n)} \tag{3.8}$$

3.4 Conclusion

On a présenté au niveau de ce troisième chapitre un bref état de l'art de la méthode PSOLA. On a également étudié les différentes étapes de la méthode TD-PSOLA. Cette méthode est utilisée pour la modification de la fréquence fondamentale du signal vocal. Les étapes d'implémentation de cette dernière ainsi que les résultats obtenus vont être exposés dans le prochain chapitre.

Chapitre 4 Résultats et simulations

4.1 Introduction

Après avoir présenté dans le chapitre précédent le principe de la méthode TD-PSOLA « *Time Domain Pitch Synchronous Overlap-Add* » pour la modification de la fréquence fondamentale, on va présenter dans ce chapitre les résultats et simulations de cette modification sur des signaux réels. On commence ce chapitre par la description des moyens et logiciels utilisés ainsi que les signaux de tests choisis. Par la suite, on va détailler les étapes d'implémentation de la TD-PSOLA en utilisant une approche symétrique. On va aussi proposer une technique de marquage du fondamental en se basant sur la forme d'onde du signal.

Dans la dernière partie de ce chapitre, on donnera une évaluation des performances de la TD-PSOLA par des tests objectifs et subjectifs. Une conclusion terminera ce chapitre.

4.1.1 Moyens et logiciels utilisés

Les outils qu'on utilise pour l'implémentation sont :

- une base de données de sons de parole pour le choix des signaux de test,
- un outil de visualisation et classification automatique des sons de la base de données, nommé « *WaveSurfer* » [61],
- le logiciel Matlab pour l'implémentation de la méthode TD-PSOLA.

4.1.2 Base de données

L'étude présentée dans ce chapitre a été réalisée sur la base de données (BD) CMU ARCTIC « *Carnegie Mellon University ARCTIC*» [62]. Cette BD contient 1138 fichiers de parole, prononcés en Anglais Américain (AA) par deux locuteurs de sexe différent (clb US Femme/awb Homme). Elle est échantillonnée à 16 KHz.

Les fichiers de sons englobent le maximum de phonèmes de toutes les classes de son étudiées au premier chapitre. La durée totale de cette BD est de 2h 52 mn.

Dans notre étude, on a choisi 36 fichiers de parole, dont 18 prononcés par un locuteur masculin et 18 par un locuteur féminin. La durée moyenne des sons est environ une minute. La durée totale des fichiers utilisés est de 38 minutes.

4.1.3 WaveSurfer

WaveSurfer (Figure 4.1) est un outil « *Open Source* » pour la visualisation et la manipulation des sons, il peut être utilisé comme un outil autonome pour un large éventail de tâches dans la recherche et l'éducation. Les applications typiques sont l'analyse de la parole, l'annotation et la transcription des sons. Il comprend un ensemble complet d'effets et de fonctionnalités d'édition et de production professionnelles. De plus, c'est un logiciel très simple à utiliser.

Dans le cadre de notre travail, on a exploité ce logiciel pour la classification automatique des sons de la parole en régions voisées et non voisées. La classification des sons par WaveSurfer repose sur la technique d'intercorrélation normalisée raffinée par la programmation dynamique [61].

La version utilisée est « *WaveSurfer v1.8.8p4* ».

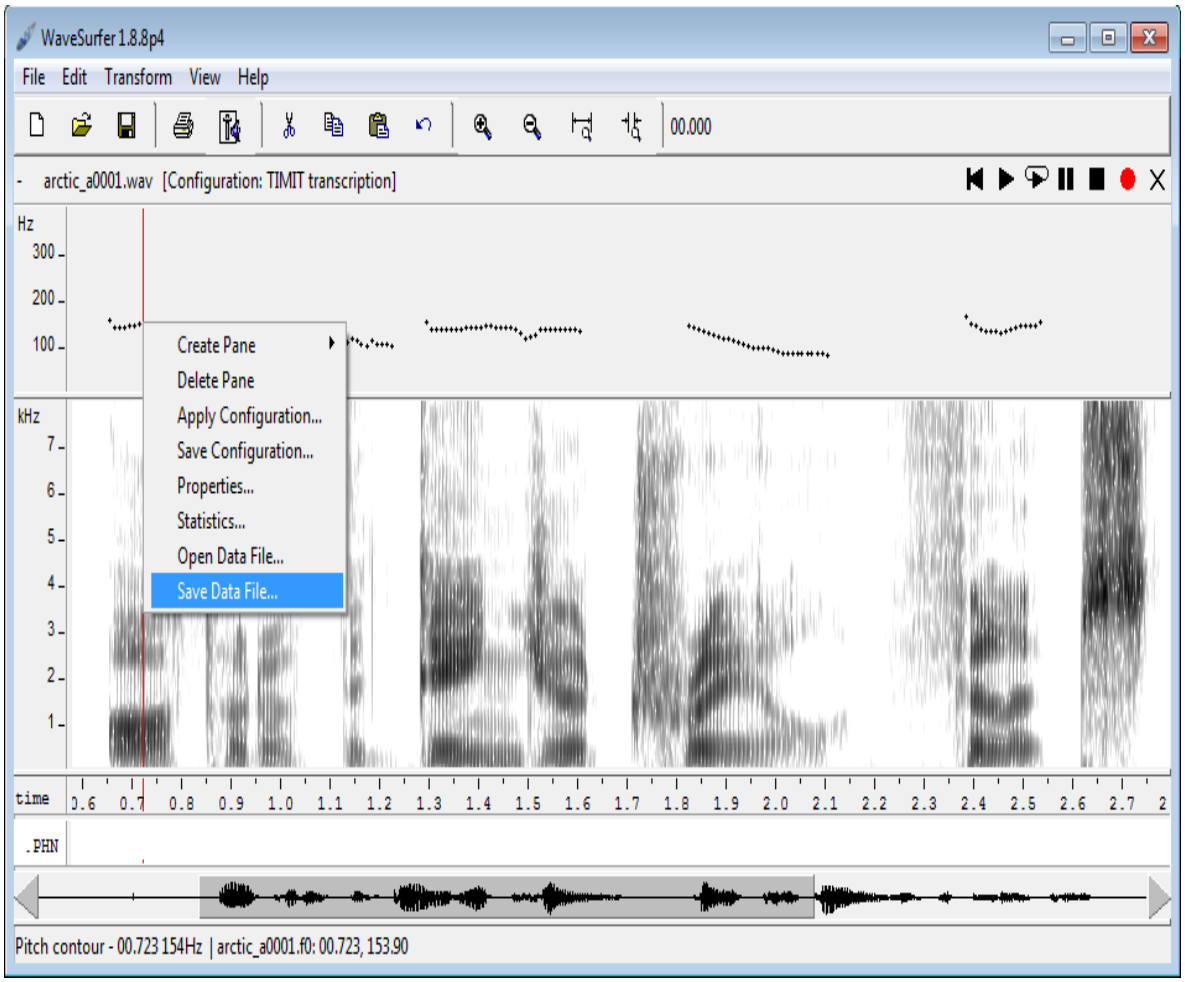


Figure 4. 1 : Environnement du logiciel WaveSurfer.

4.1.4 Matlab

Matlab est un puissant outil de calcul numérique, de programmation et de visualisation graphique. Son nom signifie *MATrix LABoratory*, c'est à dire un environnement interactif de travail avec des matrices. La facilité de développement des applications dans son langage fait de lui, pratiquement, le standard dans son domaine. Il fonctionne dans plusieurs environnements tels que Windows et Macintosh.

Matlab offre également plusieurs fonctions destinées à : la résolution (numérique) d'équations différentielles linéaires ou non-linéaires, l'intégration numérique, la recherche des solutions d'équations algébriques, la création et manipulation de polynômes, etc… (Figure 4.2).

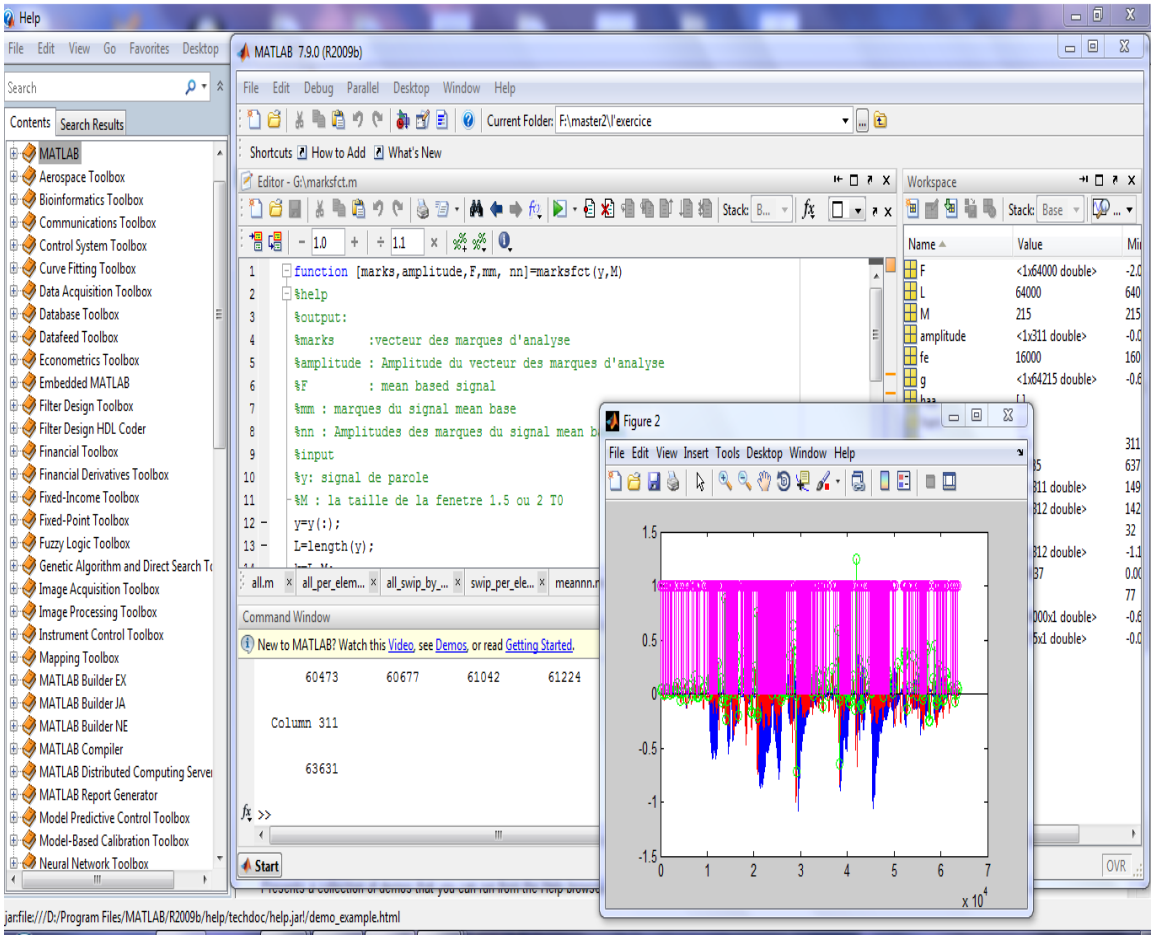


Figure 4. 2 : Environnement du Matlab.

SPTool « *Signal Processing Tool* » est une interface utilisateur graphique (GUI) de matlab, gérant quatre autres GUIs : Signal Browser, Filter Design and AnalysisTool, FVTool, and Spectrum Viewer (Figure 4.3). Ces GUIs offrent un accès à plusieurs fonctions d'analyse de signaux, filtres et spectre. Sptool est utilisée dans notre application pour l'ajustement du vecteur de classification de sons.

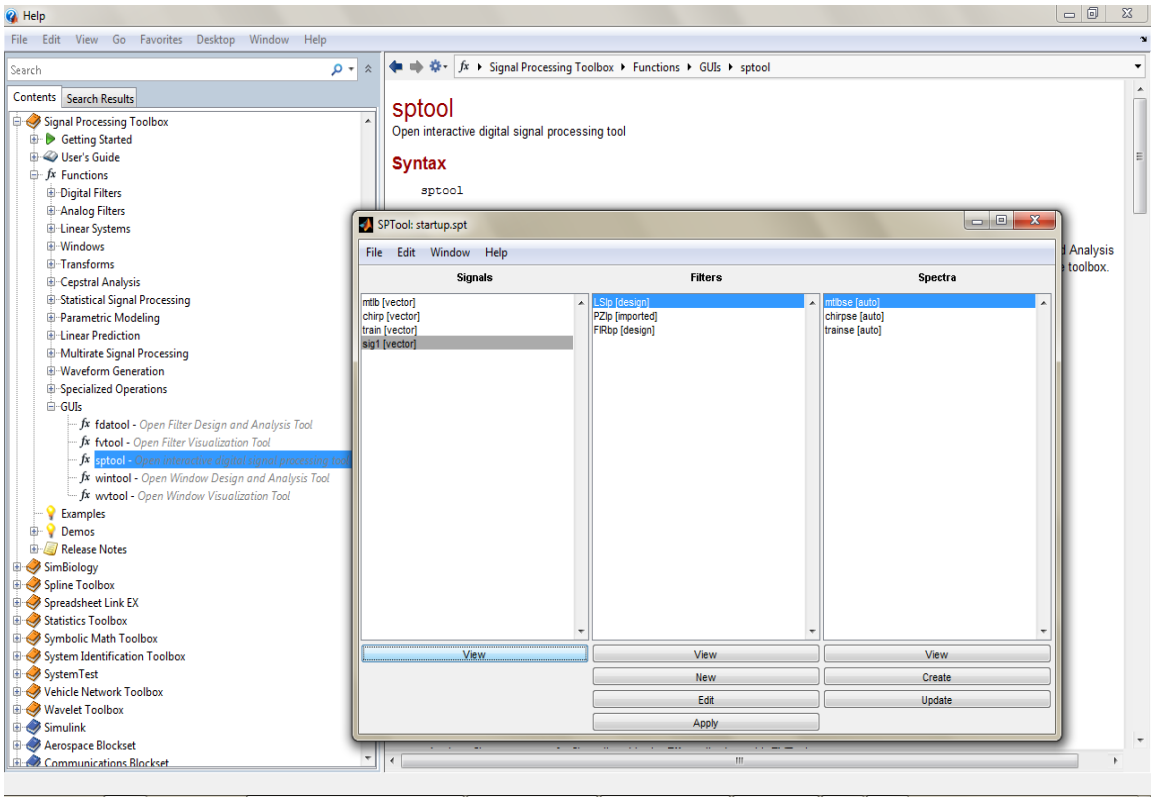


Figure 4. 3 : Environnement de SPtool.

4.2 Classification de la parole

La classification de la parole en régions (voisées et non voisées) est une étape très importante dans l'implémentation de la méthode TD-PSOLA. Les performances de cette dernière reposent sur l'exactitude de cette étape.

On a utilisé le software *WaveSurfer* pour avoir une classification préliminaire de notre BD. Les résultats obtenus par cet outil ne sont pas complètement fiables. Plusieurs erreurs de classification peuvent avoir lieu. Pour cela on était obligé de corriger ses erreurs manuellement à l'aide de l'outil SPtool de Matlab. Cette correction se base sur les caractéristiques acoustiques des sons voisés et non voisés étudiées dans le premier chapitre.

La classification est effectuée en utilisant des trames de parole d'une durée de 20 ms. On a obtenu un vecteur de classification Vs tel que :

$$V_s = \begin{cases} 1 \text{ si les zones sont voisées} \\ 0 \text{ si les zones sont non voisées, silences ou mixtes.} \end{cases} \quad (4.1)$$

La Figure (4.4) représente un exemple de classification d'un signal vocal.

