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

MEMOIRE DE MAGISTER

Présenté par :

OUNNAS Amine

Ingénieur d'Etat en Electronique

Option : Communication USD-Blida

SYNTHESE DE LA PAROLE EN ARABE STANDARD

Soutenue devant le jury :

Président	**R. AKSAS**	**Professeur à l'ENP**
Rapporteur	**M. GUERTI**	**Professeur à l'ENP**
Examinatrices	**Mme L.HAMAMI**	**Professeur à l'ENP**
	Melle N. BENBLIDIA	**MC.A à l'USD-BLIDA**

Décembre 2011

DEDECACES

A mes chers parents

A mes frères

A toute ma famille

A tous mes amis

A tous ceux qui ont contribué à la réalisation de ce Mémoire.

Amine.

REMERCIEMENTS

Je tiens à remercier ***Dieu*** *de m'avoir donné la patience de terminer ce travail de Magister.*

Je remercie vivement mon encadreuse ***Professeur GUERTI Mhania*** *pour m'avoir confié ce travail d'abord et pour son soutien constant, son rôle majeur et sa grande patience ainsi que ses encouragements durant toute la période de ce travail. Je la remercie pour ses compétences, son ouverture d'esprit et sa grande disponibilité.*

Je remercie les membres du jury, qui m'ont fait l'honneur de participer au jugement de ce travail.

J'exprime ma reconnaissance à Monsieur ***Rabia AKSAS****, Professeur à l'Ecole Nationale Polytechnique, d'avoir accepté de présider le jury de mon mémoire.*

Mes plus sincères remerciements sont adressés à Madame ***Latifa HAMAMI****, Professeur à l'Ecole Nationale Polytechnique, qui a bien voulu participer au jury.*

Je remercie également M^{elle} ***Nadjia BENBLIDIA,*** *Maitre de conférences à l'Université de Blida, d'avoir accepté de faire partie de jury.*

A.OUNNAS

ملخص

إن هذا العمل يهد ف إلى تطوير نظام تركيب الكلام في اللغة العربية الفصحى. لهذا قمنا بدراسة تقنيات تغيير العناصر العروضية المستعملة في إشارات الكلام لهدف إجراء تغييرات عروضية للإشارات الصوتية و التركيب الاصطناعي للكلام . اهتممنا في مشروعنا باعتماد تطبيق تقنيات التعديل بتغيير التردد الابتدائي و تغييرالزمن, وعليه قمنا بتصميم طريقة تغيير تسمى TD-PSOLA (تداخل وإضافة متزامنة مع التردد في المجال الوقتي) و استعملنا إشارات كلام ملفوظة عن طريق متكلمين (ذكور و إناث) باللغة العربية الفصحى, ثم قمنا بتحليليها و تقييمها لهدف اختيار أحسن عوامل التغيير الكلام ذو نوعية جيدة.

كلمات المفاتيح تركيب الكلام, اللغة العربية الفصحى, التردد الابتدائي, TD-PSOLA.

Résumé

Le but de notre travail est d'élaborer un système de synthèse de la parole en Arabe Standard. Pour cela nous avons étudié les techniques de synthèse de la parole pour effectuer des modifications prosodiques du signal vocal. Nous nous sommes intéressés dans notre cas à la modification de la fréquence fondamentale et la dilatation du temps. Pour ce faire, nous avons implémenté une technique de modification appelée TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) et nous avons utilisé un corpus constitué de phrases affirmatives en Arabe Standard (AS), prononcées par des locuteurs arabophones (masculins et féminins). Nous avons fait une évaluation de la technique utilisée, afin de tirer les meilleurs facteurs de modifications offrant une bonne qualité de la parole synthétique.

Mots clés: Synthèse de la parole, Langue Arabe Standard, Fréquence fondamentale, TD-PSOLA.

Abstract

The objective of our work is to develop a system of speech synthesis in Standard Arabic. For this we studied the techniques of speech synthesis to make changes in the speech signal prosodic. We were interested in our case to the modification of the fundamental frequency and the time dilatation. To do this, we have implemented a modification technique known as TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) and we used a corpus consisting of affirmative sentences in Standard Arabic (SA), marked by Arabic speakers (male and female). We made an evaluation of the technique used to get the best factors changes with good quality of synthetic speech.

Keywords: Speech synthesis, Standard Arabic Language, fundamental frequency, TD-PSOLA.

Liste des Abréviations

AMDF	Average Magnitude Difference Function
AS	Arabe Standard
CT	Court-Terme
F_0	Fréquence Fondamentale
FD-PSOLA	Fréquency Domain-PSOLA
Fe	Fréquence d'échantillonnage
FP	Fonction de Périodicité
FT	Fonction de Transfert
IPS	Instrumental Pitch Shifting
LPC	Linear Predictive Coding
LP-PSOLA	Linear prediction PSOLA
M I C	Modulation par Impulsions Codées
OLA	OverLap and Addition
OEAO	Oulits Enseignement Assisté par Ordinateur
PSOLA	Pitch synchronous Overlap and Add
SIFT	Simplified Inverse Filter Tracking
SRPD	Super Resolution Pitch Determination
SOLA	Synchronized Overlap-Add
T_0	Période Fondamentale
TAP	Traitement Automatique de la Parole
TD-PSOLA	Time Domain PSOLA
TDHS	Time Domain Harmonic Scaling
TFD	Transformée de Fourier discrète
TF	Transformée de Fourier
TFI	Transformée de Fourier Inverse
TPZ	Taux de Passage par Zéros
TTS	Text-To-Speech
V/NV	Voisé /Non Voisé
VODer	Voice Operation Demonstrator
WSOLA	Waveform Similarity Overlap-Add

Table des Figures

Sommaire

Sommaire

Introduction Générale

Introduction Générale

La communication par la voix est l'un des enjeux majeurs du dialogue Homme-Machine, puisque la voix véhicule à la fois un contenu linguistique explicite que l'on peut représenter sous forme écrite et un contenu non linguistique comme le type du locuteur, son attitude, ses gestes, etc. Cela rend le Traitement Automatique de la Parole (TAP) une composante fondamentale des sciences de l'ingénieur et un domaine de recherche actif, au croisement du traitement du signal numérique et du traitement symbolique du langage. Depuis les années 60, le TAP bénéficie d'efforts de recherche très importants, liés au développement des moyens et techniques de télécommunications et du traitement numérique de l'information. Ces efforts se sont concrétisés grâce à plusieurs applications du TAP, telles que le codage, la Reconnaissance Automatique et la synthèse de la parole.

Les techniques modernes de TAP tendent cependant à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions :

- les *analyseurs* de parole cherchent à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (on parle alors d'*analyseur perceptuel*), mais jamais tel qu'il est compris, ce rôle étant réservé aux reconnaisseurs. Les analyseurs sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse, soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical ou l'étude des langues ;
- les *reconnaisseurs* ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la *reconnaissance du locuteur*, dont l'objectif est de reconnaître la personne qui parle, et la *reconnaissance de la parole*, où l'on s'attache plutôt à reconnaître ce qui est dit ;
- les *synthétiseurs* ont quant à eux, la fonction inverse de celle des analyseurs et des reconnaisseurs de parole : ils produisent de la parole artificielle ;
- enfin, le rôle des *codeurs* est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole.

Dans le cadre des travaux de recherches de l'équipe du laboratoire Signal et Communication, notre travail s'inscrit dans le domaine de Traitement Automatique de la Parole, en particulier la synthèse de la parole.

Les méthodes reposant sur le principe de synchronisation avec le fondamental sont utilisées pour réaliser des modifications temporelles ou fréquentielles d'un signal de parole, ou pour mettre en œuvre des systèmes de synthèse de la parole. Ces méthodes nécessitent au préalable un marquage des périodes du fondamental. La méthode PSOLA (Pitch Synchronous OverLapp and Add), est l'une des variantes d'**OLA (**OverLapp And Add**)** qui se ramifie en plusieurs techniques (Time Domain PSOLA ou TD-PSOLA, Frequency Domain PSOLA ou FD-PSOLA, Linear Prediction PSOLA ou LP-PSOLA).

L'algorithme PSOLA consiste à concaténer, à l'aide d'un lissage, des unités de parole pré-stockées en modifiant le pitch et la durée des segments. Cette technique est associée à la méthode de synthèse par concaténation.
Le but de ce mémoire est d'élaborer un système de synthèse de la parole et d'effectuer des modifications prosodiques de signal vocal en utilisant la technique TD-PSOLA.

Pour atteindre cet objectif, nous avons structuré notre travail en quatre chapitres :

- dans le premier chapitre, nous allons décrire d'une manière générale des notions sur le traitement de la parole, des spécifications du signal vocal et des notions fondamentales sur l'arabe standard ;
- le deuxième chapitre s'articule autour des principes de la synthèse vocale, suivi d'une description bien détaillée des différentes techniques et méthodes utilisées ;
- dans le troisième, nous développerons la technique qui permet de faire la synthèse d'un signal de parole, Soit la technique PSOLA ;
- le dernier chapitre concerne la simulation et l'interprétation des résultats obtenus dans le cadre de notre application ;

Enfin, nous présentons des conclusions et des perspectives concernant la thématique abordée.

Chapitre 1 :

Généralités sur la Parole et l'Arabe Standard

1.1. Introduction

La parole est le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés. Dans ce chapitre nous allons décrire de manière générale des notions sur le traitement de la parole, des spécifications du signal vocal et des notions fondamentales sur l'Arabe Standard.

1.2. Bref historique sur le Traitement Automatique de la Parole

Depuis les temps les plus reculés, les Hommes ont toujours eu l'ambition de faire produire à des dispositifs artificiels des actions d'hommes ou d'animaux (figure1.1). Nombreuses sont les légendes qui témoignent de la persistance de ce désir. Des figurines animées (à la main) ont été fabriquées dès l'antiquité.Mais c'est vers la fin du XVIII^{ème} siècle qu'a vu maître les premières machines mécaniques capables de simuler les sons vocaux.

En 1779, l'Académie impériale de Saint-Pétersbourg organise un concours scientifique avec deux questions : qu'est ce qui différencie autant les voyelles des autres sons ? Est-il possible de faire prononcer par une machine, les sons de ces voyelles ? Le lauréat est un professeur de l'université de Copenhague, Christian Gottlieb Kratzenstein qui réalise une série de résonateurs acoustiques de dimensions et formes similaires à celles de la bouche humaine, et excités par une anche vibrante simulant le fonctionnement des cordes vocales (figure1.2).

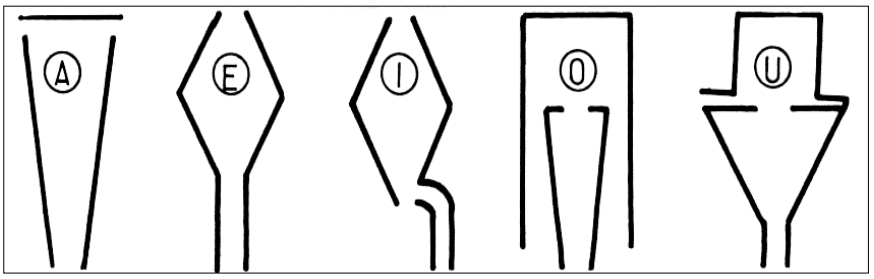


Figure 1.1 : Dispositifs artificiels [1].

En 1939, Homer Dudley, présente le Voder - "voice operation demonstrator" à l'Exposition internationale de New York. Cet appareil est excité soit par un bruit blanc, soit avec un signal très riche en harmoniques, ces sons étant modulés par une boîte de contrôle de résonance - le "conduit vocal" - contenant un banc de dix filtres passe-bande répartis entre 300 et 3000 Hz. Les recherches sur la synthèse de la parole sont motivées par le souci de transmettre la voix avec une plus grande efficacité, c'est-à-dire en réduisant la largeur de bande nécessaire aux

conversations téléphoniques. C'est la raison pour laquelle, très tôt, des recherches sont menées aux laboratoires de Bell Telephone. Ces recherches ont conduit à l'élaboration du vocoder ou « Voice Coder » qui permet d'utiliser uniquement une bande passante de 275Hz au lieu des 3100Hz requis pour le téléphone.

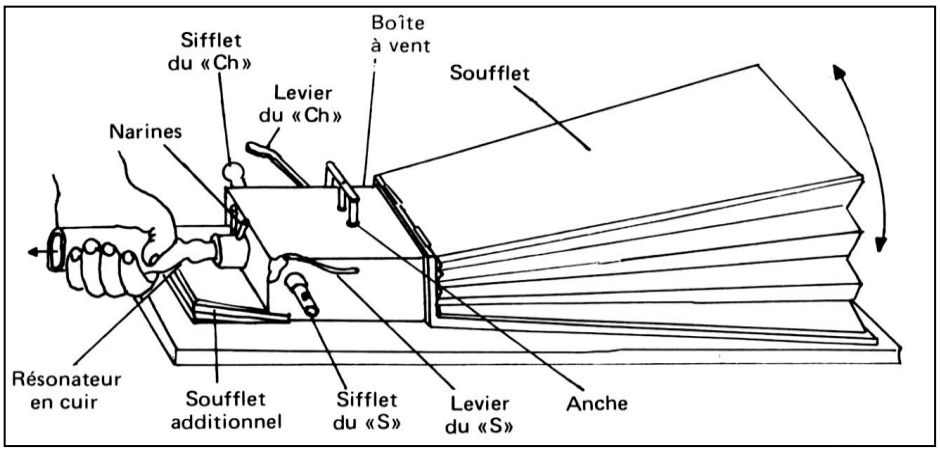


Figure 1.2 : La machine parlante de Wolfgang Von Kempelen 1791 [1].

Les années soixante et soixante dix ont vu apparaître d'autres techniques permettant de synthétiser la parole notamment la synthèse par éléments phonétiques. Cela consiste à reconstituer artificiellement des mots à partir de segmants de mots par exemple avec 'auto', 'mati','que' on peut reconstituer le mot « automatique ». Le découpage peut se faire de façon plus fine encore jusqu'à la plus petite unité phonétique : le phonème. L'assemblage des phonèmes, selon un ensemble de lois particulières à chaque langage permet de reconstituer les mots parlés.

Il existe de nombreuses autres méthodes permettant de générer de façon synthétique un signal de parole telle la synthèse par formants qui donne de meilleurs résultats mais demande une analyse fine du signal de parole.

Le développement des techniques numériques de traitement du signal a permis l'intégration de cette technique dans un circuit intégré et cela dès 1978 par Texas Instruments.

C'est à ce moment aussi que le traitement de la parole proprement dit, c'est-à-dire le traitement de l'information contenue dans le signal vocal a pris un essor considérable.

Malgré cela la recherche dans le domaine du traitement de la parole est toujours très active dans divers domaines :

- recherche de codages de plus en plus efficaces dans le but de réduire le débit binaire du signal de parole ;

- reconnaissance de la parole (Dialogue Homme Machine) ;
- synthèse de la parole à partir d'un texte écrit ;
- identification d'un locuteur ;
- l'apprentissage de langues étrangères.

Ces quelques exemples nous montrent les secteurs pouvant tirer parti de l'avancée des recherches dans le domaine du traitement de la parole. Or il se trouve qu'aujourd'hui le grand public ainsi que les industriels sont demandeurs de nouvelles technologies intégrant un module de traitement de la parole.

On peut citer par exemple les opérateurs de téléphonie mobile toujours à la recherche d'un service pouvant satisfaire des clients de plus en plus exigeant. C'est le cas de Bouygues Telecom qui propose une assistante personnelle virtuelle qui permet de gérer les appels de vos correspondants : elle filtre vos appels, prend les coordonnées des correspondants qui ont cherché à vous joindre et plus encore !

En conclusion, ces quelques applications montrent que le traitement de la parole prend une part de plus en plus importante dans notre vie quotidienne. Dans un futur proche on peut parier que tout ou presque se fera à l'aide de la parole et cela est d'autant plus vrai que les microprocesseurs chargés de faire les traitements sont plus rapides et plus petits [1].

1.3. Mécanisme de la production de la parole

Qu'est ce que le signal de parole ?

Par définition, le son est ce que l'oreille perçoit de la vibration d'un corps. Cette vibration est une sorte d'onde (produite par un objet, guitare, piano, tambour, marteau, etc.), qui se propage par et à travers des corps physiques (air, eau, métal, bois, etc.), La parole se distingue des autres sons par des caractéristiques acoustiques ayant leurs origines dans le mécanisme de production [2].

Le signal de parole est généré par l'appareil phonatoire. C'est un organe d'une grande complexité mécanique. Il se compose de deux parties anatomiquement distinctes. Les poumons et le larynx, partie supérieure de la trachée artère, constituent l'essentiel du générateur sonore. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée artère (figure1.3).

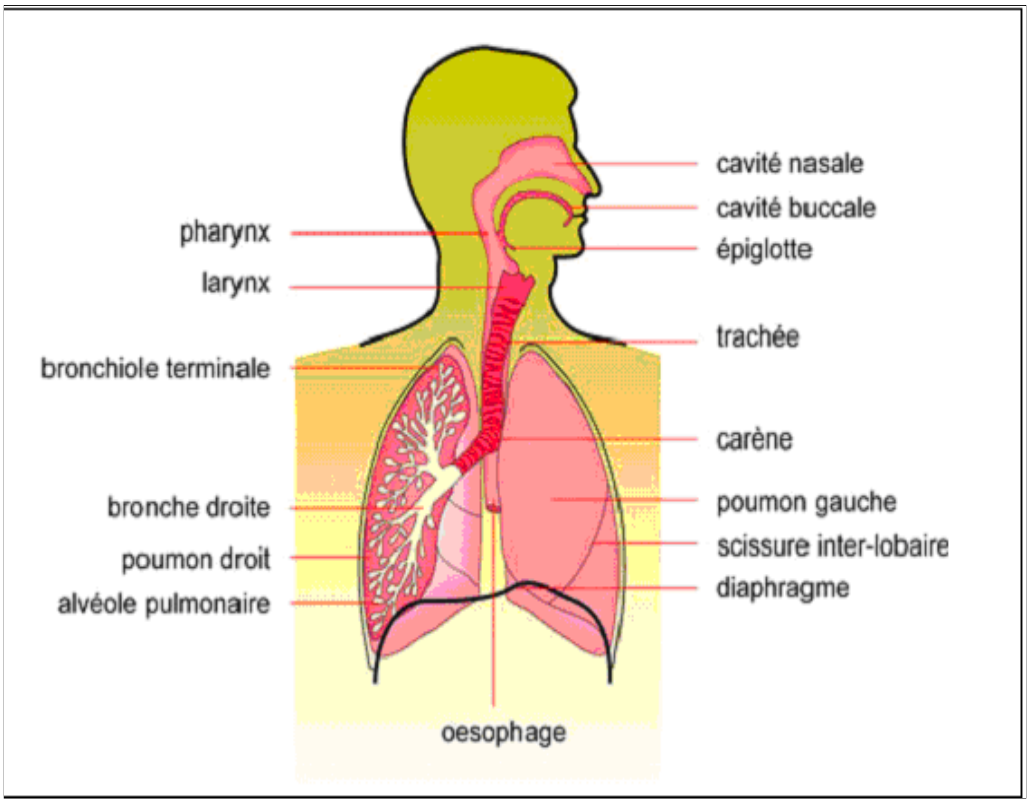


Figure 1.3 : Modèle simplifié de l'appareil phonatoire [3].

Le larynx a une fonction qui lui est propre : c'est la production des sons, ou « phonation ». Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant, déterminer une ouverture triangulaire appelée glotte. Pendant la respiration, l'air y passe librement et aussi pendant la phonation des sons sourds ou non voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales ; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal qui s'étend du pharynx jusqu'aux lèvres.

La voix résulte du fonctionnement simultané des poumons, du larynx et de la cavité de la bouche et du nez qui modifie sa forme et ses dimensions suivant le son émis et qui, avec la poitrine, jouent le rôle de caisse de résonance. Toutes les voix se ressembleraient si la voix était seulement laryngée. Or, ce sont les modifications de forme et de dimensions que subissent la bouche, le pharynx pendant l'émission de la voix, qui donnent au contraire à celle-ci un timbre qui est particulier à chacun d'entre nous.

On peut remarquer que, d'après ce que nous avons vu précédemment, nous avons deux générateurs de sons, source excitatrice les cordes vocales, quand elles vibrent le son est voisé ou sonore et quand elles ne vibrent pas le son est sourd ou non voisé, et d'un filtre (le conduit vocal) capable d'amplifier ou d'amortir certains sons [1].

En résumé, sans entrer dans les détails, un son voisé est un signal quasi périodique et un son non voisé peut être considéré comme un bruit blanc. Cette manière de modéliser la parole est un peu sommaire mais permet de réaliser des modifications satisfaisantes des paramètres prosodiques.

1.4. Les paramètres prosodiques d'un signal vocal

La prosodie est une science de la linguistique qui étudie les éléments phoniques (l'accent, l'intonation, etc.) de n'importe quelle langue, et puisque la parole est un signal réel d'énergie finie, continu, et non stationnaire ; les variations des paramètres prosodiques physiques (La fréquence fondamentale, la durée, et l'intensité) influencent de manière directe sur ces éléments phoniques.

Les recherches en linguistique ont montré que les caractéristiques prosodiques sont des composantes indispensables à la langue et à la fonction de communication. Puisqu'elles influencent directement sur l'intelligibilité de la parole synthétique. Il existe trois manières de définir les paramètres prosodiques, selon qu'on les considère sur les plans de la production, de l'acoustique, et de la perception auditive [4].

1.4.1. La Fréquence Fondamentale F_0

La Fréquence Fondamentale est la fréquence de vibrations des cordes vocales, elle varie d'une personne à une autre en fonction de la longueur et de la masse des cordes vocales de chaque personne (figure1.4). Elle permet de diviser l'ensemble des sons de parole en trois grandes macros classes [5]:

- 70 -250 Hz pour les hommes ;
- 150 - 400 Hz pour les femmes ;
- 200 - 600 Hz pour les enfants.

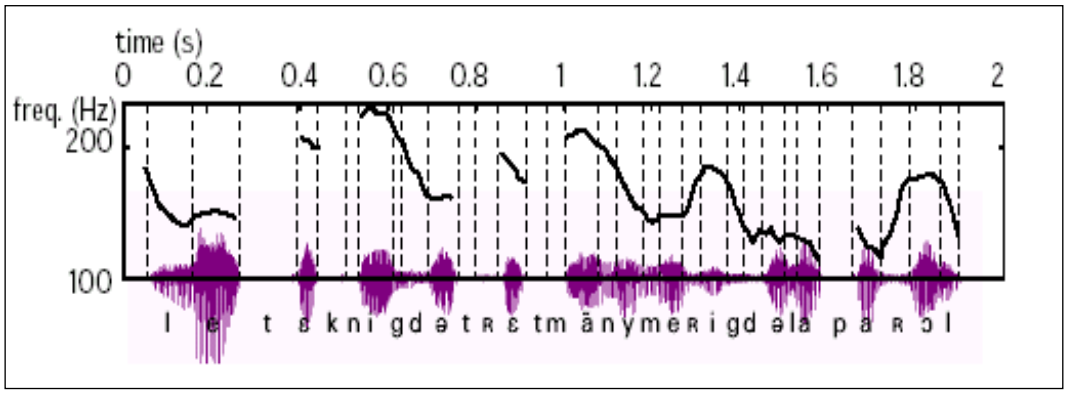


Figure 1.4 : Evolution de la fréquence de vibrations des cordes vocales de la phrase :"les techniques de traitement numérique de la parole" [5].

Les variations de la fréquence au cours de la parole constituent ce qu'on appelle la mélodie ou l'intonation. Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la F_0 [5].

1.4.2. La duré

La durée est une mesure très variable. Elle représente le temps de la prononciation d'un phonème. Il existe deux types :

- la durée observée, qui correspond à la mesure objective du temps de l'activation des organes de phonation ;
- la durée perçue, est liée au mécanisme de la perception et elle est fréquemment utilisée dans le cas des occlusives puisqu'elles sont caractérisées par une durée de réalisation non continue.

Généralement la durée d'une unité est mesurée par le nombre des trames qu'elle contient. Pour calculer la durée de chaque trame il faut fixer deux événements sur le signal de parole qui délimitent les repères initial et final de cette trame.

1.4.3. L'Intensité ou l'énergie

Elle est résultante de la pression sous glottique. Généralement elle exprime le volume sonore d'un phonème et dans le cas d'un voisement elle représente l'amplitude des vibrations des cordes vocales. Elle est exprimée pour un signal échantillonné x_n par :[5]

$$E = \frac{1}{T}\sum_{n=1}^{T} x_n^{\,2} \tag{1.1}$$

$$E_{dB} = 10 \times \log_{10}\left(\frac{1}{T}\sum_{n=1}^{T} x_n^{\,2}\right) \tag{1.2}$$

Le rythme d'élocution correspond à la vitesse du débit de parole. On peut faire varier ce paramètre de manière à ce qu'une phrase prononcée trop rapidement puisse être « ralentie » pour la rendre plus compréhensible lors de l'apprentissage d'une langue étrangère par exemple. L'intensité du son émis est liée à la pression de l'air en amont du larynx. Les figures 1.5 et 1.6 représentent l'évolution temporelle du signal vocal pour les mots cinq et parenthèse, elles donnent un exemple des parties voisées et non voisées du signal vocal.

