République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Echahid Hamma Lakhdar El-Oued
Faculté de Technologie

Mémoire de Fin d'Étude

En vue de l'obtention du diplôme de

MASTER ACADEMIQUE

Domaine : Sciences et Technologie
Filière : Télécommunications
Spécialité : Systèmes de Télécommunications

Intitulée

Amélioration d'un synthétiseur de la parole par concaténation

Présentée par :

M^me^. DEBILOU Chaima
M^lle^. BOUDAOUD Samiha

Membres du jury :

Président :	Dr.TOUHAMI Redha
Examinateur :	Dr. KHELIL Abdellatif
Rapporteur :	Dr. LAIB Ismail

Juin 2019

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Echahid Hamma Lakhdar El-Oued
Faculté de Technologie

Mémoire de Fin d'Étude

En vue de l'obtention du diplôme de

MASTER ACADEMIQUE

Domaine : Sciences et Technologie
Filière : Télécommunications
Spécialité : Systèmes de Télécommunications

Intitulée

Amélioration d'un synthétiseur de la parole par concaténation

Présentée par :

M^me^. DEBILOU Chaima
M^lle^. BOUDAOUD Samiha

Membres du jury :

Président :	Dr.TOUHAMI Redha
Examinateur :	Dr. KHELIL Abdellatif
Rapporteur :	Dr. LAIB Ismail

Juin 2019

Remerciement

Tout d’abord nous remercions Dieu de nous avoir donné la force et le courage d’accomplir ce travail.

Nous remercions vivement notre encadreure **Mr laib Ismail** pour nous 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. Nous la remercions pour ses compétences, son ouverture d'esprit et sa grande disponibilité. Et aussi nous remercier vivement **M^{lle} BETTAYEB Nedjla** pour sa gentillesse, sa disponibilitè et sa contribution générale à l’èlaboration de ce travail.

Nous remercions les membres du jury, qui nous ont fait l’honneur de participer au jugement de ce travail.

Nous souhaitent ègalement la remercie tous mes proufisseurs de la facultè sience technologie spicialement la dipartement gènie èlèctrique branche tèlècome filaire et san fil. Et tous les enseignants qui ne sont pas mentionnés, pour tous les savoir qu’il nous ont donnè.

Et puis un remerciement trè trè challeureux a **mes parents** , mes frères , mes amis et tous ceux, qui de près ou de loin, nous ont apportés leur contribution pour la réalisation de ce travail.

ملخص:

تركيب الكلام يعني محادثة نص (كتابي او رقمي) بصوت واحد . مع التطور التكنولوجي الحالي, يمكن دمج هذا النظام في العديد من المجالات مثل أنظمة المساعدة للمعاقين وأجهزة التحدث و ما إلى ذلك.

الهدف من عملنا هذا هو تطوير نظام تركيب الكلام في اللغة العربية الفصحى بالتسلسل وتحسينه بواسطة واحدة من التقنيات المناسبة

من بين الأساليب المستخدمة يعتمد التركيب على تقاطع شرائح الصوت الطبيعي في تطبيقات المفردات المحدودة (الجمل).

لكن في تطبيقات الإستخدام العامة فإنه يطرح مشاكل, خاصة في نقاط التسلسل ونسمع تخفيضات في الصوت المركب لهذا يتم تطبيق العديد من الخوارزميات و تقنيات الصقل للتغلب على هذا المشكل , مثل خوارزمية PSOLA ومتغيراتها

الكلمات المفتاحية : التركيب الآلي للكلام ,التلصيق , تقنيات الصقل ,**PSOLA**

Résumé :

La synthèse de parole signifie la conversation d'un texte (écrit ou numérique) en une voix. Avec le développent technologique de nos jours ce système peut s'intégrer dans plusieurs domaine comme des systèmes d'aide pour les handicaps, des appareilles parlante, etc.

Le but de notre travaille est d'élaborer un système de synthèse de la parole en Arabe Standard par concaténation et l'amélioré par l'une des techniques approprié. Parmi les méthodes de synthèse utilisé, la synthèse par concaténation base sur la juxtaposition des segments de son naturelle. Dans les applications à vocabulaire limité (les phrase a prononcé sont limité). Mais dans les applications d'utilisation générale elle pose des problèmes surtout dans les points de concaténation, ou nous entendons des coupures au niveau du son synthétisé.

Pour cela plusieurs algorithmes et technique de lissage sont applique pour surmonter cette limitation comme l'algorithme PSOLA et ses variantes.

Mots clé : synthèse de parole, concaténation, techniques de lissage, PSOLA.

Abstract :

The synthesis of word means the conversation of text (written or numerical) in a voice. With develop technological this system nowadays can be integrated in several field like systems of assistance for the handicaps, of install speaking…, etc

The goal of our work is to develop a speech synthessis system in arabic standard by concatination and imoroved by one of the appropriate techniques. Among the synthètic mèthode used , concatination synthise is based of the juxtaposition of segments of naturelle sound. In limited vocabulairy applications (pronanced sentences ara limited). But in application gènèral use it posse problme specially in the concatination pionts , or we hear cuts in the synthèsized sound.

For this several algorithms and smoothing technique are applied to overcome this limitation like algorithm PSOLA and its variants.

Key words: synthesis of word, concatenation, techniques of smoothing, PSOLA.

Sommaire

Liste des Abréviations

TAP	**T**raitement **A**utomatique de la **P**arole
CT	**C**ourt-**T**erme
AS	**A**rabe **S**tandard
VOD	**V**oice **O**pèration **D**èmonstration
TTS	**T**ext-**T**o-**S**peech (Un Système de Synthèse à Partir du Texte)
OCR	**O**ptical **C**haracter **R**ecgintion
LPC	**L**inear **P**redictive **C**oding
AR	**A**uto **R**égressif
ARMA	**A**uto **R**égressif à **M**oyenne **A**justée
MA	**M**oyenne **A**justée
FFT	**F**ast **F**ourier **T**ransform
RAP	**R**econnaissance **A**utomatique de la **Pa**role
TFR	**T**ransformée de **F**ourier **R**apide
PSOLA	**P**itch **S**ynchronous **O**ver**L**app and **A**dd
TD PSOLA	**T**ime **D**omaine **P**itch **S**ynchronous **O**ver**L**app and **A**dd
TDHS	**T**ime **D**omain **H**armonic **S**caling
WSOLA	**W**aveform **S**imilarity **O**ver**L**ap-**A**dd
SPR	**S**ynthèse **P**ar **R**ègle

Liste des Figures

Liste des Tableaux

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.

Tout système de synthèse de parole à partir du texte, est amené à répondre de manière plus ou moins précise et développée selon la qualité et la finalité du système, représentés par trois problèmes de natures différentes : il s'agit dans un premier temps d'analyser et de structurer le texte afin de déterminer un mode de prononciation cohérent ; par la suite, le texte analysé doit être transformé en une suite de sons de parole accompagnée d'indications concernant leur agencement ; enfin, il faut générer un signal acoustique, cette suite de sons tout en possédant les caractéristiques apparentes de la parole naturelle.