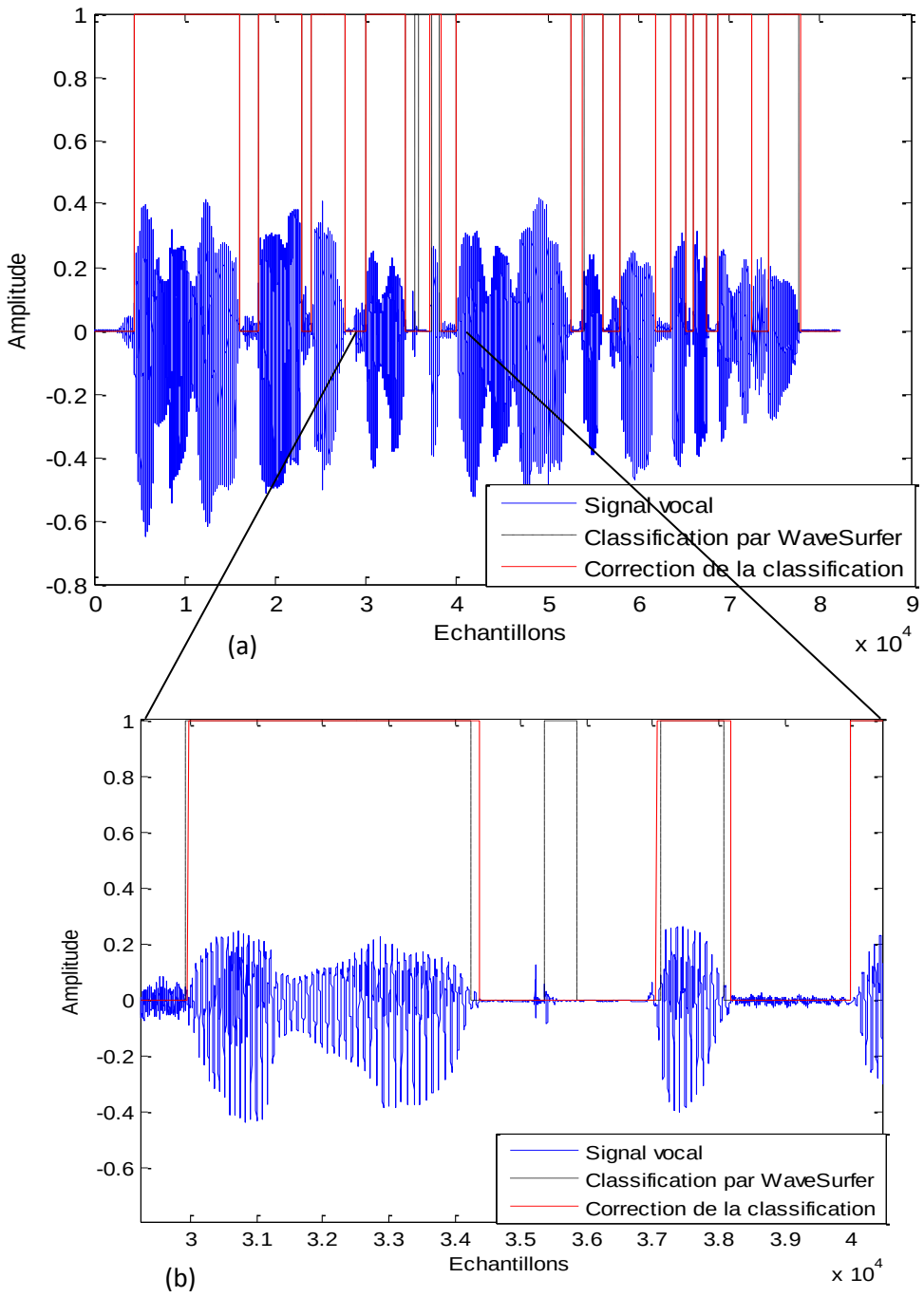


Figure 4. 4 : (a) Correction du vecteur de classification, (b) Zoom sur la région voulue**.**

4.3 Implémentation de la méthode TD-PSOLA

Comme on l'a déjà présenté dans le chapitre trois, la méthode TD-PSOLA est réalisée en trois étapes comme suit :

4.3.1 Analyse

Consiste à extraire en premier lieu les zones voisées du signal vocal, pour la modification de la F_0. Cette étape est achevée par une classification semi-automatique comme il a été expliqué au paragraphe précèdent.

L'extraction des formes d'onde élémentaires d'analyse nécessite une opération de marquage optimal du fondamental. Dans notre travail, on a exploité un signal moyenneur dit en Anglais « *Mean Based Signal* » (MBS) pour localiser les instants de marquages correspondant aux pics globaux du signal vocal (Figure 4.5).

Ce signal a été proposé par T. Drugman et T. Dutoit [59] dans un but d'extraire les GCIs du signal vocal. Les pics globaux de ce dernier sont positionnés d'une façon synchrone à ceux de la forme d'onde du signal original. La distance séparant deux marques consécutives permet de donner la période fondamentale locale.

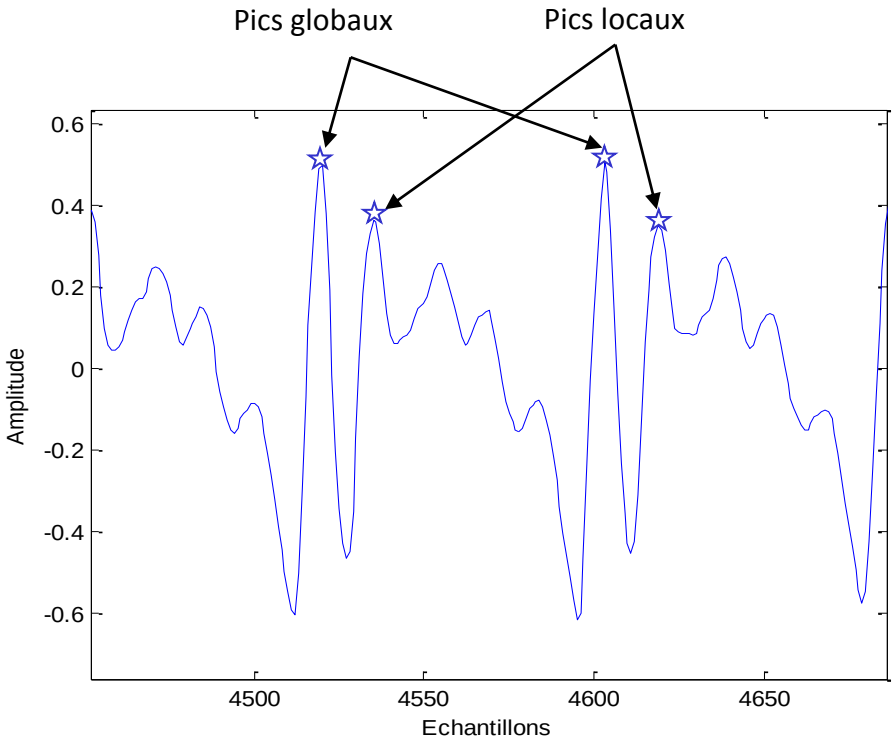


Figure 4. 5 : Pics du signal vocal.

L'organigramme de la Figure 4.6 représente les étapes de la technique de marquage utilisée pour estimer les marques d'analyse à partir du signal vocal.

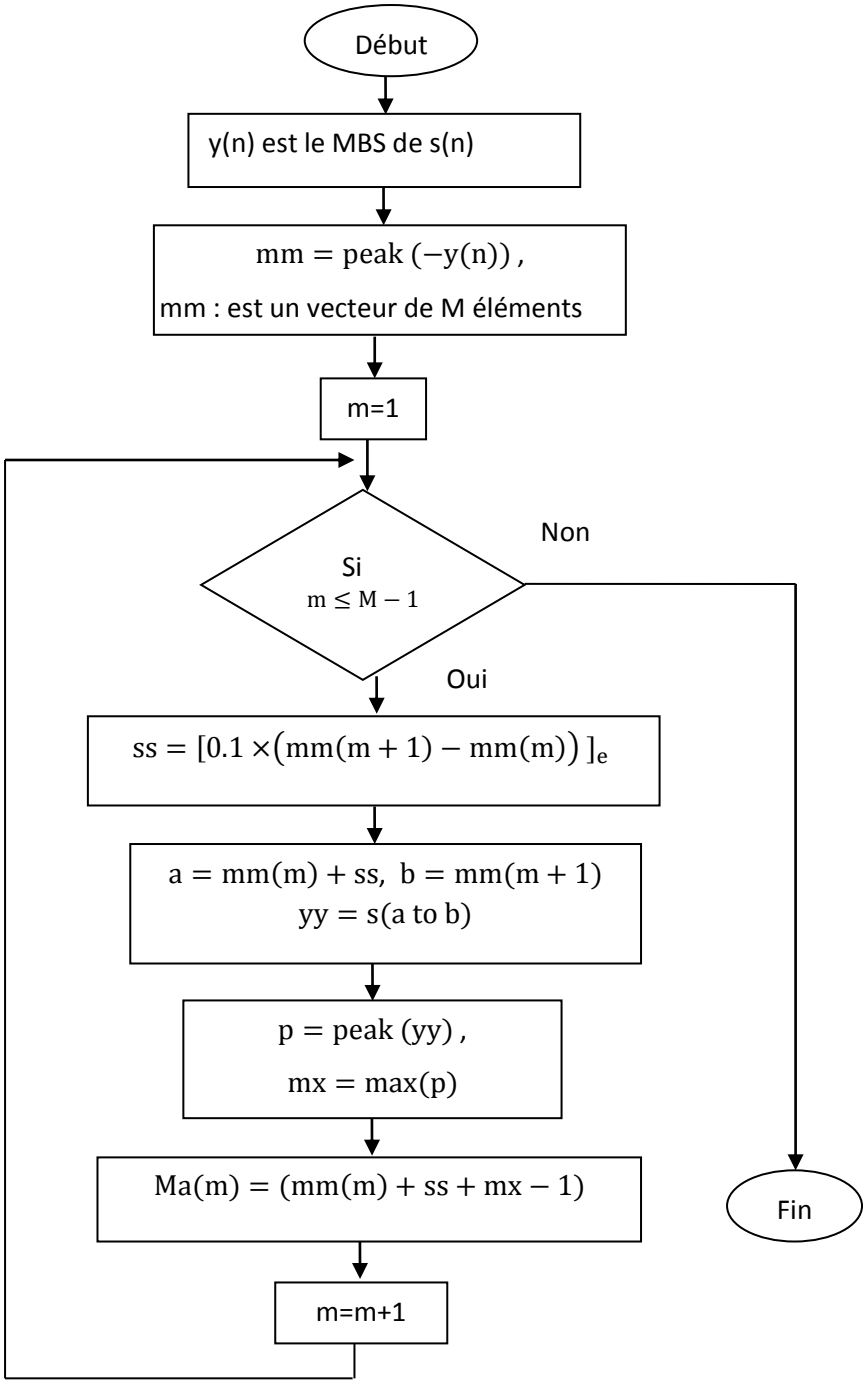


Figure 4. 6 : Organigramme d'estimation des marques d'analyse à partir du signal vocal.

Les paramètres utilisés pour cette technique sont :

y(n) : MBS de s(n) définit par la formule suivante :

$$y(n) = \frac{1}{2N+1}\sum_{m=-N}^{N} w(m)s(n+m) \qquad (4.2)$$

$s(n)$: signal vocal original,

N : est la taille de s(n),

w : Fenêtre de Blackman.

peak : est une fonction qui permet de trouver les pics du signal y(n) inversé,

ss : est une valeur de décalage. Le seuil 0.1 est choisi par des tests pratiques,

$[.]_e$: désigne la partie entière,

yy : est un signal à court terme choisi à partir de la trame d'analyse principale,

p : est un vecteur qui contient tous les pics du signal yy,

max. : est une fonction qui permet de trouver l'élément maximal d'un vecteur,

Ma : marques d'analyse.

La taille de la fenêtre est choisie entre 1.5×T et 2×T, Où T représente la période locale du signal vocal. Dans notre étude on a pris une taille fixe de 2×T.

On tient à préciser que la technique de marquage développée estime les marques d'analyse sur les zones voisées et non voisées (Figure 4.7).

A cet effet on est obligé d'éliminer les marques dans les zones non voisées à l'aide de l'information du voisement fournie par le vecteur Vs (Figure 4.8).

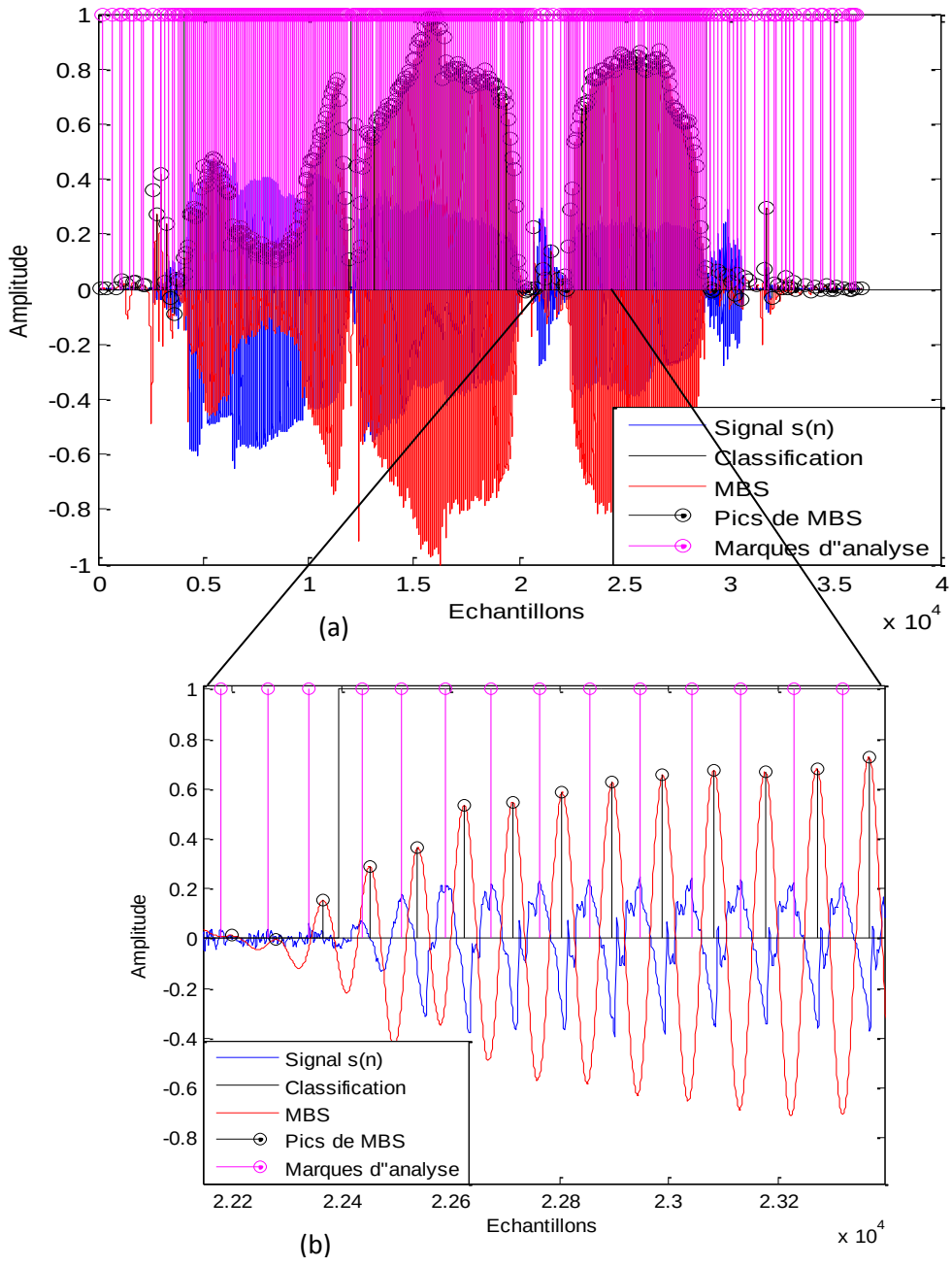


Figure 4. 7 : (a) Estimation des marques d'analyse du signal vocal, (b) Zoom sur la région voulue.

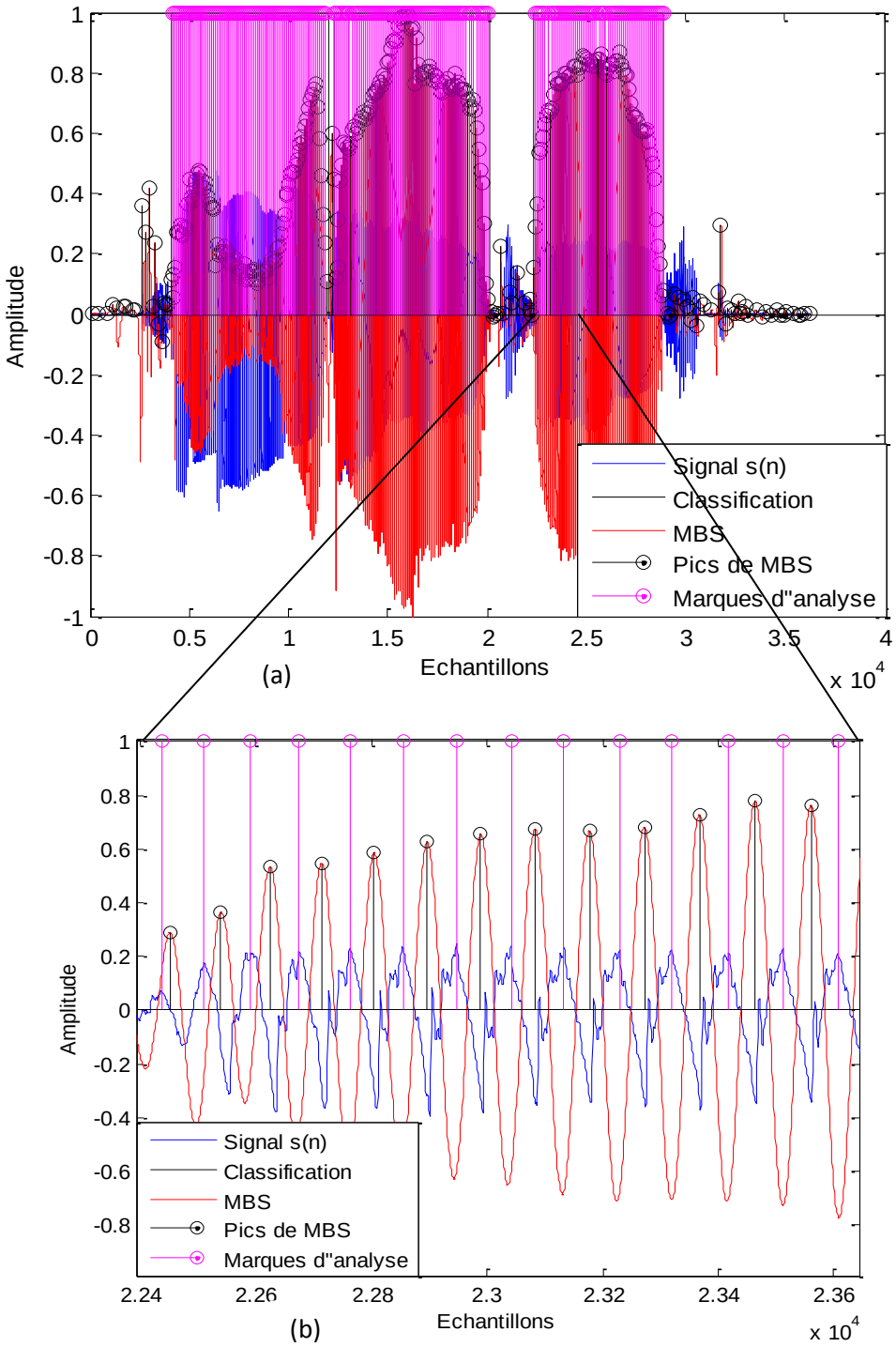


Figure 4. 8 : (a) Estimation des marques d'analyse sur une zone voisée, (b) Zoom sur la région voulue.