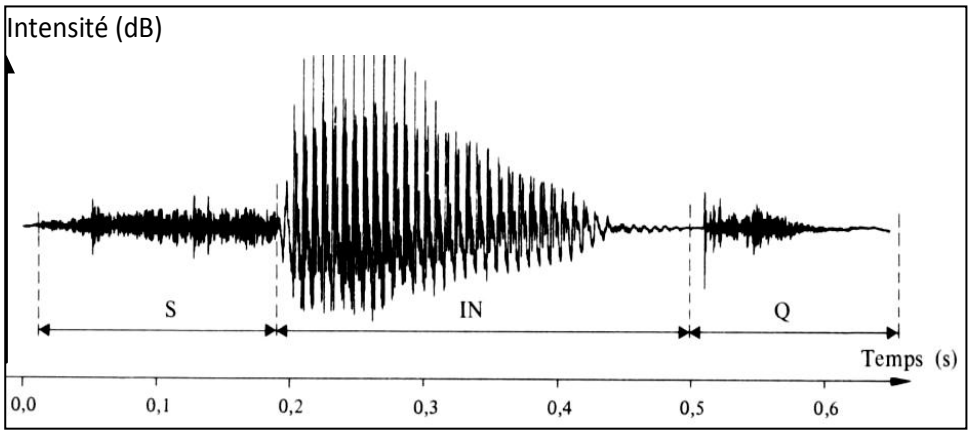


Figure 1.5 : Audiogramme du signal de parole du mot « Cinq » [1].

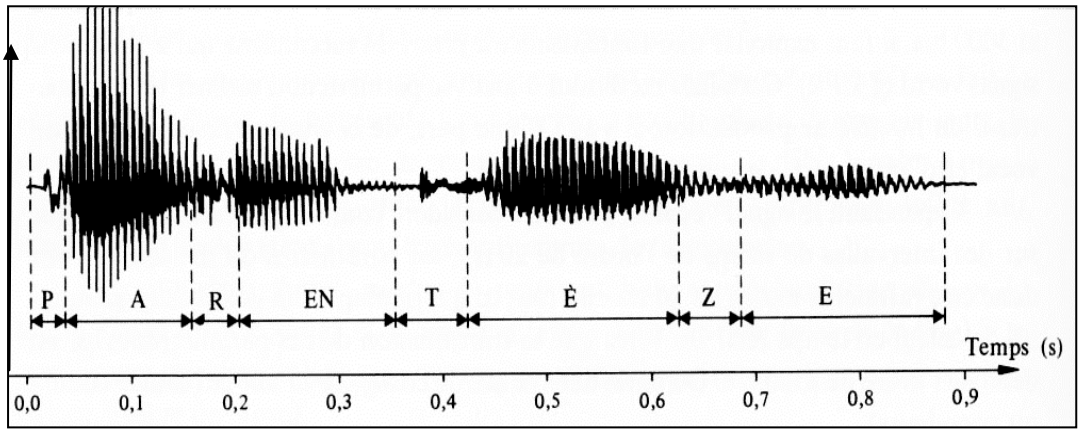


Figure 1.6 : Audiogramme du signal de parole du mot « Parenthèse » [1].

Tout l'enjeu du traitement de la parole est de modéliser l'appareil phonatoire humain de façon à créer un signal de parole synthétique aussi réaliste que l'original. Il existe plusieurs manières de le faire, notamment en utilisant un vocodeur à prédiction linéaire qui, dans un premier temps, code le signal vocal de manière à réduire le débit d'informations puis le restitue à l'aide de paramètres qui caractérisent la fonction de transfert du conduit vocal. Ces paramètres étant réactualisés toutes les 20 ms environ. En fait on part de l'hypothèse qu'un échantillon de parole peut être prédit à partir d'une pondération linéaire d'un nombre fini d'échantillons précédents.

Cette hypothèse se justifie par le fait que la forme du conduit vocal n'évolue pas rapidement. En général, on considère que l'appareil vocal est quasi stationnaire sur un intervalle de temps de l'ordre d'une vingtaine de millisecondes. On parle donc ici de statistique du signal à court terme. Cette méthode a l'avantage de donner de bons résultats au niveau du signal synthétique mais demande des capacités de calcul important pour la réalisation en temps réel [1].

1.5. Propriétés spécifiques du signal vocal

Le processus de production de la parole présente certaines caractéristiques qui sont liées au signal vocal lui-même (continuité, variabilité, conduit vocal, et encodage).

1.5.1. 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 silences au milieu d'un mot et aucun intervalle entre deux mots successifs. Par conséquent il est très difficile de déterminer le début et la fin des mots composant la phrase.

1.5.2. Variabilité

La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que ce soit pour un même ou plusieurs locuteurs. Parmi ces facteurs, les perturbations apportées par le microphone (selon le type, la distance et l'orientation) et l'environnement (bruit, réverbération).

1.5.2.1. Variabilité intra-locuteur

Elle concerne les différences de production du signal parole chez un même locuteur. Plusieurs critères peuvent être responsables de ces différences :

- la fatigue ;
- l'état émotionnel du sujet : une émotion telle que la qui affecte le timbre et le rythme de la voix ;
- les maladies affectant les organes de la voix.

1.5.2.2. Variabilité inter-locuteur

Des différences acoustiques importantes apparaissent dans un mot prononcé par plusieurs locuteurs. En effet, des contrastes considérables peuvent se manifester suivant l'âge, le sexe, l'origine géographique et *le* lieu social.

1.5.2.3. Variabilité contextuelle

Les mouvements articulatoires peuvent en effet être modifiés de façon à minimiser l'effort 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.

Les phénomènes articulatoires sont dûs au fait que chaque articulateur évolue de façon continue entre les positions articulatoires. Ils apparaissent même dans le parlé le plus soigné. Au contraire, la réduction et l'assimilation prennent leur origine dans des contraintes physiologiques et sont sensibles au débit de la parole. L'assimilation est causée par le recouvrement de mouvements articulatoires et peut aller jusqu'à modifier un des traits phonétiques du phonème prononcé.

La réduction est due au fait que les cibles articulatoires sont moins atteintes dans le parlé rapide. Ces phénomènes sont en grande partie responsables de la complexité des traitements réalisés sur les signaux de parole [6].

1.5.3. Le conduit vocal

Le conduit vocal est un tuyau tridimensionnel qui est excité par une ou deux sources acoustiques. La source laryngienne peut être considérée comme quasi périodique, avec une fréquence pouvant évoluer très rapidement. La seconde source génère du bruit de friction ou d'explosion.

1.5.4. Le codage

Le codage concerne les niveaux lexicaux, syntaxiques, sémantiques, morphologiques et phonétiques (phonèmes et leurs interactions) utilisés souvent pour assurer une meilleure qualité de la parole synthétique [3].

1.6. Analyse de la parole

Le traitement du signal vocal s'inscrit dans une succession de procédures, que ce soit pour la reconnaissance automatique ou pour la synthèse de la parole. Analyse et synthèse sont deux activités duales, l'analyse fournit une description du signal acoustique, que la synthèse utilise pour le reproduire.

L'Analyse acoustique est une partie importante dans le traitement que subit le signal sonore pour pouvoir réaliser un système de haute qualité de synthèse, de compréhension, ou de reconnaissance de la parole.

Cette opération consiste à tirer à partir du signal vocal un ensemble de paramètres pertinents, discriminants et robustes susceptibles de le représenter. Plusieurs techniques d'analyse sont utilisées parmi lesquelles on peut citer l'analyse par :

- Spectrogrammes ;
- Codage Prédictif Linéaire (Linear Predictive Coding ou LPC).

1.6.1. Analyse par spectrogrammes

Dans l'étude du phénomène acoustique, on peut réduire la description du son à trois grandeurs physiques : la fréquence (Hz), la durée (s) et l'amplitude ou l'énergie (dB). Cela signifie que les trois valeurs : durées, fréquence et énergie sont les paramètres pertinents. Une meilleure analyse consiste à les représenter de manière claire et avec précision. L'une des représentations possibles est d'associer deux à deux ces trois grandeurs et de tracer les graphes de ces associations, on obtient les trois plans suivants :

- dynamique (temps, amplitude) ;
- spectre (fréquence, amplitude) ;
- mélodique (temps, fréquence).

L'objectif principal du spectrogramme est de connaître l'évolution temporelle du spectre de parole. Pour assurer cet objectif, il faut décomposer l'onde acoustique du son en ondes sinusoïdales de différentes fréquences au moyen d'une Transformée de Fourier.

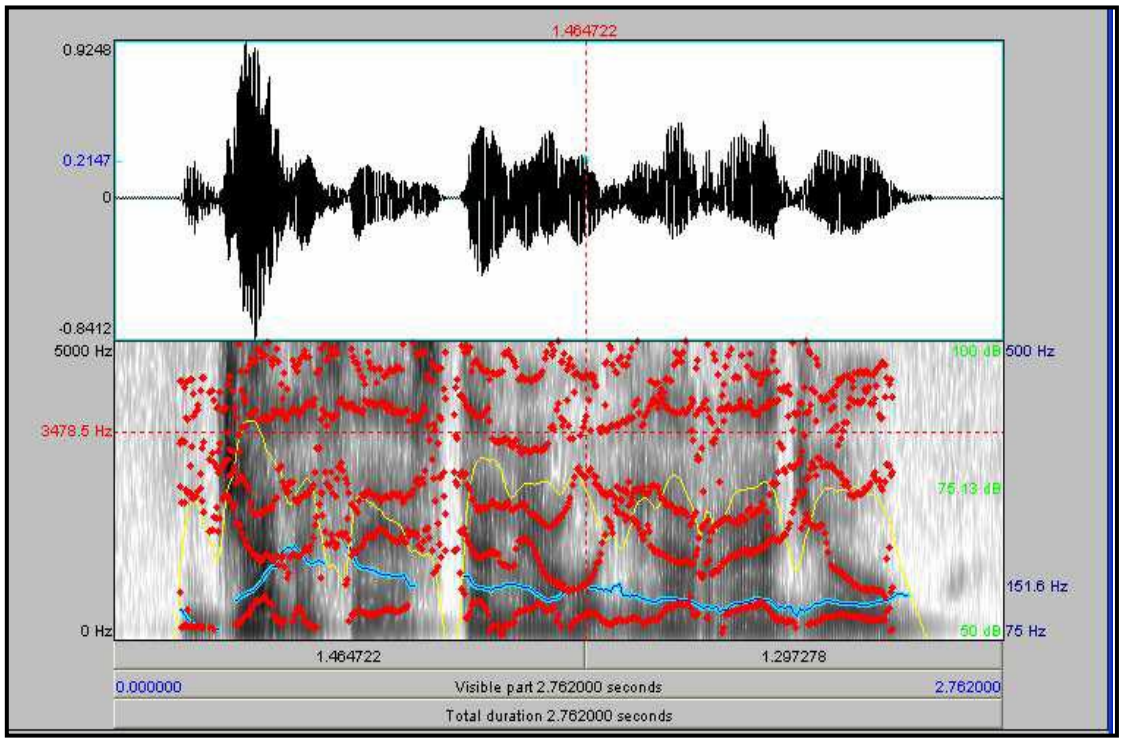


Figure 1.7 : Spectrogramme de la phrase / جلس يستمع إلى الراديو / [galasa yastami£u ilaa arraadyuu] [4].

1.6.2. Analyse par Codage Prédictif Linéaire

De la même façon qu'un signal de parole réel créé par les poumons et les cordes vocales, et produit par le passage à travers le filtre que constitue notre conduit vocal. Une

parole synthétique peut être modélisée par le passage d'un signal d'excitation à travers un filtre numérique récursif.

Cette modélisation est appelée prédictive linéaire puisqu'elle correspond à une régression linéaire entre le signal d'excitation et le signal vocal produit [4] (figure 1.8).

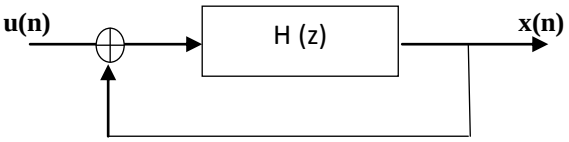


Figure 1.8 : Système d'entrée/sortie

Dans le domaine temporel nous avons :

$$x(n) + \sum_{i=1}^{p} a(i)x(n-i) = u(n) + \sum_{i=1}^{q} b(i)u(n-i) \qquad (1.1)$$

Où le couple {a(i)},{b(i)} représentent les coefficients du filtre H.

1.7. Segmentation du signal vocal

Segmenter le signal de parole, c'est effectuer une partition de ce signal en régions, telle que chacune d'entre elles possède au moins une caractéristique que n'ont pas les autres régions voisines. Les sons de la parole peuvent être classés, de manière un peu sommaire, en trois catégories : Les sons Voisés, les sons Non Voisés et les silences.

Les sons Voisés sont des signaux quasi périodiques très riches en harmoniques d'une pèriode fondamentale, appelée pitch. Ce qui leur donne un caractère assez facilement prévisible. Ils sont de forte énergie avec un faible Taux de Passage par Zéros (TPZ). Les sons Non Voisés sont des signaux qui ne présentent pas de structure périodique. Ils ont les caractéristiques spectrales d'un bruit légèrement corrélé. Ils présentent un TPZ notamment plus élevé que celui des signaux Voisés.

Les silences sont tout simplement des intervalles où le signal utile est absent. En pratique, il s'agit de bruits, d'origines diverses, d'énergie négligeable devant celle du signal utile.

A ces trois catégories s'ajoutent des segments voisés très pauvres en harmoniques (nasales Voisées), des sons plosives caractérisés par un apport instantané d'énergie, faisant passer de manière très brève du silence à un son qui peut être Voisé ou Non Voisé et sans oublier, des sons fricatifs qui sont créés par une constriction du conduit vocal au niveau du lieu d'articulation. C'est par la succession temporelle de tous ces sons que le signal de parole est constitué [5].

En ce qui concerne les modes de segmentation, nous distinguons deux catégories :

- la segmentation manuelle est assurée par des experts phonéticiens de la langue, son inconvénient majore est dû grâce à la difficulté de bien préciser les frontières des unités segmentales. De plus, une telle tâche nécessite un temps énorme et important durant l'annotation de grands corpus de parole. Il faut savoir qu'une segmentation manuelle effectuée par plusieurs experts ne fournit pas nécessairement les mêmes résultats ;
- la segmentation semi automatique / automatique, Actuellement, la segmentation complètement automatique de parole est une tâche rarement possible. En effet, étant donnée la complexité des phénomènes acoustico-phonétiques à traiter, cette tâche nécessite très souvent une intervention manuelle, que ce soit pour la préparation des données (étiquetage phonétique) du traitement automatique ou autre. Malgré l'existence des outils qui assurent cette opération, ils restent toujours non fiables puisqu'ils ne garantissent pas une très bonne qualité de parole synthétique. Pour cette raison, des vérifications manuelles faites par des experts humains sont indispensables à la segmentation de la parole [4].

1.8. Notions fondamentale sur l'Arabe standard

L'Arabe est une langue parlée par plus de 337 millions de personnes. Elle est la langue officielle d'au moins 22 pays. C'est aussi la langue de référence pour plus de 1,3 milliard de musulmans. Comme son nom l'indique, la langue Arabe est la langue parlée à l'origine par le peuple arabe. Dans le cadre de notre travail, nous parlerons de la langue Arabe en référence à ce qui est communément appelé **"l'Arabe Standard"** (AS), c'est-à-dire, la langue de communication commune à l'ensemble du Monde Arabe. Il s'agit de la langue enseignée dans les écoles, donc écrite, mais aussi parlée dans le cadre officiel.

1.8.1. Système phonétique de l'Arabe Standard

L'Arabe Standard (AS) compte 34 phonèmes: 6 voyelles et 28 consonnes. Les phonèmes arabes se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques. La graphie des lettres est différente selon leur position dans le mot. Ainsi, la lettre **ب** [b] est transcrite **بَيْتٌ** [bajtun] (une maison) en début de mot, **خُبْزٌ** [xubzun] (du pain) en milieu de mot, **كَلْبٌ** [kalbun] en fin de mot et **قُرْبَ** [qurba] (à proximité de) isolé en fin de mot.

Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscules/majuscules n'existe pas [7].

Pour les besoins de la transcription les 28 consonnes arabes ont été divisées en deux groupes:

- 14 consonnes solaires qui assimilent le « ل » de l'article ;
- 14 consonnes lunaires qui n'assimilent pas le « ل » de l'article.

Les solaires se prononcent en double, comme par exemple avec le mot « soleil » شمس [chams], au lieu de prononcer الشمس, el-chams, on prononce ech-chams, car la lettre ش [chin], est une lettre solaire.

Les lettres lunaires, se prononcent normalement et simplement pour elles-mêmes, c'est-à-dire sans les doubler. Par exemple avec le mot « lune », قمر ([qamar] - lune), on prononce القمر, [el-qamar] tout à fait normalement, parce que la lettre ق [qaf] est une lettre lunaire (Tableau 1.1).

1.8.1.1. Les voyelles

On distingue trois voyelles courtes opposées à trois voyelles longues, la durée d'une voyelle longue est environ double de celle d'une voyelle courte. Ces voyelles sont caractérisées par la vibration des cordes vocales et sont réparties comme suit :

- les voyelles courtes : [a], [u], [i] sont représentées dans un texte voyellé au-dessus ou au-dessous de la consonne, (ِ , ُ , َ) , exemple : تُرِكَ [turika] ;
- les voyelles longues [huruuf al madd] : [aa], [uu], [ii] sont écrites sous forme de caractères consonantiques (ي , و , ا) et sont obligatoirement représentées dans un texte écrite exemple : مُسَافِرُونَ [musaafiruuna].

Tableau 1.1: Transcription Orthographique Phonétique de l'AS [8].

Modes	**Type de consonnes**		**Consonnes arabes**	**Transcription des arabisants**	**Lieu d'articulation**
Occlusives	Voisées		ب د	[b] [d]	bilabiale alvéodentale
	Non-Voiseés		ق ت ك ء	[q] [t] [k] [ʾ]	uvulaire alvéodentale postpalatale glotal
	Voisée	Emphatiques	ض	[$\underline{d}$]	Alvéolaire
	Non-Voiseé		ط	[$\underline{t}$]	Alvéodentale
Fricatives	Voisées		ز ذ غ ع	[z] [ð] [ġ] [ɛ]	sifflante dorsoalvéolaire interdendale uvulaire pharyngale
	Non-Voisées		س ث ف ش خ ه ح	[s] [$\underline{t}$] [f] [$\tilde{s}$] [$\underline{h}$] [h] [ḥ]	sifflante dorsoalvéolaire interdendale labiodentale chuinchante palatale vélaire glottale Pharyngale
	Voisées	Emphatiques	ص	[ṣ]	doralveolaire sifflante
	Non-Voisées		ظ	[ḍ]	Interdentale
Nasales	Voisées		م ن	[m] [n]	bilabiale Alvéodentale
Liquide	Voisée		ل	[l]	Dentale
Affriquée	Voisée		ج	[$\tilde{g}$]	Alvéodentale
Vibrante	Voisée		ر	[r]	apicvoalvéolaire
Semi-voyelles	Voisées		و ي	[w] [y]	bilabiale Palatale

1.8.1.2. Les consonnes

Les consonnes de l'Arabe peuvent être classées suivant plusieurs critères:

- les consonnes articulées avec une vibration des cordes vocales sont dites sonores (voisées), sinon elles sont dites sourdes (non voisées) ;
- le franchissement de l'air à travers le conduit vocal:
  - ✓ les fricatives qui sont caractérisées par un frottement sur les parois du conduit vocal. comme س [s] et ز [z] ;
  - ✓ les occlusives qui sont caractérisées par un passage de l'air momentanément arrêté en un point quelconque de l'articulation, l'échappement de l'air s'effectue avec une petite explosion. On rencontre des dentales, des labiales et des glottales ;
  - ✓ une liquide caractérisée par un passage de l'air sur les deux côtés de la langue : (latérale) ل [l] ;
  - ✓ deux nasales caractérisées par un échappement de l'air en même temps par la bouche et par le nez : م [m], ن [n] ;
  - ✓ une vibrante caractérisée par le deplacement de la langue au passage de l'air: ر [r] ;
  - ✓ deux semi-consonnes (ou semi-voyelles) caractérisées par un passage rapide de l'air à travers la bouche accompagné de frottements consonantiques : ي [j], و [w].
- le mode d'articulation: suivant le mode d'articulation, on distingue les consonnes géminées et emphatiques. Toute consonne géminée est formée par l'assemblage de deux consonnes identiques fortement articulées. La gémination est indiquée par un signe graphique spécifique appelé chadda (ّ). Les consonnes emphatiques ط [t], ض [ð], ص[ʂ], ظ [d̲] sont caractérisées par une forte tension des différents organes du conduit vocal [9].

1.8.1.3. Le tanwin

Le signe du tanwin est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination ال placé en début de mot. Les symboles du tanwin sont au nombre de trois et sont constitués par le dédoublement des signes diacritiques ci-dessus, ce qui se traduit par l'ajout du phonème [n] au niveau phonétique :

[an] : ً [un] : ٌ [in] : ٍ

1.8.1.4. La chadda

Le signe de la chadda peut être placé au-dessus de toutes les consonnes en position non initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiques : Signe (كَلَّمَ [kallama] ″il a parlé à″).

1.8.2. Particularités de l'Arabe Standard

Le système phonétique de la langue arabe diffère de celui des autres langues par la présence : de voyelles longues (huruuf al madd), phonèmes arrières, de phénomènes d'emphase et de la gémination. Ces caractéristiques donnent une valeur particulière à cette langue.

1.8.2.1. Voyelles longues

En arabe standard les voyelles longues présentent une caractéristique très importante au niveau sémantique. Par exemple, les deux mots g̃amal (chameau) et g̃amāl (beauté) ne différent que par l'allongement de la voyelle finale.

Sur le plan articulatoire, il existe une similitude entre les voyelles [i] et [i̅], [u] et [u̅] cependant une différence existe entre les voyelles [a] et [a̅] car la position de la langue est plus basse pour le [a] que pour le [a̅].

Sur le plan acoustique, les niveaux des formants entre chaque voyelle brève et son opposée longue sont assez rapprochés. L'allongement temporel effectué par les voyelles longues n'influe pas sensiblement sur les niveaux formantiques de ces derniers. La partie stable de la voyelle longue est beaucoup plus allongée par rapport à son opposée brèves [7].

1.8.2.2. La gémination

Au niveau graphique, elle est symbolisée par le signe de la chadda qui signifie le dédoublement de la consonne. Sur le plan phonétique, l'opposition simple/géminée peut se résumer de la manière suivante : pour une consonne non-occlusive, l'opposition se réduit essentiellement à l'opposition temporelle brève/longue ; pour une occlusive, elle réside au niveau de la durée du silence. Ce rallongement entraîne l'accentuation des propriétés de la consonne (augmentation du caractère emphatique). Une consonne géminée est un son unique pour lequel les organes de phonation ne changent pas de position (les lèvres ne se referment pas après le premier [b] dans [ka**bb**ara]. Dans beaucoup de langues, ce phénomène permet de mettre en relief un mot dans son contexte, alors qu'il s'avère être un élément distinctif sur les plans morpho-sémantiques en langue Arabe حضَرَ : [hadara] ″ il a assisté″ est différente de حضَّرَ [haddara] ″il a préparé″ où la deuxième consonne est géminée [8].

1.8.2.3. Phonèmes arrières

Le système phonétique de l'AS possède quatre phonèmes arrières spécifiques, et n'ont leurs équivalents exacts dans aucune autre langue européenne :

- les spirantes pharyngales [h] ,[ε] qui ont comme point d'articulation la partie médiane du pharynx ;
- l'occlusive uvulaire [q] qui a pour point d'articulation la partie la plus reculée de la langue et la région du palais supérieure ;
- l'occlusive glottale [ء], les Grammairiens Arabes indiquent pour ce phonème la partie la plus reculée du pharynx [9].

1.8.2.4. Emphase

Basculant entre plusieurs vocables tantôt relevant du domaine perceptif tantôt domaine fonctionnel, la définition de l'emphase et la description des consonnes emphatiques lors du processus articulatoire a suscité beaucoup de controverses. Parmi les diverses définitions existant nous citerons quelques-unes :

- basé sur de nombreuses mesures spectrographiques, S. Al Ani [14] définit ce phénomène comme étant produit dans la région vélaire et non la pharyngale. Cependant R. Jackbson proclame que les emphatiques sont réalisées par la pharyngalisation qui se produit lors de la contraction de la partie supérieure du pharynx ;
- Les phonèmes emphatiques sont caractérisés par une tonalité plus pleine et grave car ils exigent la dépense d'un volume d'air important et une tension organique supérieure par rapport aux autres consonnes. L'intérêt porté par ce phénomène remonte jusqu'aux Grammairiens Arabes du moyen-âge. Attirés par le système phonétique de leur langue, ils ont pu déterminer par de simples constatations ciblées, le fonctionnement et les positions des principaux organes entrant dans la production d'un son emphatique que se soit sur le plan auditif ou physiologique [8].

1.9. Conclusion

Dans ce chapitre nous avons exposé des notions de base sur le traitement de la parole, des spécifications du signal vocal et quelques caractéristiques de la langue Arabe Standard.

Les objectifs de ce chapitre sont de définir les notions que nous utiliserons dans notre travail. Cette partie théorique sera complétée dans le chapitre suivant par une étude approfondie des systèmes de synthèse de la parole et ses variantes.

Chapitre 2 :

Synthèse de la Parole

2.1. Introduction

La qualité d'un système de synthèse vocale dépend du naturel, de l'intelligibilité de la parole générée et des caractéristiques propres à la voix produite. Ces caractéristiques dépendent des techniques et des méthodes de synthèse, mais également du soin apporté à la modélisation linguistique et prosodique. Plusieurs travaux soulignent le fait que des structures linguistiques entretiennent des liens étroits avec les réalisations prosodiques. Dans ce chapitre, nous allons introduire le cadre technique de notre étude : la synthèse de la parole. Le chapitre s'articule autour des principes de la synthèse vocale, suivi d'une description bien détaillée des différentes techniques et méthodes utilisées.