Le but de notre travail qui s'inscrit dans le domaine de Traitement Automatique de la Parole, en particulier la synthèse de la parole est d'élaborer un système de synthèse de la parole et d'effectuer des modifications prosodiques de signal vocal en utilisant la technique 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. Nous avons structuré notre travail en quatre chapitres :

- Dans le premier, nous allons décrire d'une manière générale des notions sur le traitement de la parole ainsi que sa production, les appareils phonatoires et auditifs de l'être humain, des spécifications du signal vocal et des notions fondamentales sur l'Arabe Standard ;
- Le deuxième, nous donne une brève définition de la synthèse de la parole, et du traitement linguistique du texte, En outre, nous étudions les différentes

techniques d'analyse du signal vocal. Puis nous expliquons les méthodes de la synthèse de la parole ainsi que ses différentes applications.

- Dans le troisième chapitre, nous nous intéressons à faire une analyse acoustique de notre corpus en étudiant les caractéristiques et les paramètres pertinents de ce signal vocal (fréquence fondamentale, formants et intensité). Nous introduisons les étapes de l'élaboration de notre corpus et son traitement. Nous expliquons le logiciel Praat et finissons par une étude comparative pour quelques signaux vocaux avant et après la concaténation des unités sonores.

- Dans 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

Chapitre 1

Généralités sur la parole

1.1 Introduction

La parole est un atout que seul nous, êtres humains, possédons dans tout le monde animal. La générer naturellement résulte d'une combinaison complexe de phénomènes physiques et d'interprétations psycho acoustiques. Donc, Le but de ce chapitre est de présenter le mécanisme de la production de la parole puis nous allons décrire de manière générale des notions sur des notions fondamentales sur la langue arabe [1].

1.2 Définition de la parole

La parole est une succession de séquences sonores et de silences, et 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, et il est le résultat d'une variation de la pression produite par l'émission d'un son par un locuteur [2].

1.3 La production de la parole

La production de la parole est l'opération la plus complexe de l'activité biologique humaineest un système dynamique, dont le comportement à un moment donné dépend de ses états antérieurs. Le système est donc dépendant d'une variable paramétrable.

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.

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 : Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée artère, se trouve au sommet

supérieur de trachée-artère, où lapression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles. 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 progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer [3].

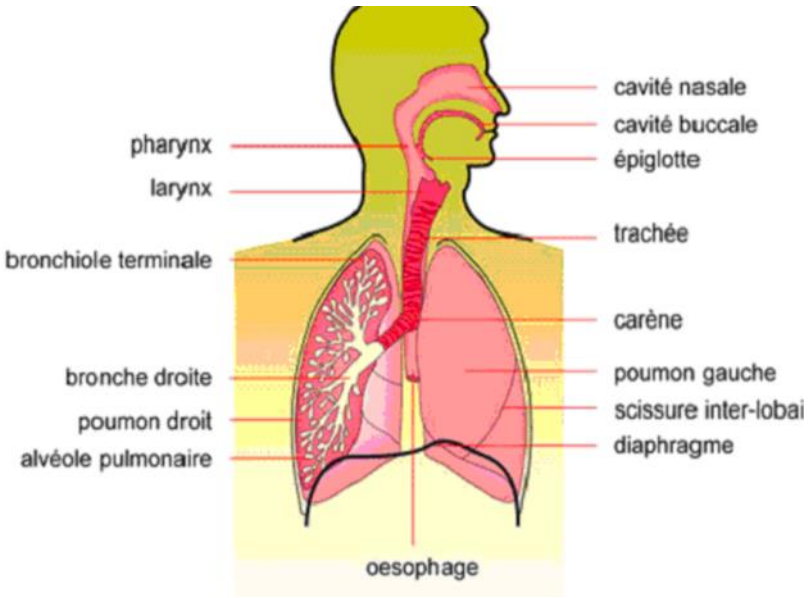

Figure 1.1 : Modèle simplifié de l'appareil phonatoire

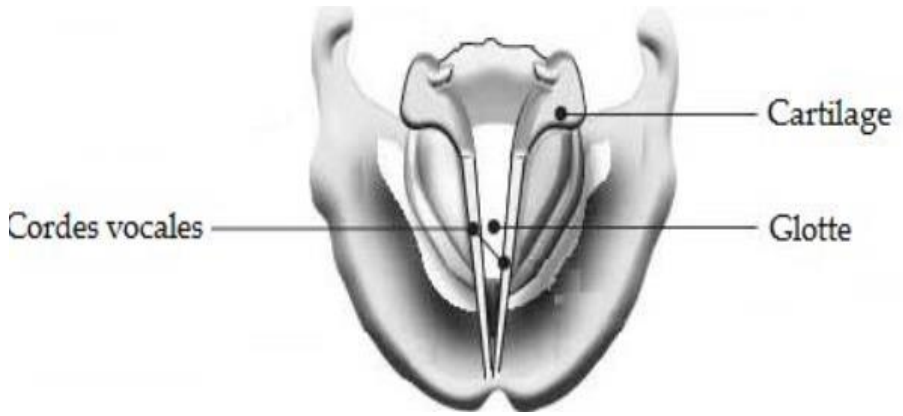

Figure 1.2:Le larynx

1.4Analyse 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 il y a Plusieurs techniques d'analyse sont utilisées parmi les quelles on peut prend l'analyse par les spectrogrammes [4].

1.4.1 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).

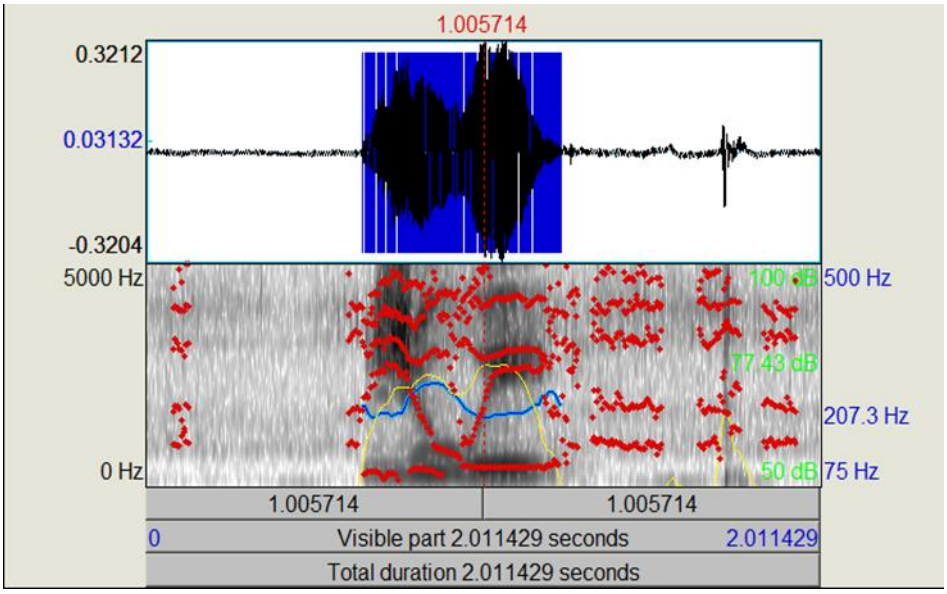

Figure 1.3 : Spectrogramme du mot 'jouer'

1.5 Les paramètres prosodiques d'un signal vocal par spectrogramme :

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

1.5.1La Fréquence Fondamentale F0 :

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. Les variations de la fréquence au cours de la parole constituent cequ'on appelle l'intonation.