La Figure 4.9 représente l'opération de suppression (élimination) des marques d'analyse dans la zone non voisée (les marques situées dans la région non voisée du signal de la Figure 4.7 (b) ont été éliminées).

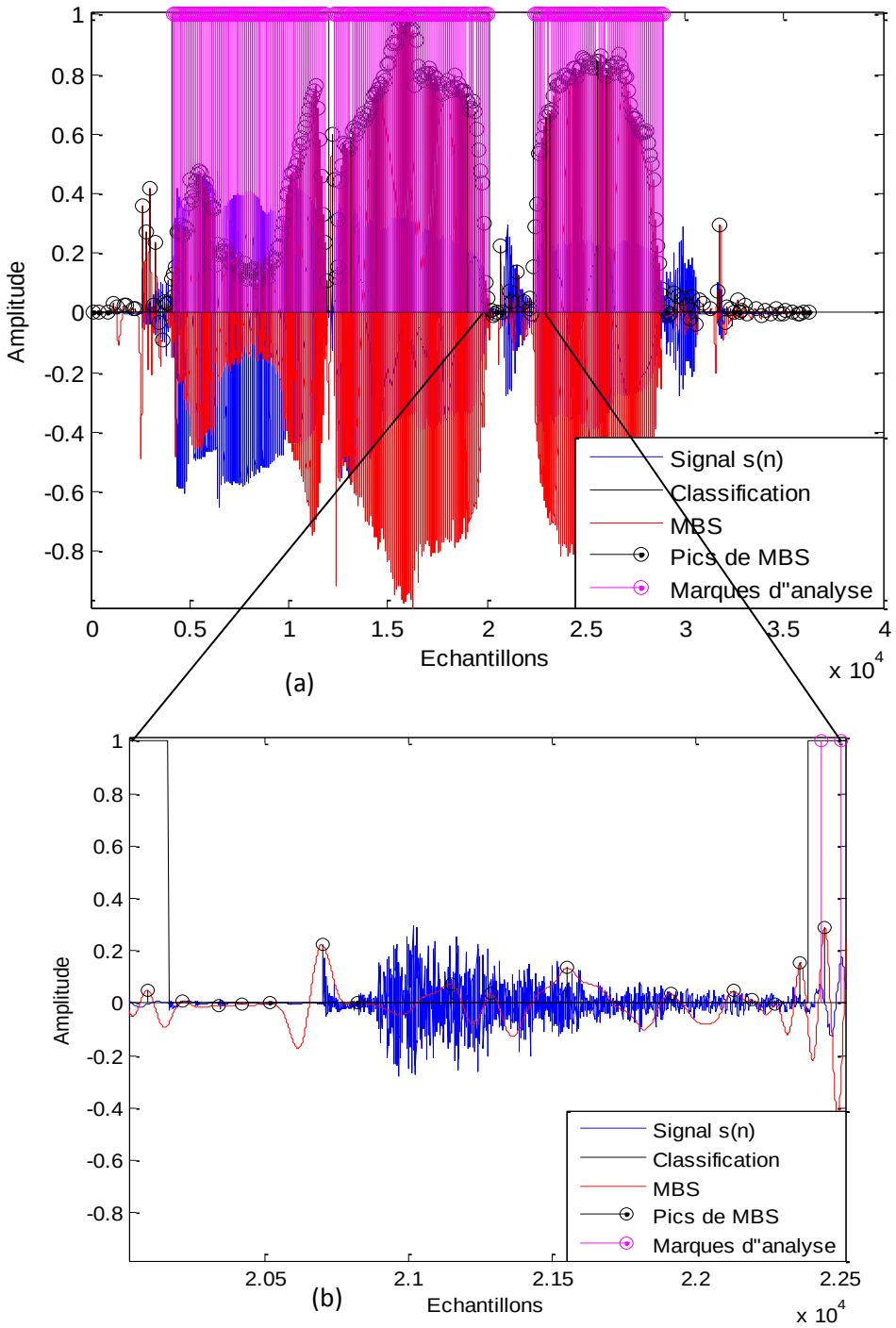


Figure 4. 9 : (a) Suppression des marques d'analyse sur la zone non voisée, (b) Zoom sur la région voulue.

La Figure 4.10 montre le positionnement final des marques d'analyse sur les pics globaux du signal vocal.

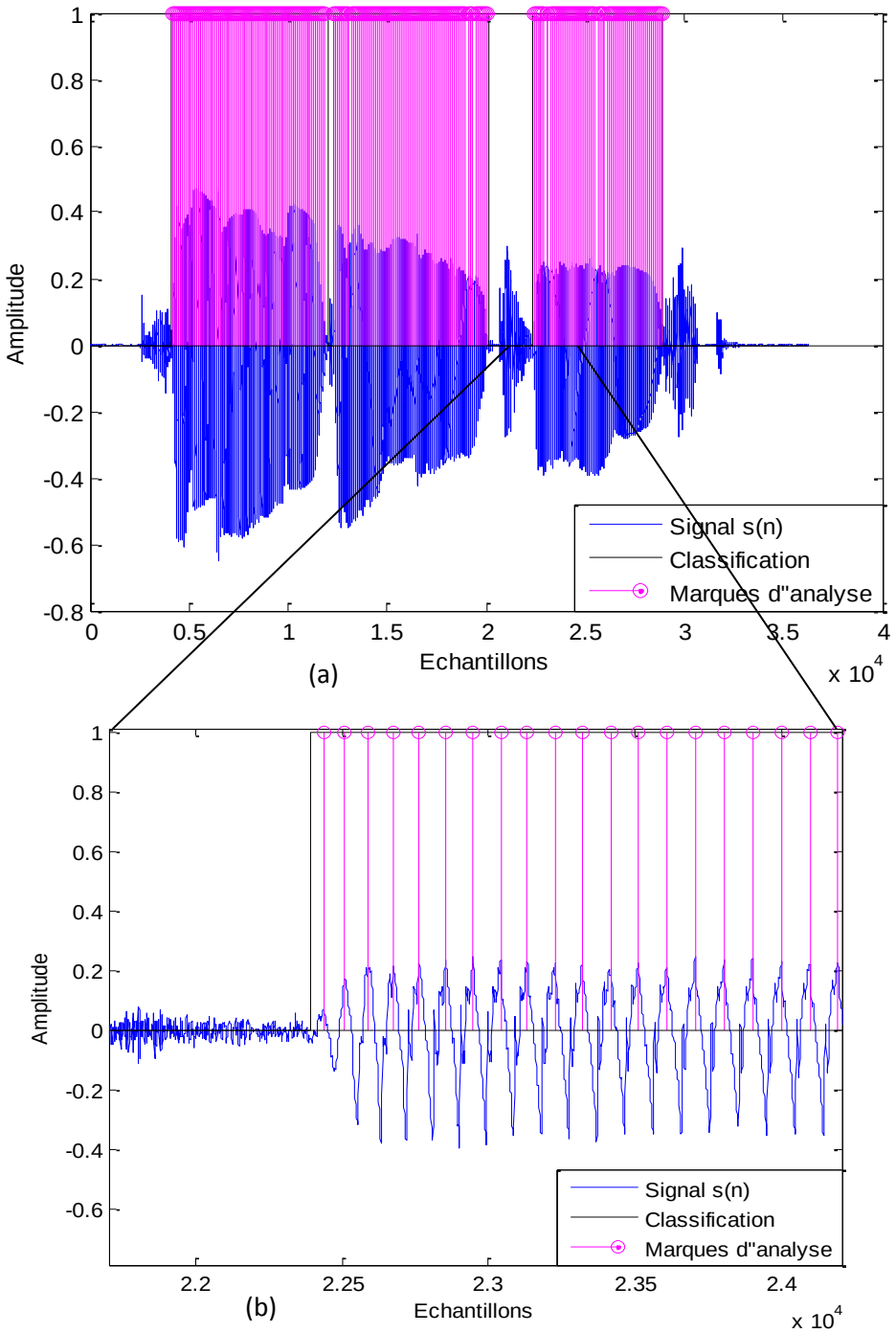


Figure 4. 10 : (a) Positionnement des marques d'analyse aux pics globaux du signal, (b) Zoom sur la région voulue.

Les formes d'onde élémentaires situées autour de chaque marque d'analyse sont alors extraites par l'utilisation d'une fenêtre temporelle de type Hanning (voir Annexe B). La durée de cette dernière est égale à deux fois la période fondamentale (Figure 4.11).

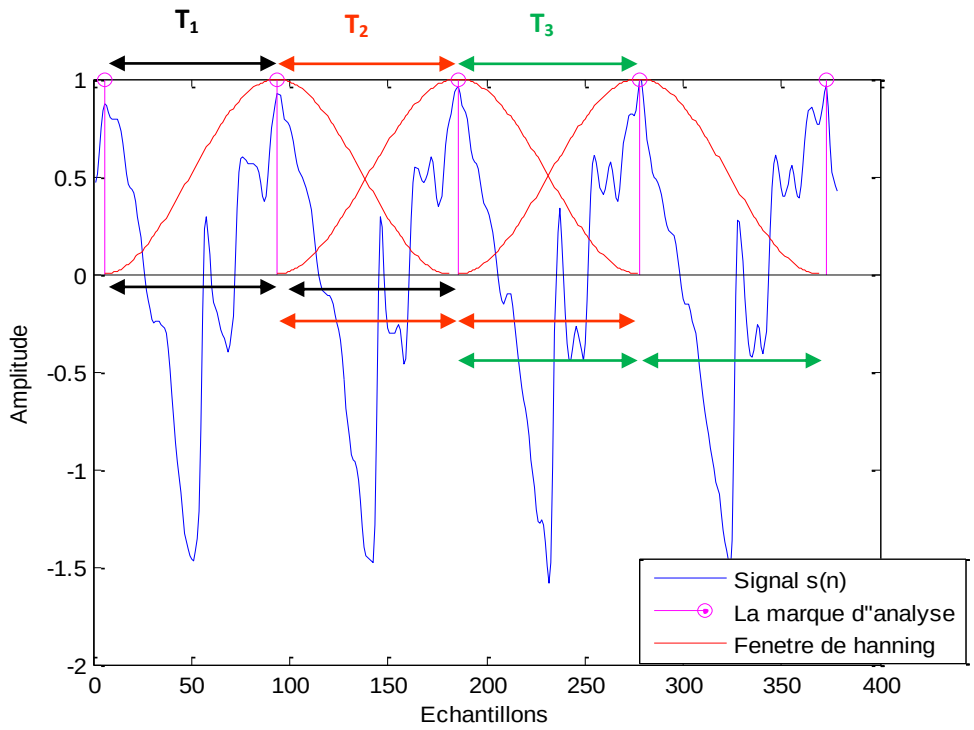


Figure 4. 11 : Extraction des formes d'onde élémentaires d'analyse. T_1, T_2 et T_3 représentent les périodes fondamentales du signal

4.3.2 Modification de la F_0

La modification de l'échelle fréquentielle par la TD-PSOLA consiste à modifier la F_0 d'un segment vocal sans altérer sa durée (voir chapitre 3, paragraphe 3.3.2).

En fonction du facteur de modification souhaité α, le calcul des marques de synthèse est effectué à partir de celles d'analyse. Il dépend des règles suivantes :

- si $\alpha = 1$, aucune modification.
- si $\alpha > 1$, certaines marques d'analyse seront dupliquées .
- si $\alpha < 1$, certaines marques d'analyse seront éliminées .

Le facteur de modification varie entre 0.1 et 2 (Equation 3.3).

On va citer maintenant quelques exemples de modification selon les valeurs de α.

La Figure 4.12 montre un exemple de calcul des marques de synthèse pour $\alpha = 1$ (aucune modification des marques d'analyse n'est observée).

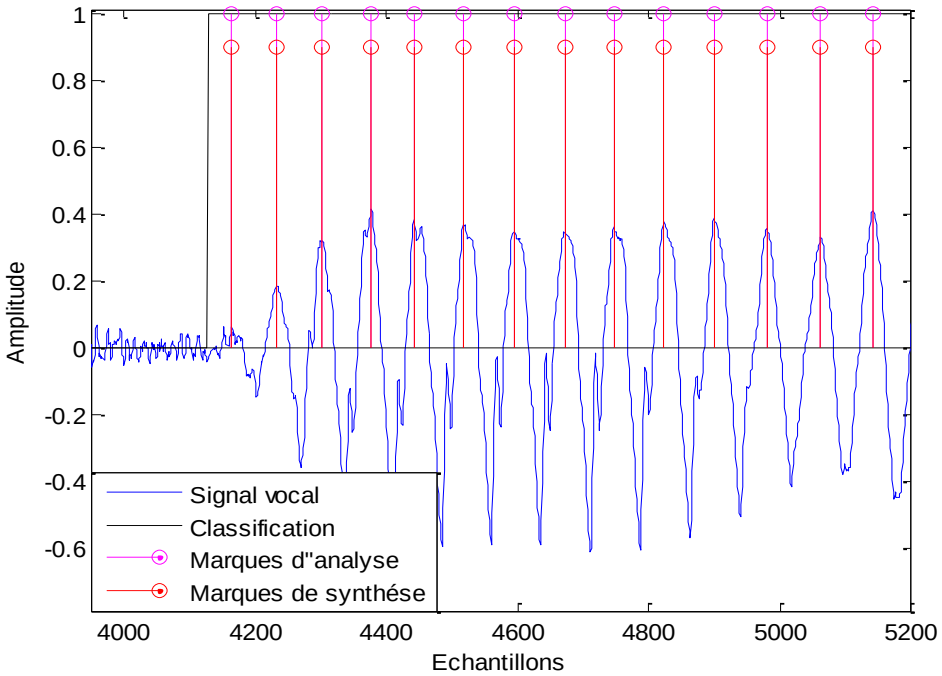


Figure 4. 12 : Positionnement des marques analyse-synthèse pour $\alpha = 1$.

La Figure 4.13 montre un exemple de calcul des marques de synthèse pour $\alpha = 1.5$. On observe que chaque deux périodes à l'étape d'analyse deviennent trois périodes après la modification.

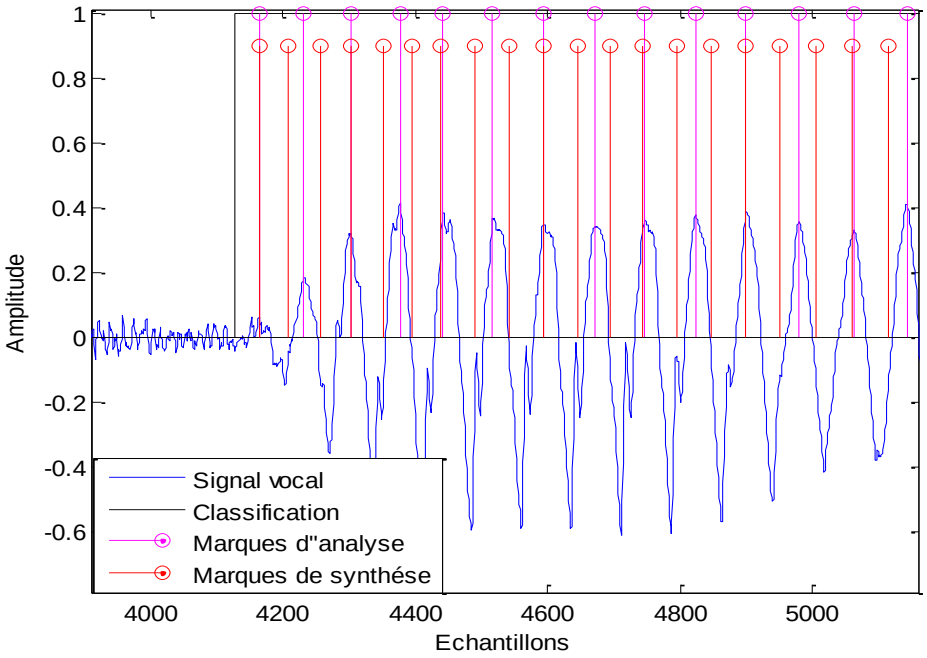


Figure 4. 13 : Positionnement des marques analyse-synthèse pour $\alpha = 1.5$.

La Figure 4.14 montre un exemple de calcul des marques de synthèse pour$\alpha = 0.5$. On observe que chaque deux périodes à l'étape d'analyse deviennent une période après la modification.