2.2. Historique de la synthèse de la parole

À plusieurs reprises au cours de l'histoire, on a tenté de reproduire la voix humaine. Au XVIII^{ème} siècle, on met au point à cet effet des dispositifs mécaniques équipés de soufflets et d'anches vibrantes. Au XX^{ème} siècle, l'apparition de l'électricité et de l'électronique autorisent des tentatives plus ambitieuses : en 1922, J.C. Stewart fabrique une machine capable de reproduire des voyelles, des diphtongues et quelques mots simples ; plusieurs années plus tard en 1939, H. Dudley présente, à l'occasion de l'exposition universelle de New York, le VODer (Voice Operation Demonstrator), appareil mis au point par les laboratoires Bell.

Mais ce n'est que dans les années cinquante que les premiers véritables synthétiseurs de la parole font leur apparition, avec, par exemple, le Pattern Playback, système mis au point par les laboratoires Haskins aux USA, qui se présente comme un lecteur de sonagraphe (un faisceau de lumière produit, après amplification, des sons à partir de la représentation de leur durée, de leur fréquence et de leur intensité).

Depuis les années soixante-dix, des progrès considérables ont été accomplis, avec notamment le développement de l'utilisation des calculateurs numériques. Aujourd'hui encore, ces progrès se poursuivent, dans plusieurs directions (perfectionnement des synthétiseurs à formants, des synthétiseurs à prédiction linéaire, etc.) [10].

2.3. Principe de synthèse de la parole

Qu'est-ce que la synthèse de la parole ?

Une simple réponse à cette question pourrait être : « la production de la parole par une machine ». Mais chacun sait qu'un magnétophone peut produire de la parole sans que l'on n'ait jamais songé à l'appeler « synthétiseur ».

Une meilleure définition serait alors : « la production par une machine de sons ou de mots qui n'ont jamais été prononcés auparavant par un être humain ». Mais cette définition est trop restrictive car elle ne tient pas compte des techniques de synthèse par assemblage d'éléments préenregistrés.

Si l'on peut simplement définir cette technique en fonction de la sortie, considérons alors le type d'entrée qui va engendrer une parole de synthèse. Deux cas peuvent se présenter : ou bien l'entrée est une succession de concepts, ou bien c'est une chaîne de caractères orthographiques. Dans un cas comme dans l'autre, l'émission de la parole sera déterminée par une représentation phonétique de ce qui doit être dit. Nous adoptons donc la définition suivante :

« La synthèse de la parole permet de produire des sons de la parole à partir d'une représentation phonétique du message » [11].

Le message vocal est un continuum acoustique dans lequel il n'y pas de frontière marquée entre les mots ni entre les sons élémentaires (ou phonèmes) du langage. En synthèse, la reproduction de ce message résulte de l'encodage d'information au niveau :

- segmental par le choix des unités phonétiques et de leurs enchaînements ;
- suprasegmental par la génération automatique de la prosodie donnant à ces unités une importance de nature linguistique et expressive.

A cette étape, il est important de bien distinguer la différence qui existe entre « synthèse de la parole » (on l'appelle parfois synthèse de la parole à partir du texte) et un « synthétiseur de parole », ainsi nous nommons :

- un système de synthèse de la parole comme étant capable de reproduire des sons « parlés » à partir d'un texte ou d'une entrée conceptuelle (Figure 2.1) ;
- un synthétiseur de parole comme étant la dernière étape de la transformation d'un certain nombre de paramètres de contrôle en parole.

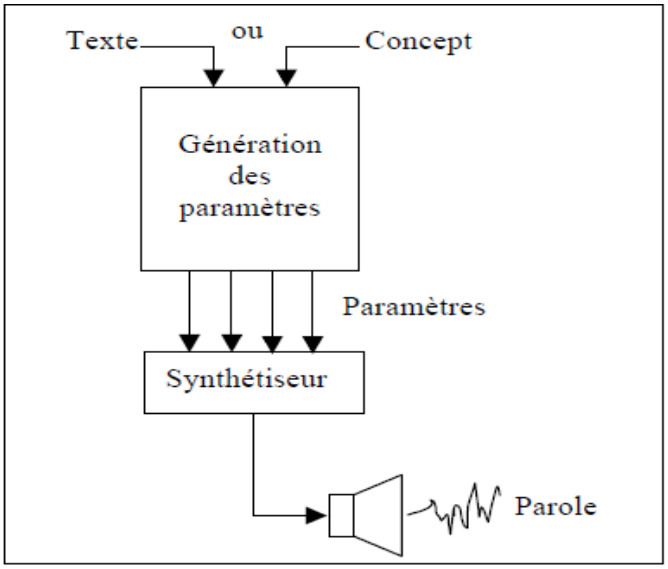


Figure 2.1 : Système de synthèse de la parole [11].

Les synthétiseurs ont quant à eux la fonction inverse de celle des analyseurs et des reconnaisseurs de parole : ils produisent de la parole artificielle. On distingue fondamentalement deux types de synthétiseurs :

- les synthétiseurs de parole à partir d'une représentation numérique, inverses des analyseurs, dont la mission est de produire de la parole à partir des caractéristiques numériques d'un signal vocal telles qu'obtenues par analyse.
- les synthétiseurs de parole à partir d'une représentation symbolique¸ inverse des reconnaisseurs de parole et capables en principe de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, on classe également les synthétiseurs en fonction de leur mode opératoire :
    - ✓ les synthétiseurs à partir du texte reçoivent en entrée un texte orthographique et doivent en donner lecture ;
    - ✓ les synthétiseurs à partir de concepts, appelés à être insérés dans des systèmes de dialogue homme-machine, reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue [7].

2.4. Architecture d'un système de synthèse de la parole

Tout système TTS (Text To Speech) est généralement constitué de deux blocs de traitements principaux: un bloc de traitements **linguistiques** et un bloc de traitements

acoustiques. Le premier bloc vise à analyser et à structurer le texte afin de déterminer un mode de prononciation cohérent, puis à transformer le texte analysé en une séquence de descripteurs symboliques décrivant les unités cible. Le deuxième bloc consiste à générer un signal acoustique adapté à cette séquence symbolique.

La Figure 2.2 présente l'architecture générale d'un système de synthèse de la parole à partir du texte. Les deux premières parties qui concernent les traitements de *haut niveau* permettent le passage de la représentation orthographique du texte en entrée à une représentation phonétique munie d'une description prosodique. La dernière partie englobe les traitements de bas niveau du synthétiseur qui permettent la génération proprement dite du signal acoustique [7].

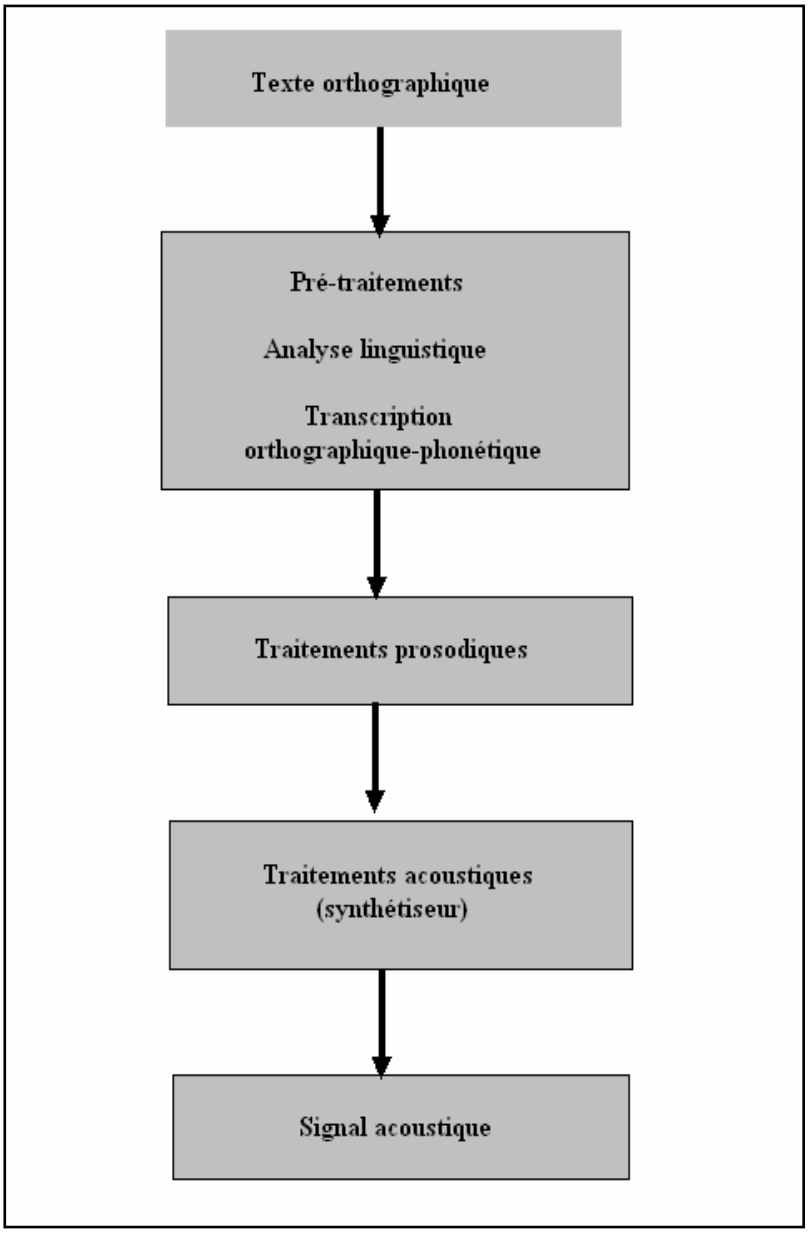


Figure 2.2 : Architecture générale d'un système de synthèse de la parole à partir du texte [7].

2.5. Les diverses techniques de synthèse de la parole

Il est possible de rassembler les différentes techniques offertes par les synthétiseurs de parole. En première approximation nous pouvons les regrouper comme suite :

2.5.1. Dans le domaine spectral

Cette technique englobe l'ensemble des vocodeurs (Voice Coder) à canaux, à formants, et à prédiction linéaire.

2.5.1.1. Vocodeur à canaux

Le vocodeur à canaux est un appareil destiné à transmettre la parole à un faible débit d'information. Il se compose de deux parties l'analyseur et le synthétiseur. La fonction d'analyse de l'enveloppe spectrale est effectuée à l'aide de canaux dont le nombre peut varier, suivant les réalisations, entre 10 et 20. Chaque canal traite une bande de fréquence déterminée. Le signal de parole issue d'un microphone est analysé au moyen d'un banc de filtres passe bande contigus, couvrant l'étendue de la bande téléphonique 300 à 3400 Hz.

Le signal délivrer par chacun des filtres subit une détection puis traverse des filtres passe bas, dont les fréquences de coupures sont de l'ordre de 20 à 50 Hz, parce que les variations énergétiques dans les canaux sont lentes (à l'image de variation lente de l'articulation).
L'analyse du vocodeur comporte également un détecteur de voisement. Ce dernier permet de différencier les sons Voisés et de donner la valeur de la F_0.
La synthèse est effectuée à l'aide d'un banc de filtres passe-bande analogue à celui de l'étage d'analyse. Pour chaque canal, le signal basse fréquence issue du filtre d'analyse est multiplié par le signal d'excitation dans un modulateur. Selon que la parole à reproduire est détectée comme Voisée ou Non Voisée, le signal d'excitation attaque les modulateurs provient soit d'un générateur périodique soit d'un générateur de bruit. Le signal de sortie est obtenu par addition des sorties des filtres de synthèse. L'intelligibilité est assez bonne, bien que l'agrément et le naturel de la voix soient dégradés par le traitement. La partie délicate est constituée par le détecteur de pitch [7].

2.5.1.2. Vocodeur à formants

Les formants sont les fréquences propres du conduit vocal lors de la production d'un son voisé. Dans cette technique, le filtre du conduit vocal est composé d'un certain nombre de

résonateurs similaires au nombre de formants de la parole naturelle. Les résonances formantiques naturelles du conduit vocal sont simulées par des filtres résonants du deuxième ordre caractérisés par une fréquence centrale et une largeur de bande spécifiques. Une synthèse de qualité est obtenue par la simulation des quatre premiers formants. L'implantation de ces filtres peut se faire soit de façon cascade, soit de façon parallèle, soit de façon mixte (Figure 2.3). Pour les sons voisés, ce système est excité par une onde périodique dont la forme est aussi proche que possible de l'onde glottale. Pour les sons non voisés, l'excitation est un bruit blanc [7].

Où :

- F_0 et A_0 sont respectivement la fréquence fondamentale et l'amplitude de la composante voisée ;
- F_n et Q_n sont respectivement les fréquences de formants et leur bande passante;
- V_L et V_H sont respectivement l'amplitude basse et haute de la composante voisée ;
- F_L et F_H sont respectivement l'amplitude basse et haute de la composante non voisée ;
- QN est la valeur de la bande passante du formant nasal à 250 Hz.

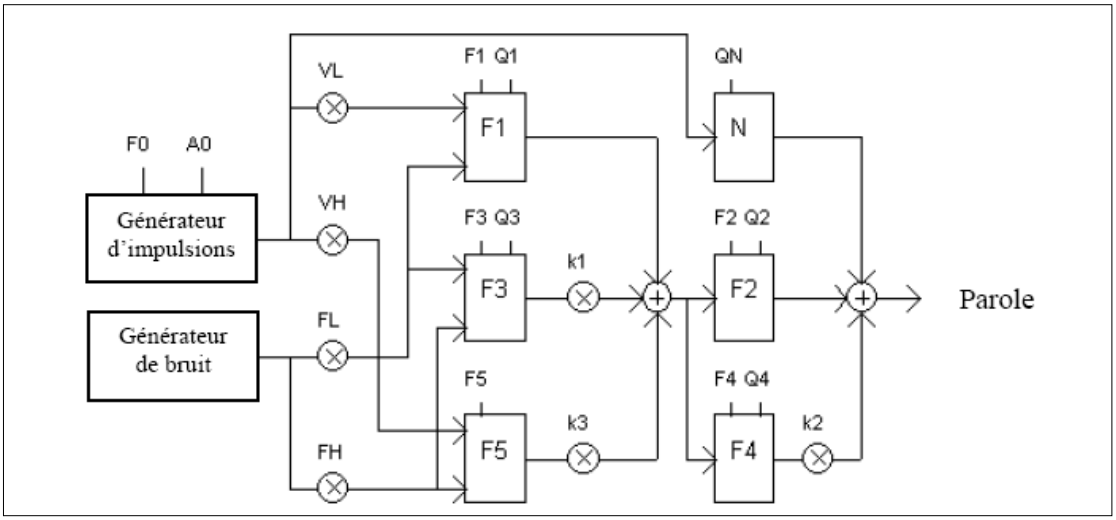


Figure 2.3 : Synthétiseur à formants qui combine les deux structures (Cascade et parallèle) [11].

2.5.1.3. Vocodeur à prédiction linéaire

La synthèse par prédiction linéaire est une méthode qui s'inscrit dans le cadre de la théorie source/filtre. Elle a largement été utilisée dans les systèmes par concaténation, car elle permet un codage rapide des unités à concaténer contrairement à la synthèse par formants. La parole de synthèse produite en utilisant la synthèse LP est loin d'être parfaite. Klatt [19] a

montré que la synthèse LP fondée sur la méthode d'autocorrélation ne reproduit pas correctement les fréquences et les bandes passantes des formants lors de la synthèse avec une fréquence fondamentale différente de la fréquence initiale. La synthèse du signal de parole avec sa fréquence fondamentale d'origine conduit aussi à une dégradation du signal due à l'excitation utilisée. En effet, cette dernière est de nature très simplifiée par rapport au signal d'erreur réel. Plus particulièrement, dans le cas de sons voisés, d'autres informations ne sont pas prises en compte, conduisant à la dégradation du signal.

Pour pallier ce problème, une technique dite de **prédiction linéaire par impulsions multiples** est mise en œuvre. Elle consiste à construire une excitation composée de plusieurs impulsions pour chaque trame de parole analysée. La synthèse avec la combinaison de cette excitation et les coefficients LP produit un signal de parole très proche du signal naturel. Cette technique est très intéressante pour les vocodeurs et le codage de la parole à bas débit.

2.5.2. Dans le domaine articulatoire

La synthèse articulatoire est potentiellement considérée comme la technique la plus performante car elle reflète théoriquement le processus physiologique. Cette technique est basée sur une modélisation géométrique du conduit vocal. Elle consiste à représenter le conduit vocal comme un tube de section variable, avec des embranchements et des sections parallèles, puis à y simuler le trajet des ondes produites au niveau de la glotte. Les modèles d'écoulement d'air (mécanique des fluides), de sources et de propagation acoustique (phénomènes physiques), en association avec des modèles articulatoires (mécaniques), permettent de constituer un synthétiseur articulatoire complet, contrôlé par deux jeux de paramètres : les paramètres supra–laryngés qui commandent le modèle articulatoire, et un jeu de paramètres qui pilotent les cordes vocales (pression sub-glottique, longueur des cordes vocales et hauteur de la glotte au repos) [7].

2.5.3. Dans le domaine temporel (Synthétiseur par formes d'ondes)

Cette technique englobe l'ensemble des synthétiseurs telle que la compression de la parole numérisée M.I.C (Modulation par Impulsions Codées) et les autres basé sur la synthèse par concaténation.

2.6. Les méthodes de la synthèse de parole

Il y a deux approches principales pour convertir un texte en parole la synthèse par concaténation et la synthèse par règles.

2.6.1. Synthèse par règles

La synthèse par règles est une méthode qui a eu beaucoup de succès dans le contexte de la synthèse de la parole à partir du texte. Des règles sont utilisées pour estimer les paramètres nécessaires. Cette approche est fondée sur un modèle de production du signal vocal, modèle commandé par un nombre restreint de paramètres. La synthèse se décompose alors en deux étapes : une transformation des informations phonético prosodiques, à l'aide de règles contextuelles, en commandes permettant de spécifier l'évolution temporelle des paramètres du modèle de synthèse; les paramètres ainsi déterminés sont utilisés pour synthétiser le signal acoustique.

Dans ce type de synthèse, les caractéristiques supra-glottiques sont modélisées à l'aide d'un filtre linéaire dont la fonction de transfert varie au cours du temps. Les paramètres utilisés pour le contrôle du filtre sont les paramètres formantiques, à savoir la fréquence centrale, la bande passante et l'amplitude des maxima significatifs de la fonction de transfert du conduit vocal. Pour obtenir une parole intelligible, il suffit de spécifier les paramètres des 3 à 4 formants les plus importants, d'où la dénomination de synthèse par formants couramment employée pour ce type de synthèse.

Une telle approche ne permet pas de restituer un signal de parole apparaissant naturel. La qualité médiocre obtenue résulte d'une part de la difficulté à modéliser suffisamment finement les trajectoires acoustiques et d'autre part de la modélisation trop grossière du signal glottique [7].

Parmi les grands avantages de cette méthode, nous pouvons citer notamment la grande souplesse d'utilisation, la facilité d'extension, et surtout la grande portabilité de ces systèmes facilitant leur intégration dans une large gamme de produits.

Les synthétiseurs par règles sont organisés comme à la (figure 2.4).

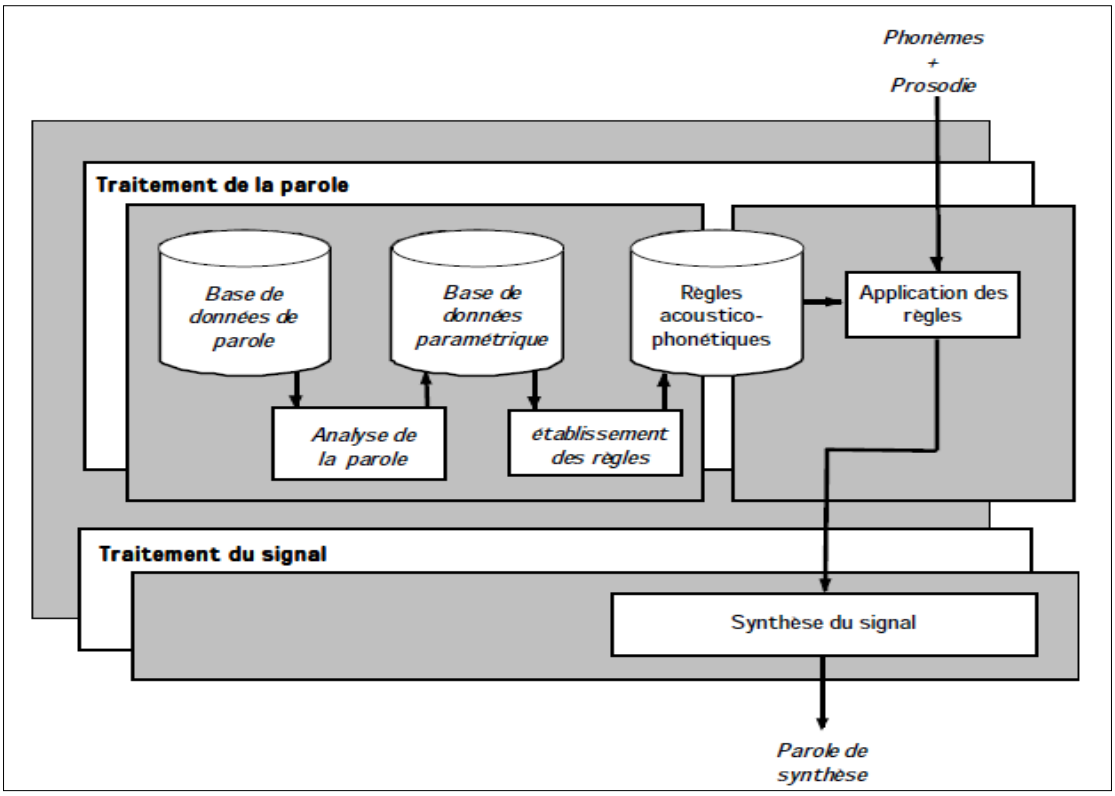


Figure 2.4 : Schéma de conception et fonctionnement typique d'un système de synthèse par règles [2].

2.6.2. Synthèse par concaténation d'unités pré-stockées

La synthèse par concaténation d'unités pré-stockées est la génération des sons à partir de la juxtaposition d'un ensemble d'unités préenregistrées, ces dernières sont obtenues par une opération d'analyse du signal qu'on veut produire. Elle consiste à choisir dans une large base de données les unités sonores les plus appropriées pour construire, par concaténation la phrase à produire (Figure 2.5).

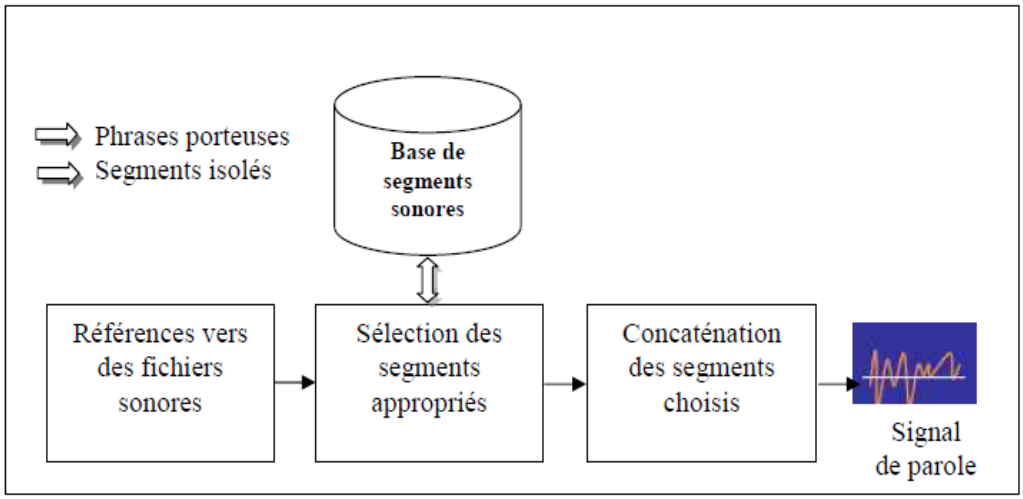


Figure 2.5 : Principe de base de la méthode de synthèse par concaténation [4].

En réalité dans cette approche on peut trouver plusieurs types d'unités (phonèmes, diphones, syllabes, polysons, mots, etc.). Parmi les méthodes de synthèse par concaténation on a :

2.6.2.1. La concaténation de phrases

Est une simple opération d'enregistrement et de restitution des phrases à synthétiser en vue d'une réalisation bien précise. Cette précision limite leur utilisation à un petit vocabulaire, ainsi qu'à des applications très restreintes telles que les jouets pour enfants, les répondeurs téléphoniques, l'horloge parlante, etc. Cette dernière se compose de deux parties, une stable qui correspond à la phrase «il est : … heures ….minutes….secondes », et une autre variable où on trouve les nombres qui correspondent à la valeur actuelle de l'heure, des minutes, et des secondes ; **Base de segments sonores** Références vers des fichiers sonores Sélection des segments appropriés Concaténation des segments choisis Signal de parole.

2.6.2.2. La concaténation de mots

Il s'agit de juxtaposer un ensemble de mots l'un à côté de l'autre pour générer une phrase avec une qualité moins bonne par rapport aux phrases qui sont obtenues à travers l'utilisation de type précédant de concaténation.

2.6.2.3. La concaténation de phonèmes

Puisque les phonèmes représentent les éléments atomiques dans n'importe quelle langue, il suffit de les juxtaposer pour synthétiser un mot ou une phrase. Malgré la simplicité de cette méthode, elle présente l'inconvénient de discontinuité du signal généré et cela à cause du problème de la coarticulation qui est dû grâce à l'influence d'un son sur ses voisins. Pour résoudre ce problème la solution est de changer le phonème par une autre unité plus coûteuse en information qui est le diphone.