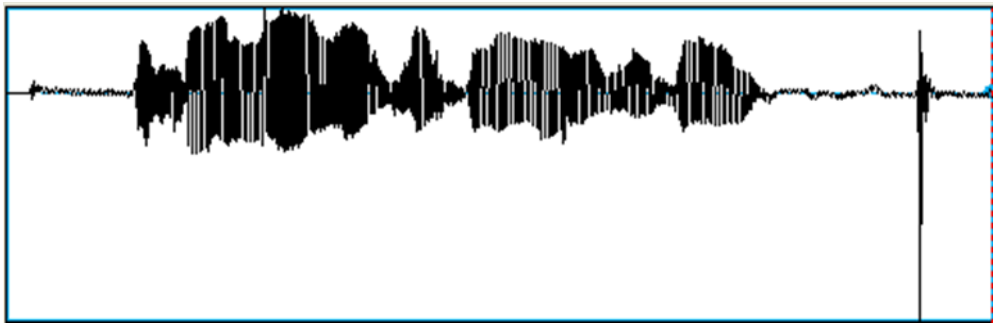

Figure 1.4 : Évolution de la fréquence de vibrations des cordes vocales de la phrase : "بسم الله الرحمان الرحيم"

1.5.2La durée:

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. Elle représente le temps de la prononciation d'un phonème.

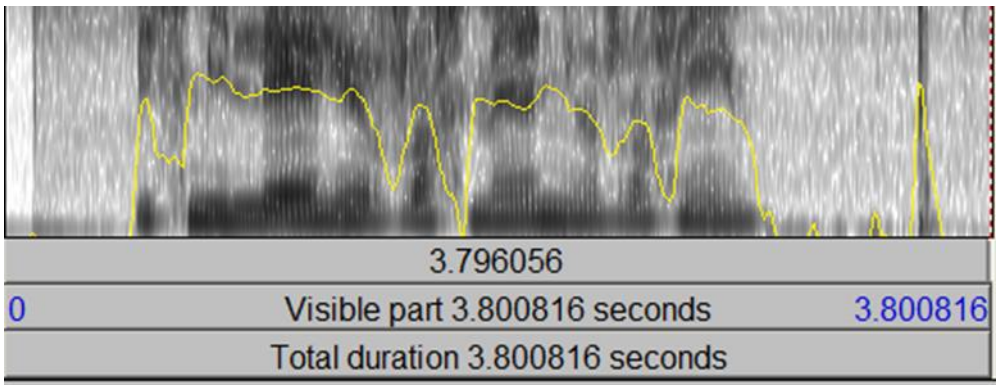

Figure 1.5 : L'évolution temporelle du signal vocal pour "بسم الله الرحمان الرحيم"

1.5.3. L'Intensité :

L'Intensité elle exprime le volume sonore d'un phonème et dans le cas d'un voisement elle représente l'amplitude des vibrations des cordes vocales.

1.5.4. Formants

Les formants sont des zones fréquentielles de forte énergie, correspondent à une résonancedans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants représentent les maximas de la courbe de réponse en fréquences du conduit vocal.Chaque son à ses formants caractéristiques.

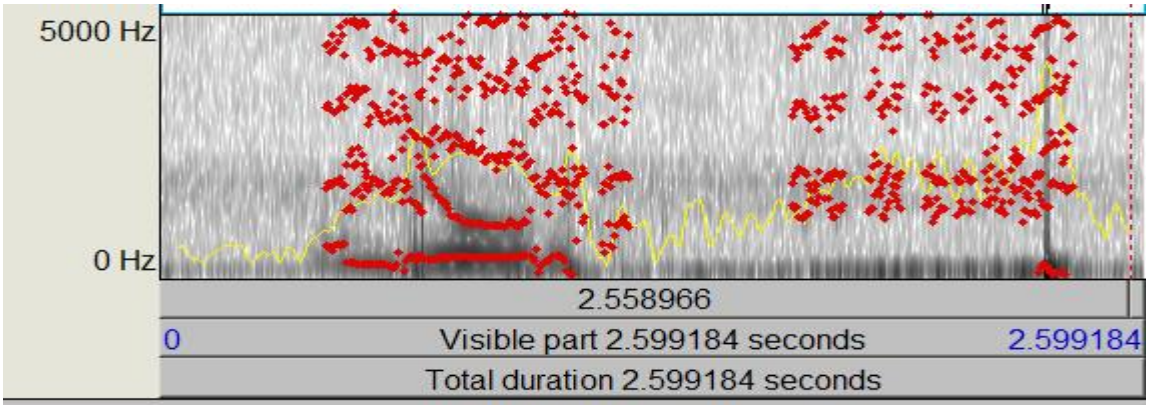

Figure 1.6 :L'Intensité et les formants de parole du mot 'bonjour'

1.6 Propriétés spécifiques du signal vocal

1.6.1. 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.

1.6.2. 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).

1.6.3. 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.

1.6.4. Le codage :

Le codage concerne les niveaux lexicaux, syntaxiques, sémantiques, morphologiqueset phonétiques (phonèmes et leurs interactions) utilisés souvent pour assurer une meilleure qualité de la parole synthétique [5].

1.7. Classification des sons du langage :

La production des sons ou d'un mot réside dans la production en série de tous les phonèmes constituant ce mot. Ces phonèmes forment les unités phonétiques qui sont classées en voyelles, consonnes et semi-voyelles. Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode et lieu d'articulation. Dans la cavité buccale, le point d'articulation est l'endroit où se trouve un obstacle au passage de l'air. D'une manière générale, le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air [6].