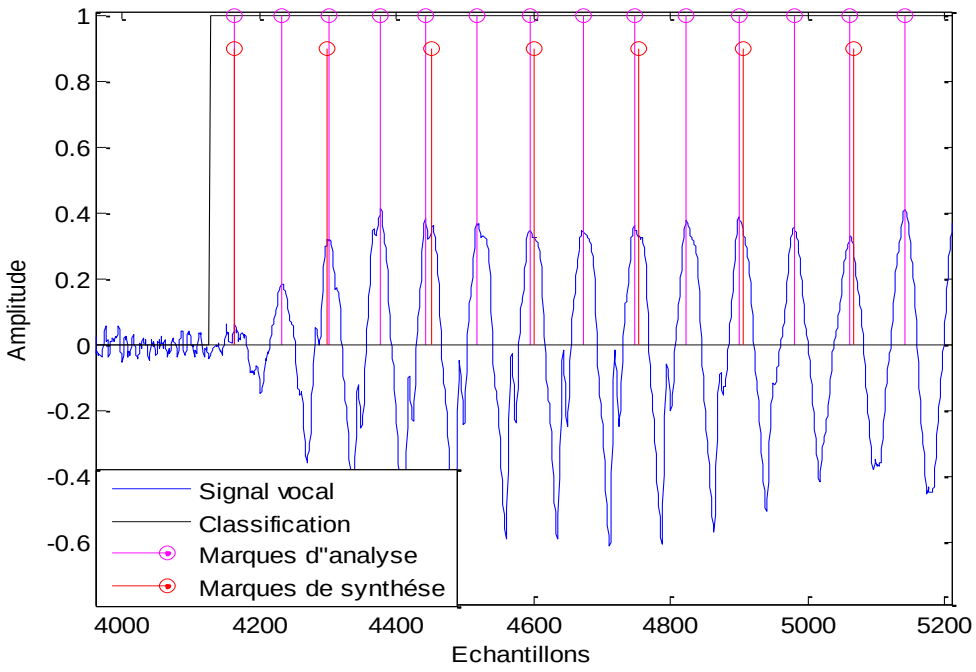


Figure 4. 14 : Positionnement des marques analyse-synthèse pour $\alpha = 0.5$.

4.3.3 Synthèse

a Correspondance des marques de synthèse

L'organigramme de la Figure 4.15 représente les étapes de correspondance des marques de synthèse à celles d'analyse.

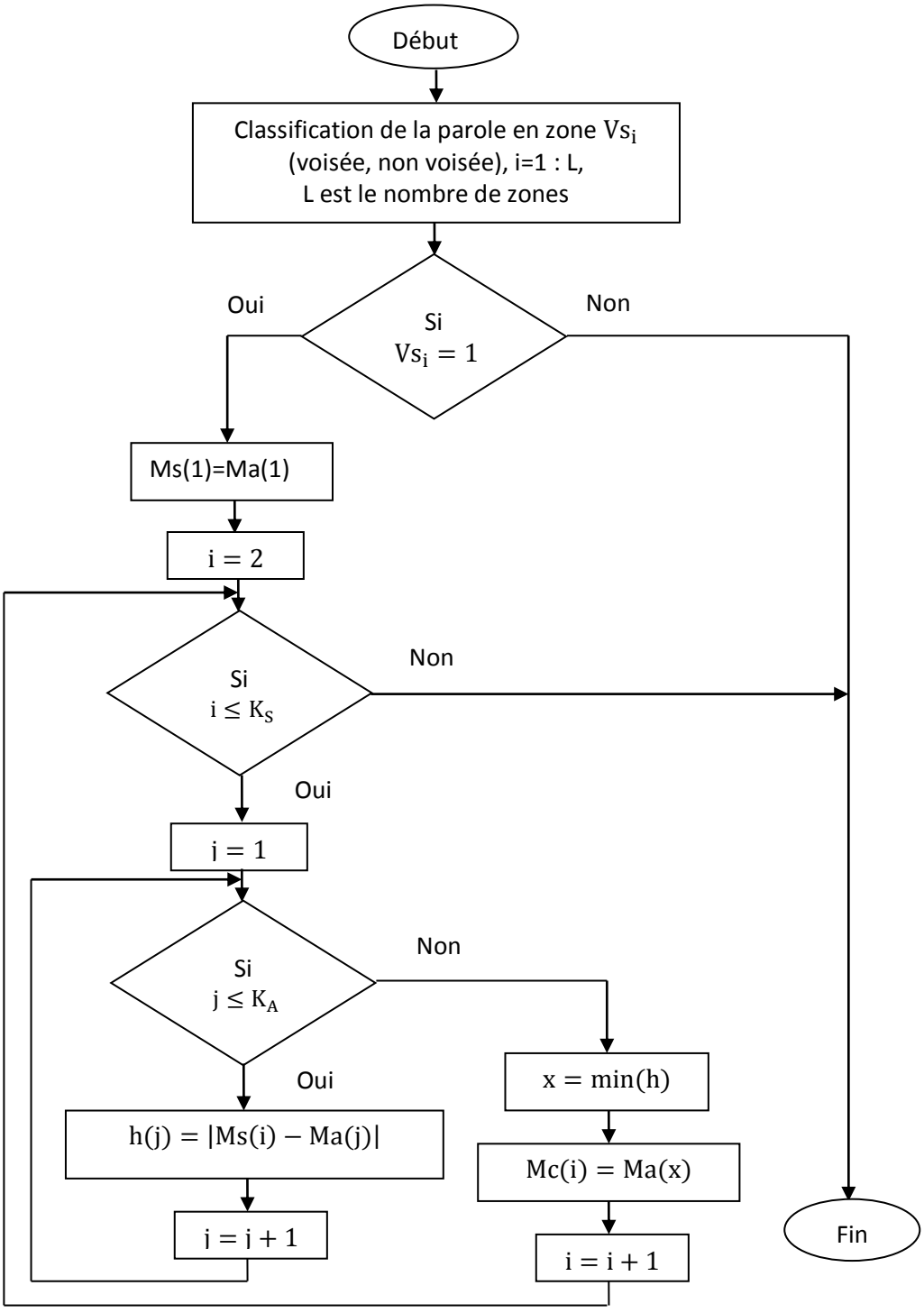


Figure 4. 15 : Organigramme de correspondance des marques de synthèse.

Avec :

Ma : Les marques d'analyse,

Ms : Les marques de synthèse,

K_A : Le nombre de marques analytiques,

K_S : Le nombre de marques synthétiques,

h : vecteur de distances,

min : est une fonction qui permet de trouver le point minimal parmi tous les éléments du vecteur h,

Mc : Les marques de correspondance.

Les marques de correspondance sont des positions de liaison entre les marques d'analyse et celles de synthèse. La Figure 4.16 donne un exemple qui représente les étapes de calcul des marques de correspondance pour $\alpha = 1.5$. Dans cet exemple, le nombre de marques d'analyse est de 5 marques. La correspondance des marques analytiques à celles de la synthèse (7 marques) est obtenue par le bais des marques Mc.

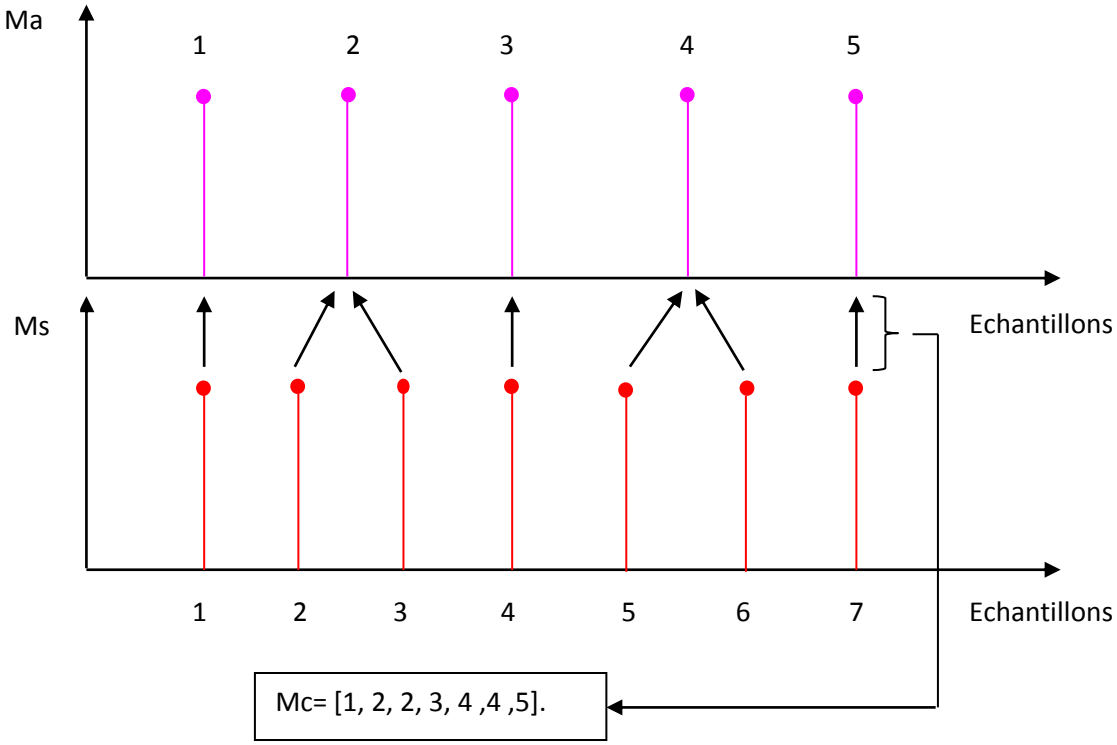


Figure 4. 16 : Calcul des marques de correspondance.

b Superposition/Addition (Overlap/Add)

Le signal de synthèse est construit par superposition/addition (Overlap-Add) des formes d'onde d'analyse placées sur les marques de synthèse (équation 3.6, chapitre 3).

La Figure 4.17 représente les étapes d'implémentation de la TD-PSOLA pour la modification de la F_0 du signal vocal.

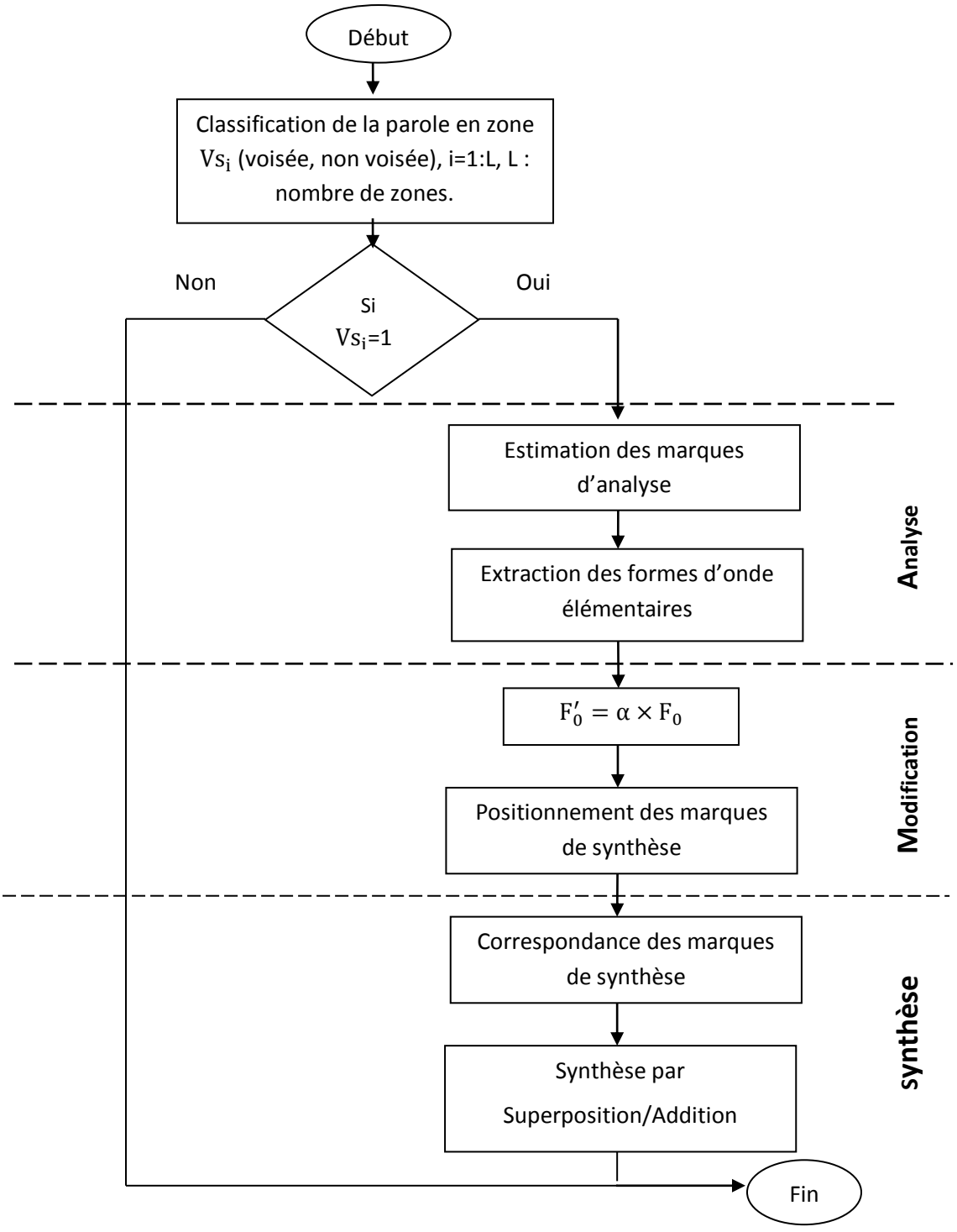


Figure 4. 17 : Organigramme de la méthode TD-PSOLA pour la modification de la F_0.

4.4 Construction du signal modifié par le bais des marques estimées

Le signal vocal est modifié en fonction de la valeur de α comme suit:

- si $\alpha = 1$, aucune modification de la voix,
- si $\alpha > 1$, la voix devient aigue,
- si $\alpha < 1$, la voix devient grave.

On va citer maintenant quelques exemples de modification. La Figure 4.18 représente un exemple de construction du signal synthétique pour α=1. Le signal d'analyse (original) est superposé sur le signal de synthèse (modifié).

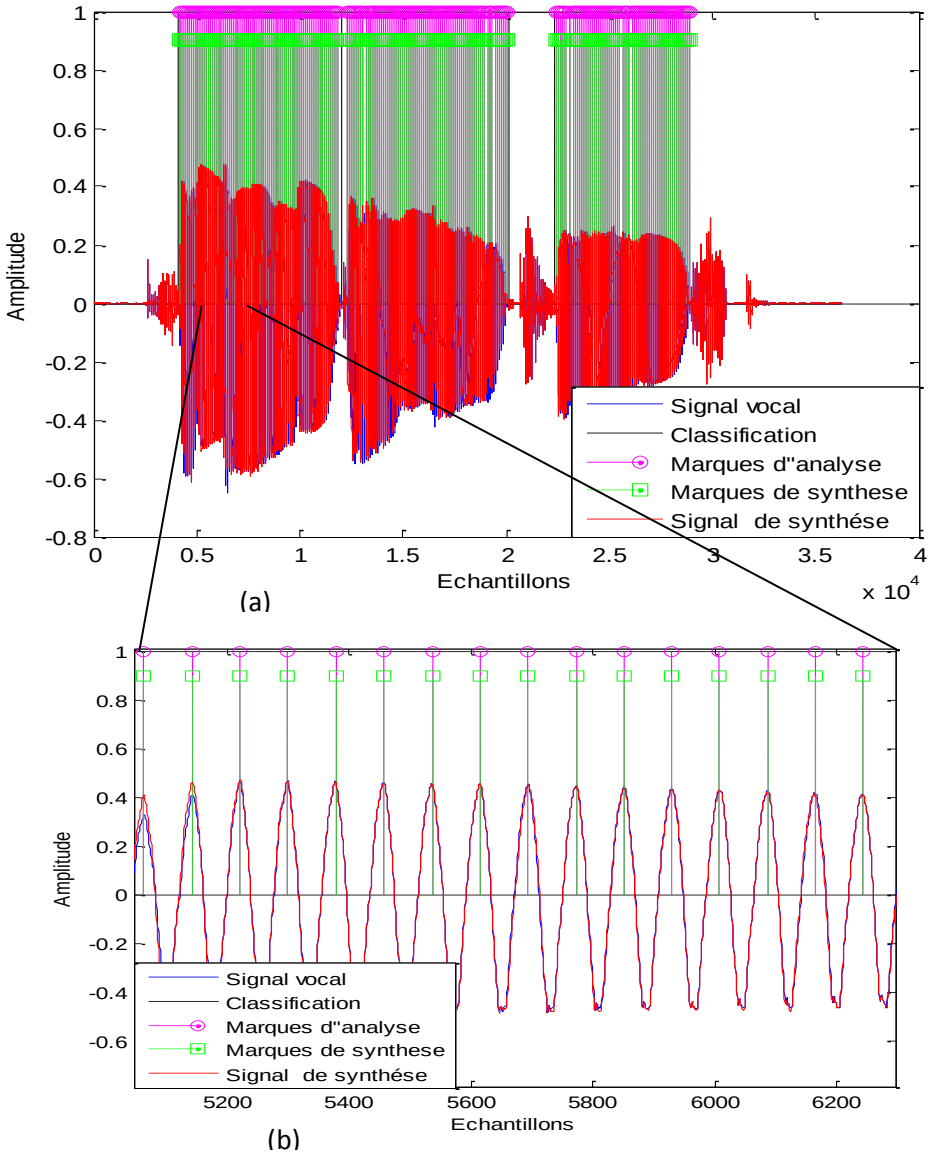


Figure 4. 18 : (a) Construction du signal synthétique pour α=1, (b) Zoom sur la région voulue.

La Figure 4.19 représente un exemple de modification de la F_0 pour α=1.5. D'après les tests d'écoute, la voix a tendance à devenir plus aigüe.