2.6.2.4. La concaténation par diphones

Consiste à enregistrer dans la base de données sonore les diphones nécessaires pour produire la parole. Chaque diphone représente le segment qui est compris entre deux parties stables de deux phonèmes consécutifs en prenant toute la transition [4].

2.7. Quelques critères d'évaluation des systèmes TTS

Les systèmes d'évaluation des synthétiseurs TTS représentent un domaine de recherche très vivant qui évolue côte à côte avec la synthèse vocale. En effet, si l'on savait

évaluer précisément et diagnostiquer les défauts de qualité des synthétiseurs, on saurait aussi comment y remédier, ou au moins comment chercher les solutions. Ces systèmes se basent sur les critères suivants :

- la qualité de la parole générée, parmi les systèmes de synthèse qui utilise ce critère on peut citer TD-PSOLA (brevet CNET en 1988 déposé par France Télécom) qui est souvent applicable à des systèmes de synthèse par concaténations ;
- l'enregistrement ou synthèse à partir du texte : Malgré ces progrès manifestes, le naturel de la parole de synthèse reste encore aujourd'hui nettement inférieur à celui de la parole des êtres humains. Cet écart de qualité n'est accepté par les usagers que pour des services nouveaux, qui ne pourraient pas leur être fournis d'une autre manière (lecture de FAX ou des E-mails, par exemple). Pour réduire cet écart de qualité, certains prototypes récents d'application combinent la souplesse de la synthèse à partir du texte pour produire les parties variables des messages avec l'utilisation de patrons prosodiques naturels, spécifiques aux messages de l'application [13];
- l'intelligibilité est un facteur crucial qui permet de vérifier si la phrase générée a été bien perçue, par rapport à son niveau linguistique (phrase affirmative, négative, interrogative, etc.) ;
- la fiabilité, de nos jours les systèmes de synthèse vocale sont utilisés dans des services grand public. Il est clair qu'ils doivent être robustes pour assurer une très grande durée de vie, et une meilleure publicité de synthétiseur vocal lui-même ;
- l'Interface Homme machine (l'interactivité), un système de synthèse de bonne qualité doit assurer une meilleure interaction entre l'utilisateur et le système (la machine de synthèse). Cette nouvelle notion intègre la reconnaissance de parole, la communication intelligente et la synthèse vocale sophistiquée qui utilise des fondements de langage parlé, connus a priori [4].

2.8. Les applications de synthèse de la parole

Les applications actuelles de synthèse de la parole à partir du texte peuvent être regroupées en cinq grands domaines :

- aides pour personnes handicapées
  - ✓ lecture d'écrans ou de documents écrits pour non-voyants ;
  - ✓ aides à la communication vocale pour personnes muets, laryngectomisés ou à infirmité motrice cérébrale ;

- ✓ journaux vocaux, etc.

- Outils d'Enseignement Assisté par Ordinateur (OEAO)
  - ✓ système de dictées automatiques ;
  - ✓ système d'apprentissage des langues.
- applications industrielles
  - ✓ serveurs d'alerte, de surveillance de sites et de supervision de réseaux ;
  - ✓ télémaintenance ;
  - ✓ fonctions d'aide dans les postes de pilotage ;
  - ✓ fonction de vérification vocale dans les postes d'édition (correction des épreuves) ou de saisie d'informations écrites (bases de données), etc.
- applications grand public non téléphoniques
  - ✓ domotique (alarmes, appareils domestiques parlants, etc.) ;
  - ✓ micro-informatique (jeux et CDROMs parlants, bureautique, etc.).
- télématique vocale
  - ✓ serveurs vocaux d'informations (la synthèse remplaçant la parole naturelle enregistrée pour des informations rapidement évolutives et disponibles sous forme textuelle) ;
  - ✓ serveurs de lecture vocale de FAX ou de messages électroniques (e-mails);
  - ✓ Automatisation de services de renseignements (Annuaires, standards d'entreprises, etc.) [7].

2.9. Conclusion

Nous avons exposé dans ce chapitre les principales méthodes et techniques utilisées dans la synthèse de la parole. La première génération des systèmes de synthèse de la parole avait pour objectif de minimiser le volume de la base de données pour réduire le coût de stockage et de rendre le système de synthèse flexible et facile à adapter pour une autre voix ou une autre langue.

Cette flexibilité dépend de l'ensemble des règles qui doivent être élaborées soigneusement, ce qui induit à une complexité très élevée. Avec la génération actuelle, le problème de synthèse s'est réduit à un problème de base de données et d'optimisation de la sélection d'unités.

L'objectif est donc de réduire au maximum la modification du signal des unités de synthèse afin de préserver l'aspect naturel de la parole. Nous présentons donc dans la partie suivante la technique PSOLA qui, comme nous allons le voir, est également efficace pour certaines modifications de la prosodie.

Chapitre 3 :

Synthèse de la Parole par la Technique PSOLA

3.1. Introduction

Ce troisième chapitre représente une étude de la technique qui permet de faire la synthèse d'un signal de parole, Soit la technique PSOLA. Nous nous intéressons principalement au principe de fonctionnement de cette technique, à l'algorithme de synthèse de TD-PSOLA et les méthodes de détection de pitch.

3.2. La technique PSOLA

Les méthodes reposant sur le principe de synchronisation avec le fondamental sont utilisées pour réaliser des modifications temporelles ou fréquentielles d'un signal de parole, ou pour mettre en œuvre des systèmes de synthèse.

Ces méthodes nécessitent au préalable un marquage des périodes du fondamental. La méthode PSOLA (Pitch Synchronous OverLapp and Add), est une des variantes d'**OLA** qui se ramifie en plusieurs techniques (Time Domain PSOLA ou TD-PSOLA, Frequency Domain PSOLA ou FD-PSOLA, Linear Prediction PSOLA ou LP-PSOLA).

L'algorithme PSOLA consiste à concaténer, à l'aide d'un lissage, des unités de parole pré-stockées en modifiant le pitch et la durée des segments. Cette technique est associée à la méthode de synthèse par concaténation.

3.2.1. Etablissement de la formulation OLA

La représentation la plus adéquate du signal est celle du spectrogramme. Celui-ci donne la distribution de l'énergie en fonction du temps et de la fréquence. C'est une représentation bi-dimensionnelle et sa version discrète traitable sur ordinateur est une matrice dont le nombre de colonnes est le nombre de spectres calculés sur toute la durée du signal ; et le nombre de lignes est la moitié de la fréquence d'échantillonnage du signal (Fréquence de Nyquist). Nous donnons ici une autre démonstration de la formule OLA que celle donnée en [12]. La distribution de l'énergie dans le plan temps-fréquence est :

$$X(\omega,\tau) = \int_T x(t)h(t-\tau)\exp(-j\omega t)dt \tag{3.1}$$

Cette formule peut être inversée par la transformation de Fourier Inverse et nous obtenons :

$$x(t)h(t-\tau) = \int_\tau X(\omega,\tau)\exp(j\omega t)dw \tag{3.2}$$

En intégrant les deux membres de l'égalité par rapport à τ nous obtenons :

$$\int_{\tau} x(t)h(t-\tau)d\tau = \iint_{\tau} X(\omega,\tau)\exp(j\omega t)d\omega d\tau \tag{3.3}$$

Comme x(t) ne dépend pas de τ , nous avons finalement :

$$x(t) = \frac{1}{A}\iint_{\tau} X(\omega,\tau)\exp(j\omega t)d\omega d\tau \tag{3.4}$$

où $A=\int h(t-\tau)d\tau$ et une constante indépendante de ω et τ (elle représente l'aire délimitée par l'axe des temps et la fenêtre $h(t)$)

$$x(t) = \frac{1}{A}\int_{\tau} x(t)h(t-\tau)d\tau \tag{3.5}$$

Nous pouvons retrouver le signal x(t) à partir du spectrogramme $X(\omega,\tau)$. Si on désire effectuer des modifications sur le signal (à long terme), il suffit de modifier le spectrogramme de ce signal et utiliser ensuite la synthèse OLA. Notons aussi qu'en faisant dans (3,5) $x(t)\equiv 1$, nous obtenons :

$$\int_{\tau} h(t-\tau)d\tau = A = Cte = cons\tan te \tag{3.6}$$

Les formules 3.5 et 3.6 sont des formulations continues de OLA et sont par conséquent incommodes en pratique. Elles peuvent être remplacées par des formules équivalentes discrètes où la variable τ qui prend des valeurs discrètes avec un pas $P=L-R$ où L est la taille de la fenêtre et R est le taux de recouvrement.

$$x(t)=\sum_{\tau}\int X(\omega,\tau)\exp(j\omega t)d\omega \tag{3.7}$$

$$\boldsymbol{x(t) = \sum_{\tau} h(t-\tau)x(t)} \tag{3.8}$$

Ou encore :

$$\sum_{\tau} h(t-\tau)=Cte \tag{3.9}$$

Nous avons vérifié l'exactitude de cette équation 3.9 avec une fenêtre de Hamming de 256 points et un recouvrement de ¾ (75%) et nous avons obtenu un signal quasi-constant oscillant

entre 2.1520 et 2.1543, soit une précision de moins de 0.3%. Réécrivons l'équation 3.9 sous la forme :

$$\sum_{-\infty}^{\infty} h(t-nP)=Cte \tag{3.10}$$

Nous pouvons continuer à considérer la variable t continue mais si nous la prenons discrète, il faut alors remplacer l'intégrale de Fourier par sa version discrète (Transformation de Fourier discrète ou TFD) :

$$x(n)h(n-\tau)=\sum_{k=0}^{k=L-1} X(k,\tau)\exp(\frac{2\pi j}{L}nk) \tag{3.11}$$

Nous pouvons établir la formulation OLA en utilisant la théorie de Shannon comme suit : Au signal périodique $h(t-nP)$ avec $P=L$ correspond à une suite de raies sur le spectre continu $H(\omega)$. Si les périodes de ce signal se recouvrent (c'est à dire si P diminue), il se produit un espacement dans les raies du spectre ou en d'autres termes, celui-ci devient sous-échantillonné. Avec un taux de recouvrement suffisamment grand (et donc un pas P assez petit) le sous-échantillonnage du spectre $H(\omega)$ se réduit à une seule raie (un pic de Dirac) d'où l'équation 3.12.

$$TF\{\sum_{n} h(t-nP)\}=H(0)\delta \tag{3.12}$$

où TF désigne la Transformation de Fourier et δ la fonction (distribution) de Dirac, c'est à dire que :

$$\sum_{-\infty}^{\infty} h(t-nP)=H(0) = \text{constante.} \tag{3.13}$$

C'est à dire que :

$$\sum_{-\infty}^{\infty} x(m)h(t-nP)=H(0)x(m) \tag{3.14}$$

Ou encore : $$x(m)=\frac{\sum_{-\infty}^{\infty} x(m)h(m-nP)}{\sum_{-\infty}^{\infty} h(m-nP)} \tag{3.15}$$

Les formules (équation 3.14 et équation 3.15) sont illustrées par la figure ci-dessous (Figure 3.1) ou la somme infinie dans équation 3.15 est remplacée par une somme temporelle finie de fenêtres (Time Domain OLA ou TDOLA).

Ainsi, si nous réalisons une modification dans le spectre Court Terme $X(\omega,\tau)$, nous obtenons un signal modifié donné par :

$$y(t)=\sum\int Y(\omega,\tau)\exp(j\omega t)d\omega \tag{3.16}$$

où $Y(\omega,\tau)$ représente le spectre court terme modifié.

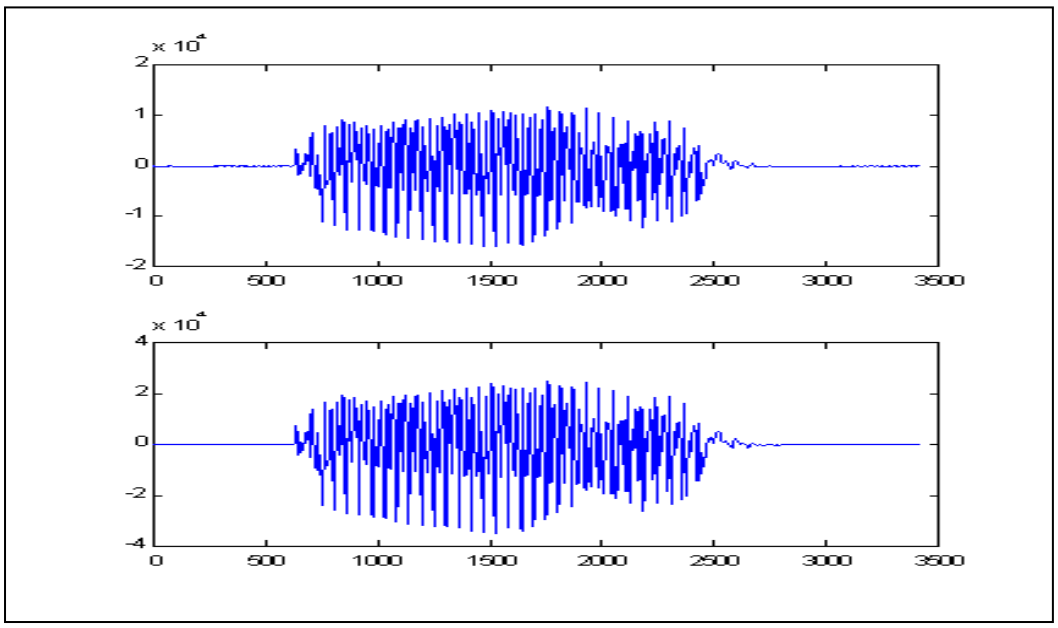


Figure 3.1 : Analyse synthèse par la méthode TD_OLA.

Le recouvrement est donc un paramètre qui dépend de la fenêtre utilisée lors de l'analyse spectrographique. Il dépend en fait de la largeur de son spectre. Dans l'exemple ci-dessus, la fenêtre utilisée est une fenêtre de Hamming avec un recouvrement de 75% c'est à dire un pas $P=L/4$ [13].

3.2.2. Principe de fonctionnement de la technique 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 (Time Domain Harmonic Scaling), SOLA (Synchronized Overlap-Add), WSOLA (Waveform Similarity Overlap-Add).

La méthode PSOLA est une des variantes d'OLA, dans ces techniques, le fenêtrage ne se fait pas, à pas constant mais de manière synchrone de la fréquence fondamentale, ce qui exige un marquage précis de la fréquence fondamentale. Le taux de recouvrement est d'une période locale ($\approx 50\%$) et chaque sommet d'une fenêtre (fenêtre de Hamming) coïncide avec un pic glottique dont la taille est le double de la période locale. Les pics sont alors déplacés suivant l'axe des temps de façon à épouser la forme du nouveau contour $F(\tau)$ (contour lissé des fréquences) et leurs positions sont calculées par la formule.

$$t_{j+1} = t_j + \frac{1}{t_{i+1} - t_i} \int_{t_i}^{t_{i+1}} 1/F(\tau) d\tau \qquad (3.17)$$

L'indice i renvoie aux instants avant modification alors que l'indice j renvoie aux instants après modification du contour. Dans les zones non voisées, les instants t_j sont régulièrement espacés d'une durée de l'ordre de 10ms [11].

La méthode PSOLA se distingue de ces méthodes par une synchronie à la période fondamentale tant à l'analyse qu'à la synthèse. Ceci permet un contrôle à la fois du déroulement de l'axe temporel et de la hauteur du signal.
Les différentes versions de PSOLA existantes fonctionnent selon le même principe. Le segment de signal de parole naturelle est subdivisé en un ensemble de signaux dits à Court-Terme (CT) en utilisant un fenêtrage synchronisé avec le pitch (trame voisée) et à intervalles fixes (trame non voisée). Le pitch est augmenté ou diminué en agissant sur la distance entre les signaux à CT durant le processus de synthèse. La durée est gérée par suppression ou duplication des signaux à CT.

3.2.3. La technique TD-PSOLA

La technique dite d'addition recouvrement des fenêtres temporelles synchrones avec le pitch **TD-PSOLA**; applique le principe de la réharmonisation spectrale directement sur le signal de parole. Appliquée à la synthèse par concaténation, elle conduit à une base de données « paramétriques » où les seuls paramètres stockés sont les marqueurs du pitch indiquant le milieu des fenêtres d'OLA dans les signaux de base de données des segments. Ces marqueurs sont positionnés en synchronisme avec le pitch à l'aide d'un algorithme d'extraction de pitch, et régulièrement espacés sur les zones Non Voisées où aucune modification de pitch ne devra de toute façon être effectuée.

La fenêtre utilisée doit garantir l'atténuation des lobes secondaires, car elles seront candidates à une sommation ultérieure, et elles portent des informations sur l'identité des fenêtres voisines du signal.

On choisit souvent une fenêtre de Hamming ou une fenêtre Triangulaire, avec une longueur égale à deux fois la période du pitch du signal (Figure 3.2).

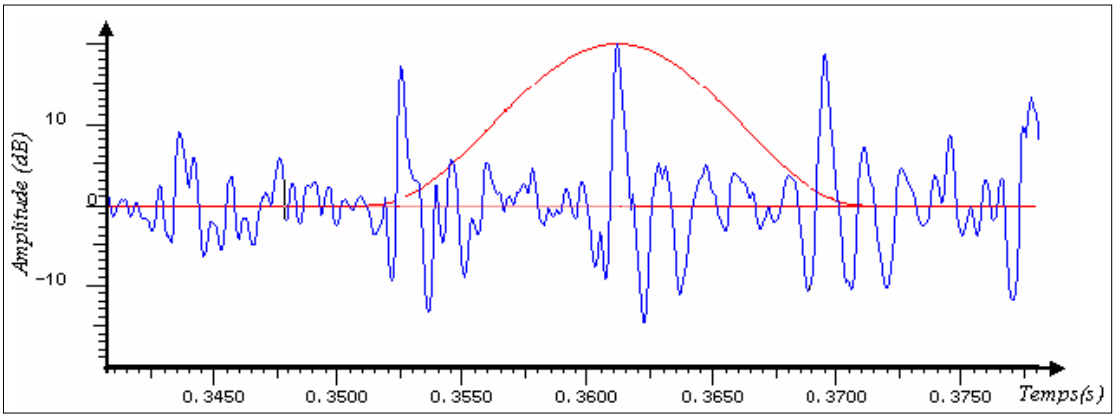


Figure 3.2 : Fenêtrage du signal de parole [7].

Une fenêtre plus large fait apparaître des harmoniques dans le spectre de signal synthétisé; une fenêtre plus courte n'approxime que très grossièrement l'enveloppe spectrale de signal original [7].

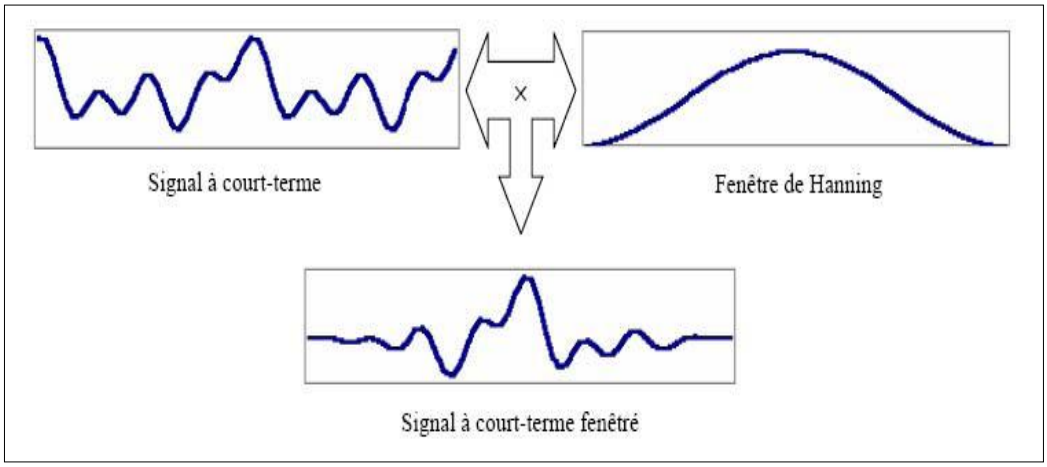


Figure 3.3 : Exemple de signal à Court-Terme [7].

Les signaux à CT sont recombinés pour produire le signal de synthèse à l'aide d'une technique d'addition/recouvrement OLA (Figure 3.4).

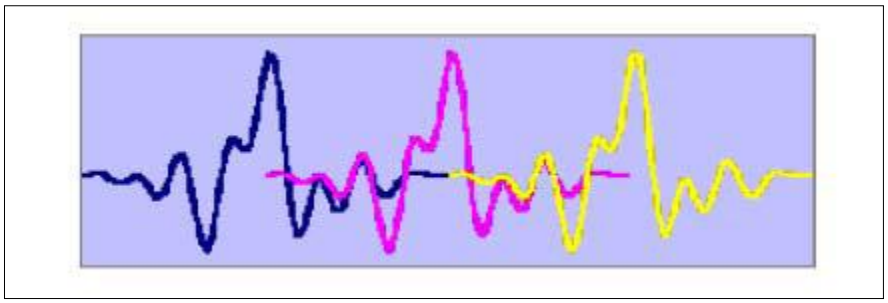


Figure 3.4 : Etape d'addition et recouvrement OLA [7].

Après la recombinaison des signaux à cour terme par addition recouvrement OLA nous trouvons le signale synthétique (figure 3.5).

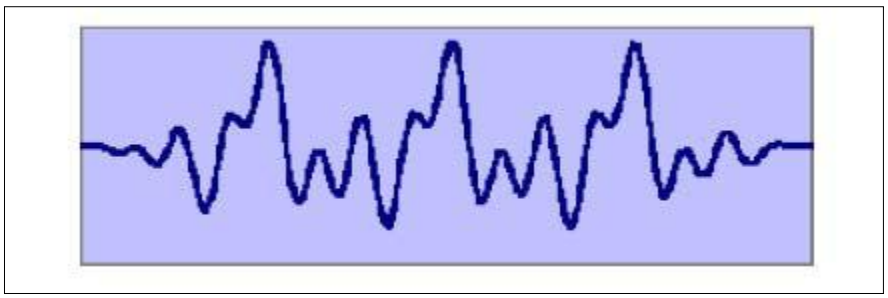


Figure 3.5 : Signal synthétisé avec PSOLA [7].

3.3. Algorithme de synthèse de la technique TD-PSOLA

Après avoir présenté le principe de la technique de synthèse TD-PSOLA, nous pouvons décrire l'algorithme correspondant. Celui-ci requiert trois étapes principales:

- analyse du signal d'origine ;
- modification prosodique apportée à ces signaux à CT ;
- synthèse du signal modifié par recouvrement addition des signaux à CT.

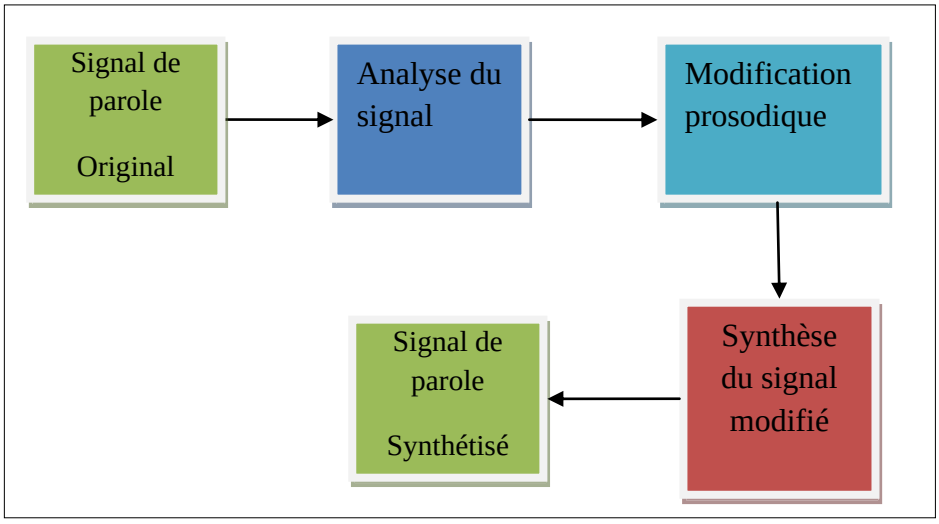


Figure 3.6 : Algorithme de synthèse TD-PSOLA.

3.3.1. Analyse du signal de parole

Les opérations d'analyse à effectuer pour le marquage du signal sont les suivantes :

- prétraitement : séparation des composantes périodiques du signal (dites Voisées dans le cas de la voix) et des composantes aléatoires (dites Non Voisées) ;
- détection et détermination de la F_0 avec une méthode appropriée ;
- détermination des signaux à CT d'analyse.