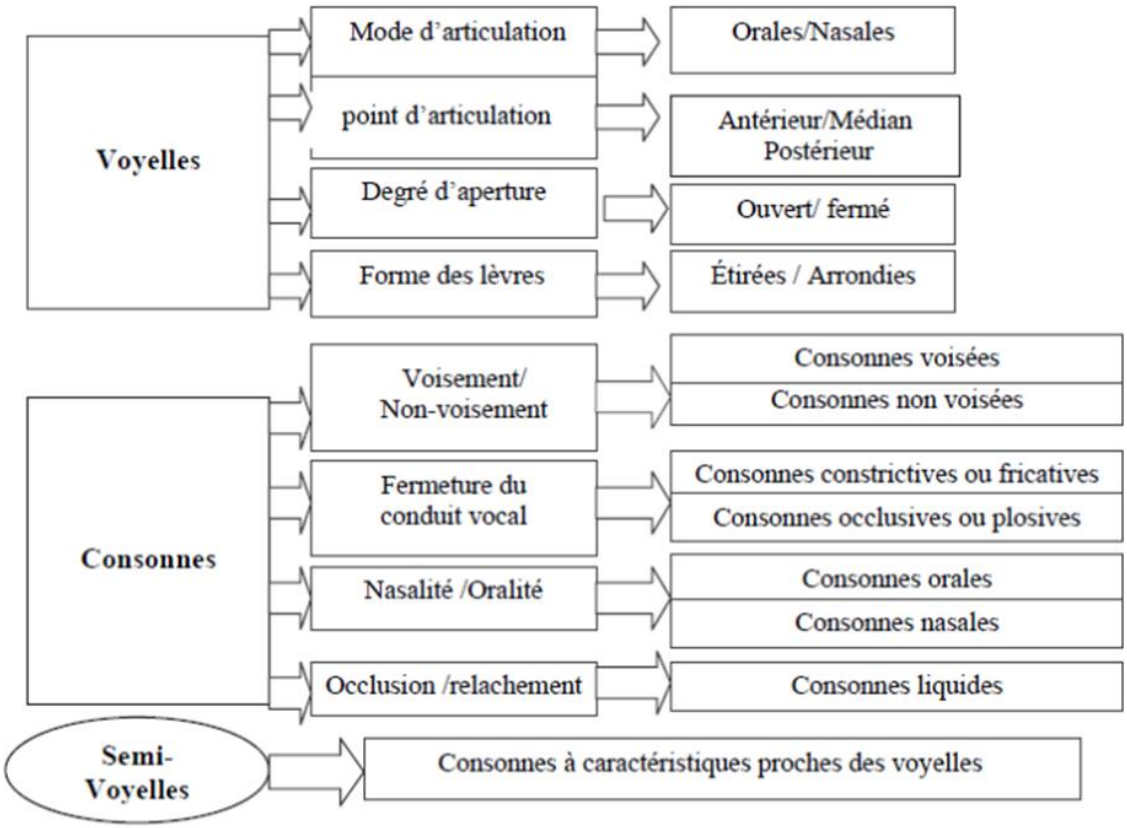

Tableau 1.1 : Classification des sons du langage

1.7.1. Sons voisés :

Les sons voisés, tels que les voyelles, semi-voyelles et les consonnes nasales, sont produits par le passage de l'air des poumons à travers la trachée qui met en vibration les cordes vocales. Ce mode, qui représente 80% du temps de phonation, est caractérisé en général par une quasi-périodicité et une énergie élevée.

1.7.2. Les sons non voisés :

Le second mode d'excitation est obtenu par divers bruits produits par le passage de l'air en un point de resserrement du canal vocal ou par des bruits d'occlusion, provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais.

1.7.3. Les voyelles :

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Elles sont toutes voisées (sonores) car produites par la vibration des cordes vocales qui donne naissance au flux laryngé. Ce courant d'air ne rencontre aucun obstacle dans les cavités supra glottiques. Les différences de timbre entre les voyelles sont le résultat des modifications de forme, de volume, du nombre de résonateurs, des mouvements de la langue,de la mâchoire, des lèvres de la luette, de vibrations engendrées par le passage de l'air au niveau des différents organes situés dans les cavités supra glottiques, constitués de tissus durs ou mous.

1.7.4. Les consonnes :

Les consonnes se caractérisent par une fermeture partielle du conduit vocal ou constriction (constrictives ou fricatives) ou totale du conduit vocal (occlusion) : occlusives ou plosives. Nous classons principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré. Les consonnes liquides combinent une occlusion et une ouverture simultanée du conduit vocal. Elles sont caractérisées par un degré de sonorité proche de celui des voyelles. Enfin, les consonnes nasales font intervenir la cavité nasale par abaissement du voile du palais. Elles sont produites par l'écoulement de l'air phonatoire dans le conduit nasal.

1.8. Notions fondamentales sur l'Arabe standard :

La langue Arabe est la langue parlée à l'origine par le peuple arabe. 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. Dans le cadre de notre travail, nous parlerons de la langue Arabe en référence à

cequi est communément appelé ″l'Arabe Standard″ (AS), c'est-à-dire, la langue de communication commune à l'ensemble du Monde Arabe [7].

1.8.1. Système phonétique de l'Arabe Standard :

L'Arabe Standard (AS) compte 34 phonèmes : 28 (26 consonnes + 2 semi consonnes (و + ي) + 6 voyelles (3 courts + 3 langues). La graphie des lettres est différente selon leur position dans le mot. Ainsi, lalettre [a] est transcrite [k a t a b a] (كتب) en début de mot et [a] en milieu et en fin de mot.

Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscules/majuscules n'existe pas.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, et l'autre consonne lunaire qui n'assimilent pas le «ل » de l'article.Les solaires se prononcent en double, comme par exemple avec le mot « soleil » [شمس], aulieu de prononcer el-chams, on prononce ech-chams, car la lettre [ش], 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] on prononce [elqamar] tout à fait normalement, parce que la lettre[ق] est une lettre lunaire.

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 ouau-dessous de la consonne, (َ , ُ , ِ) , exemple : [قُرأَ] ;
- ✓ les voyelles longues [المد حروف] : [aa], [uu], [ii] sont écrites sous forme de caractères consonantiques [ي , و , ا] .

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	[d̪ˤ]	Alvéolaire
	Non-Voiseé		ط	[t̪ˤ]	Alvéodentale
Fricatives	Voisées		ز ذ غ ع	[z] [ðˤ] [ɣ] [ʕ]	sifflante dorsoalvéolaire interdendale uvulaire pharyngale
	Non-Voisées		س ث ف ش خ ه ح	[s] [θ] [f] [ʃ] [χ] [h] [ħ]	sifflante dorsoalvéolaire interdendale labiodentale chuinchante palatale vélaire glottale Pharyngale
	Voisées	Emphatiques	ص	[sˤ]	doralveolaire sifflante
	Non-Voisées		ظ	[ðˤ]	Interdentale
Nasales	Voisées		م ن	[m] [n]	bilabiale Alvéodentale
Liquide	Voisée		ل	[l]	Dentale
Affriquée	Voisée		ج	[ʒ]	Alvéodentale
Vibrante	Voisée		ر	[r]	apicvoalvéolaire
Semi-voyelles	Voisées		و ي	[w] [y]	bilabiale Palatale

Tableau 1.2 : Transcription Orthographique Phonétique de l'AS

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] et [ز] 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] deux nasales caractérisées par un échappement de l'air en même temps par la bouche et par le nez : [ن] ,[م] une vibrante caractérisée par le déplacement de la langue au passage de l'air : [ر] deux semi-consonnes (ou semi-voyelles) caractérisées par un passage rapide del'air à travers la bouche accompagné de frottements consonantiques : [و] ,[ي].

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],[ظ] ,[ص] ,[ض] sont caractérisées par une forte tension des différents organes du conduit vocal.