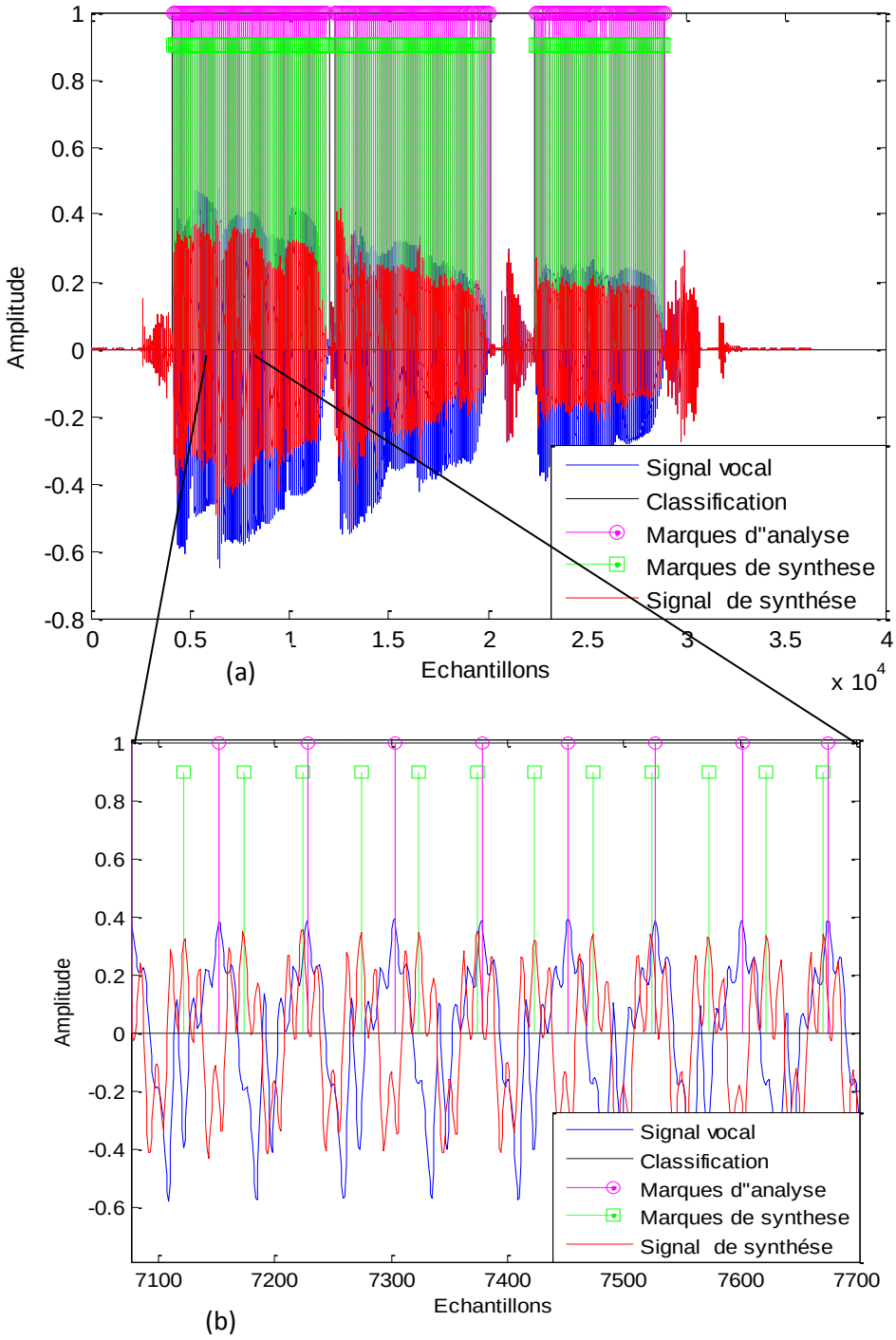


Figure 4. 19 : (a) Construction du signal synthétique pour α=1.5, (b) Zoom sur la région voulue.

La Figure 4.20 représente un exemple de modification de la F_0 pour α=0.5. D'après les tests d'écoute, la voix a tendance à devenir plus grave.

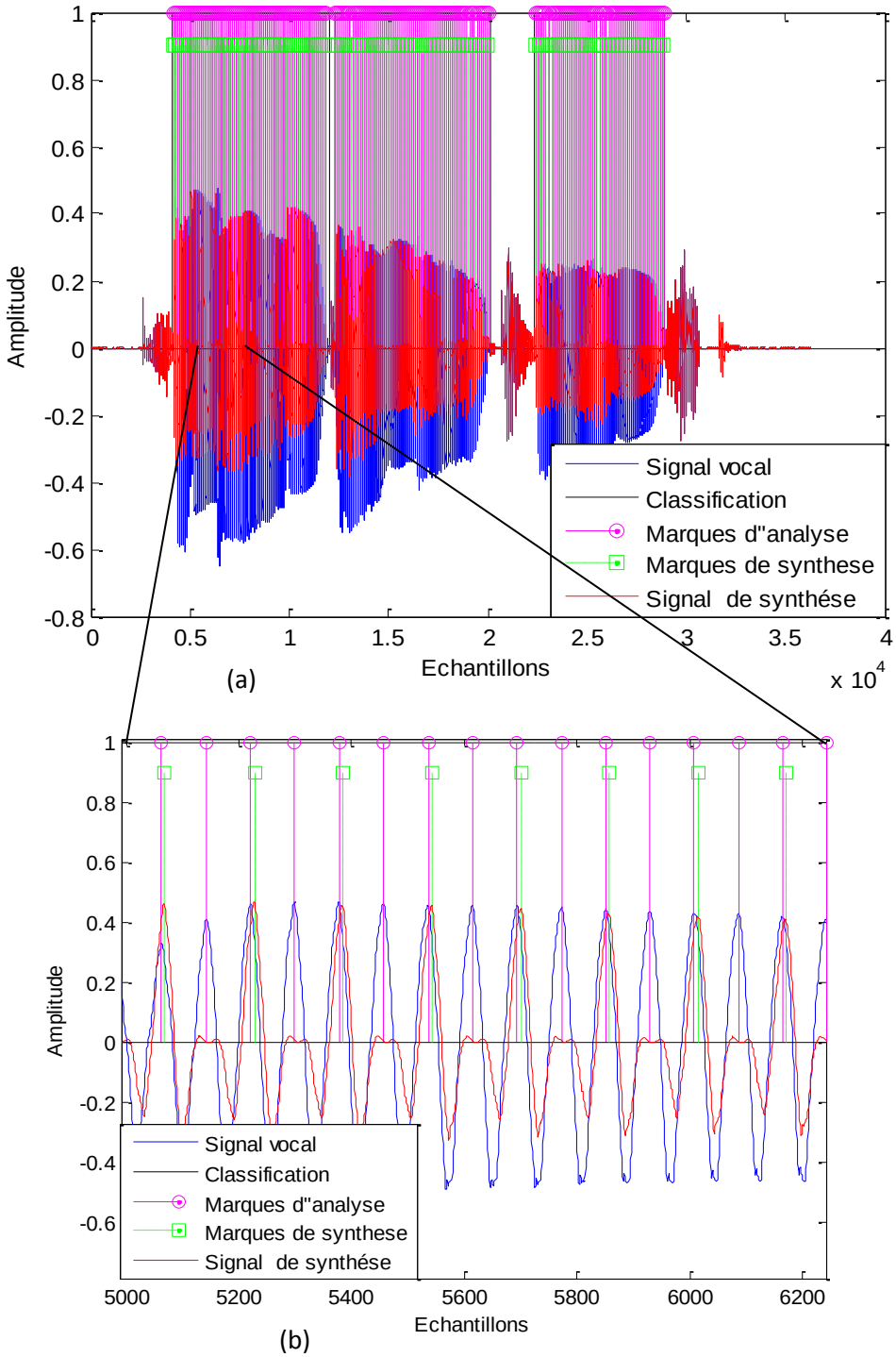


Figure 4. 20 : (a) Construction du signal synthétique pour α=0.5, (b) Zoom sur la région voulue.

La technique de marquage proposée dans notre étude ne donne pas des résultats optimaux. On a observé que certaines marques sont ajoutées (Figure 4.21) ou ratées (Figure 4.22), donnant ainsi des erreurs de marquage qui peuvent affecter le signal synthétisé. Pour résoudre ce problème, une étape de correction des marques est nécessaire pour améliorer la précision.

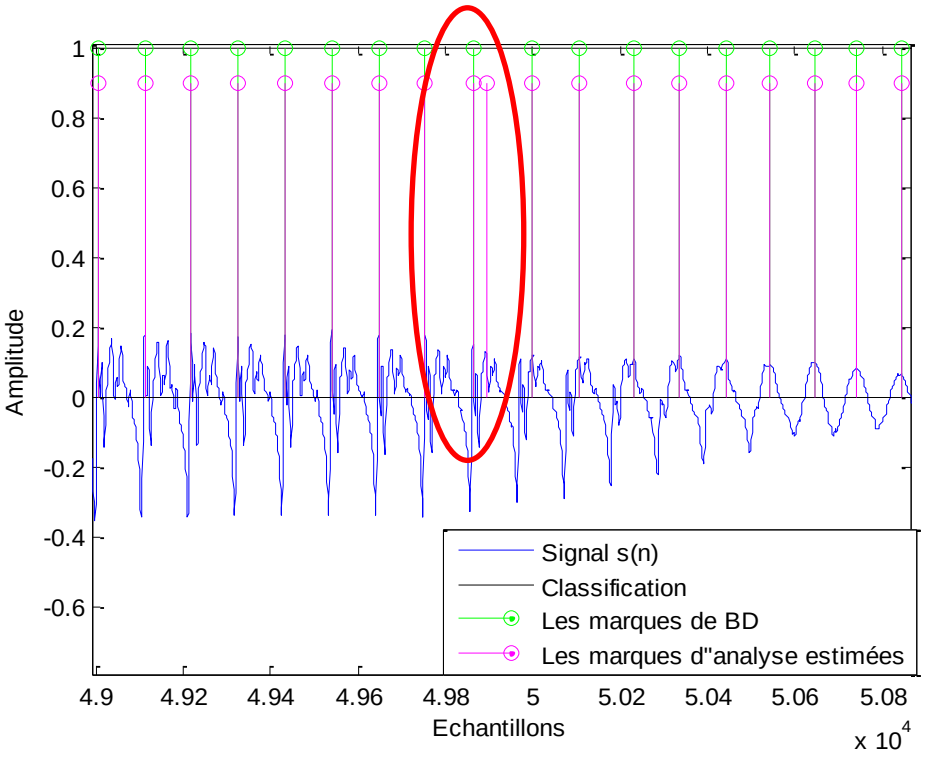


Figure 4. 21 : Marques ajoutées.

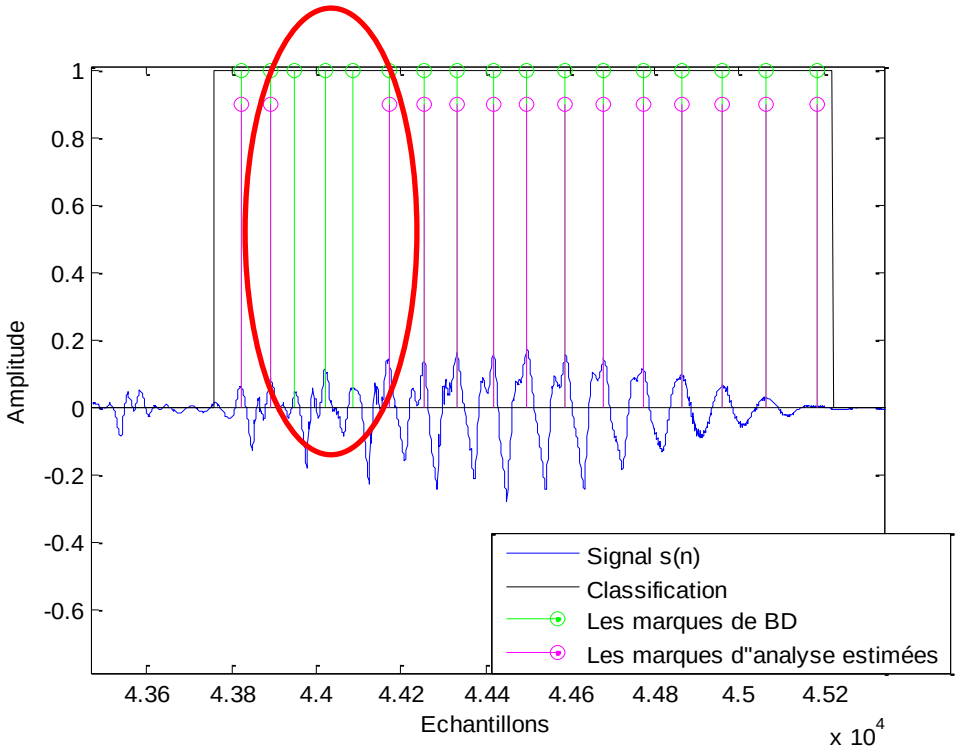


Figure 4. 22 : Marques ratées.

4.5 Correction des marques estimées

L'opération de correction des marques d'analyse (estimées à base du MBS) consiste à :

- éliminer les marques estimées à l'intérieur de l'intervalle où l'algorithme de marquage a ajouté des marques en plus.
- créer des marques virtuelles à l'intérieur de l'intervalle où l'algorithme de marquage a échoué d'estimer les marques réelles.

Dans le but de créer ou d'éliminer les marques à l'intérieur des intervalles voulus, deux seuils nominatifs sont pris : d_{min} et d_{max} . Ils définissent les distances séparant deux marques d'analyse consécutives.

- si la distance entre deux marques analytiques est inférieure ou égale à d_{min} : on élimine la marque qui possède l'amplitude minimale (Figure 4.23),
- si la distance entre deux marques analytiques est supérieure ou égale à d_{max} , on ajoute des marques (Figure 4. 24).

Comme il a été mentionné dans le premier chapitre, la plage de variation moyenne de la F_0 varie d'un locuteur à un autre en fonction de son sexe :

- 80 à 200 Hz pour les voix masculines,
- 150 à 350 Hz pour les voix féminines,

Les seuils $[d_{min}, d_{max}]$ sont calculés respectivement en tenant compte de l'intervalle de variation de la F_0 pour les deux types de voix :

$$d_{min} = \frac{F_e}{F_{0\,max}} \tag{4.3}$$

$$d_{max} = \frac{F_e}{F_{0\,min}} \tag{4.4}$$

En replaçant les intervalles de la F_0 données ci-dessus, on obtient les valeurs numériques correspondantes :

- [80, 200] échantillons pour les voix masculines,
- [46, 107] échantillons pour les voix féminines,

Pour des raisons de minimisation d'erreurs de marquage, on a élargis la plage de variation moyenne des intervalles de la F_0 de la manière suivante :

- de 72 à 200 Hz pour les voix masculines,
- de 107 à 400 Hz pour les voix féminines,

Les nouveaux seuils qui correspondent à la plage fréquentielle élargis sont donnés comme suit :

- [80, 220] échantillons pour les voix masculines,
- [40, 150] échantillons pour les voix féminines,

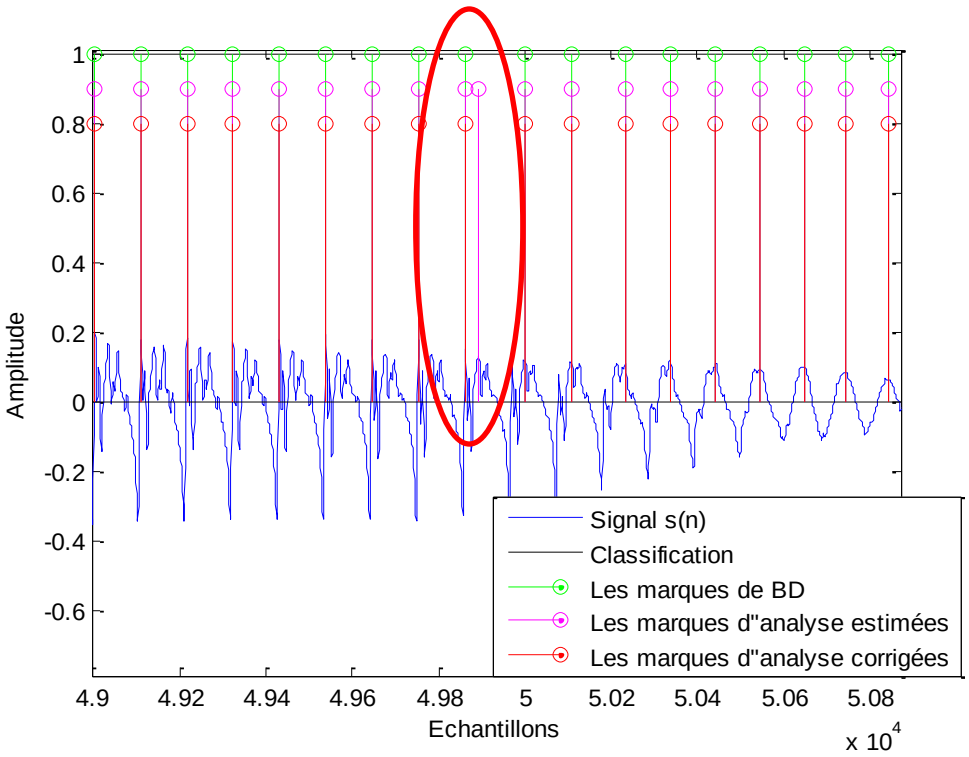


Figure 4. 23 : Correction des erreurs de marquage par l'élimination d'une marque.