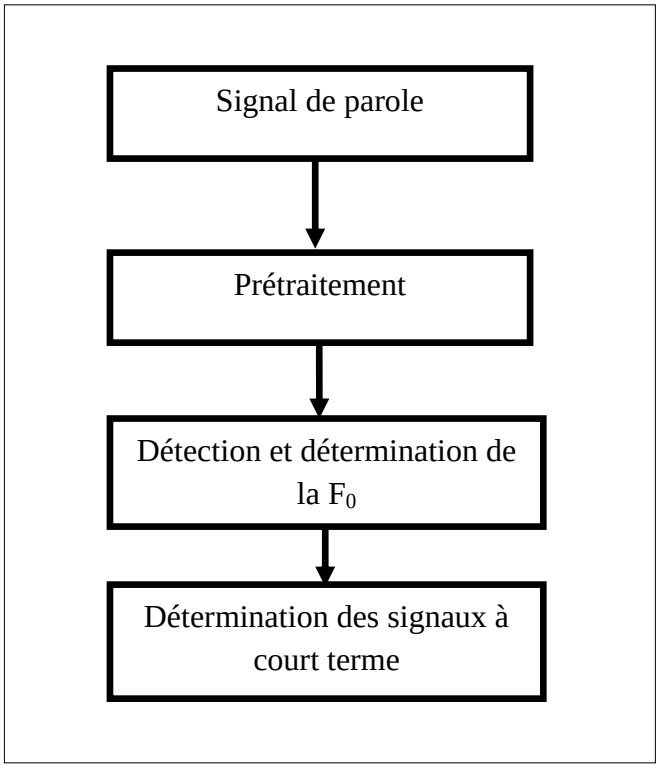


Figure 3.7 : Analyse du signal de parole**.**

3.3.1.1. Prétraitement

Cette phase est en général réservée à la préparation du signal issue d'un microphone. Elle consiste à choisir la durée de la trame d'analyse et du recouvrement afin de moins de compromettre :

- la condition de stationnarité souvent exigée par les algorithmes de traitement ;
- l'effet de bord lié aux fenêtres de pondération appliquées, et d'assurer ainsi la présence d'au moins une période du fondamental. La durée de la trame est généralement choisie entre 20 et 30ms avec un recouvrement de 30 à 50% ;
- la détection de silence garantissant la présence du signal utile ;

- la détection du voisement puisque la F0 n'est présent que sur une séquence Voisée ;
- une préaccentuation afin de rehausser l'énergie des hautes fréquences.

Dans cette phase de prétraitement, nous trouvons souvent d'autres techniques permettant d'améliorer la rapidité d'extraction. Il s'agit de toutes sortes de techniques de filtrage permettant d'atténuer ou même d'éliminer les formants d'ordres supérieur ou égal à 2, et de minimiser l'effet du bruit sur la détection. La décimation dans le rapport de 5 à 1 est souvent utilisée par des détecteurs à temps réels [5]. Cependant la décimation ne peut être utilisée dans les systèmes où une grande précision est requise (telle que la reconnaissance, l'identification, la modification de la F0 etc.).

3.3.1.2. Détection et détermination de Pitch

La phase de traitement est réservée à l'extraction de la fréquence fondamentale et dépend donc de l'algorithme utilisé.

3.4. Les Méthodes de détection du Pitch

Les méthodes de détection de pitch, sont souvent classées en trois catégories principales : temporelles, spectrales et hybrides (combinatoires).

3.4.1. Méthodes temporelles

Les méthodes temporelles sont dites à décalage, Elles sont destinées à exploiter la forte corrélation existant en général entre deux périodes fondamentales successives d'un signal voisé.

Lors de la mise en œuvre de ces méthodes, le signal est divisé en fenêtres temporelles d'une longueur variable, selon les auteurs et les procédés, entre 10 et 30 ms.

Théoriquement la fenêtre doit être suffisamment courte pour que le paramètre à mesurer soit considéré comme constant, et suffisamment long pour qu'il soit mesurable. Notons que ces deux conditions ne sont pas toujours faciles à réaliser.

3.4.1.1. Algorithmes de type corrélation

Les algorithmes de type corrélation travaillent dans le domaine temporel à court terme: le signal est extrait trame par trame, aucune transformation n'a été appliquée sur ces

trames dont la taille est un paramètre important, lorsqu'elle a une valeur fixe elle contient généralement deux à trois périodes du signal. Pour la plupart des algorithmes, il s'agit de trouver un extremum d'une fonction de la période appelée Fonction de Périodicité (FP). Les méthodes de type corrélation se basent sur la similarité du signal entre deux périodes. Il est possible de corréler le signal de départ avec une version décalée de ce signal, décalage correspondant à la période cherchée. Cette corrélation peut aussi être remplacée par une fonction de dissemblance.

3.4.1.2. Fonction d'auto-corrélation

Dans le cas de l'auto corrélation, les deux séquences en entrée sont dérivées du même signal. On introduit un décalage qui constitue le paramètre de la fonction d'auto corrélation :

$$FP_{AutoC}(\tau) = \frac{1}{N}\sum_{i=1}^{N-\tau} x_i x_{i+\tau} \tag{3.18}$$

$x_i\big|_{i=1,N}$ Suite finie d'échantillons du signal.

Les maxima de la FP correspondent à des multiples de la période fondamentale (Figure 3.8). Le premier pic (le décalage donnant la meilleure corrélation) indique la valeur de la F_0. Cette méthode suppose que le signal soit stationnaire, tout au moins dans la trame utilisée or, cette hypothèse est rarement valide sur les signaux étudiés. Avec un signal de parole, l'extraction de la période est donc moins simple car la moindre irrégularité du signal peut provoquer l´apparition de pics dont le décalage est inférieur à T_0. Dans ces conditions, pour améliorer la recherche du pic correspondant à la période, il est possible de choisir d'autres heuristiques, par exemple, en ne retenant que les pics d'amplitude supérieure à un certain seuil (par exemple 50% du maximum de la fonction d'autocorrélation) [14].

L'intérêt de cette méthode, est qu'elle permet le calcul du pitch directement sur le signal surtout pour un signal transmis sur une ligne téléphonique ou dans le cas d'un signal bruité. Les inconvénients liés à cette méthode sont :

- le choix de la fenêtre adéquate pour le calcul à court terme, afin d'atténuer son influence sur la fonction d'autocorrélation. Cependant, la fenêtre idéale doit contenir 2 à 3 périodes de pitch. Sa durée doit être située entre 5 et 20 ms pour des valeurs élevées de la F_0 et entre 20 et 50 ms pour des valeurs plus faibles ;

- le premier maximum peut être lié à la structure des formants surtout pour les voix féminines ou enfantines.

Le problème lié à la structure formantique peut être atténué en proposant un filtrage préalable (filtre passe bas-sélectif aux environs de 800Hz) afin d'éliminer les formants d'ordre supérieur à deux.

L'avantage de cette méthode est qu'elle est très simple, ne nécessite pas un temps de calcul trop coûteux et donne des résultats relativement satisfaisants [5].

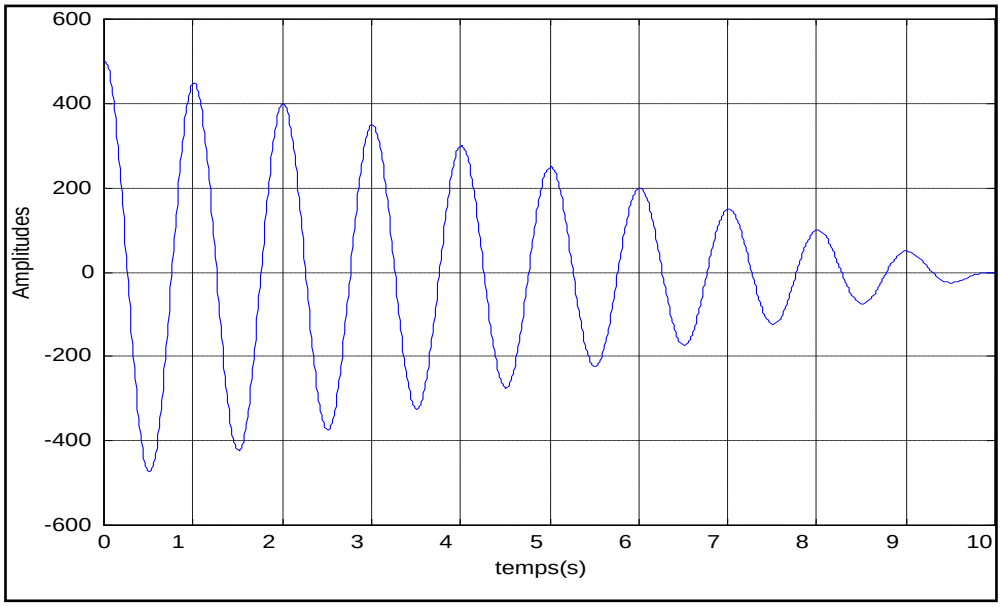


Figure 3.8 : Fonction d'Autocorrelation d'un signal périodique (sinus) [5].

3.4.1.3. **Fonction de distance AMDF** (Average Magnitude Difference Function)

Le critère de variation d'amplitude à court terme au lieu de calculer la corrélation entre deux signaux, utilise la valeur absolue des différences point par point. La FP de l'AMDF est :

$$FP_{AMDF}(\tau) = \sum_{i=1}^{N} \left| x_i - x_{i+\tau} \right| \qquad (3.19)$$

Cette fonction est ensuite normalisée par N ou par $\sum_{i=1}^{N} x_i$ pour que la valeur de l'AMDF puisse être comparée à un seuil absolu dans le but de décider si le signal est périodique ou non.

La FP présente un minimum au niveau des multiples de la période. Cette méthode n'utilise pas l'hypothèse de stationnarité du signal. D'ailleurs, l'ambiguïté entre les pics T_0, $2T_0$, $3T_0$, . . . nT_0 est souvent atténuée par la non-stationnarité du signal analysé : plus le décalage est grand, plus le signal, de part sa non-stationnarité, intègre des différences par rapport à la trame de départ; le signal présente alors plus de différences pour un décalage de T_0 que pour un décalage de $2T_0$ [14].

Une conséquence de la non utilisation de l'hypothèse de stationnarité est que le choix de la taille des fenêtres de signal et celui décalé est libre.

L'AMDF utilise des fenêtres de taille fixe, mais il est possible de concevoir des algorithmes avec des tailles de fenêtre variable, par exemple égale au décalage testé. Cette résistance au problème des erreurs grossières et la rapidité de calcul font de l'AMDF une méthode couramment employée.

3.4.1.4. **Super résolution SRPD** (Super Resolution Pitch Determination)

Le SRPD est un algorithme proposé par Medan, Yair et Chazan [14], avec comme objectif initial de réduire le plus possible les "erreurs fines" d'estimation de la F_0. L'idée de l'algorithme est de comparer selon une mesure de ressemblance deux fenêtres de signal décalées de la valeur de la période test. Cela ressemble fort aux algorithmes du type AMDF, la différence essentielle étant que la taille des fenêtres est ici variable, et plus exactement égale à la période de test. Ainsi l'algorithme vise à positionner au mieux deux fenêtres successives représentant deux périodes successives du signal. La FP de l'algorithme SRPD s'écrit alors :

$$FP_{SRPD}(\tau) = \frac{\sum_{i=1}^{\tau} x_i x_{i+\tau}}{\sum_{i=1}^{\tau} x_i^{\,2} \sum_{i=1}^{\tau} x^{2}_{i+\tau}} \qquad (3.20)$$

Soquet [14] a trouvé que le SRPD, malgré sa grande simplicité donne des taux d'erreurs très acceptables, seules les erreurs de sous-harmoniques présentent un score relativement élevé.

3.4.1.5. Algorithmes basé sur le filtre inverse

La modélisation LPC (Linear Predictive Coding), est en effet aussi applicable en détection de pitch. Markel en 1972 a proposé une méthode de détection qui pourrait être considérée comme temporelle [15]. Elle est basée sur l'examen de la fonction d'autocorrélation du résidu LPC. Cette particularité de la méthode permet en fait de travailler directement sur la source évitant ainsi l'interaction source-conduit vocal, cette méthode est connue sous le nom de méthode SIFT (Simplified Inverse Filtrer Tracking).

Comme nous l'avons vu, la fonction d'autocorrélation présente un maximum à chaque période du fondamental. Le but à atteindre par cette méthode est de calculer le maximum de la fonction d'autocorrélation du résidu de prédiction.

Le signal microphonique capté à la sortie des lèvres est, en fait le résultat de différents filtres mis en cascade. Chacun de ces filtres apporte une certaine déformation au signal de parole. Il suffit d'appliquer à x(n) un autre filtre h'(n) qui est l'inverse de h(n), c'est- à- dire :

H'(f)=1/H(f), pour obtenir le signal d'excitation e(n) (déconvolusion de la sortie) (Figure.3.9)

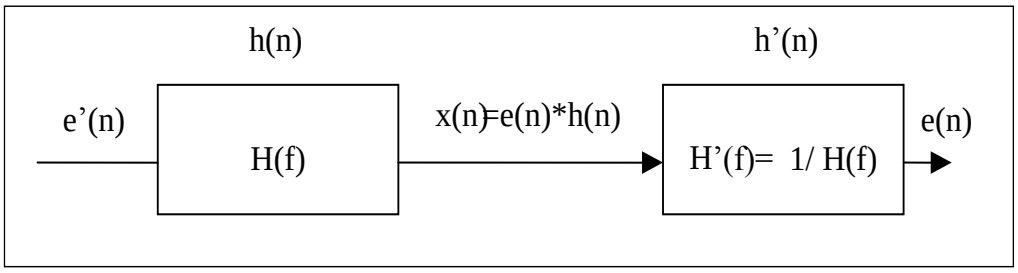


Figure 3.9 : Modélisation du filtre inverse

H(f) étant la FT du filtre direct, ayant la structure d'un modèle AR. Le signal après avoir été filtré par H'(f), ressemble plus à l'excitation glottique.

Afin de faciliter le calcul en temps réel, on opère une décimation dans le rapport 5 à 1 après passage par un filtre passe-bas. On procède après à une analyse LPC (P=4) pour définir le filtre inverse. L'ordre 4 a été choisi comme étant l'ordre optimal suffisant pour la gamme de fréquence utilisée, soit 0-900 Hz et il permet une bonne vitesse de traitement [16]. Les résultats pratiques ont montré que cette méthode est bien adaptée aux applications en temps réel. Cependant la précision atteinte avec ce détecteur est assez médiocre.

3.4.2. Méthodes spectrales

Dans ces méthodes, l'analyse porte sur le spectre instantané du signal obtenu à partir d'une fenêtre temporelle. Le but à atteindre, est de mettre en évidence la structure harmonique des spectres correspondants à des séquences voisées, afin de mesurer l'intervalle fréquentiel entre deux raies harmoniques.

En effet, le spectre d'un signal contient toutes les informations relatives à la source et au conduit vocal. Le spectre d'un signal vocal est le produit du spectre de la source par la FT du conduit vocal. Les variations rapides du spectre sont dues à la source, tandis que les lentes sont liées au conduit vocal. Le problème qui se pose est de trouver un moyen d'isoler les deux phénomènes.

3.4.2.1. Méthode du cepstre

Avec la méthode du cepstre, on arrive à séparer la source du conduit vocal, en prenant logarithme du spectre. On passe ainsi d'un produit à une somme. On calcule ensuite la Transformée de Fourier Inverse (TFI), et on obtient le cepstre dans le domaine des fréquences La figure (3.10).

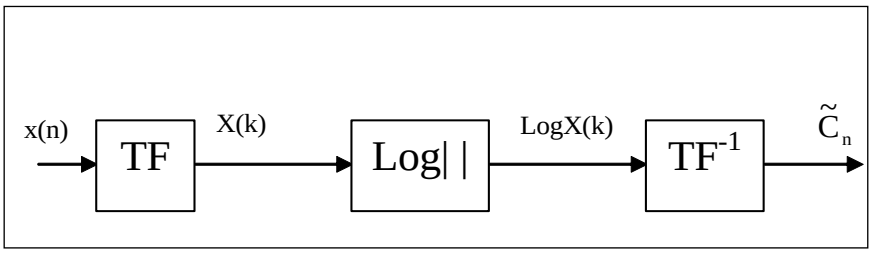


Figure 3.10 : Méthode du cepstre.

$$Cepstre(x(n)) = TF^{-1}(Log\,|X(f)|) \tag{3.21}$$

Avec X(f) la TF de x(n)

Le signal à la sortie du microphone est donné par équation 3.2 Par TF on obtient

$$X(f) = E(f).H(f) \tag{3.22}$$

Afin de transformer le produit en une somme, on utilise l'opérateur logarithme :

$$Log(X(f)) = Log(E(f)) + Log(H(f)) \tag{3.23}$$

Par Transformé de Fourier Inverse on obtient :

$$TF^{-1}(Log(X(f))) = TF^{-1}(Log(E(f))) + TF^{-1}(Log(H(f))) \quad (3.24)$$

Enfin

$$Cepstre(x(n)) = Cepstre(e(n)) + Cepstre(h(n)) \quad (3.25)$$

Les maxima du cepstre correspondent à ceux du spectre en fréquence, c'est-à-dire à la fréquence fondamentale et aux formants. A l'aide d'un filtrage adéquat dans le domaine des fréquences, on arrive à isoler soit les formants (filtrage passe bas), soit le pitch (filtrage passe haut) [19].

3.4.2.2. Méthode d'intercorrelation avec la fonction peigne

La méthode d'intercorrelation avec une fonction peigne proposée par Martin [14] utilise une représentation fréquentielle à court terme. Les algorithmes utilisant cette représentation exploitent généralement la structure harmonique des signaux périodiques.

L'algorithme est fondé sur l'intercorrelation entre le spectre du signal et une série d'harmoniques d'impulsions de Dirac d'amplitude normalisée ("peigne"). Cela revient à cumuler toutes les valeurs des amplitudes du spectre à des positions multiples de la fréquence de test f. Le spectre étant noté S(f) et ses coefficients $\alpha_{i=1...n(f)}$ la Fonction de Périodicité (FP) s'écrit alors :

$$FP_{peigne}(f) = \sum_{i=1}^{n(f)} \alpha_i \left| S(i*f) \right| \quad (3.26)$$

Où n(f) désigne le nombre d'harmoniques : il s'agit d'une fonction dépendante de la fréquence fondamentale de test f et qui est à valeurs inférieures à F_{max} où F_{max} représente la plus haute fréquence prise en compte dans le spectre.

Pour résoudre le problème des sous-harmoniques, Martin applique une fonction de décroissance exponentielle sur les amplitudes des pics du peigne ($\alpha_i = e^{-\beta i}$). De cette façon les sous-harmoniques ont une intercorrelation inférieure à celle du F_0.

Dans ces conditions, la FP présente un maximum pour la F_0 même si cette fréquence est la seule composante du spectre dépourvu de toute harmonique, ou si seules deux harmoniques

consécutives sont présentes. Le volume de calcul de cette méthode est du même ordre que celui du cepstre.

3.4.3. Méthodes combinatoires

Il existe un très grand nombre de méthodes pour extraire la F_0, chacune présentant des avantages et des inconvénients, mais aucune ne permet d'évaluer la fréquence fondamentale avec une précision absolue. Ces observations ont conduit Hess [17] à suggérer de combiner différentes approches pour augmenter les performances globales du système d'extraction. L'idée est d'appliquer différents analyseurs simultanément sur le signal et de combiner les différentes estimations ainsi obtenues. Dans cette troisième catégorie, on effectue des traitements fréquentiels sur le signal de parole dans le but d'aplanir le spectre d'amplitude. Le signal obtenu après ce traitement est ensuite analysé par des méthodes de type autocorrélation, afin d'estimer la périodicité.

3.5. Détermination des signaux à court terme

Comme nous avons vu, la méthode PSOLA repose sur le découpage d'un signal x(n) en des fenêtres successives $s_i(n)$ en fonction des périodes fondamentales du signal.

Les signaux à court terme sont donnés par l'équation suivante :

$$s_i(n) = x(n)w_i(n - iT_0) \tag{3.27}$$

$$s(n) = \sum_i s_i(n - i(T - T_0)) \tag{3.28}$$

On effet, cette opération résulte d'après le théorème de la somme de Poisson, en une réharmonisation du spectre de $s_i(n)$ (qui, si nous supposons le signal de départ purement périodique, et indépendant de i) avec une nouvelle F_0 $\frac{1}{T}$. Si

$$s_i(n) \overset{\Im}{\longleftrightarrow} S_i(\omega) \quad \text{alors} \quad s(n) \overset{\Im}{\longleftrightarrow} \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} S_i(n\frac{2\pi}{T})\delta(\omega - n\frac{2\pi}{T}) \tag{3.29}$$

Il s'ensuit que si la fenêtre de pondération w(n) est choisie de façon à ce que le spectre de $s_i(n)$ approxime l'enveloppe spectrale de x(n), l'équation 3.28 fournit un moyen très simple de modifier la F_0 d'un signal périodique.

Rappelons le Théorème de **POISSON**

Suivant la formule de Poisson, la somme d'une infinité de versions décalées d'un même signal f(t) conduit à un signal périodique dont les raies spectrales viennent se positionner exactement sur le spectre du signal du départ [14] :

$$Si \quad f(t) \overset{\Im}{\longleftrightarrow} F(\omega),$$

$$alors \quad \sum_{n=-\infty}^{\infty} f(t-nT_0) \overset{\Im}{\longleftrightarrow} \frac{2\pi}{T_0} \sum_{n=-\infty}^{\infty} F(n\frac{2\pi}{T_0})\ \delta(\omega - n\frac{2\pi}{T_0}) \tag{3.30}$$

Ces fenêtres successives sont obtenues par placement de marques appelées marques de lecture t_r^i de manière synchrone au pitch du signal (la différence entre deux marques de lecture successive est égale à T_0 locale) (Figure.3.11). Le signal est alors découpé à l'aide de fenêtres d'analyse centrées sur ces marques de lecture.

$$s_i(n) = x(n)w(n - t_r^i) \tag{3.31}$$

Le signal de synthèse s(n) sera donc obtenu par Superposition/Addition des signaux élémentaires centrés en de nouvelles positions t_w^i que nous appelons marques d'écriture. Ce sont ces positions qui déterminent le pitch et la durée du signal de synthèse.

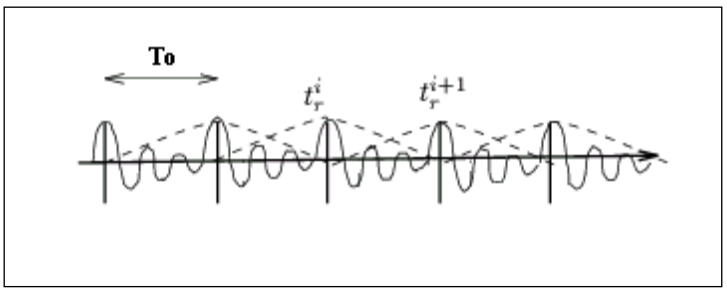


Figure 3.11: Placement des marques de lecture.

3.5.1. Opération de marquage de la fréquence fondamentale

La précision du placement des marques d'écriture détermine en grande partie la qualité du signal de synthèse obtenu. Les marques doivent être placées non seulement de manière synchrone au pitch du signal, mais également de manière à ce que le fenêtrage

préserve au maximum les caractéristiques temporelles du signal. Elles doivent aussi se trouver prés des maxima locaux d'énergie [5].

Un algorithme de marquage vise à optimiser les contraintes suivantes :

- marques synchronisées à la période locale ;
- un espacement entre deux marques successives égales à la période locale vraie ;
- un éloignement minimum des marques par rapport aux maxima d'énergie locaux.

La position des marques est optimisée sur l'ensemble des marques appartenant à une région Voisée. Pratiquement l'optimisation des différentes contraintes est difficile à réaliser, ainsi il existe plusieurs façons d'élaborer un algorithme de marquage. Le principe est d'essayer d'avoir une méthode moins complexe avec des résultats acceptables [14].

3.6. Modifications prosodiques des signaux à court terme

La séquence des signaux à court terme analysée est reprise pour reproduire une nouvelle séquence de signaux synthétiques synchronisés avec un nouvel ensemble de marques de pitch de synthèse.

Les nouvelles marques de pitch sont déterminées en fonction des spécifications prosodiques, ainsi elles seront plus ou moins écartées si une modification de pitch est demandée. Par ailleurs, il n'y pas une correspondance exacte entre les marques de pitch de synthèse et celle d'analyse, car on peut être amené à éliminer ou à dupliquer quelques marques. Ceci est effectué puisque le nombre de marque de pitch détermine la durée du signal synthétique qui est aussi spécifié par le module prosodique.

Comme nous venons de voir, la modification de la F_0 entraîne une modification de la durée totale du signal résultant, alors dupliquer ou supprimer des marques veut dire que certaines fenêtres doivent être répétées ou éliminées selon qu'on augmente ou on diminue le pitch.

Il existe plusieurs façons d'élaborer des règles de duplication/élimination des fenêtres recouvrantes. Le but est de ne pas perdre des informations et d'avoir une qualité acceptable du signal synthétique avec la même durée de départ.

En l'absence de modifications, les instants de synthèse correspondent aux instants d'analyse, et les signaux à court terme de synthèse sont égaux aux signaux à court terme d'analyse

3.7. Synthèse du signal modifié par recouvrement /addition des signaux à court terme

La synthèse est effectuée par Superposition/Addition des signaux élémentaires (obtenue à partir des $s_i(n)$ placés en de nouvelles positions t_w^i). Ces positions sont déterminées selon la hauteur voulue (Figure 3.12).

$$s(n) = \sum_{1}^{N} s_i(n - t_w^i) \tag{3.32}$$

Ou N est le nombre de signaux à court terme.