1.9. Conclusion :

Dans ce chapitre nous avons exposé des notions de base sur la généralité et la production 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 par concaténation.

Chapitre 2

Synthèse de la parole

Chapitre2:

Synthèse de la parole

2.1 Intoduction

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.

2.2 Définition de la synthèse de la parole

La synthèse de parole présente plusieurs avantages, elle est d'une part plus naturelle pour le grand public, elle est plus rapide et efficiente qu'un message écrit court et le champ de vision reste libre pour effectuer une autre tâche de lecture.

Les deux principaux critères exigés par la synthèse de la voix sont l'intelligibilité et l'aspect naturel. Si de nos jours, le premier critère est atteint, le deuxième est encore au stade de développement. En effet, si les synthétiseurs reproduisent une voix tout à fait intelligible, les intonations et l'expressivité ne sont pas encore au point.

2.3 Historique de la synthèse de la parole

À plusieurs reprises au cours de l'histoire, on a tenté de reproduire la voix humaine. Au 18 ème siècle, on met au point à cet effet des dispositifs mécaniques équipés de soufflets et d'anches vibrantes. Au 20 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 Opération Démonstration), 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.

2.4 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 aupar avant 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 » [8].

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.

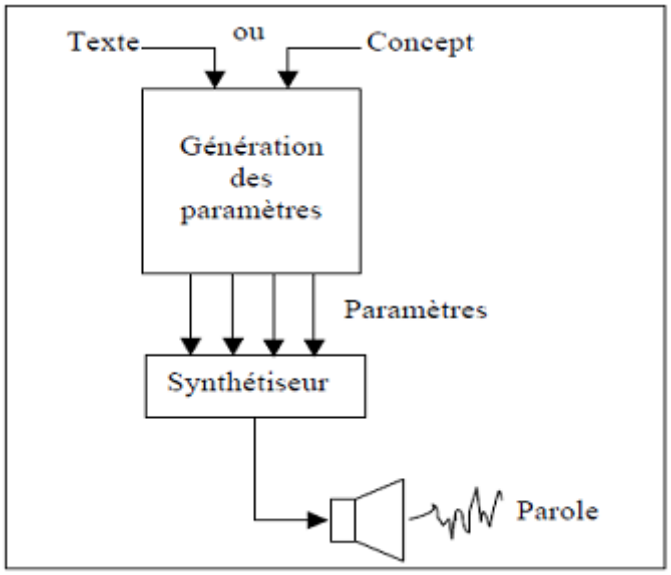

Figure 2.1 : Système de synthèse de la parole [8].

Les synthétiseurs ont quant à eux la fonction inverse de celle des analyseurs et des reconnaissais 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 reconnaissais 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.

2.5 LE SYSTEME TEXT-TO-SPEECH (TTS)

Un Système de Synthèse à Partir du Texte (TTS : Text-To-Speech) est une machine capable de lire a priori n'importe quel texte à voix haute, que ce texte ait été directement introduit par un opérateur sur un clavier alphanumérique, qu'il ait été scanné et reconnu par un système de reconnaissance optique des caractères (OCR : Optical Character Recognition), ou qu'il ait été produit automatiquement par un système de Dialogue Homme-Machine. On définira donc plutôt la synthèse TTS comme la production automatique de phrases par calcul de leur transcription phonétique [9].

2.6 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 [10].

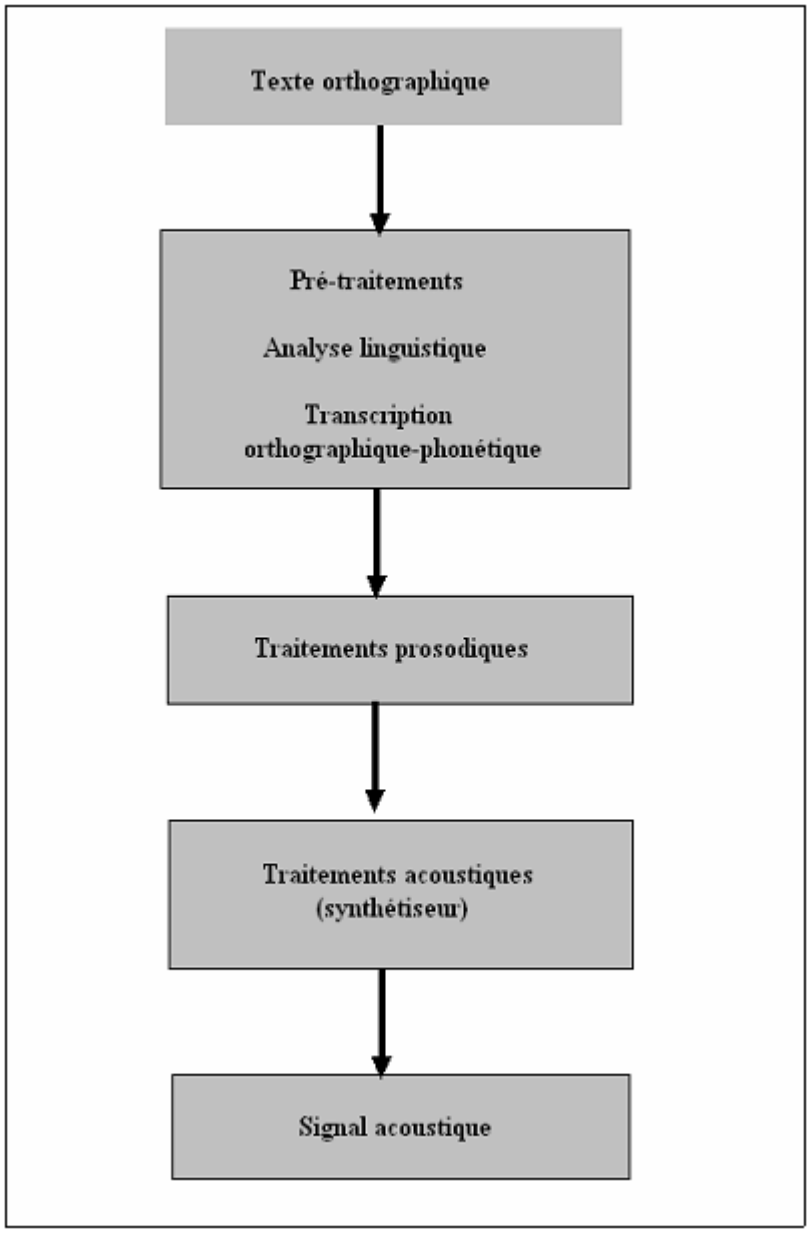

Figure 2.2 : Architecture générale d'un système de synthèse de la parole à partir du texte [10]

2.7 TECHNIQUES D'ANALYSE DU SIGNAL VOCAL

Le signal vocal peut être analysé soit, en tenant compte des mécanismes de production en utilisant les méthodes paramétriques, soit en utilisant les méthodes non paramétriques.