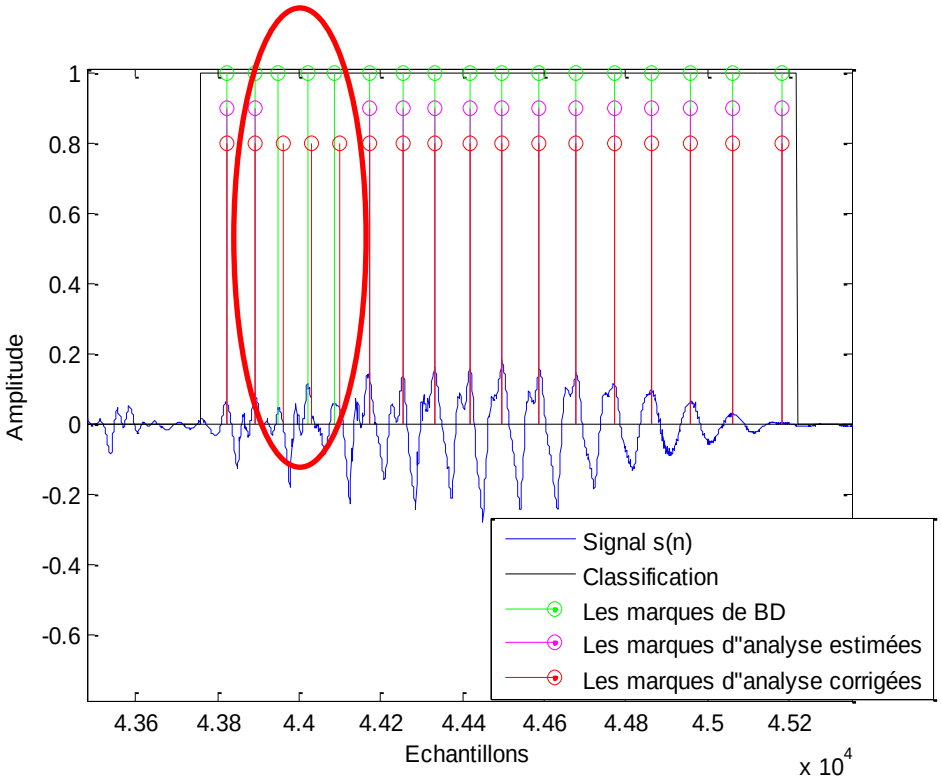


Figure 4. 24 : Correction des erreurs de marquage par l'ajout de trois marques.

4.6 Evaluation des performances de la TD-PSOLA

En général, l'évaluation des techniques de modification prosodiques est effectuée par des méthodes objectives et subjectives. L'évaluation objective repose sur le calcul d'erreurs de marquage, et l'évaluation subjective consiste à faire des tests d'écoute pour apprécier la qualité sonore.

4.6.1 Evaluation objective

Pour évaluer les performances de la technique de marquage proposée, on a employé les mesures utilisées dans [49], à savoir :

- le taux d'identification IR « *Identification Rate* »,
- le taux de marques ratées MR « *Miss Rate* »
- le taux de fausses alarmes FA « *False Alam* »

Principalement, ces mesures ont été utilisées pour évaluer les performances de la technique de détection des GCIs nommée DYPSA, « *Dynamic Programming Projected Phase-Slope Algorithm* » [49]. La Figure 4.25 représente un schéma synoptique des erreurs de marquage qui peuvent apparaitre.

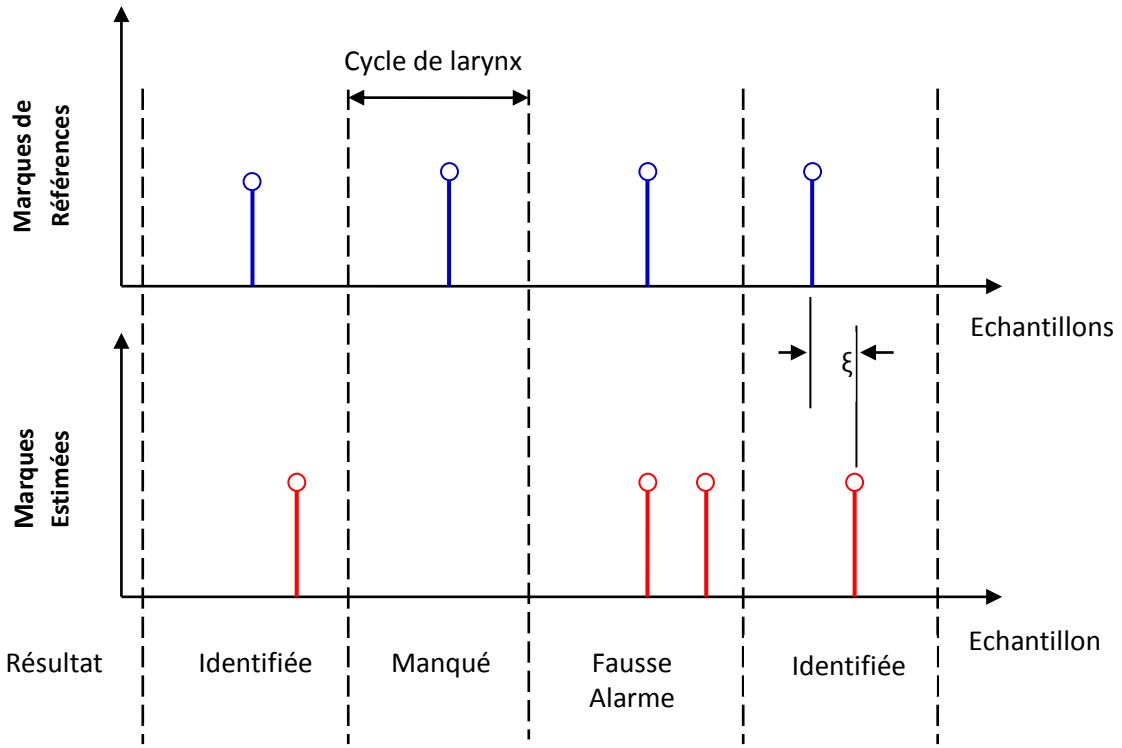


Figure 4. 25 : Types d'erreurs.

L'équation suivante permet de définir le cycle de larynx :

$$\left(\frac{1}{2}\right)(\tilde{n}_{r-1} + \tilde{n}_r) \leq n < \left(\frac{1}{2}\right)(\tilde{n}_r + \tilde{n}_{r+1}) \quad (4.5)$$

Avec :

- n : est le n^{eme} échantillon dans un cycle de larynx.
- $\tilde{n}_r$, $\tilde{n}_{r-1}$ et $\tilde{n}_{r+1}$: sont les marques de référence (BD).

Le taux d'identification : est le pourcentage de cycles de larynx pour le quel exactement une seule marque est détectée. Deux paramètres secondaires sont calculés à partir de cette mesure :

- l'erreur d'identification ξ, « *Identification Error* »,
- l'exactitude d'identification σ, « *Identification Accuracy* ».

ξ est l'erreur de synchronisation entre les marques de références et celles estimées dans les cycles du larynx, σ est l'écart type de ξ. Les petites valeurs de σ indiquent une grande précision d'identification.

Le taux de marques ratées : est le pourcentage de cycles de larynx pour lequel aucune marque n'est détectée.

Le taux de fausses alarmes : est le pourcentage de cycles de larynx pour lequel plus d'une marque est détectée.

Cette procédure a été adoptée dans notre travail pour évaluer les performances de la technique de marquage développée. Les erreurs sont calculées pour :

- les marques d'analyse estimées à base du MBS,
- les marques d'analyse estimées à base du MBS après correction,

Le calcul d'erreur est effectué sur la BD de test (voir Annexe C). Les erreurs moyennes sont données dans les tableaux 4.1 et 4.2.

Types d'erreurs	FA (%)	MR (%)	σ
Homme	2.21	5.53	5.41
Femme	0.37	2.41	3.41
Moyenne	1.29	3.97	4.41

Tableau 4. 1 : Taux d'erreurs de marques estimées à base du MBS.

Types d'erreurs	FA (%)	MR (%)	σ
Homme	0.81	2.89	5.56
Femme	0.34	0.94	4.07
Moyenne	0.575	1.91	4.81

Tableau 4. 2 : Taux d'erreurs de marques estimées à base du MBS après correction.

D'après les résultats du Tableau 4.1, on peut dire que la technique de marquage proposée a donné de bons résultats. On constate aussi une amélioration des résultats obtenus après la correction des marques (Tableau 4.2). L'évaluation subjective va donner une comparaison entre les qualités des voix construites par les deux types de marques.

4.6.2 Evaluation subjective

Pour évaluer les performances de la TD-PSOLA dans le cas de la modification de la F_0, on fait une série de tests d'écoute à base du MOS en utilisant les trois types de marquages :

- les marques de la BD,
- les marques d'analyse estimées à base du MBS,
- les marques d'analyse estimées à base du MBS après correction,

Les auditeurs sont nous-même et notre promoteur. Le score global est donné sous forme de la moyenne des évaluations.

On a choisi quatre phrases aléatoires à partir de notre BD, deux prononcées par un locuteur masculin et deux autres prononcées par un locuteur féminin. Les phrases choisies sont :

Locuteur féminin :

- Phrase 4: "*She turned in at the hotel*",
- Phrase 5: "*It was a curios coincidence*",

Locuteur masculin :

- Phrase 23: "*It is the aurora borealis*",
- Phrase 25: "*He moved away as quietly as he had come*".

Les facteurs de modification de la F_0 utilisés sont : 0.5, 0.8, 1.2, 1.6 et 2.

Les résultats trouvés sont mentionnés sur les tableaux suivants (du Tableau 4.3 au Tableau 4. 8).

Facteur de modification (α)	Phrase 4	Phrase 5	Moyenne
0.5	4	3.33	3.66
0.8	4	4.66	4.33
1.2	4.66	4	4.33
1.6	4	3.66	3.83
2	3	3.33	3.16

Tableau 4. 3 : Test MOS en utilisant les marques de la BD.

Facteurs de modification (α)	Phrase 4	Phrase 5	Moyenne
0.5	4	3.33	3.66
0.8	4	4.66	4.33
1.2	4.66	4	4.33
1.6	4	3.66	3.83
2	3	3.33	3.16

Tableau 4. 4 : Test MOS en utilisant les marques d'analyse estimées à base du MBS.

Facteurs de modification (α)	Phrase 4	Phrase 5	Moyenne
0.5	4	3.33	3.66
0.8	4	4.66	4.33
1.2	4.66	4	4.33
1.6	4	3.66	3.83
2	3	3.33	3.16

Tableau 4. 5 : Test MOS en utilisant les marques d'analyse estimées à base du MBS après correction.

D'après les tests d'écoute des voix féminines, on peut dire que la qualité vocale perçue (Tableaux 4.3, 4.4 et 4.5) est bonne pour l'ensemble des marques utilisées. Elle varie en fonction du facteur de modification choisi.

Facteur de modification (α)	Phrase 23	Phrase 25	Moyenne
0.5	4.33	4.33	4.33
0.8	4.66	4.66	4.66
1.2	4.66	4.66	4.66
1.6	4.33	4.33	4.33
2	3.83	3.83	3.83

Tableau 4. 6 : Test MOS en utilisant les marques de la BD.

Facteur de modification (α)	Phrase 23	Phrase 25	Moyenne
0.5	4	4.33	4.16
0.8	4.66	4.66	4.66
1.2	4.66	**4**	4.33
1.6	**4**	**3.83**	3.91
2	3.83	3.33	3.58

Tableau 4. 7 : Test MOS en utilisant les marques d'analyse estimées à base du MBS.

Facteur de modification (α)	Phrase 23	Phrase 25	Moyenne
0.5	4.33	4.33	4.33
0.8	4.66	4.66	4.66
1.2	4.66	**4.66**	4.66
1.6	**4.33**	**4.33**	4.33
2	3.83	3.83	3.83

Tableau 4. 8 : Test MOS en utilisant les marques d'analyse estimées à base du MBS après correction.

D'après les tests d'écoute des voix masculines, on peut dire que la qualité vocale perçue est bonne. Elle varie en fonction du facteur de modification choisi. Cependant, on remarque dans le Tableau 4.7 une réduction du MOS pour les facteurs de modification 1.2 et 1.6 (en comparant ce MOS à celui du Tableau 4.6). Cette réduction est due aux erreurs de marquages obtenues par la technique proposée.

Si on analyse les résultats de modification de la F_0 par l'application d'un facteur de 1.6 pour les cas suivants :

- marques d'analyse de la BD : Tableau 4.6, phrase 25, MOS=4.33,
- marques d'analyse estimées à base du MBS : Tableau 4.7, phrase 25, MOS=3.83.

Une réduction du MOS est observée en comparant ces deux cas. Cette réduction est due aux deux pics aléatoires qui sont visualisés sur la Figure 4.26 (résultats des erreurs de marquages).

L'utilisation des marques corrigées (Tableau 4.8, phrase 25, avec le même facteur de modification « 1.6 ») a permis d'améliorer la qualité vocale (MOS=4.33). Les deux pics qui sont apparus dans le cas de l'utilisation des marques estimées (Figure 4.26)

sont éliminés en utilisant les marques corrigées pour la construction du signal synthétique (Figure 4.27).

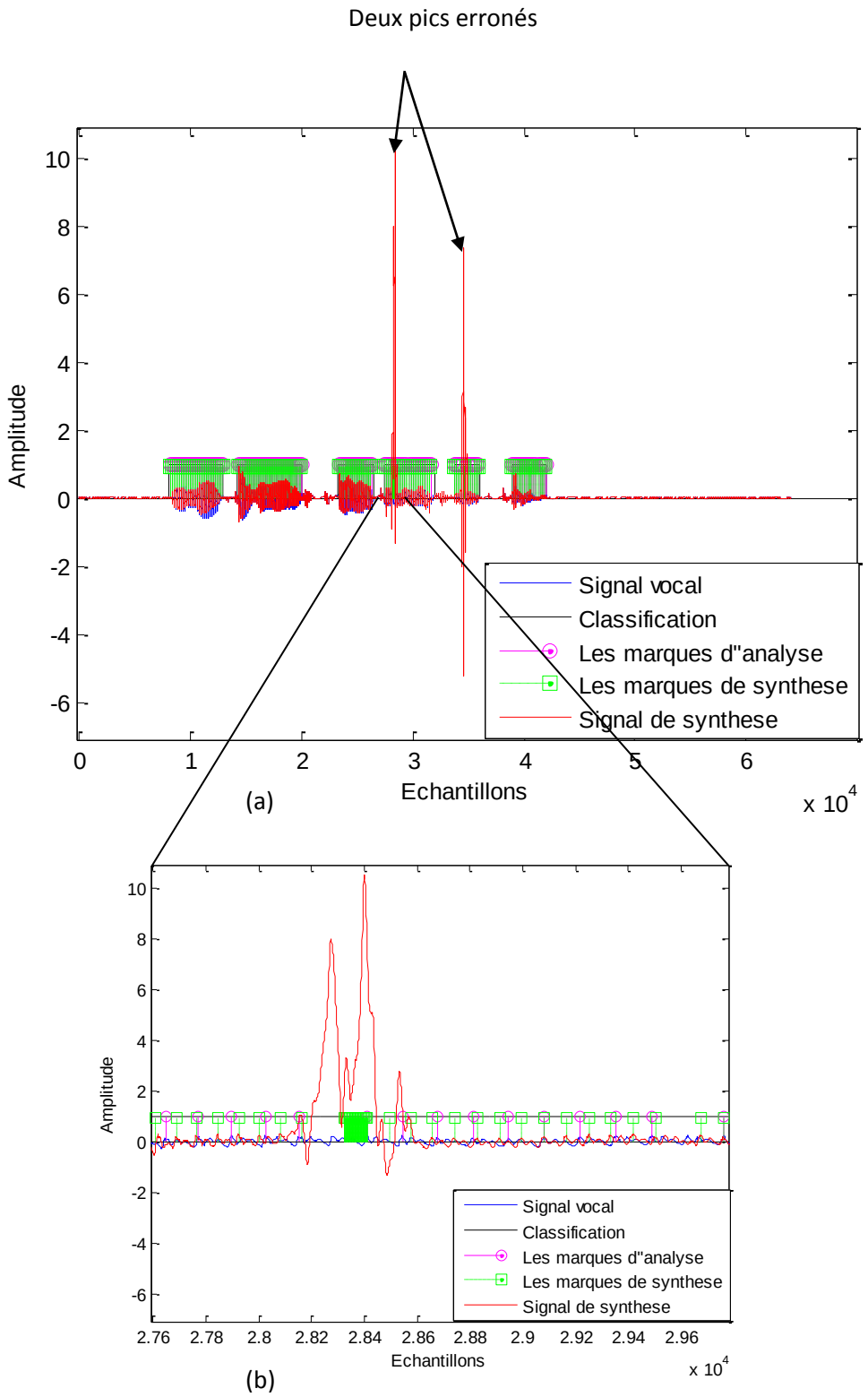


Figure 4. 26 : (a) Modification de la F_0 « phrase 25» en utilisant les marques estimées à base du MBS pour α=1.6, (b) Zoom sur la région voulue.