Pour tenir compte du fenêtrage de l'analyse, il est nécessaire de normaliser le signal obtenu par simple sommation des signaux élémentaires, et on obtient le signal de synthèse donné par :

$$s(n) = \frac{\sum_{1}^{N} s_i(n - t_w^i)}{\sum_{1}^{N} w_i(n - t_w^i)} \tag{3.33}$$

Nous utilisons souvent la méthode de Giffin et Lim [19] des moindres carrées donnée par.

$$s(n) = \frac{\sum_{1}^{N} r_i(n) w_i(n - t_w^i)}{\sum_{1}^{N} w_i^{\,2}(n - t_w^i)} \tag{3.34}$$

Avec $r_i(n) = s_i(n - t_w^i)$

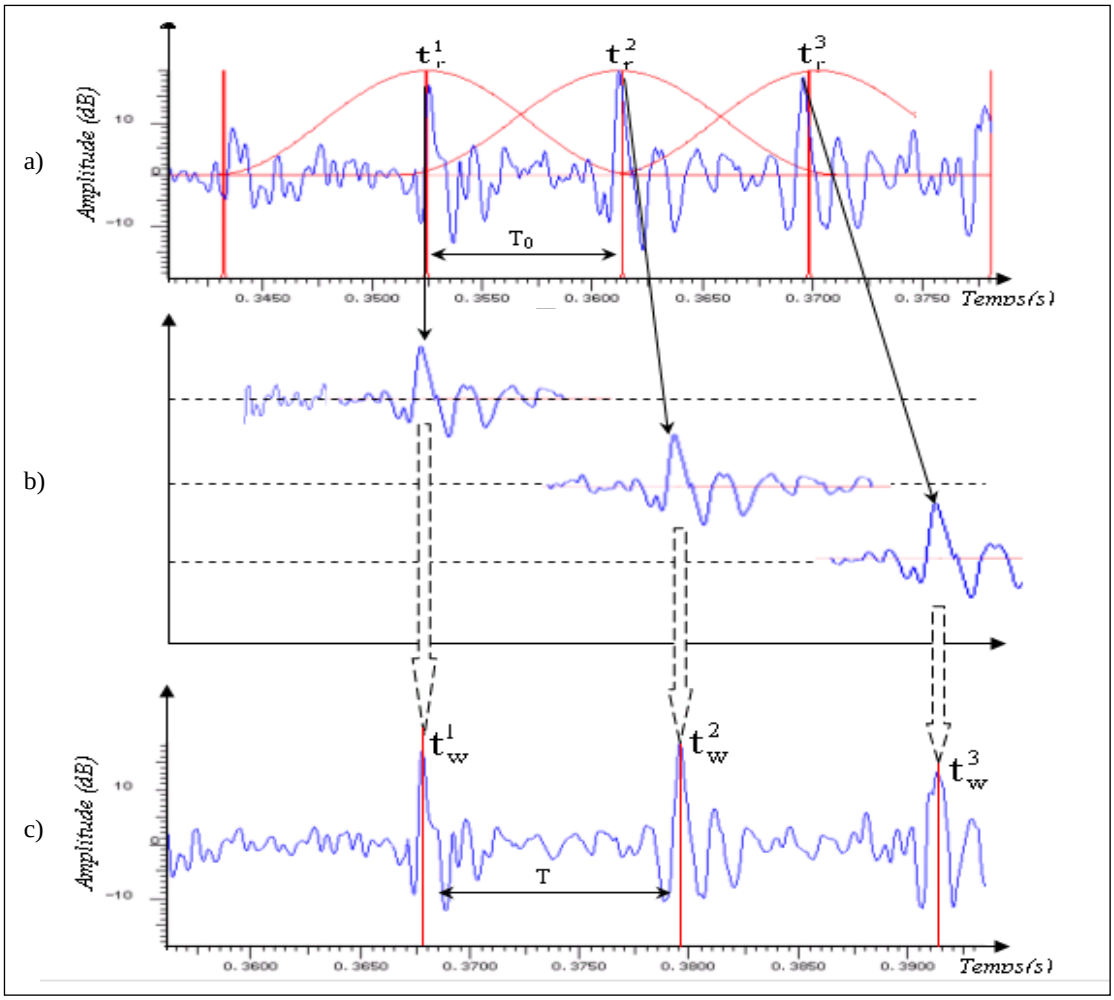


Figure 3.12 : Modification de la F_0 par un facteur 1.2 avec TD – PSOLA.

a) : Signal de parole original ainsi que les positions centrales des signaux à CT ;

b) : Signaux à court terme décalés ;

c) : Signal modifié obtenu par addition des signaux à court terme décalés.

Le facteur de normalisation variable tient compte des variations d'énergie liées à la cadence irrégulière de l'analyse et de la synthèse. On remarque qu'en l'absence de modifications le signal de synthèse correspond exactement au signal d'analyse. Nous pouvons vérifier cette équation comme suite [20]:

D'après l'équation 3.28 et 3.29. Si nous posons

$$r_i(n) = s_i(n - i(T - T_0)) \tag{3.35}$$

Le signal de synthèse s'écrit :

$$s(n) = \sum_{1}^{N} r_i(n) \tag{3.36}$$

s(n) peut s'écrire :

$$s(n) = \sum_{1}^{N} x(n - i(T - T_0))w_i(n - iT) \tag{3.37}$$

Si $T=T_0$ c.à.d on veut retrouver le signal original, on obtient :

$$s(n) = \sum_i x(n)\ w_i(n - iT) \tag{3.38}$$

Donc

$$s(n) = x(n)\sum_i w_i(n - iT) \tag{3.38}$$

$$s(n) = cst \times x(n) \tag{3.39}$$

Ce qui résulte que le signal synthétique est égal au signal d'origine multiplié par une constante.

Si on veut utiliser la méthode des moindres carrées pour $T=T_0$ on obtient :

$$s(n) = \frac{\sum_i r_i(n)\, w_i(n - iT)}{\sum_i w_i{}^2(n - iT)} \tag{3.49}$$

On utilisant les équations 3.27, 3.28 et 3.35, nous obtenons :

$$s(n) = \frac{x(n)\sum_{1}^{N} w^2{}_i(n - iT)}{\sum_{1}^{N} w_i{}^2(n - iT)} = x(n) \tag{3.50}$$

On retrouve le signal d'origine.

3.8. Conclusion

La technique PSOLA est une approche pour la manipulation de la parole, elle représente une étape importante dans le développement des techniques du traitement de la parole. Le développement des techniques de synthèse de la parole reflète une attention croissante à la nature physique de production de la parole. Les travaux actuels sont acheminés vers les traitements des sons qui réfléchissent et incluent la large complexité, nuance, expressivité et la richesse en informations de la voix humaine.

Chapitre 4 :

Implémentation des Algorithmes et Résultats de la Simulation

4.1. Introduction

Après avoir présenté, au chapitre précédent, la technique de synthèse et de modification utilisées pour les signaux de parole, nous allons présenter dans ce chapitre les étapes utilisées pour simuler cette technique ainsi que les résultats obtenus par l'application de ces algorithmes.

4.2. Description du corpus utilisé

Dans le cadre de cette thèse, nous avons utilisé un corpus des phrases affirmatives en AS, prononcées par des locuteurs arabophones (masculins et féminins). Ces phrases ont été enregistrées et ont subi une analyse sonagraphique grâce au logiciel de transcription et d'analyse phonétique PRAAT. Afin de pouvoir effectuer des modifications prosodiques du signal de parole (la modification de la durée et la fréquence fondamentale F_0), nous avons choisi comme un signal d'entrée la phrase « [naam kataba kalima] » énoncée en langue arabe et prononcée par un locuteur masculin (Figure 4.1), où toutes les algorithmes seront appliqués sur la dite phrase.

Les moyens informatiques dont nous disposons sont constitués d'un Micro Ordinateur Portable type Dell Vostro de RAM 3Go et d'un logiciel dénommé MATLAB (Version 2009a) dont le but est de faciliter les calculs intermédiaires.

Phrase : [naam kataba kalima] نعم كتب كلمة

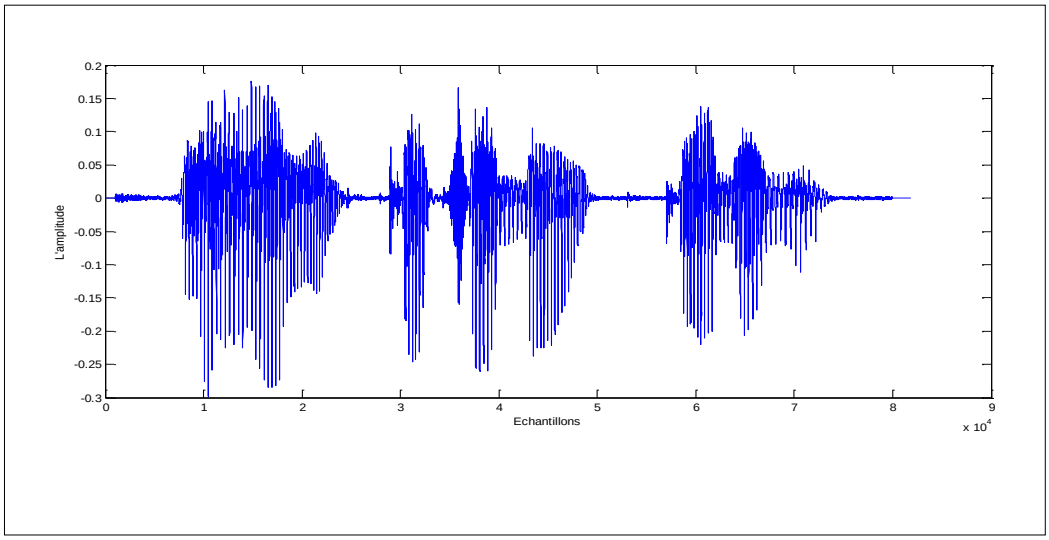


Figure 4.1 : Représentation temporelle de la phrase [naam kataba kalima].

4.3. Organigramme de TD-PSOLA

Les méthodes reposant sur le principe de synchronisation avec le fondamental sont utilisées pour réaliser des modifications temporelles ou fréquentielles d'un signal de parole, ou pour mettre en œuvre des systèmes de synthèse de la parole. Cet organigramme de premier niveau nous montre l'organisation générale du notre programme. Dans ce chapitre vous aurez une explication plus détaillée des différentes fonctions mises en œuvre pour la synthèse de la parole et la modification prosodique. Cet organigramme est chargé d'appeler les différentes fonctions nécessaires à la réalisation de notre objectif tel que

- le chargement du son en mémoire ;
- le filtrage et l'élimination du bruit ;
- le découpage ou la segmentation, etc.

Les dites fonctions seront détaillés tout au long de ce chapitre.

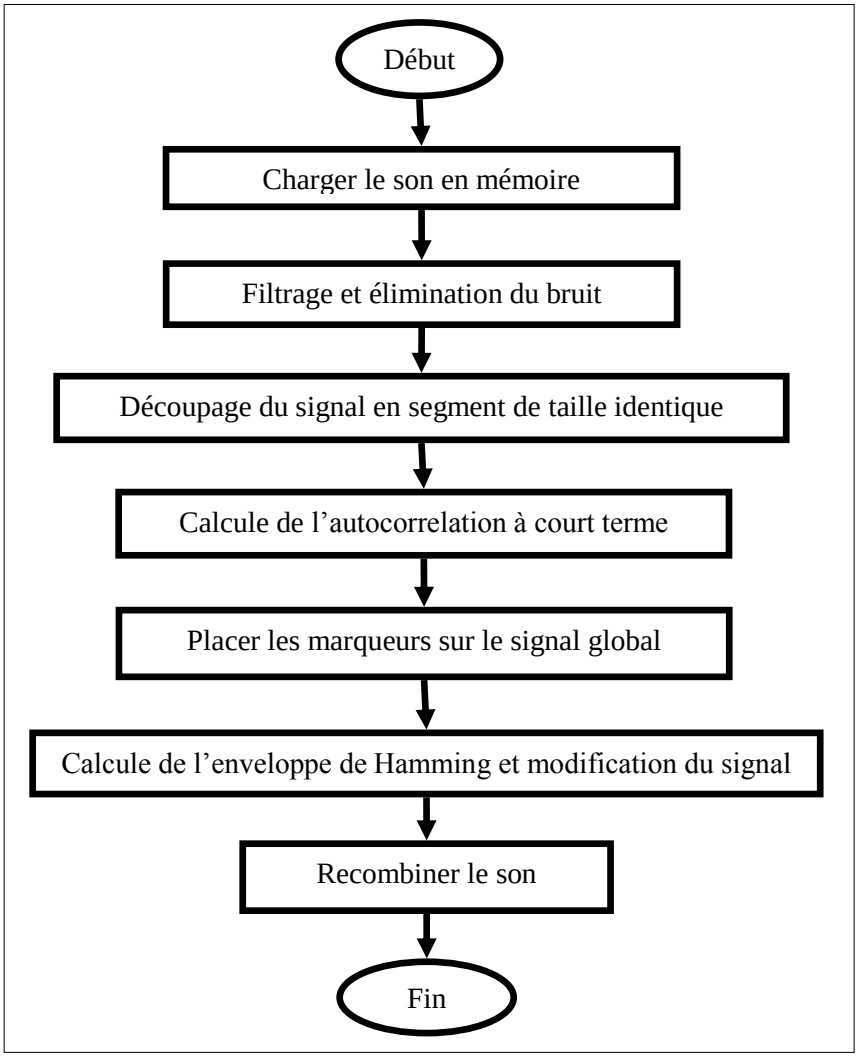


Figure 4.2 : Organigramme représente l'organisation générale du programme.

4.3.1. Chargement de son en mémoire

En effet comme indiqué précédemment nous avons enregistré les sons sous format « .wav », or Matlab n'accepte pas ce format. Les sons ont donc été convertit à l'aide de la fonction « wavread ».

4.3.2. Filtrage du signal vocal

La fonction de filtrage permet de réaliser un prétraitement du signal en le filtrant par trois filtres passe-bas du premier ordre mis en cascade de façon à créer un filtre du troisième ordre. Cette manière de procéder a été plus efficace qu'un filtre de troisième ordre classique. Il est ainsi préférable, avant de poursuivre l'analyse, de commencer par éliminer les fréquences supérieures à 600 Hz, car ce filtrage est nécessaire pour éliminer le bruit haute fréquence présent sur le signal (dû à un mauvais enregistrement) avant d'utiliser la fonction d'autocorrélation pour la détermination de la fréquence fondamentale.

Concernant les paramètres utilisées pour la programmation sont comme suite :

- Gain statique =1 ;
- Filtre du premier ordre ;
- Fréquence de coupure d'environ 600 Hz.

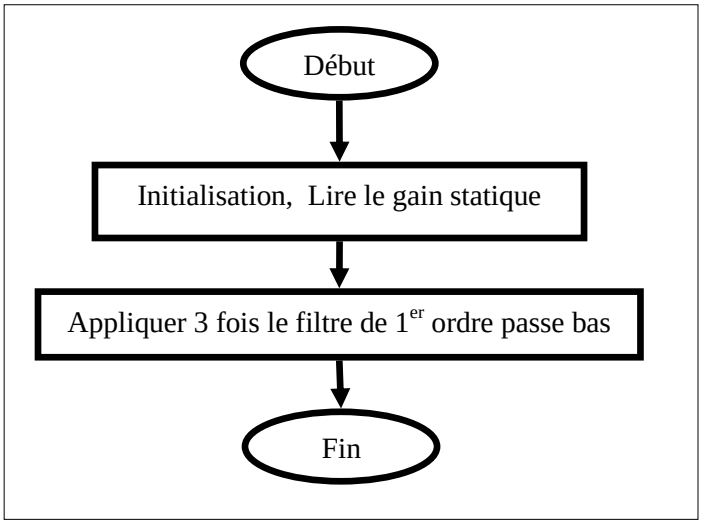


Figure 4.3 : Organigramme de la partie de filtrage.

4.3.3. La fonction de segmentation

Dans un système de synthèse de la parole, le premier traitement à faire, après le filtrage, est de segmenter le signal en des suites d'unités élémentaires, la segmentation fait référence aux notions de différences et de similitudes.

Le but de la segmentation est de fournir une résolution avec le plus que possible de précision, dans la mesure où tous les traitements qui vont suivre reposent sur les résultats de cette segmentation. Cette fonction fragmente le signal filtré en N trames de durée égale à 20 ms pour respecter la condition de stationnarité d'un tel signal. Elle organise le résultat de la fragmentation en un tableau dont les colonnes représentent les trames et les lignes les échantillons de chaque trame. De plus cette fonction calcul le nombre d'échantillons de chaque trame, le nombre d'échantillons total du signal en entrée donc non fragmenté et enfin le nombre de trames que nécessite le signal original. Ces paramètres sont envoyés comme arguments à la fonction suivante qui les utilise pour déterminer la valeur de la fréquence fondamentale pour chacune des trames.

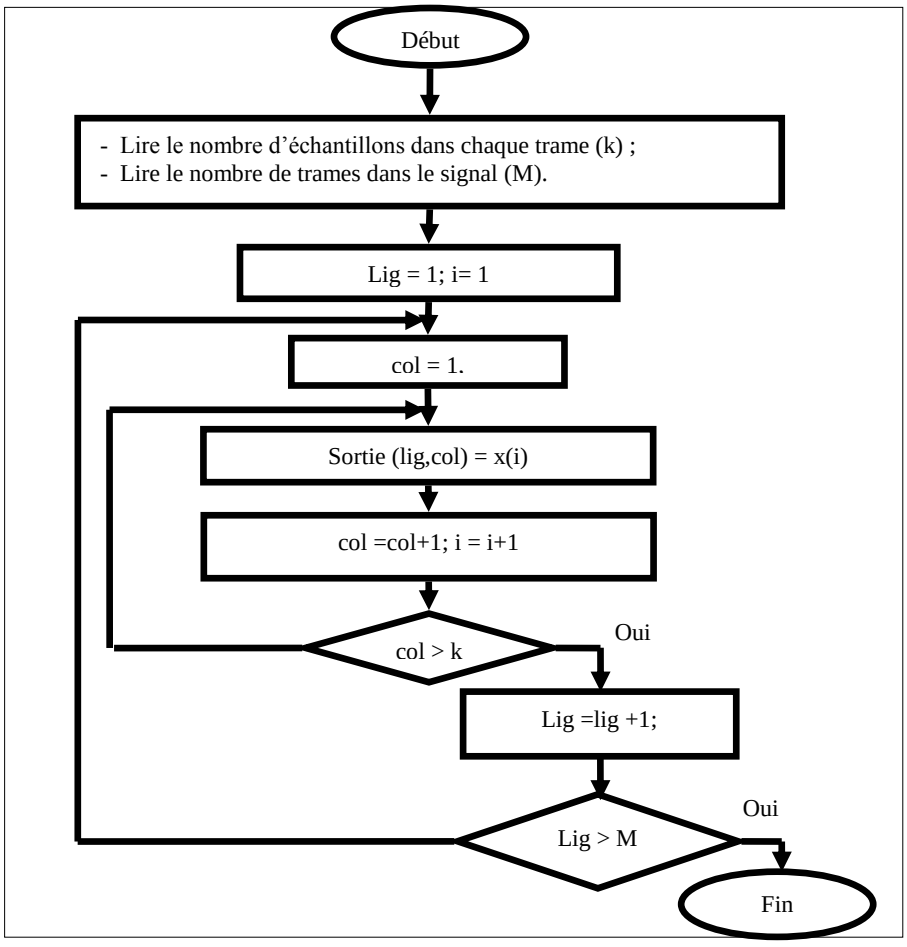


Figure 4.4 : Organigramme représentant le principe de fonctionnement de la segmentation.

4.3.4. Détermination du pitch

La période du fondamental (appelé communément le pitch) est un paramètre très important pour la synthèse de la parole, en effet l'oreille est très sensible à ses variations, qui constituent la prosodie ou timbre du locuteur. L'extraction du pitch a été une tâche particulièrement difficile pour trois raisons :

- premièrement, les vibrations des cordes vocales n'ont pas nécessairement une périodicité complète, particulièrement au commencement des sons voisés ;
- Deuxièmement, il est difficile de séparer le pitch des effets des paramètres vocal ;
- Troisièmement, la plage dynamique de la F_0 est très grande.

Dans notre cas, nous avons utilisé une technique temporelle qui est souhaitable pour l'analyse de la micromélodie, elle est basée sur le calcul de la fonction d'autocorrelation. C'est une technique de base pour la plupart des techniques utilisées à présent et qui donne des résultats satisfaisants avec moins de complexité.

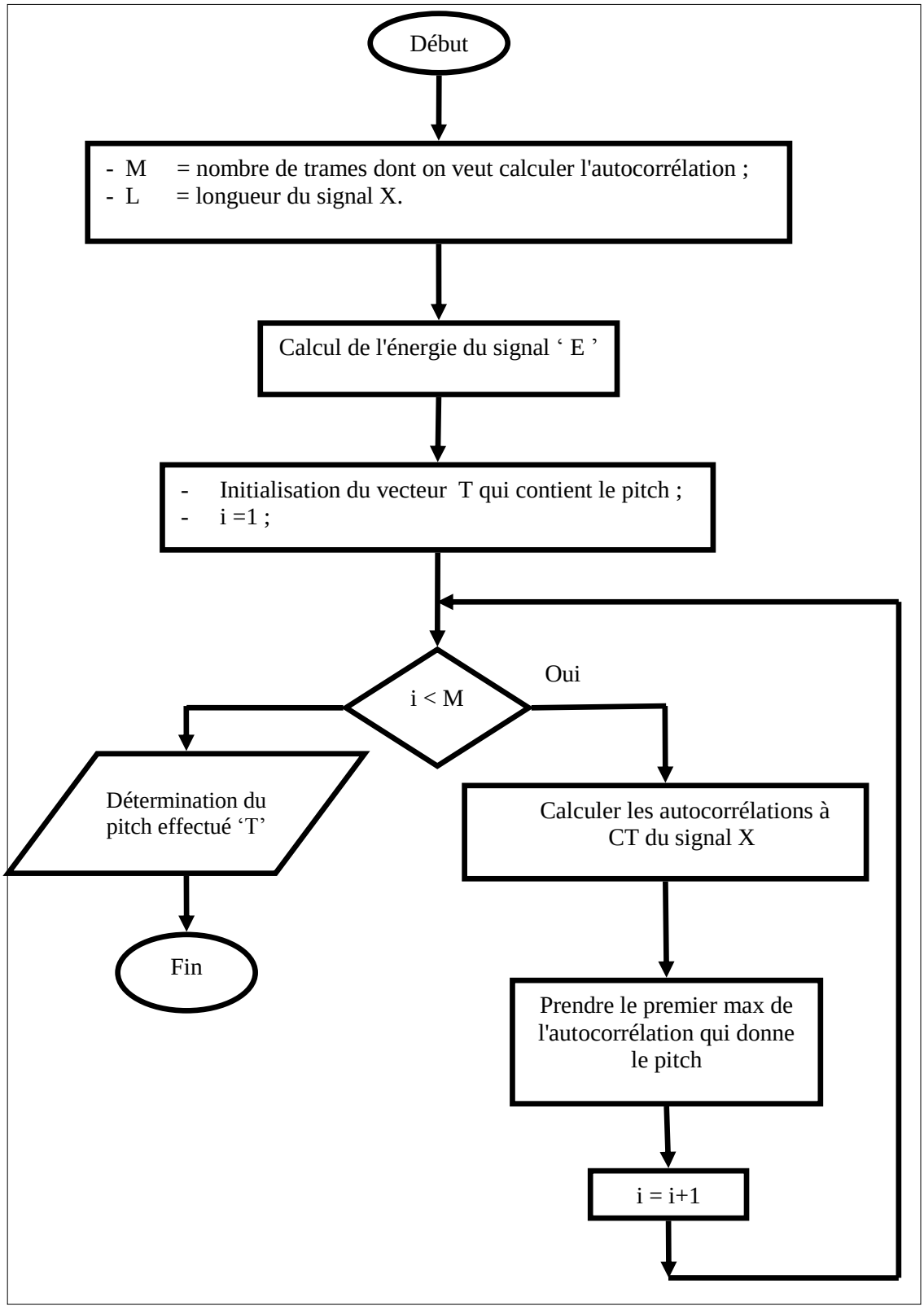


Figure 4.5 : Organigramme de la fonction qui détermine le pitch.

Dans cette première partie du programme nous calculons l'énergie totale du signal de manière à pouvoir pondérer le résultat de l'autocorrélation par la valeur trouvée. Normalement, pour un homme, la fréquence fondamentale peut varier de 70 à 250 Hz, nous avons choisit de ne prendre en compte que les valeurs de pitch comprises entre 70 et 400 Hz. La boucle ci-dessus permet de calculer les autocorrélations à court terme de chaque trame, nous effectuons ensuite le seuillage de l'autocorrélation en prenant 50% de la valeur maximum trouvée dans la trame. C'est en fait le premier échantillon dont on divise la valeur par 2 qui sert de seuil de détection. On recherche ensuite entre les limites indiquées plus haut (70 – 400 Hz) un maximum. Une fois ce maximum trouvé on vérifie qu'il s'agit bien d'un maximum (si l'échantillon précédent et le suivant sont bien inférieurs à celui trouvé) et qu'il est supérieur ou égal au seuil fixé, alors ce maximum correspond à la valeur du pitch de la trame en cours d'analyse. Si une de ces conditions n'est pas réalisée on considère que la trame est non voisée.

Le vecteur T contient les valeurs de pitch pour chaque trame. Il contient la valeur 0 s'il s'agit d'une trame non voisée et la valeur du pitch trouvée si la trame est voisée.