2.7.1 Méthodes non paramétriques

Le signal de parole peut être analysé dans le domaine temporel ou dans le domaine spectral par des méthodes non paramétriques, sans faire l'hypothèse d'un modèle pour rendre compte du signal observé. Les méthodes spectrales sont fondées sur la décomposition

fréquentielle du signal sans connaissance a priori de sa structure fine. Une analyse spectrale du signal permet de mettre en évidence certaines caractéristiques de la production de la parole qui peuvent contribuer à l'identification phonétique. L'articulation des phonèmes a une influence directe sur la forme du conduit vocal et des cavités, et donc sur les résonances qui apparaissent dans l'enveloppe du spectre [9].

2.7.2 Méthodes paramétriques

Les méthodes paramétriques appelées aussi méthodes d'identification sont fondées sur une connaissance des mécanismes de production de la parole (Exemple : le conduit vocal). Les plus utilisées sont celles basées sur l'analyse prédictive linéaire et l'analyse cepstrale. L'hypothèse de base est que le conduit buccal est constitué d'un tube cylindrique de section variable. L'ajustement des paramètres de ce modèle permet de déterminer à tout instant sa fonction de transfert. Cette dernière fournit une approximation de l'enveloppe du spectre du signal à l'instant d'analyse. Ces méthodes consistent à ajuster un modèle aux données observées.

Les paramètres du modèle, en nombre faible, caractérisent le signal, nous pouvons ainsi injecter des connaissances a priori sur le processus physique qui a engendré ce signal.

Les avantages de cette approche sont la souplesse de l'analyse, l'introduction naturelle de l'information et les choix variés des espaces de représentations paramétriques. Dans le cas de la modélisation du signal parole, nous n'avons accès qu'à une seule sortie du système alors que l'entrée n'est pas mesurée. Il en résulte un problème d'estimation non linéaire car nous ne disposons pas d'observation de l'onde glottique d'excitation. En conséquence, nous en sommes limités à faire quelques hypothèses relativement neutres sur l'entrée ; par exemple, bruit blanc à moyenne nulle et reporter tout l'effort de modélisation sur le système.

2.7.2.1 Codage Prédictif Linéaire (LPC)

Cette méthode connue de la production sous le sigle LPC (Linear Prédictive Coding) se fonde sur les connaissances de la production de la parole et suppose que le modèle de production de la parole est linéaire selon le schéma (Fig.2.3).

Globalement, ce modèle peut se décomposer en deux parties : la source active, le conduit passif de manière plus détaillée, il peut se décrire de la maniéré suivante : l'onde est modélisée comme la sortie d'un filtre passe bas à deux pôles de fréquence de coupure

d'environ 100 Hz (glotte), l'entrée en de ce filtre est un train d'impulsions de période T0 pour les sons voisés ou un bruit blanc pour les sons non voisés (source).

Le modèle du conduit vocal est un filtre tout pôle (AR : auto - Régressif) d'ordre 2M décomposable en une cascade de résonateurs à 2 pôles en série (tuyaux résonants). Le modèle du conduit nasal est un filtre pôle zéro ARMA (Auto Régressif à Moyenne Ajustée) et le rayonnement aux lèvres peut se modéliser par un filtre tout zéro (MA : Moyenne Ajustée).
L'ensemble des conduits se comporte donc comme un système linéaire ARMA.

Modèle glottale:

$$\boldsymbol{G(z) = \frac{1}{(1-e^{-2\pi f g T z-1})^2}} \text{ avec } \boldsymbol{f_g = 100MZ} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{2.1})$$

Modèle du conduit vocal:

$$\boldsymbol{V(z) = \prod_{i=1}^{M} \left(\frac{1}{(1 - 2e^{-2\pi B_i T}.cos(2\pi F_N T)z^{-1} + e^{-4\pi BiT} z^{-2}} \right)} \ldots\ldots\ldots (\mathbf{2.2})$$

.

F_i: Fréquence du formant n° i, B_i sa bande passante
Modèle du conduit nasal

$$\boldsymbol{N(z) = \frac{(1 - 2e^{-2\pi B'_N T}.cos(2\pi F'_N T)z^{-1} + e^{-4\pi B_N T} z^{-2}}{(1 - 2e^{-2\pi B_N T}.cos(2\pi F_N T)z^{-1} + e^{-4\pi B_N T} z^{-2}}} \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{2.3})$$

Avec F_N et F'_N formant nasal ou anti formant nasal et respectivement, B_N et B'_N leurs bandes passantes.
Si on suppose qu'une partie α du signal gn est dérivée vers le conduit nasal le modèle du conduit peut se mettre sous la forme :

$$\mathbf{H(Z)= G(Z).\ [1-\alpha).V(Z)\ L(Z) + \alpha N(Z)]} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\mathbf{(2.4)}$$

Avec $0 \leq \alpha \leq 1$ pour un son nasal $\alpha=1$; pour un son non nasal $\alpha=0$.

H(z) Est en tout généralité un modèle ARMA d'ordre p:$H(z) = \frac{B(z)}{A(z)}$
Dans le domaine temporel on aura :

$$\boldsymbol{y_n + \sum_{i=1}^{p} a_i y_{n-p} = e_n + \sum_{i=1}^{q} b_i e_{n-p}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{2.5})$$

Caractériser le signal $\mathbf{y_n}$ revient donc a estimé les coefficients {ai; bi}.

Pour une source connue $\mathbf{e_n}$(séquence d'impulsions ou bruit blanc). Souvent pour simplifier la résolution de ce problème, on suppose que bi = 0, i ≥ 1 ce qui rend le modèle AR [9].

2.7.2.2 Analyse cepstrale

Le défaut majeur des méthodes d'analyse, comme la FFT, pour le calcul du spectre réside dans l'intermodulation source/conduit vocal qui rend difficile la mesure du fondamental F0 et des formants.

Le lissage cepstral est une méthode qui vise à séparer la contribution du conduit vocal de l'excitation glottique. Cette séparation est réalisée par un homomorphisme qui transforme la convolution des signaux dans le domaine temporel en une addition dans le domaine cepstral. Entoure, cette méthode permet de fournir un vecteur spectral des MFCC pour des fins de la RAP et de lisser le spectre de parole pour trouver les formants.

Pour cela, nous faisons l'hypothèse que le signal vocal y_n est produit par le signal excitateur un traversant un système linéaire de réponse impulsionnelle b_n.