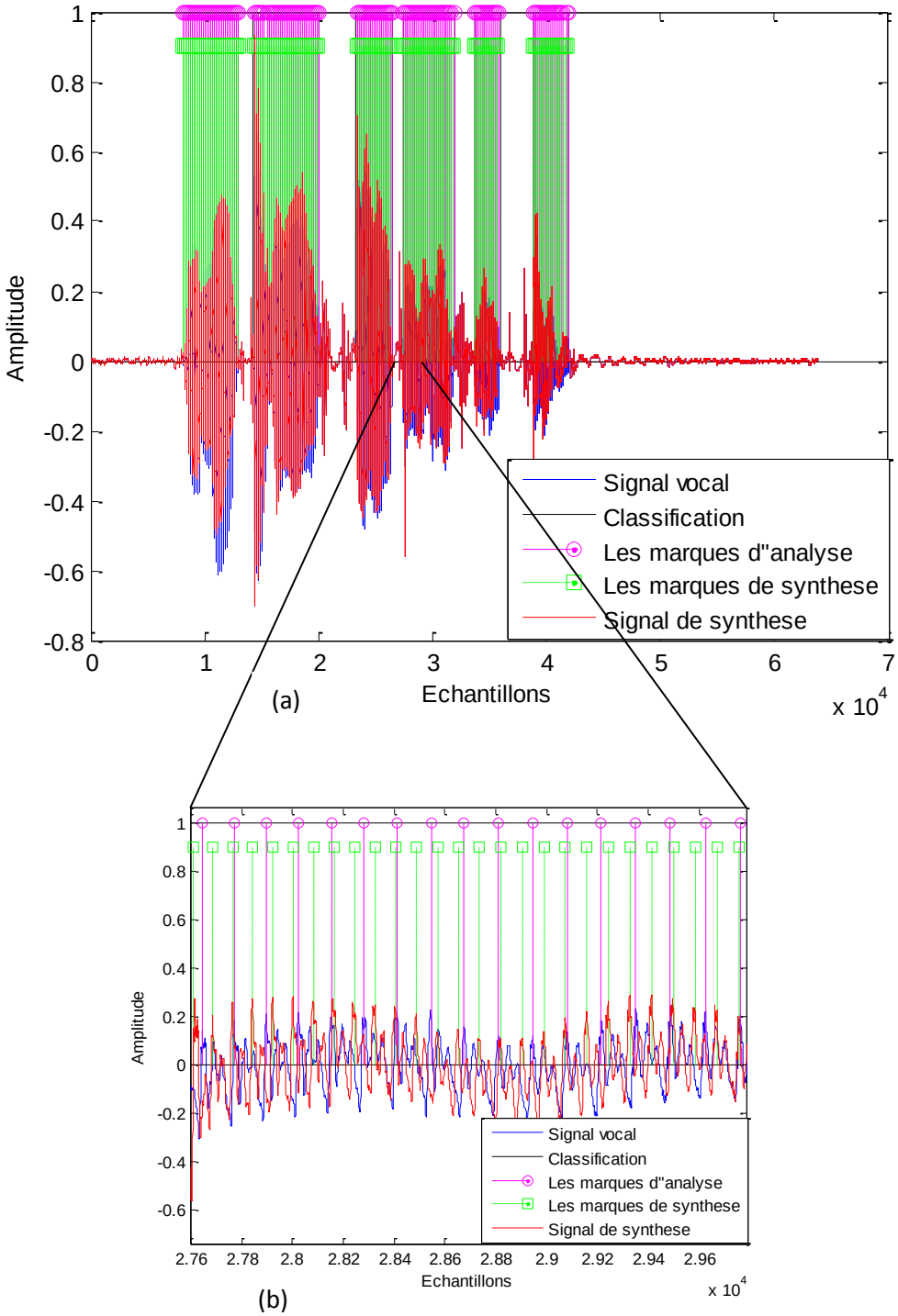


Figure 4. 27 : (a) Modification de la F_0 « phrase 25», en utilisant les marques à base du MBS corrigées pour α=1.6, (b) Zoom sur la région voulue.

4.7 Conclusion

Dans ce chapitre on a implémenté la méthode TD-PSOLA « *Time Domain Pitch Synchronous Overlap-Add* » pour la modification de la fréquence fondamentale. La technique de marquage est basée sur « *Mean Based Signal* ». Le positionnement des marques d'analyse correspondent aux instants des pics globaux du signal vocal.

Le marquage proposé ne donne pas des instants optimaux. A cet effet, une étape de correction des marques a été introduite.

D'après les évaluations objectives et subjectives à base de la CMU ARCTIC, on peut conclure que l'opération de correction des marques analytiques a clairement amélioré les performances de la TD-PSOLA.

Conclusion générale

Dans le cadre de notre projet de master, on s'est intéressé à la transformation de la voix qui correspond à des besoins courants dans divers domaines du traitement du son et de la parole. Plusieurs méthodes de transformations existent dans la littérature. La majorité de ces méthodes s'intéresse à la modification de la prosodie et de l'enveloppe spectrale de la parole.

L'objectif principal de notre projet de fin d'étude est la modification de la fréquence fondamentale par la TD-PSOLA «*Time Domain Pitch Synchronous Overlap-Add* ».

La méthode TD-PSOLA est réalisée en trois étapes qui sont : l'analyse, la modification et la synthèse. L'étape d'analyse consiste à extraire les formes d'onde élémentaires des régions voisées du signal vocal. La durée de ces formes d'onde est égale à deux fois la période fondamentale locale. Ces signaux élémentaires sont positionnés autours des instants où l'énergie du signal est maximale. Ils sont dits « *marques d'analyse* ».

Généralement, les marques d'analyse peuvent être positionnées sur trois instants différents :

- les instants de fermeture de la glotte « GCIs, *Glottal Closure Instants*»,
- les instants qui correspondent aux pics globaux du signal,
- et les instants qui correspondent aux vallées globales du signal,

Dans notre travail, on s'est intéressé au deuxième type de positionnement. La technique de marquage développée se base sur le calcul d'un signal moyenneur dit « *Mean Based Signal* : MBS». Les marques estimées ne donnent pas des instants optimaux. Ainsi, la correction des instants estimés est nécessaire pour réduire les erreurs de marquage. Les performances de cette technique sont mesurées par le calcul des erreurs.

L'étape de modification consiste à imposer un facteur de modification de la fréquence fondamentale qui varie entre 0.1 et 2. La synthèse est effectuée par addition/recouvrement des formes d'ondes élémentaires.

D'après les résultats de simulations obtenus, on peut conclure que la correction des marques a été bien effectuée.

La méthode TD-PSOLA pour la modification de la fréquence fondamentale est évaluée sur la base de données CMU ARCTIC, constituée d'un corpus de 36 phrases prononcées par deux locuteurs de sexe différent. La méthode d'évaluation consiste à faire une suite de tests objectifs (calcul d'erreurs) et subjectifs (test MOS).

On peut conclure que la modification de la fréquence fondamentale en utilisant les marques analytiques corrigées conduit à un signal de synthèse de bonne qualité.

On espère que la technique de marquage développée dans notre travail peut être utilisée pour des modifications simultanées des paramètres prosodiques à base de la TD-PSOLA.

Ce projet fût une expérience enrichissante pour nous en termes de connaissances théoriques et pratiques acquises durant la période de préparation de notre PFE.

Annexes

Le but de ces annexes est de réunir un certain nombre de résultats analytiques et quelques exemples sur la prononciation de l'Anglais Américain.

Annexe A

Modes		Sons	Exemples	API
Voyelles	Avant	**I** **I** **e** **æ** **ɛ**	**e**ve, **ea**t **i**t, **i**nvite h**a**te, **ei**ght **a**t, gl**a**ss m**e**t, **e**xtra	i I e @ E
	Centrée	**ɜ** **ʌ**	bi**r**d, f**ur** **u**p, s**u**n	R A
	Arrière	**A** **c** **o** **U** **u**	f**a**ther, r**aw** **a**ll, j**aw** **o**bey, **o**ver f**oo**t dr**ew**	a ɔ o U U
Semi-Voyelles	Liquides	**R** **l**	**r**ead, ba**rr**on **l**ate, fa**ll**	r l
	Glissées	**W** **y**	**w**e, **w**ish **y**ou, **y**ellow	w y
Diphtongues		**aI** **aU** **ɔI**	H**i**de, tr**y** **ou**t, pl**ow** B**oy, oi**l	Y W O
Consonnes	Chuchoté	**H**	**h**e, **h**appy	H
	Fricatives	**V** **ð** **z** **ʒ** **f** **θ** **s** **ʃ**	**v**ote , **v**oise **th**en, **th**at **z**oo, **z**ipper a**z**ure, vi**s**ion **f**or, **f**ood **th**in, wi**th** **s**ee, mi**ss** **sh**e, wi**sh**	v D z Z f T s S
	Plosives	**B** **d** **g** **p** **t** **k**	**b**e, **b**all **d**ay, **d**eer **g**o, a**g**o **p**ay, to**p** **t**o, loo**t** **k**ey, wa**k**e	b d g p t k
	Nasales	**M** **n** **ŋ**	**m**e, **m**ask **n**o, **kn**ob si**ng,** ri**ng**	m n G
	Affriquées	**tʃ** **dʒ**	**Ch**ew, **ch**op ca**g**e, **j**ob	tS J

Tableau A. 1 : Transcription Orthographique Phonétique [16], [18].

Annexe B

Fenêtrage

Au traitement du signal, le fenêtrage est utilisé dès que l'on s'intéresse à un signal de longueur volontairement limité. En effet, un signal réel ne peut qu'avoir une durée limitée dans le temps. De plus, un calcul ne peut se faire que sur un nombre de points finis.

Au lieu d'étudier le signal s(n), on étudie le signal tronqué : $s_h(n) = s(n)\ h(n)$; en passant dans le domaine fréquentiel via une Transformée de Fourier « *TF* », on obtient le produit de convolution $S_h(f) = S(f) * H(f)$, où H(f) est la TF de la fenêtre.

Les formules des fenêtres de pondération courantes :

Fenêtre rectangulaire : $h(t) = \begin{cases} 1 & \text{si } t \in [0, T-1] \\ 0 & \text{si non} \end{cases}$

Fenêtre triangulaire: $h(t) = \begin{cases} \frac{2t}{T} & \text{si } t \in \left[0, \frac{T}{2}\right[\\ \frac{2(T-t)}{T} & \text{si } t \in \left[\frac{T}{2}, T\right[\\ 0 & \text{si non} \end{cases}$

Fenêtre de Hanning: $h(t) = \begin{cases} 0.5 - 0.5\cos\left(2\pi\frac{t}{T}\right) & \text{si } t \in [0, T-1] \\ 0 & \text{si non} \end{cases}$

Fenêtre de Hamming : $h(t) = \begin{cases} 0.54 - 0.46\cos\left(2\pi\frac{t}{T}\right) & \text{si } t \in [0, T-1] \\ 0 & \text{si non} \end{cases}$

T est la taille de la fenêtre.

Les fenêtres de pondération sont représentées respectivement dans les deux domaines temporel et fréquentiel (Figures B.1 et B.2).

- **Domaine temporel**

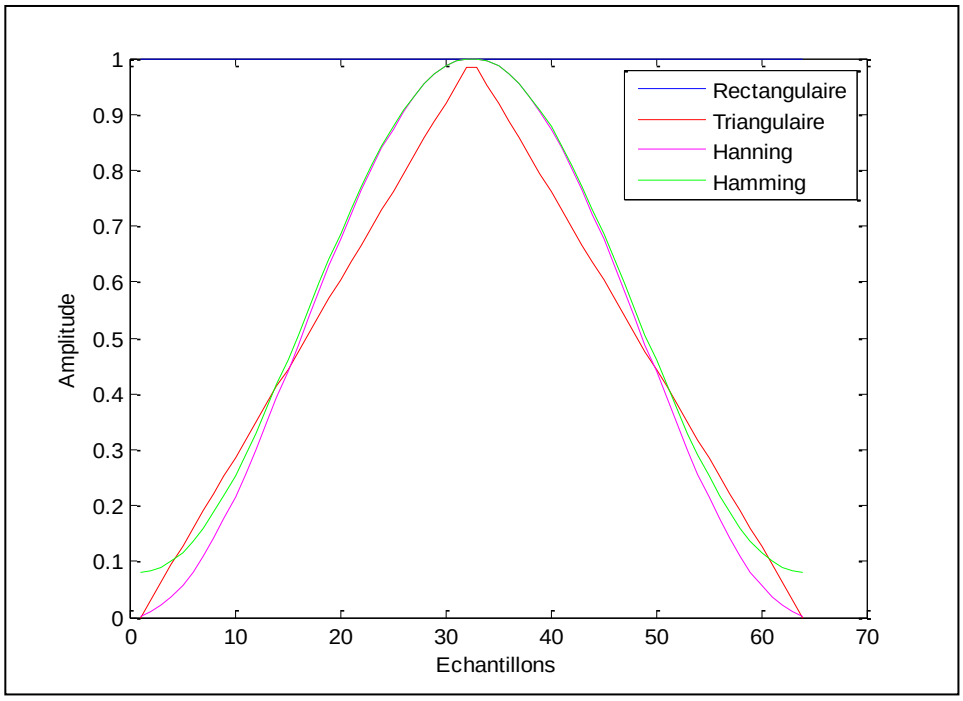


Figure B. 1 : Fenêtres de pondération dans le domaine temporel.

- **Domaine fréquentiel**

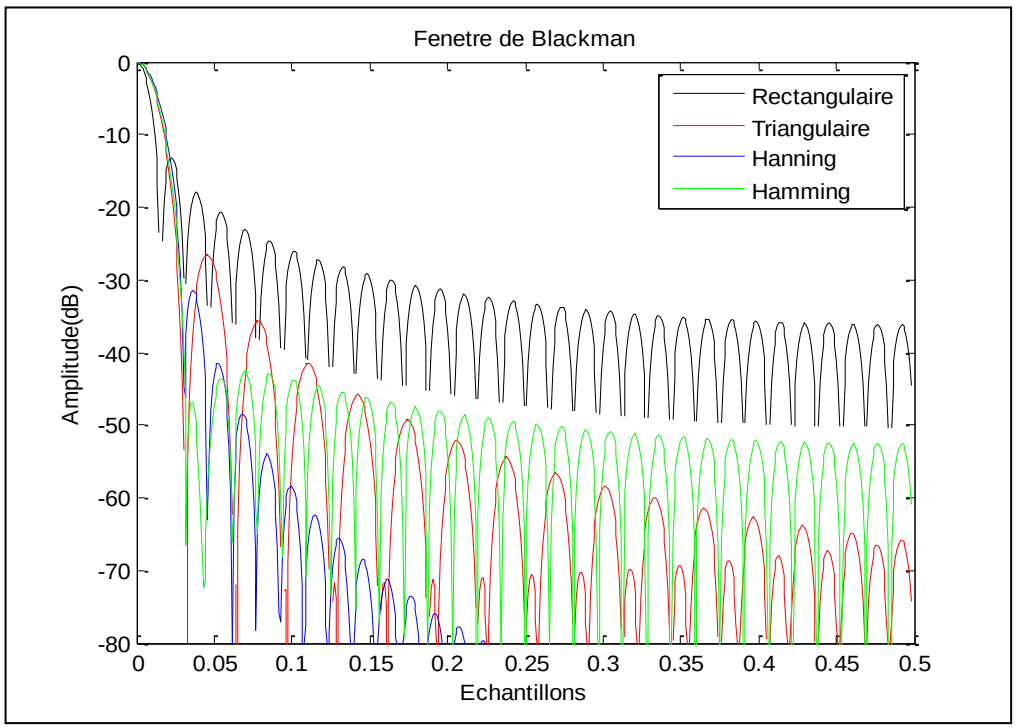


Figure B. 2 : Fenêtres de pondération dans le domaine fréquentiel en dB.

- **Le choix de la fenêtre**

Le choix de la fenêtre est selon les paramètres suivants :

- La largeur du lobe principal que caractérise la résolution en fréquence.
- L'amplitude maximum des oscillations des lobes secondaires qui caractérise la dynamique du spectre utile.
- La décroissance des lobes secondaires en décibels/octave ($\log_2(f)$) que donne une idée de l'erreur en amplitude et en position que l'on commet sur l'analyse d'une sinusoïde [63] (Tableau B.1).

Fenêtres	Largeur du Lobe principal (N est la taille de la fenêtre)	Amplitude des lobes secondaires en dB	Décroissance des lobes secondaires en dB/octave
Rectangulaire	2/N	-13	-6
Triangulaire	4/N	-27	-12
Hanning	4/N	-32	-18
Hamming	4/N	-43	-6

Tableau B. 1 : Paramètres qui caractérisent les fenêtres de pondération.

D'après les paramètres qui caractérisent les fenêtres. Nous utilisons dans notre étude la fenêtre de Hanning qui représente le meilleur choix.

Annexe C

Les signaux de test choisis sont donnés dans les tableaux C.1 et C.2

- ***Femme***

Numéro de phrase	Les phrases
Phrase 1	Author of the danger trail Philips steel exetra
Phrase 2	God bless them, I hope I will go and see them forever
Phrase 3	Hardly were plans made public before we were met by powerful opposition
Phrase 4	She turned in at the hotel
Phrase 5	It was a curious coincidence
Phrase 6	Skarcly had he uttered the name when Pierre's closing eyes shut open
Phrase 7	He waited into the edge of the water and began scrubbing himself
Phrase 8	Two years ago I gave up civilization for this
Phrase 9	Of course , that is uninteresting, she continued
Phrase 10	Don't you see I'm chewing this thing in two
Phrase 11	Fresh cases, still able to walk ,they clustered about the spokesman
Phrase 12	I was brought up the way most girls in Hawaii are brought up
Phrase 13	Wash your hand of me
Phrase 14	Between him and all domestic animals, there must be no hostilities
Phrase 15	Man could not conquer them
Phrase 16	At that moment I got the impression that she was really weak
Phrase 17	Earnest saw in the affair the most sinister import
Phrase 18	"come on" Delmar challenged

Tableau C. 1 : Phrases prononcées par un locuteur féminin.