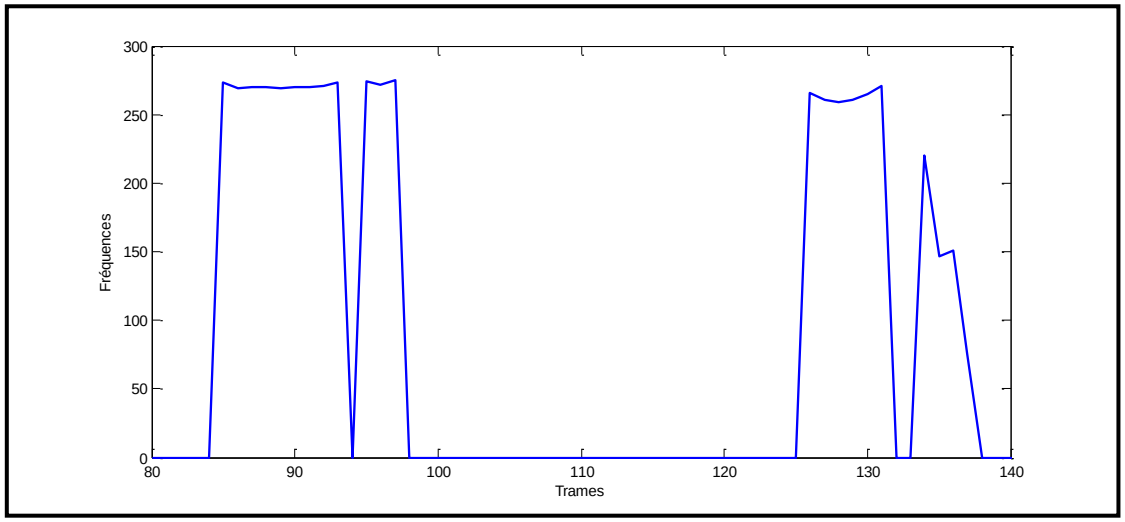


Figure 4.6: Evaluation de la F_0 du signal d'entré en fonction des blocs d'analyse par la méthode d'autocorrélation.

4.3.5. Détermination des marqueurs sur le signal

Cette fonction permet de créer un tableau de « marqueurs » dont la première colonne indique la fin de la période coupé entre deux trames, la seconde colonne indique le nombre, entier, de périodes que peut contenir une trame, la dernière colonne donne la valeur du pitch de la trame.

Pour pouvoir bien repérer le début et la fin d'une période nous calculons le dépassement du pitch P1 de la trame 1 sur la trame 2 car une période pitch peut chevaucher la trame suivante. Donc, on dispose ainsi d'un tableau regroupant, le dépassement, le nombre de période pitch contenue dans la trame actuelle et la valeur du pitch. On peut ainsi grâce à ce tableau connaître le début et la fin d'une période pitch.

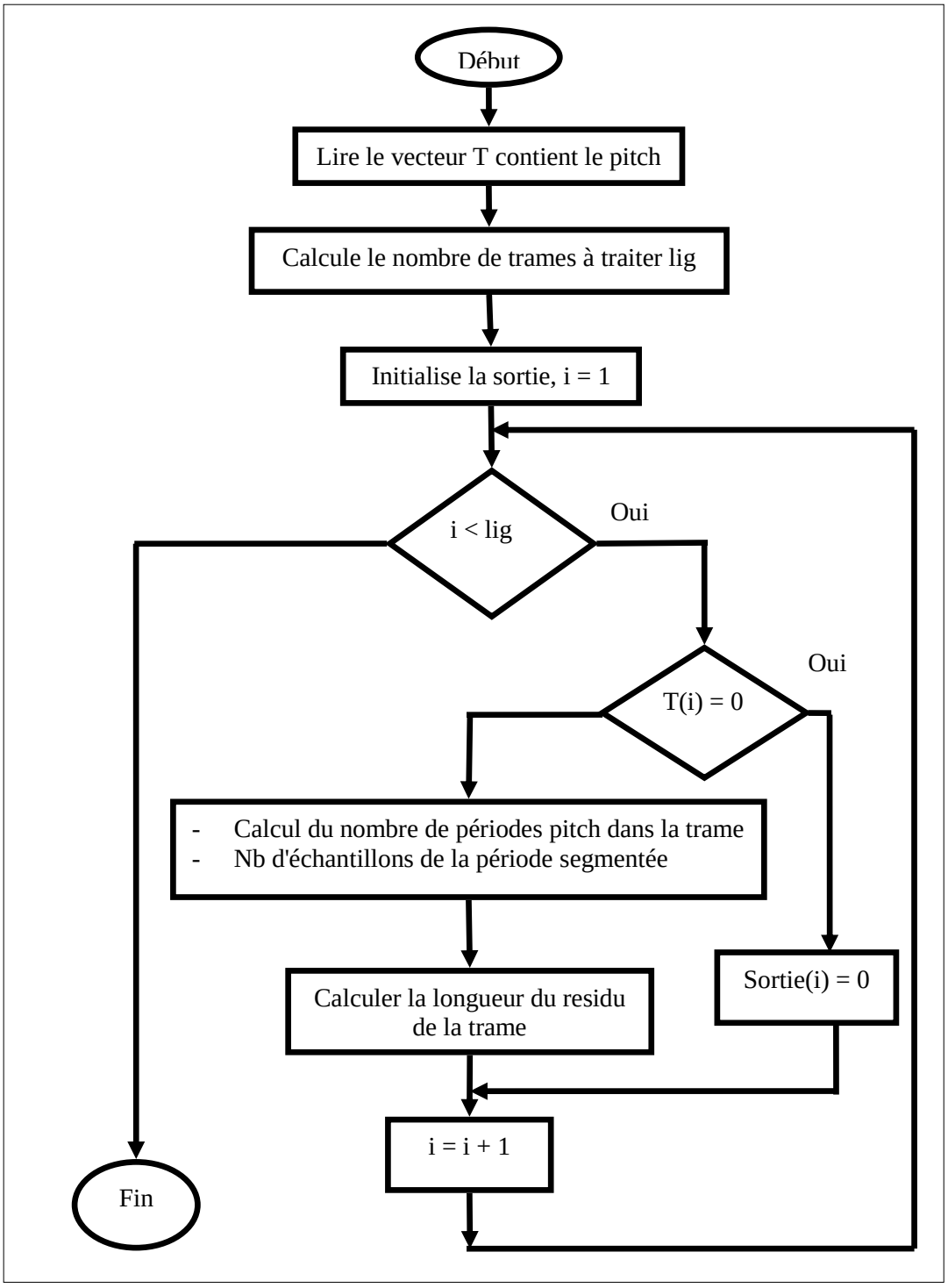
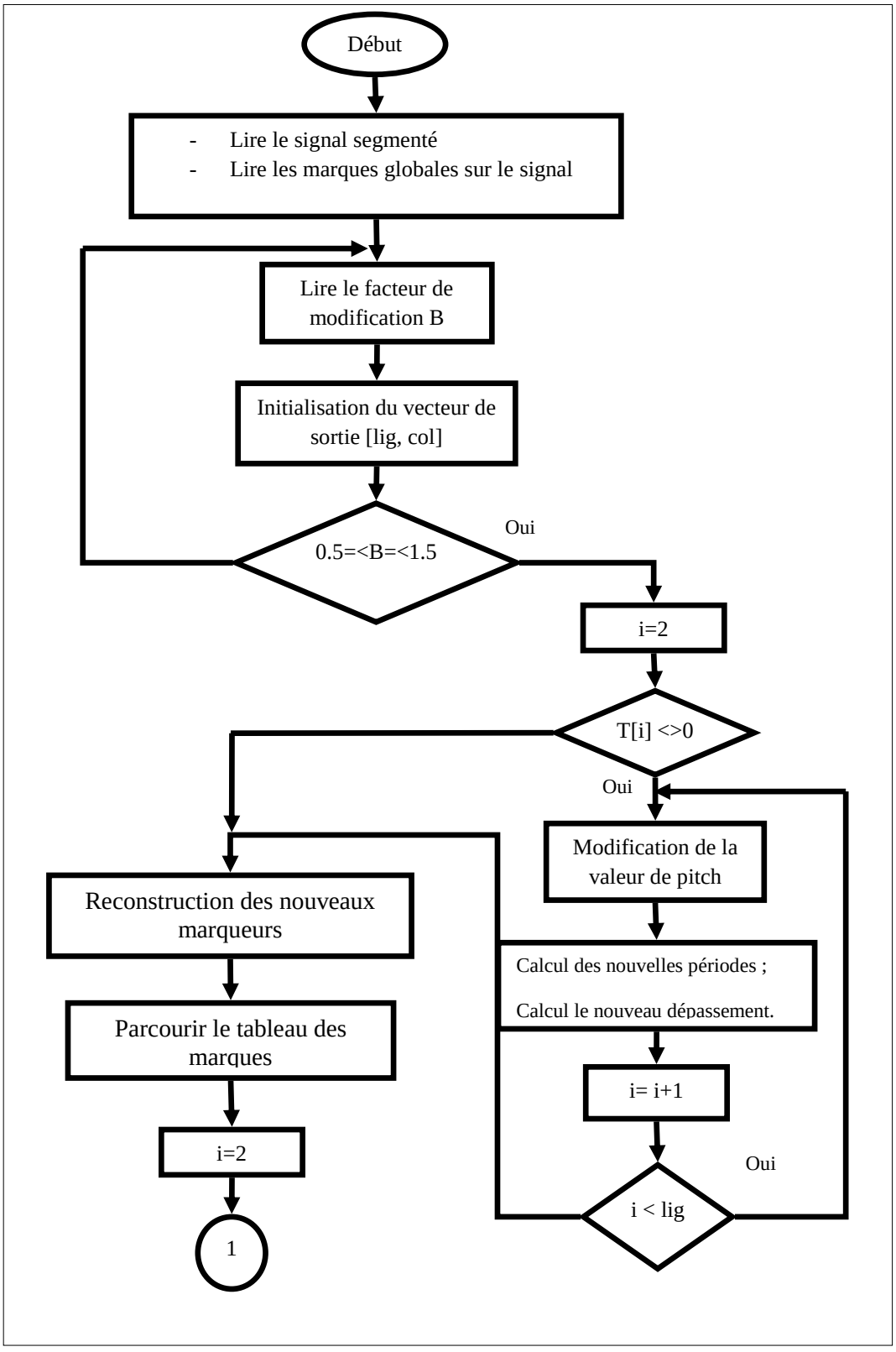


Figure 4.7 : Organigramme de calcul des marqueurs sur le signal global.

4.3.6. Calcul de l'enveloppe de Hamming et modification du signal vocal

Dans cette première partie du programme on réalise différentes initialisations, puis on modifie les marques suivant le facteur de modification entré par l'utilisateur.



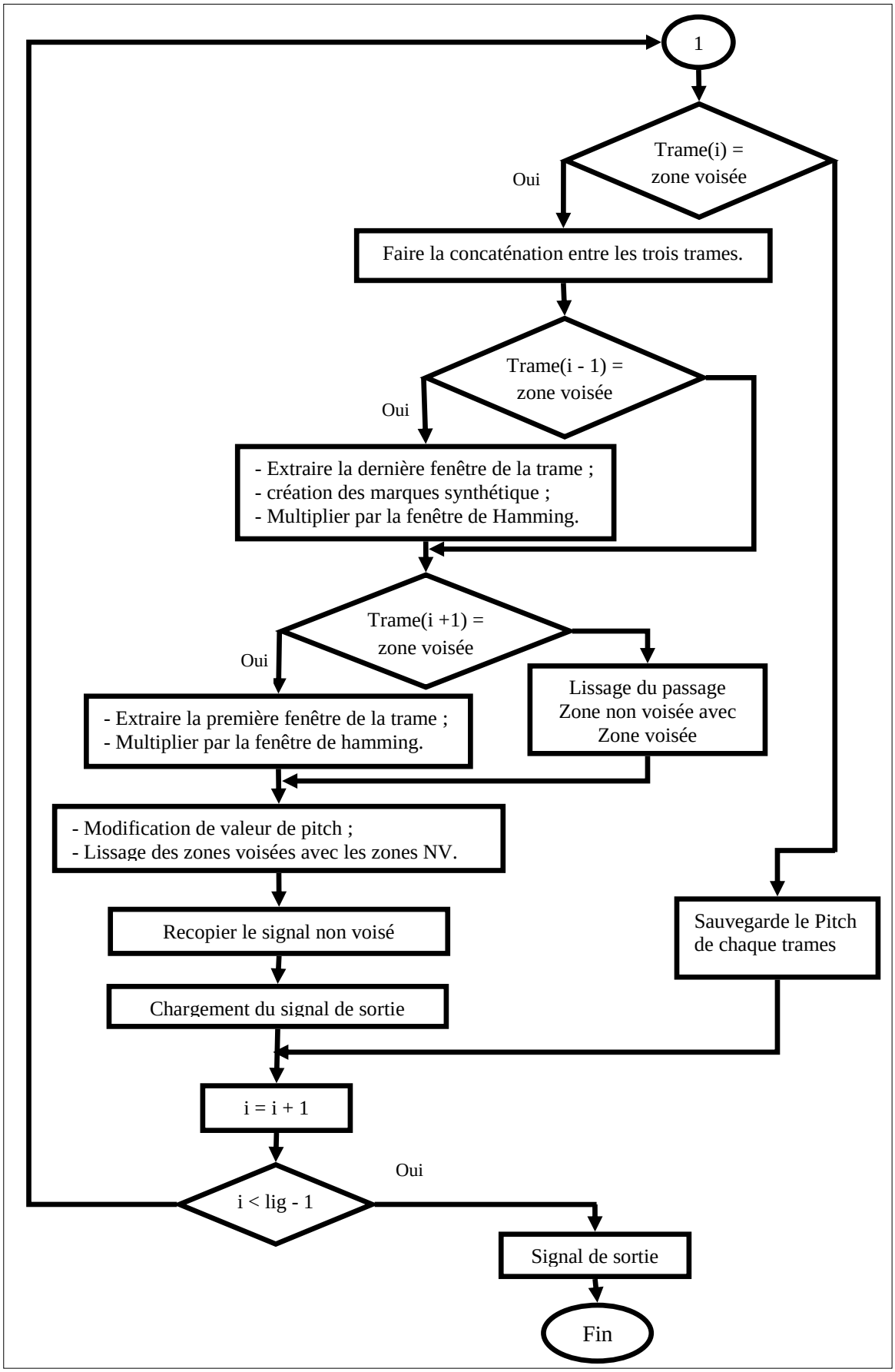


Figure4.8 : Calcul du l'enveloppe de Hamming et modification du signal vocal.

La première condition Si (Détection des zones voisées) indique si la trame actuelle est voisée ou non, elle permet si la condition est réalisée de concaténer les trames adjacentes.

La deuxième condition Si permet d'extraire la dernière fenêtre de la trame précédente.

La troisième condition Si permet d'extraire la première fenêtre de la trame suivante ainsi que les fenêtres de la trame actuelle qui dépendent du nombre de périodes pitch contenues dans la trame actuelle. La trame sur laquelle on effectue le traitement étant la trame centrale comme indiqué plus haut.

Dans cette dernière partie du programme on applique la méthode TD-PSOLA en utilisant en plus une fenêtre de Hamming ou la partie qui se chevauche entre les deux signaux à court terme sont additionnées, Les trames non voisées ne sont pas modifiées elles sont simplement recopiées sur le signal synthétique.

4.4. Résultats de la simulation TD PSOLA

Nous présentons dans ce paragraphe, des tests réalisés sur le signal de parole de notre corpus pour l'obtention d'un signal synthétique avec modification prosodique pour trois facteurs de modifications différents, le facteur 1.5 (facteur supérieur à 1), le facteur 0.5 (facteur inférieur à 1) et enfin pour un facteur égal à 1 qui normalement donne le même signal de départ.

4.4.1. Modification de la fréquence fondamentale

Le facteur de modification égal à 0.5 correspond à un abaissement de la F_0 c'est à dire à une augmentation de la période pitch, et le facteur égal à 1.5 correspond à une diminution de la période du pitch, c'est à dire une augmentation de la F_0.

La figure ci-dessous donne une représentation temporelle du signal décrivant un intervalle temporel du premier phonème [a] extrait du signal original à partir du mot **« Kataba»**, Ce dernier est superposé à un autre signal obtenu par l'algorithme TD PSOLA au même intervalle temporel du même phonème mais possédant une fréquence fondamentale multipliée par un facteur égal à 0.5 par rapport à la fréquence du signal original.

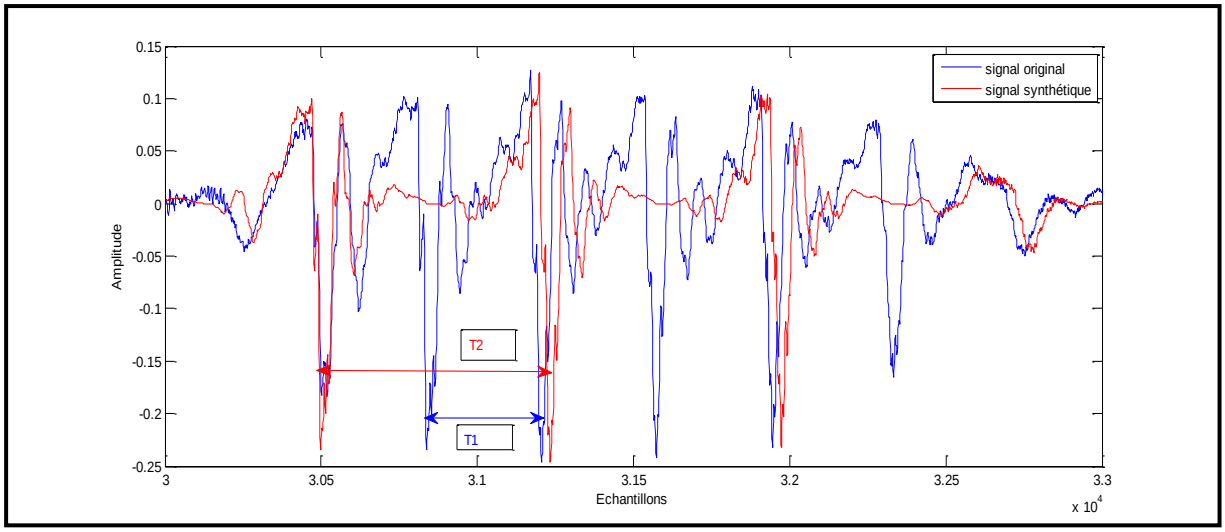


Figure 4.9 : Représentation des signaux original et synthétique pour un facteur de modification de 0,5.

De même, si nous voulons obtenir un signal synthétique d'une période fondamentale inferieur à celle d'origine, il suffit de préciser la valeur du facteur de modification et d'appliquer l'algorithme TD-PSOLA. La figure 4.10 donne la représentation temporelle du signal analytique et synthétique pour le même intervalle temporel et avec un facteur de modification égal à 1,5.

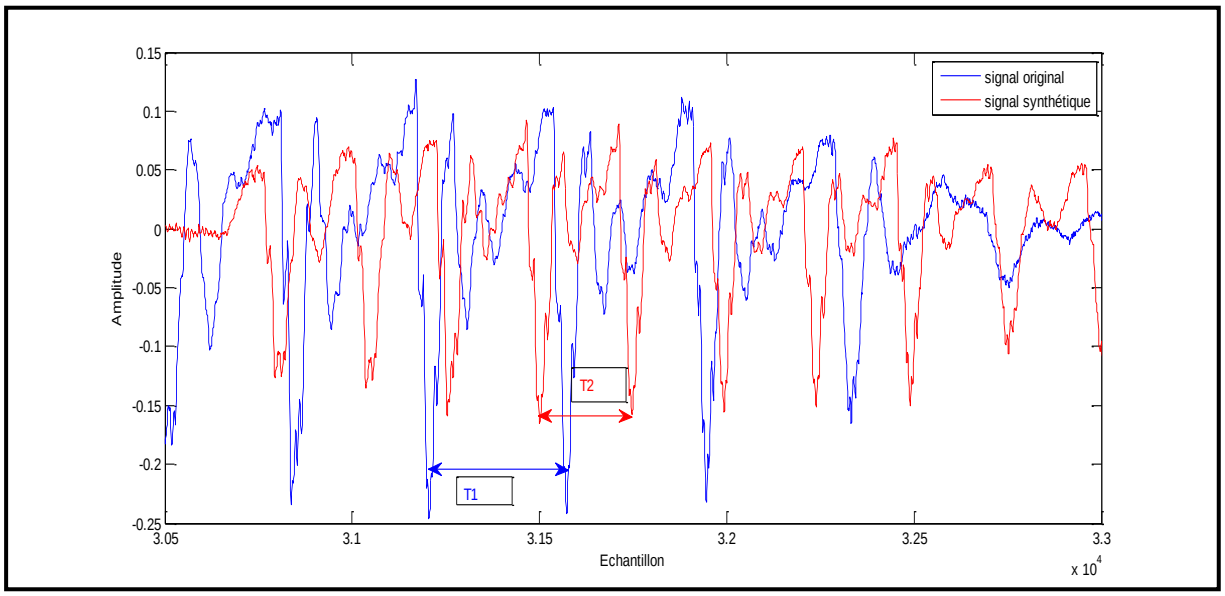


Figure 4.10 : Représentation des signaux original et synthétique pour un facteur de modification de 1,5.

Nous remarquons que le signal reconstitué a subit une distorsion d'amplitude. Cela est dû essentiellement au positionnement des marques proposées sur le signal original qui ne sont pas très précises pour pouvoir déterminer rentablement les périodes de pitch et de vérifier toutes les conditions de marquage décrites au chapitre précèdent.

Pour tester l'efficacité de l'algorithme et vérifier la validité des approximations effectuées, nous sommes amenés à expérimenter l'algorithme avec un facteur de modification égal à 1, qui normalement reproduit le signal sans changement et préserve ses caractéristiques de départ (Figures 4.11 et 4.12).

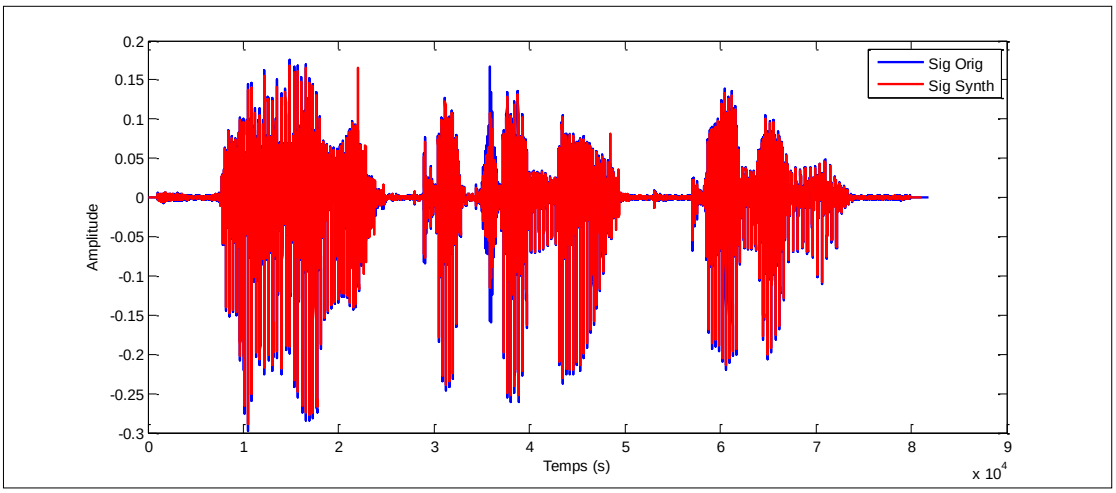


Figure 4.11 : Représentation des signaux original et synthétique pour un facteur de modification de «1».

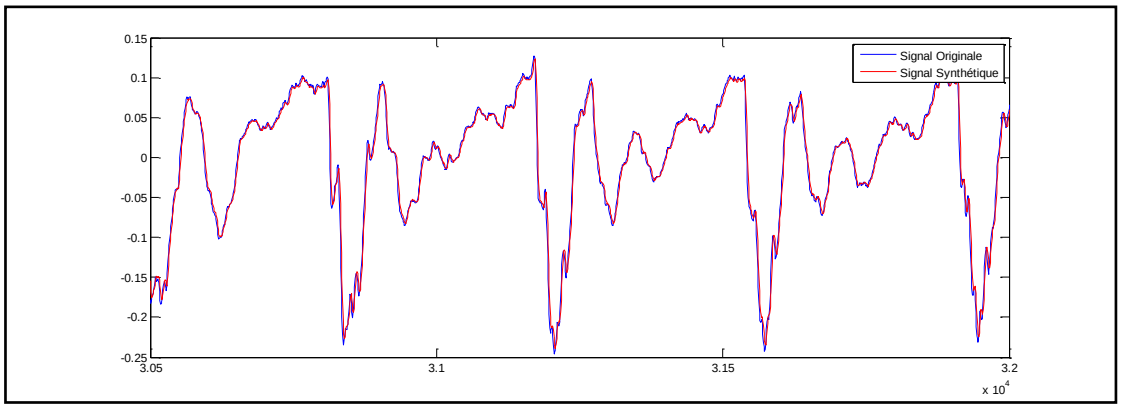


Figure 4.12 : Représentation des signaux original et synthétique pour un Facteur de Modification de «1».

4.4.2. Comparaison des spectrogrammes des signaux modifiés

La figure 4.14 représente une comparaison entre les spectrogrammes des deux signaux modifiés soit par un facteur de 0.5 ou 1.5 et le spectrogramme du signal original.

Les résultats des spectrogrammes sont obtenus à l'aide de logiciel PRAAT, En effet, le logiciel dispose d'une fenêtre principale permettant d'effectuer des traitements importants et

de fenêtres annexes affichant le résultat des divers traitements comme la stylisation de F0 ou le tracé du spectre.

PRAAT a été développé par Paul Boersma et par David Weenink de l'Institut de Phonétique d'Amsterdam. Il permet de mener des analyses phonétiques, de manipuler des données (analyses statistiques, construction de grammaires, etc.). Avec ce logiciel, il est possible (figure 4.13) [6] :

- d'enregistrer des fichiers audio qui pourront ensuite être analysés ;
- de transcrire, d'étiqueter et de segmenter des données audio (enregistrements effectués sous PRAAT ou provenant d'autres fichiers, au format WAV, par exemple) ;
- d'effectuer des analyses phonétiques et acoustiques au niveau segmental (spectrogramme, analyse de formants, sonagramme, etc.) et au niveau suprasegmental (pitch, courbe de Fo, intensité et durée) …etc.