Le but du cepstre est de séparer ces deux contributions par déconvolution. Il est fait l'hypothèse qu'un signal excitateur est soit une séquence d'impulsions (périodiques, de période T0, pour les sons voisés), soit un bruit blanc pour les sons non voisés, conformément au modèle de production de la parole. Une transformation en Z permet de transformer la convolution en produit.

$$\mathbf{Y(z)=B(z).U(z)} \quad \mathbf{(2.6)}$$

Le logarithme du module uniquement (car nous ne s'intéressons pas à l'information de phase) transforme le produit en somme. Nous obtenons alors:

$$\boldsymbol{log|Y(Z)|= log|U(Z)|+ log\ |B(Z)|} \quad \mathbf{(2.7)}$$

Par transformation inverse, nous obtenons le cepstre. Dans la pratique, la transformation en Z est remplacée par une TFR. L'expression du cepstre est donc:

$$\mathbf{C(n)=FT^{-1}\{log\ (FT\ \{y(n)\})\}} \quad \mathbf{(2.8)}$$

Le cepstre qui ne fait appel à aucune information a priori sur le signal acoustique, est basé sur une connaissance du mécanisme de production de la parole. L'espace de représentation du cepstre ou espace fréquentiel est homogène par rapport au temps. Les premiers coefficients cepstraux contiennent l'information relative au conduit vocal. Cette contribution devient négligeable à partir d'un échantillon n_0 qui correspond à la fréquence

fondamentale F_0 Les pics périodiques visibles au-delà de n_0, reflètent les impulsions de la source.

Le spectre du cepstre pour les indices inférieurs à n_0 permet d'obtenir un spectre lissé, en éliminant les lobes secondaires dû à la contribution de la source. Ces deux contributions peuvent être séparées par une simple fenêtre temporelle notée F (liftrage) telle que le filtre adouci ou le filtre rectangulaire.

La présence d'un pic important dans le cepstre renseigne d'une part sur le caractère voisé ou non du son et d'autre part constitue une bonne indication sur la fréquence fondamentale. L'enveloppe spectrale du conduit vocal (structure formantique) est obtenue par une transformation supplémentaire (Fig.2.3).

Le spectre lissé débarrassé théoriquement de la contribution de la source ne contient que des informations sur le conduit vocal et en particulier sur ses extrema (Formants).

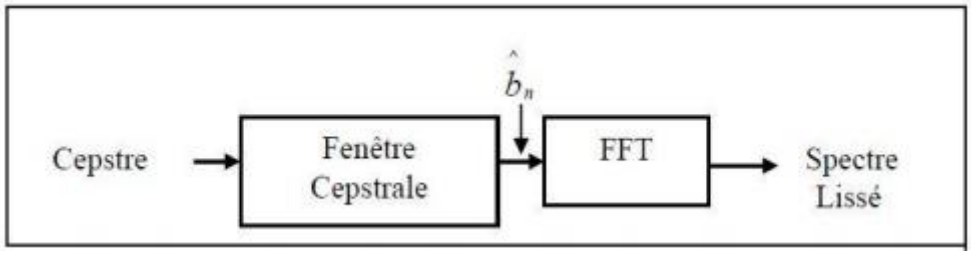

Figure 2.3 : Obtention de la structure formantique à partir du spectre

2.8 LES METHODES DE LA SYNTHESE DE PAROLE

On peut établir une analogie fonctionnelle entre le rôle joué par le module de traitement du signal et celui du système phonatoire humain, qui contrôle en permanence l'activité de tous ses muscles (y compris de ceux qui règlent la fréquence de vibrations des cordes vocales) de façon à produire le signal voulu. Pour y arriver, il est clair que ce module doit, dans une certaine mesure, prendre en compte les contraintes articulatoires. En effet, on sait depuis longtemps que les transitions phonétiques contribuent plus à l'intelligibilité du signal vocal que les zones stables des phonèmes. On peut alors envisager de le faire de façons.

- ✓ Explicite, sous la forme d'une série de règles décrivant formellement l'influence des phones les uns sur les autres.

- ✓ Implicite, en enregistrant des exemples de transitions entre phones dans une base de données de segments de parole, et en les utilisant tels quels comme unités de parole (en lieu et place des phones).

Cette alternative a donné lieu à deux grandes familles de synthétiseurs : la synthèse par règles et la synthèse par concaténation.

2.8.1 La Synthèse Par Règles (SPR)

La Synthèse Par Règles (SPR) a connu un essor considérable dans les années 60-70. Elle n'est plus guère utilisée aujourd'hui que lorsque les contraintes de mémoire et de temps de calcul sont très importants. La qualité des voix disponibles n'est, en effet, pas aussi bonne qu'en synthèse pas concaténation, pour un coût de développement supérieur.

Les synthétiseurs par règles ont principalement la faveur des phonéticiens et des phonologistes. Ils permettent une approche cognitive, générative du mécanisme de la phonation. Ils sont basés sur l'idée que, si un phonéticien expérimenté est capable de « lire » un spectrogramme, il doit lui être possible de produire des règles permettant de créer un spectrogramme artificiel pour une suite de phonèmes donnés. Une fois le spectrogramme obtenu, il ne reste plus alors qu'à générer l'audiogramme correspondant [11].

2.8.2 Synthèse par concaténation d'unités acoustiques

Cette technique, qui repose sur l'utilisation de segments de signaux extraits de la parole naturelle, est la seule qui permet à ce jour de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain.

2.8.2.1 Mise en œuvre

La synthèse proprement dite comprend trois étapes distinctes :

- ✓ **Sélection des unités acoustiques** : cette première étape consiste à choisir dans le répertoire d'unités acoustiques les unités qui seront effectivement utilisées pour synthétiser la succession de sons désirée. Cette étape est à peu près évidente quand les unités sont régulières (à l'instar des phonèmes et des diphones) : seule la présence de plusieurs versions pour le même segment est à prendre en considération. Cette étape est en revanche plus délicate pour les systèmes d'unités de taille variable. Pour une suite de sons donnée, plusieurs choix d'unités sont en général possibles. Il faut alors arbitrer entre les différentes décompositions avec des critères composites.

- ✓ **Ajustement des paramètres prosodiques :** les unités acoustiques préenregistrées possèdent une prosodie intrinsèque (les sons qui la composent ont une certaine durée et la fréquence fondamentale décrit un certain contour). Bien sûr, cette prosodie intrinsèque n'a que très peu de chances d'être conforme à la prosodie de synthèse, spécifiée par le module prosodique. Il va donc falloir utiliser une technique de traitement de signal pour ajuster aux valeurs cibles définies les paramètres prosodiques des unités de synthèse.
- ✓ **Concaténation des unités :** les unités acoustiques, quelles que soient les précautions prises lors de la sélection et de l'enregistrement des unités, ne possèdent pas exactement à leur frontière les mêmes caractéristiques acoustiques (en particulier énergétiques).

En l'absence de traitement, ces discontinuités vont engendrer des artefacts perceptibles et gênants. Il est donc important de lisser ces discontinuités en interpolant les trajectoires des différents paramètres caractéristiques de l'unité [9].

2.8.2.2 Synthèse fondée sur l'algorithme PSOLA

L'algorithme PSOLA (Pitch Synchronous OverLap and Add) 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. L'algorithme PSOLA permet la synthèse d'une parole de haute qualité [10].

2.9 CONCLUSION

Dans cette partie, nous avons abordé les principales méthodes et techniques de la synthèse de la parole. Les deux principaux critères exigés par la synthèse de la voix sont l'intelligibilité et l'aspect naturel, d'où elle vise à améliorer le quotidien, mais n'oublions pas que si elle atteint le niveau de conversation d'un être humain, elle engendrerait aussi sa substitution dans certains domaines en augmentant ainsi l'emprise de la machine sur l'homme.