- ***Homme***

Numéro de phrase	Les phrases
Phrase 19	Author of the danger trail ,Philips steels exetra
Phrase 20	Lord, but I'm glad to see you again Phil
Phrase 21	I'm playing a single hand and what looks like a losing game
Phrase 22	Kraysen shift back his chair and rose to his feet
Phrase 23	h's the Aurora Borealis
Phrase 24	Men of self and stamp don't stop at women and children
Phrase 25	He moved away as quietly as he had come
Phrase 26	The men stood into each other's face
Phrase 27	Shall I carry you
Phrase 28	Now these things have been struck dead within him
Phrase 29	Sometimes her dreams were filled with visions
Phrase 30	Give them their choice between a fine or an official flipping
Phrase 31	This taset promess of continued acquaintance gives Sax in a little joy through
Phrase 32	Stand-off butcher and baker and all the rest
Phrase 33	How could I answer the question on the spot of the moment
Phrase 34	Sandel will never become a world champion
Phrase 35	King took every advantage he knew
Phrase 36	You were making them talk sharp, Ruth charged him

Tableau C. 2 : Phrases prononcées par un locuteur masculin.

Le calcul d'erreur est effectué sur les marques estimées à base du MBS et celles trouvées après la correction. Les Tableaux C.1 à Tableau C.4 résument les résultats obtenus.

- ***Femme***

Type d'erreurs	FA (%)	MR (%)	σ
Phrase 1	0	2.11	0.59
Phrase 2	0.23	3.17	3.24
Phrase 3	0.76	3.18	0.88
Phrase 4	0	0.41	0.32
Phrase 5	0	3.35	0.98
Phrase 6	0.18	1.06	2.43
Phrase 7	0.69	2.60	5.29
Phrase 8	0	1.44	3.99
Phrase 9	0.31	3.40	1.45
Phrase 10	0	3.79	2.37
Phrase 11	0.45	2.25	8.62
Phrase 12	0.89	4.21	8.00
Phrase 13	1.05	1.05	7.33
Phrase 14	0.87	2.78	6.46
Phrase 15	0.37	0.74	0.98
Phrase 16	0.69	3.72	5.88
Phrase 17	0.25	3.48	2.58
Phrase 18	0	0.78	0
La moyenne	0.37	2.41	3.41

Tableau C. 3 : Taux d'erreurs obtenues en utilisant les marques estimées à base du MBS.

Types d'erreurs	FA(%)	MR (%)	σ
Phrase 1	0	0.47	1.44
Phrase 2	0	1.13	3.26
Phrase 3	0.15	1.21	2.31
Phrase 4	0	0	1.26
Phrase 5	0	2.09	3.73
Phrase 6	0.53	0.35	2.51
Phrase 7	0	1.21	5.80
Phrase 8	0.24	1.44	3.77
Phrase 9	0.62	1.85	1.97
Phrase 10	0.25	1.77	3.78
Phrase 11	1.12	0.45	8.52
Phrase 12	0.22	1.33	8.51
Phrase 13	1.05	0	7.38
Phrase 14	0.69	1.39	6.61
Phrase 15	0	0	1.27
Phrase 16	0.69	1.40	6.18
Phrase 17	0.50	0.75	2.89
Sentence18	0	0	2.12
La moyenne	0.34	0.94	4.07

Tableau C. 4 : Taux d'erreurs obtenues en utilisant les marques estimées à base du MBS après correction.

- ***Homme***

Types d'erreurs	FA (%)	MR (%)	σ
Phrase 19	1.77	9.77	9.44
Phrase 20	1.44	3.86	9.03
Phrase 21	4.31	5.31	6.33
Phrase 22	2.98	6.03	9.24
Phrase 23	4.08	4.08	7.50
Phrase 24	5.10	7.65	6.82
Phrase 25	2.65	5.31	2.53
Phrase 26	0	3.06	9.51
Phrase 27	1.90	4.76	9.16
Phrase 28	1.08	4.69	4.22
Phrase 29	1.31	3.07	0.74
Phrase 30	0.71	3.58	2.72
Phrase 31	1.38	6.37	0.35
Phrase 32	2.75	5.96	5.37
Phrase 33	1.71	6.43	0.068
Phrase 34	4.76	7.61	9.73
Phrase 35	0.50	5.55	0
Phrase 36	1.29	6.03	4.62
Moyenne	2.21	5.53	5.41

Tableau C. 5 : Taux d'erreurs obtenues en utilisant les marques estimées à base du MBS.

Types d'erreurs	FA (%)	MR (%)	σ
Phrase 19	1.33	7.55	5.28
Phrase 20	2.41	1.44	5.79
Phrase 21	0.66	0.99	9.74
Phrase 22	0.99	0.99	6.72
Phrase 23	1.36	2.04	3.82
Phrase 24	0.85	0.85	8.41
Phrase 25	0	1.06	4.41
Phrase 26	0.61	2.45	7.29
Phrase 27	0.95	0	4.86
Phrase 28	0.36	4.69	8.44
Phrase 29	1.31	1.75	3.07
Phrase 30	0	2.50	4.57
Phrase 31	1.38	3.32	7.67
Phrase 32	0	5.50	4.82
Phrase 33	0.42	5.15	4.26
Phrase 34	0.95	2.38	1.94
Phrase 35	1.01	5.05	2.91
Phrase 36	0	4.31	6.09
Moyenne	0.81	2.89	5.56

Tableau C. 6 : Taux d'erreurs obtenues en utilisant les marques estimées à base MBS après correction.

Bibliographie

[1] T. Dudoit : 'Introduction au Traitement Automatique de la Parole ', Notes de cours, Première édition, Faculté polytechnique de Mons, TCTS Lab, France ,2000.

[2] Calliope : 'La Parole et Son Traitement Automatique', Edition Masson, 1989.

[3] D. G. Childers, M. Hahn and J. N. Larar: 'Silent and Voiced/ Unvoiced/ Mixed Excitation (Four-Way) Classification of Speech ', IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 31, n°11, November 1989.

[4] C. D'Alessandro : ' La Parole ', Ecole d'été Ondelettes, 1993.

[5] R. Boite : ' Traitement Automatique de la Parole ', Edition Masson, 1989.

[6] F. Ykhlef : ' Modification de la Fréquence Fondamentale en Vue de la Synthèse de la Parole à Partir de Texte de l'Arabe Standard ', Mémoire de Magister, Université de Saad Dahleb de Blida, Algérie, Septembre 2007.

[7] Jean Laroche : ' Cours sur le Traitement des Signaux Audio-Fréquences ', Département du Signal, Groupe Acoustique-Télécom Paris, Février 1995.

[8] G. Blanchet et M. Charbit : ' Traitement Numérique du Signal ', Technique de l'ingénieur, Traité Electronique, E-3 087, 2001.

[9] J .Hernandez : 'Algorithme d'Acquisition et Compression à Vitesse Variable, Etude et Mise en Place ', Mémoire de Master, ENSEA, France, Avril 1995.

[10] LE Manh Tuan : ' Analyse des Voyelles Spéciales du Vietnam ', Rapport final, Institut de la Francophonie pour l'Informatique, Hanoi, France, 2008.

[11] R. Amiar, A. Hecini et M. Kessas : 'Etude et Simulation des Algorithme de Détection de la Fréquence Fondamentale d'un Signal Vocal ', Mémoire d'ingénieur d'Etat, Département d'Electronique, Université Saad Dahlab de Blida, Octobre 2008.

[12] Van Loo Jonathan : 'Analyse du Signal Vocal: Détermination de la Fréquence Fondamentale ', Rapport de Stage, Ecole Nationale supérieure d'Electricité et de mécanique, France, 2008.

[13] L. Buniet : 'Traitement Automatique de la Parole en Milieu Bruité : Etude de Modèles Connexionnistes Statiques et Dynamiques ', Thèse de Doctorat, Spécialité Information, Université Henri Poincaré-Nancy 1, UFR STMIA, France, Février 1997.

[14] M. André : 'Eléments de Linguistique Générale ', Librairie Armand Colin, Paris, 1970.

[15] J. Benzerrouk et JA. Djebbar : ' Etude et Simulation d'un Banc de Filtre QMF à Base des Filtres RII pour le Codage de la Parole par la Méthode des sous-Couches ', Mémoire d'Ingénieur d'Etat, Département d'Electronique, Université de Saad Dahleb, Blida, 1996.

[16] T. F. Quatieri:' Discrete Time Speech Processing Principles and Pratice', Prentice Hall PTR, Upper Saddle River, 2001.

[17] Site web: IOWA State University, http://www.uiowa.edu/~acadtech/phonetics/english/frameset.html, [20/05/2012].

[18] Dr. C. George: ' Phonetics ', Tutorial, Shippensburg University, Copyright 2005, Site web, http://webspace.ship.edu/cgboer/phonetics.html, [10/05/2012].

[19] T. Dutoit : ' Je Parle Donc Je Suis ', bilan des développements récents en traitement automatique de la parole, Faculté Polytechnique de Mons, TCTS Lab, 2000.

[20] M. Kabache : ' Application des Réseaux des Neurones à la Reconnaissance Automatique des Phonèmes Spécifiques en Arabe Standard ', Mémoire de magister, CRSTDLA, Alger, Algérie, Mai 2005.

[21] Dr. AndrezjDrygajlo : ' Traitement de la Parole ', Notes de cours, ELE 233, Ecole Internationale des Sciences de Traitement de l'Information, France, 2011.

[22] G. Droua-Hamdani : ' Prédiction de la Durée Segmentale des Phonèmes de l'Arabe Standard ', Mémoire de Magister, CRSTDLA, Alger, Algérie, Février 2004.

[23] Ch. Meunier : ' Parole et Langage ', Notes de cours, Laboratoire Parole et Langage, Université de Provence, France, 2007.

[24] M. Brahim : ' Analyse du Signal de Parole par les Ondelettes, Application à la Reconnaissance des Mots Isolés ', Mémoire de Magister, Département d'Electronique, Université de Batna, Algérie, 2009.

[25] G. Peeters : ' Modèles et Modification du Signal Sonore Adaptés à ses Caractéristiques Locales', Thèse de Ph Doctorat, Université Paris 6, Juillet 2001.

[26] E. Moulines and F. Charpentier: 'Pitch-Synchronous Waveform Processing Techniques For Text-to-Speech Synthesis Using Diphones' , Speech Communication, Vol. 9, n° (5/6), pp. 453-467, December 1990.

[27] E. Moulines and J. Laroche: ' Non-Parametric Techniques for Pitch-Scale and Time-Scale Modification of Speech', Speech Communication, Vol. 16, pp. 175-205, 1995.

[28] M. Narendranath, H. Murthy, S. Rajendran and B. Yegnanarayan: ' Transformation of Formants for Voice Conversion Using Artificial Neural Networks ', Speech Communication, Vol. 16, Issue 2, pp. 207–216, 1999.

[29] F. Ykhlef, M. Bensebti and L. Bendaouia: 'A New Method for Time-Frequency Modification of Speech Signal Using a Combining Cosine Modulated Pseudo-QMF-Bank & SOLA Algorithm for Listeners with Hearing Impairment', The 2nd International Conference on Advanced Computer Theory and Engineering (ICACTE '9), Egypt, September 2009.

[30] Site web: http:// www.dspdimension.com, [10/02/2012].

[31] M. R. Portnoff: ' Time-Scale Modification Based on Short-Time Fourier Analysis ', IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. Assp-29, No. 3, pp. 374-390, June 1981.

[32] D. Malah: ' Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Speech Signals ', IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. ASSP-27, No.2, pp. 121-133, 1979.

[33] D.W. Griffin and J. S. Lim; 'Signal Estimation from Modified Short-Time Fourier Transform', IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. Assp-32, No.2, pp. 236-243, April 1984.

[34] S. Roucos and A. M. Wilgus: ' High Quality Time –Scale Modification for Speech ', IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, FL, pp.493-496, March 1985.

[35] S. Patrick-André : ' Méthode Hybride de Modification de Durée d'un Signal Audio ', Mémoire de Maîtrise et Sciences Appliquées, Département de Génie Electrique et Génie Informatique, Université de Sherbrooke, Canada, Mai 2008.

[36] P. Grégory : ' Dilatation et Transposition sous Contraintes Perceptives des Signaux Audio : Application au Transfert Cinéma-Vidéo ', Thèse de Doctorat, Département de Mécanique, Physique et Modélisation, Université de la Méditerranée - Aix-Marseille ii, Juin 2003.

[37] D.J. Hejna, B.R. Musicus, and A.S. Crowe: ' Method for Time-Scale Modification of Signals ', United States Patent No: 5 175 769, December 1992.

[38] R. Muralishankar, A.G. Ramakrishnan and P. Prathibha: ' Modification of Pitch Using DCT in the Source Domain ', Speech Communication, Vol. 42, Issue 2, pp. 143–154, February, 2004.

[39] J. Laroche and M. Dolson: 'New Phase-Vocoder Techniques for Pitch-Shifting, Harmonizing and Other Exotic Effects ', IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 91-94, 1999

[40] K. Kumar and J. Jain; ' Speech Pitch Shifting using Complex Continuous Wavelet Transform ', IEEE Annual India Conference, pp. 1-4, India, September 2006.

[41] J. Laroche : ' Traitement des Signaux Audio-Fréquences', Notes de cours, Département Signal, Groupe Acoustique-TELECOM Paris, Février 1995.

[42] J. D. Markel and A.H. Jr. Gray: ' Linear Prediction of Speech', Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[43] B. Atal and S. Hanauer: ' Speech Analysis and Synthesis by Linear Prediction of the Speech Wave ', Journal of the Acoustical Society of America, Vol. 50, No 2, pp. 637–655, 1971.

[44] E. Moulines and W. Verhelst: ' Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech ', Elsevier Science, Book Chapter in Speech Coding and Synthesis, 1995.

[45] Y.Lifu, T.Jing and S. Jingcheng: ' Applying Source-Filter Model in Chinese Speech Synthesis', International Symposium on Chinese Spoken Language Processing (ISCSLP), China, 2002.

[46] R. J. McAulay and T. F.Quatieri: ' Speech Transformation Based on a Sinusoidal Representation ', IEEE Transactions on Acoustics, Speech, And Signal Processing, Vol. ASSP-34, No. 6, pp. 1449 – 1464, 1986.

[47] D. David: ' Audio Time-Scale Modification ', PhD Thesis, School of Control Systems and Electrical Engineering, Dublin Institute of Technology, Dublin, 2005.

[48] R. Badeau : ' Modification Temporelle et Spectrale ', Notes de cours, Télécom-Paris, Ecole Nationale Supérieure des Télécommunications, Décembre 2007.

[49] P. A. Naylor, A. Kounoudes, J. Gudnason and M. Brookes: ' Estimation of Glottal Closure Instants in Voised Speech Using the DYPSA Algorithm ', IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, n°1, January 2007.

[50] CCITT, Recommendations of the P Series: 'Method for the Evaluation of Service from the Stand Point of Speech Transmission Quality'. CCITT Red book, Vol. V-VIII the Plenary Assembly, 1984.

[51] A. Amehraye : ' Débruitage Perceptuel de la Parole ', Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications de Bretagne, Bretagne, Mai 2009.

[52] Site web: 'Speech Quality and Evaluation ', http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap10.html, [15/05/2012].

[53] S. Lemmetty: 'Review of Speech Synthesis Technology', Master thesis, Helsinki University of Technology, Departement of Elecrical and Communications Enginneering, March 1999.

[54] AM. Engebretson: 'Benefits of Digital Hearing Aids ', IEEE Engineering in Medicine and Biology Magazine, Vol. 13, Issue 2, pp. 238-248, April 1994.

[55] Dj. Van Tasell: ' Hearing Loss, Speech and Hearing Aids', Journal of Speech and Hearing Research, Vol. 36, No 2, pp. 228-244, 1993.

[56] Y. Nejime, T. Aritsuka, and T. Imamura: ' A Portable Digital Speech Rate Converter for Hearing Impairment', IEEE Transactions on Rehabilitation Engineering, Vol. 4, pp. 73-83, 1996.

[57] R. Olivier et C. Didier: 'Procédé et Dispositif de Modification d'un Signal Audio', European Patent Office, EP1 970 894 A1, Paris, 2008.

[58] G. Baudoin et J. Cernocky : ' Codage de la Parole à Bas et Très Bas Débit', Annales des Télécommunications, 2000.

[59] T. Drugman and T. Dutoit: ' Glottal Closure and Opening Instant Detection from Speech Signals', Proceeding of Interspeech, pp. 2891–2894, Brighton, UK, 2009.

[60] M. Legàt, D. Tihelka and J. Matoušek: ' Pitch Marks at Peaks or Valleys?', Lecture Notes in Artificial Intelligence, pp. 502–507, Springer, 2007.

[61] K. Sjolander and J. Beskow: 'Wavesurfer— An Open Source Speech Tool', Proceedings of International Conference on Spoken Language Processing (ICSLP), pp. 464–467, China, October 2000.

[62] Site web, CMU ARCTIC speech synthesis databases, http://festvox.org/cmuarctic/, [17/10/2011].

[63] J. L. CROWLEY :' Filtrage Numérique', Notes de cours, Traitement du Signal, Octobre 2000.