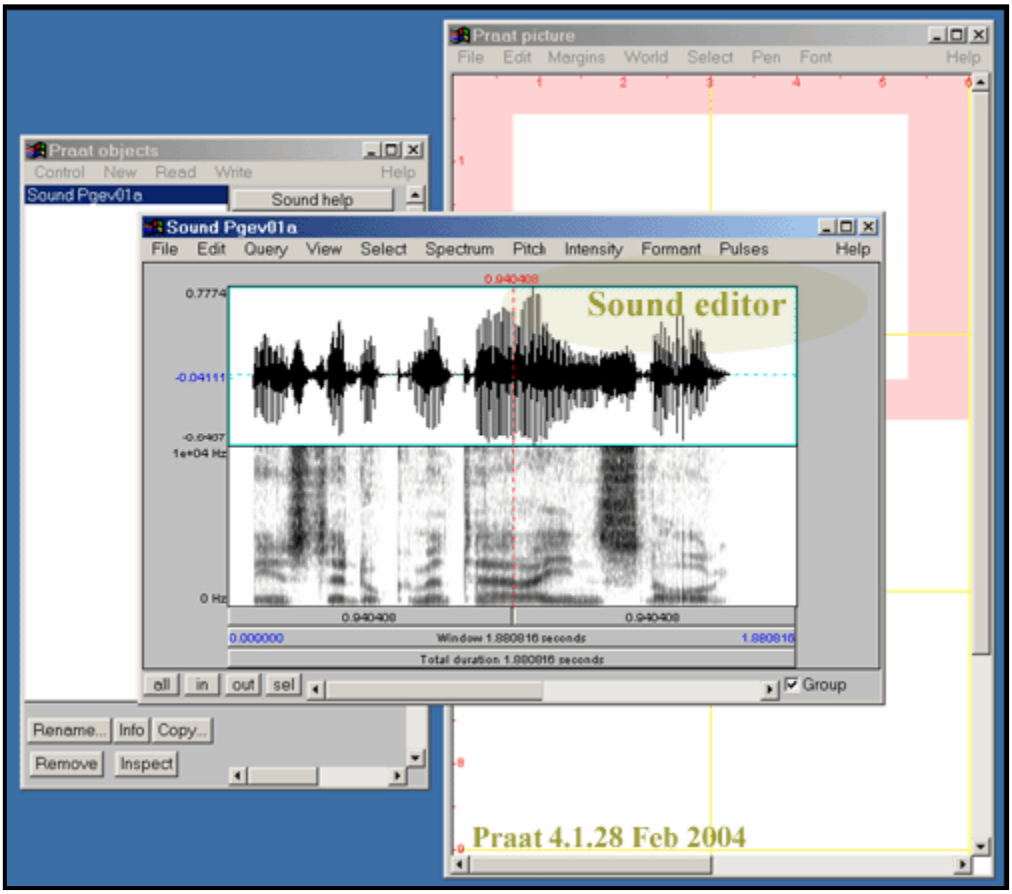


Figure 4.13 : Interface du logiciel PRAAT [7].

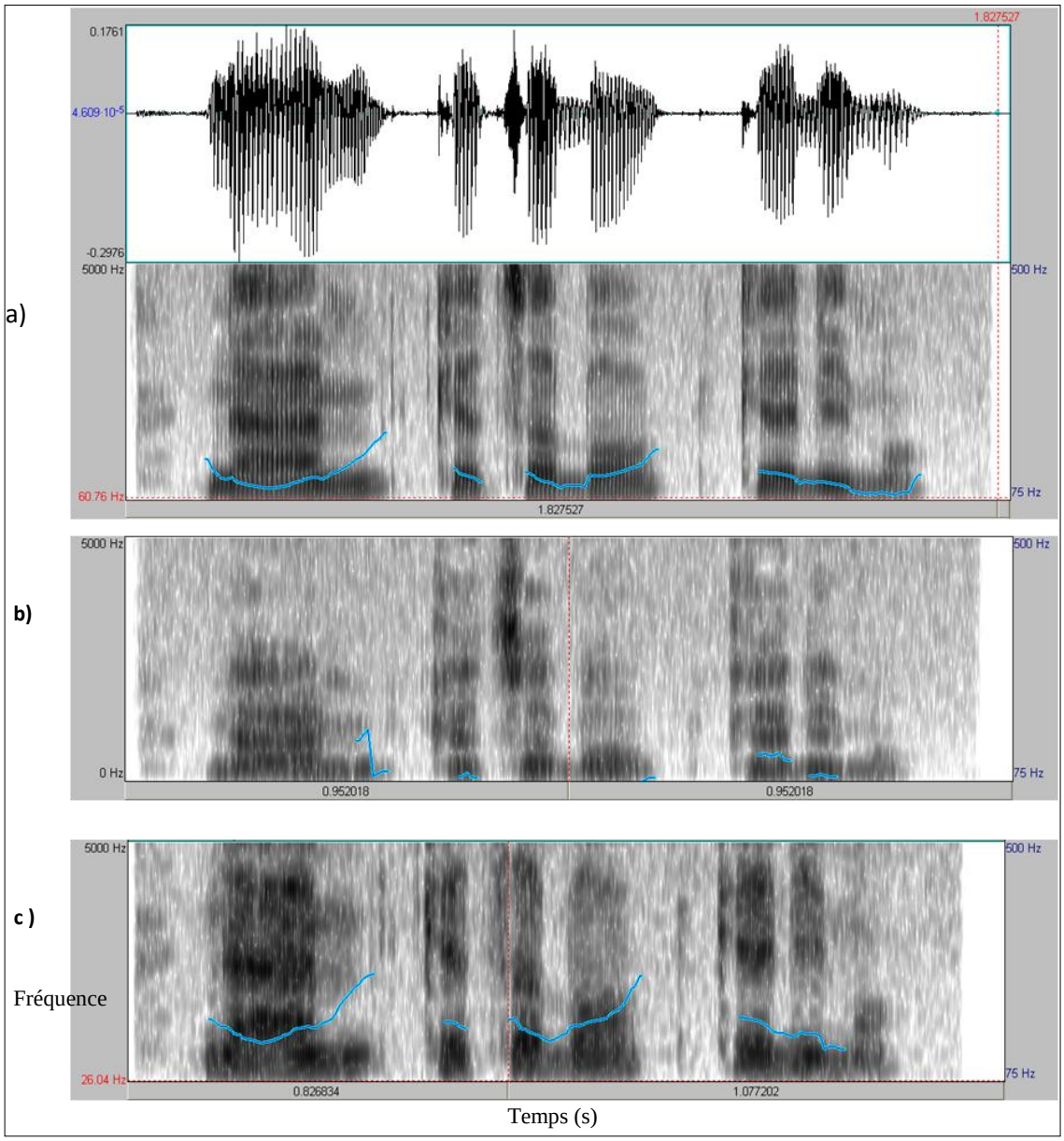


Figure 4.14 : comparaison des spectrogrammes des signaux modifiés :

- **a)** Signal original
- b) Signal synthétique avec FM= 0,5
- c) Signal synthétique avec FM= 1,5

Il est bien clair dans cette figure que la valeur et la trajectoire des formants sont maintenues le long du signal, ainsi nous pouvons conclure que nous avons pu aboutir à des signaux de fréquences fondamentales différentes (tracé en bleu) à partir d'un signal de référence, en maintenant l'enveloppe spectrale inchangée et en préservant ainsi le timbre de la voix.

4.4.3. Modification de la durée du signal

Nous présentons dans cette partie, des tests réalisés sur le signal de parole de notre corpus pour l'obtention d'un signal synthétique avec modification de l'axe de temps pour trois facteurs de modifications différents, le facteur 1.5 (facteur supérieur à 1), le facteur 0.5 (facteur inférieur à 1) et enfin pour un facteur égal à1 qui normalement donne le même signal de départ.

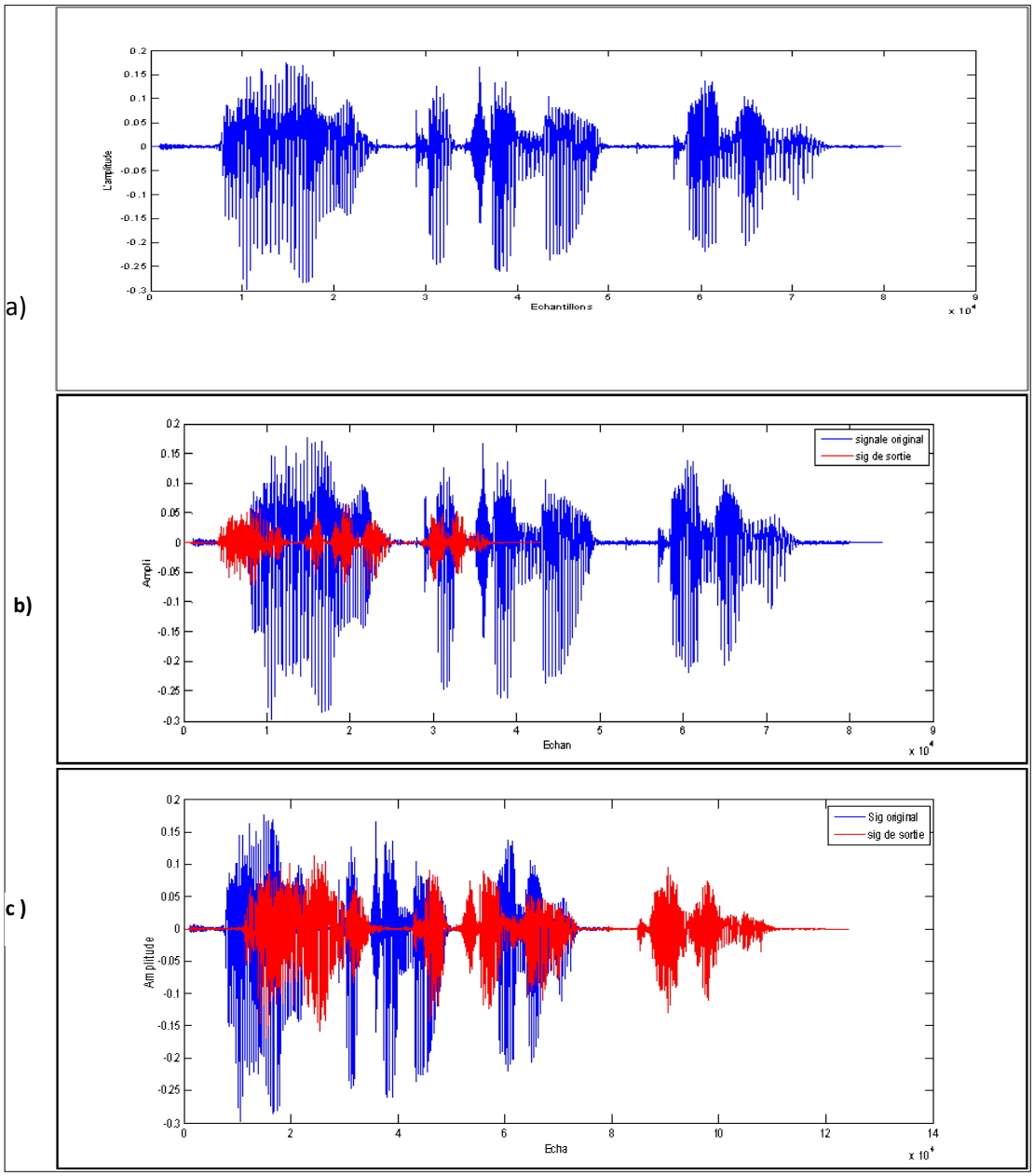


Figure 4.15 : Comparaison des représentations des signaux modifiés

a) Signal original ;
b) Signal synthétique avec Facteur de Modification = 0.5 ;
c) Signal synthétique avec Facteur de Modification = 1.5.

4.4.4. Comparaison des spectrogrammes des signaux modifiés

La figure 4.16 représente une comparaison entre les spectrogrammes des deux signaux synthétiques modifiés avec le signal original.

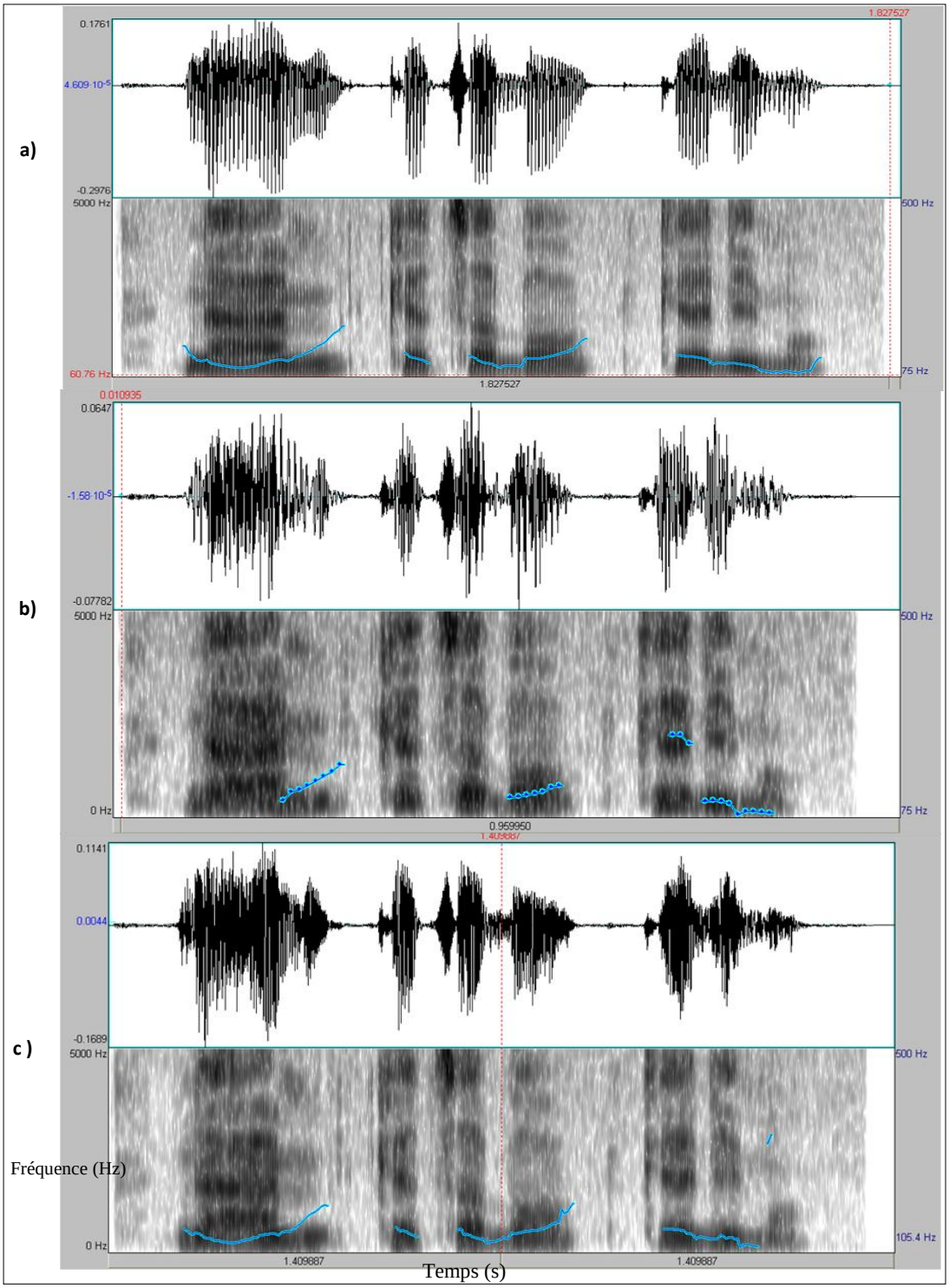


Figure 4.16 : comparaison des spectrogrammes des signaux modifiés.

a) Signal original ;
b) Signal synthétique avec Facteur de Modification = 0.5 ;
c) Signal synthétique avec Facteur de Modification = 1.5.

4.5. Evaluation de la technique

Vu l'importance de l'évaluation dans le développement des systèmes de synthèse de la parole à partir du texte (TTS), nous allons essayer dans ce paragraphe de donner une idée très simple et sommaire sur la manière d'évaluation des systèmes ou plus précisément sur les techniques de synthèse utilisées, puisque les progrès de l'évolution, de l'analyse de la qualité des voix de synthèse, sont fortement liés à ceux des algorithmes de synthèse.

- Dans un premier temps, nous allons donner une appréciation sur la qualité globale des deux techniques de synthèse employées en se basant sur des tests d'écoute qui visent à juger l'intelligibilité et la qualité du message parlé obtenu à la sortie de notre système sans se préoccuper de son fonctionnement interne et sans chercher la source des défauts éventuels ;
- Le second test vise à estimer l'aptitude des techniques à effectuer des modifications tout en conservant l'intelligibilité et la qualité de la parole synthétisée.

Dans notre cas, nous avons utilisé une technique temporelle qui est souhaitable pour l'analyse de la micro mélodie, elle est basée sur le calcul de la fonction d'autocorrelation. TD-PSOLA n'est pas à proprement parlée une technique de synthèse du signal de la parole. Il s'agit d'une technique de traitement du signal de parole - qu'il soit naturel ou de synthèse - dont l'objectif est de modifier les paramètres prosodiques de celui-ci. Si maintenant on la considère comme une technique de synthèse de parole (si le facteur de modification de F_0 est égal à 1) ; elle va sans aucun doute avoir les meilleurs échelles puisque c'est une reproduction excellente de signal naturel. L'évaluation est résumée comme suite :

D'une façon globale, c'une étape préliminaire pour tout système d'analyse - synthèse - modifications prosodiques est la détection du fondamental qui peut être dans le cas d'une erreur de détection, la source des différentes dégradations éventuelles qui peut surgir et affecter le signal synthétisé.

Le problème se trouve dans le cas de la modification de F_0 où la dégradation de la qualité du message vocal obtenu lors d'une modification est remarquable et augmente en élevant ou en diminuant suffisamment le facteur de modifications. Les problèmes ou les raisons de cette dégradation peuvent être résumés comme suit :

- Erreurs de détection du voisement : Il existe plusieurs méthodes de décision du voisement qui ont en commun d'analyser un intervalle de parole précis pour assurer la

stationnarité de ces segments. C'est à partir de cette décision que nous allons donner une estimation de la période du fondamentale ;

- Erreurs dues à l'estimation de la fréquence fondamentale : l'approximation faite sur la valeur du pitch par la fonction d'autocorrélation ne permet pas d'estimer les fréquences des trames du signal avec une grande précision.

Il est pratiquement remarqué dés qu'on est en dehors d'une extrémité maximale d'un facteur de modification qui est égal à 2.5 et une extrémité minimale de 0.4 le message vocal obtenu par la TD-PSOLA a tendance de devenir de plus en plus robotique alors une perte du naturel. Donc c'est une technique de base pour la plupart des techniques utilisées à présent et qui donne des résultats satisfaisants avec moins de complexité. D'une façon générale, il est difficile d'évaluer une technique puisque chaque technique présente des avantages et des inconvénients.

4.6. Conclusion

Nous avons pu concevoir dans ce chapitre une méthode de modifications prosodiques qui se démarquent fortement des autres par le contrôle des divers paramètres qui définissent le timbre de la voix. Son principe est basé sur la ré-harmonisation spectrale ; nommé TD-PSOLA et qui appartient aux familles des synthétiseurs acoustiques dans le domaine temporel. L'Algorithme TD-PSOLA peut donc être utilisé pour changer le pitch d'un signal vocal et de préserver son format. En préservant le format du signal, nous sommes effectivement préserver l'identité vocale. .

Conclusions Générales et Perspectives

Conclusions générales et perspectives

Les simulations et les testes que nous avons réalisées tout au long de ce Mémoire de Magister nous ont permis d'avoir des notions sur le traitement de la parole. Ils nous ont permis également de comprendre le fonctionnement d'un système de synthèse de la parole. Il est logique de chercher à modifier la durée et la fréquence fondamentale du signal sans faire appel à un modèle de représentation, c'est-à-dire en utilisant des techniques non paramétriques. Parmi ces techniques de modifications prosodiques étudiées jusqu'à présent, la plus efficace est l'algorithme PSOLA. Nous nous sommes concentrés sur une variante dénommée la technique TD-PSOLA. C'est une technique de traitement du signal de parole dont l'objectif est de modifier les paramètres prosodiques de celui-ci. Si maintenant on la considère comme une technique de synthèse de parole (si le facteur de modification de F_0 est égal à 1) ; elle va avoir des résultats meilleurs car c'est une reproduction excellente de signal original (d'après des tests).

La caractéristique la plus remarquable de la technique TD-PSOLA est qu'elle opère directement sur la forme d'onde du signal de parole. L'idée de base est d'extraire du signal des grains de sons élémentaires, représentant les caractéristiques locales du signal, et de jouer avec ces grains élémentaires pour réaliser les modifications désirées. L'analyse par TD-PSOLA pour changer le pitch est identique à l'analyse pour étirer le temps, la différence est visible dans la partie synthèse où, au lieu d'ajouter ou de retirer des segments et donc d'étirement du temps, donc préserver la durée du signal tout en changeant son pitch. La méthode décrite dans le présent travail offre un outil de base pour la manipulation de la tonalité, et en raison de sa faible complexité de calcul, elle est un outil efficace pour le traitement des signaux en temps réel.

Les résultats obtenus, par le biais de cette analyse, sont motivants car nous ont permis d'obtenir des meilleurs résultats de synthèse et de modifications prosodiques par plusieurs facteurs appliquées sur la langue Arabe Standard du corpus choisi. Ainsi, nous avons déduit que les dits résultats sont acceptables et satisfaisants.

Comme suite à ce travail, il serait très intéressant de faire une étude comparative de toutes les variantes de la technique PSOLA telle que la FD-PSOLA et la LP-PSOLA, etc. et de tester leurs performances afin de montrer la bonne qualité de synthèse et de modifications prosodiques.

Références bibliographiques

Références bibliographiques

[1] R. Benslimane, Transformation de voix en temps réel, Département de Traitement Du Signal, université de la Marne la Vallée, France, juillet 2000.

[2] www.ircam.fr/ equipes/ analyse synthése / tassart, 1998-1999.

[3] Z.A Benslama, Pathologie du Langage Parlé Arabe : Cas des Sigmatismes Occlusifs et Constrictifs, These de doctorat en Electronique, Ecole Nationale Polytechnique, Alger, Algérie, 15 / 12 / 2007.

[4] H. Tebbi, Transcription Orthographique Phonétique vue de la synthèse de la parole à partir du texte en l'Arabe Standard, Mémoire de Magister Spécialité : Ingénierie des systèmes et des connaissances, USD-Blida, Juin 2007.

[5] R. Boite, H. Bourlard, T. Dutoit, Traitement de la parole, Collection électronique, Presses Polytechniques et Université Romandes, 1999.

[6] M. Aissiou, Application des Algorithmes génétiques en vue de la Reconnaissance Automatique des voyelles de l'Arabe Standard, Mémoire de Magister, CRSTDLA, Alger, Algérie, Février 2004.

[7] A. Chentir, Etude de la Microprosodie en vue de la Synthèse de la parole en Arabe Standard, Thèse de Doctorat en Electronique, Ecole Nationale Polytechnique, Alger, Algérie, 01 Octobre 2009.

[8] G. Droua-Hamdani, Prédiction de la durée segmentale des phonèmes de l'Arabe Standard , Mémoire de Magister, CRSTDLA, Alger, Algérie, Février 2004.

[9] M. Kabache, Application des Réseaux de Neurones à la Reconnaissance Automatique des phonèmes spécifiques en Arabe Standard, Mémoire de Magister, CRSTDLA, Alger, Algérie, Mai 2005.

[10] P. Yves Le Meur, Synthèse de la parole par unités de taille variable, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, Paris, France, 1996.

[11] Calliope, La parole et son traitement automatique, Collection Techniques et Scientifiques des Télécommunications. Préface de G. Fant, CNET/ENST, Ed. Masson, 1989.

[12] T. Dutoit, Introduction au traitement automatique de la parole notes de cours /DEC2, Collection électronique, Faculté Polytechnique de Mons, 2000.

[13] K. Benbellil, Synthèse par polysons de l'Arabe Standard, Mémoire de Magister Spécialité : Electronique acoustique et physiologique de la parole, Septembre 2005.

[14] J. Farina, La prosodie pour l'identification des langues, Cours Doctorale en Informatique, Université Sabatier & Inpt par Pr R.Caubet, France, 1998.

[15] T. Dutoit, Je parle donc je suis, Un bilan des développements récents en Traitement Automatique de la Parole, Faculté Polytechniques de Mons.

[16] J. wiley & S,Baffine Lane, DAFX Digital Audio Effects, Copyright 2002. www.wiley.co.uk

[17] S. Baloul, Développement d'un système automatique de synthèse de la parole à partir du texte Arabe Standard voyellé , Thèse de Doctorat d'université, Le Mans, France, 27 Mai 2003.

[18] http://www.bibliotheque.refer.org/ html/ parole/ sorin/ sorin.htm.

[19] L. Rabiner, R.W. Schaffer, Digital Processing of Speech Signal, Prentice Hall Engelwood Cliffs, New Jersey 07632, 1978.

[20] V. Collotte, Y. Laprie, Amélioration de la précision de la resynthèse avec TD-PSOLA, Actes des XXIVémes JEP, Journées d'Etude sur la Parole, 377 – 380, Nancy, France, 2002.

[21] G. Peeters, Analyse et synthèse des sons musicaux par la méthode PSOLA, Actes des XXI^{émes} JIM, Journée d'Informatique Musicale, Agelonde, France, 302 – 306,1998.