CHAPITRE 3

Synthèse de la parole par la technique PSOLA

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 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 (≈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.

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 [12].

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.

3.2.2 Fenêtrage du signal de parole de la technique PSOLA

La technique dite d'addition recouvrement des fenêtres temporelles synchrones avec le pitch PSOLA; applique le principe de la réharmonisation spectrale directement sur le signal de parole. 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 3.1).

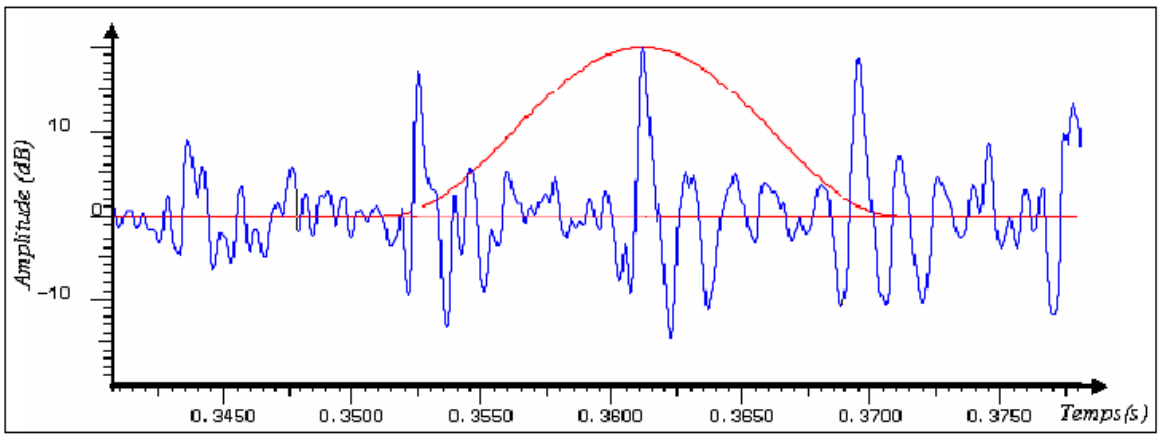

Figure 3.1 : Fenêtrage du signal de parole [10]

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].

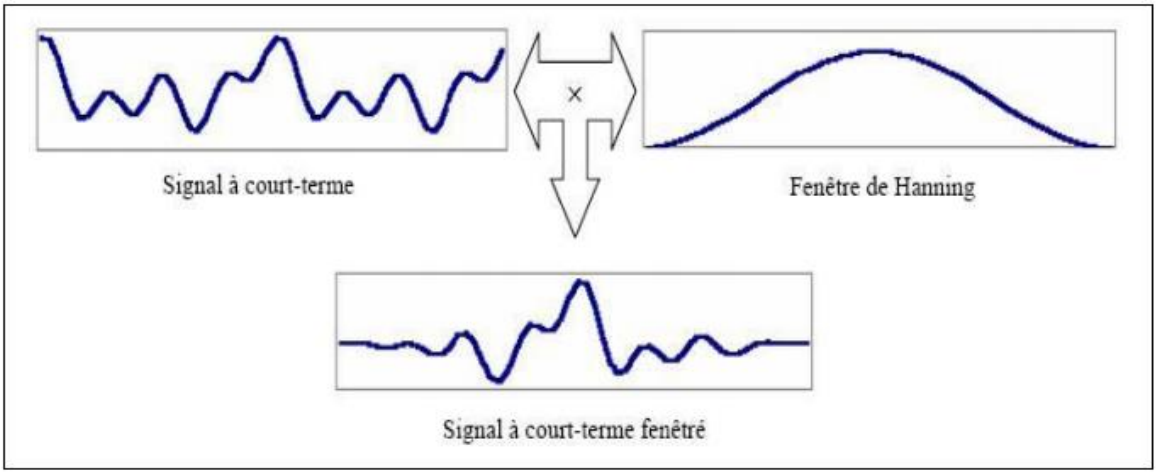

Figure 3.2 : Exemple de signal à Court-Terme [10].

Les signaux à CT sont recombinés pour produire le signal de synthèse à l'aide d'une technique d'addition/recouvrement OLA (Figure 3.3).

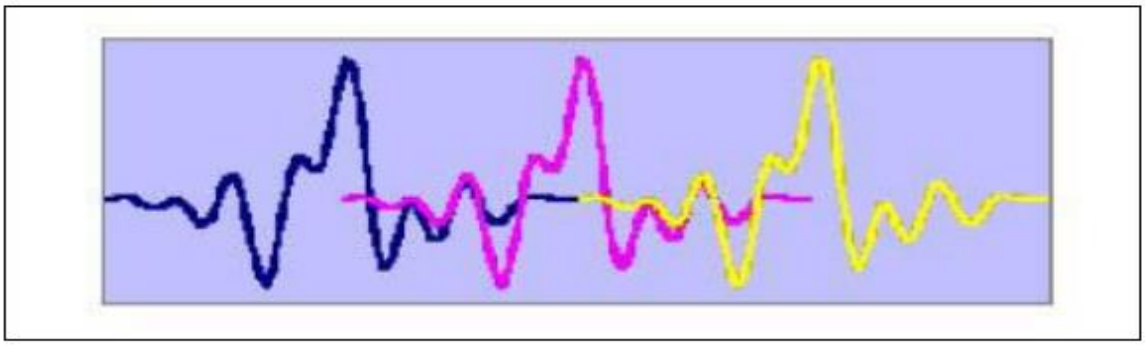

Figure 3.3 : Etape d'addition et recouvrement OLA [10].

Après la recombinaison des signaux à cour terme par addition recouvrement OLA nous trouvons le signale synthétique (figure 3.4).

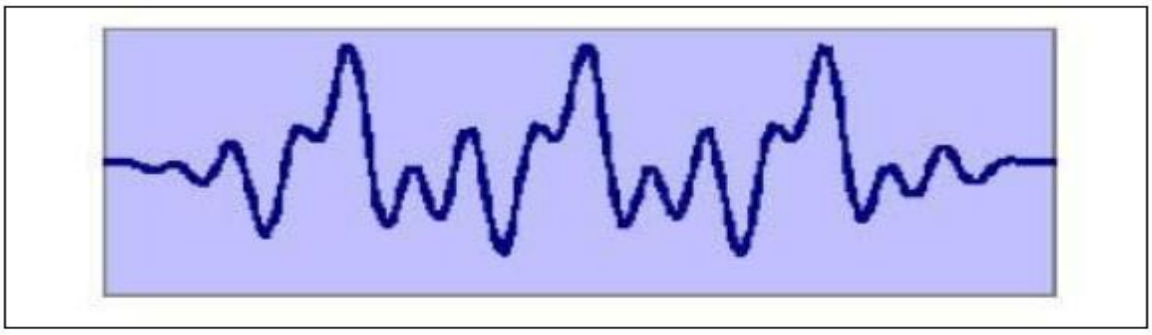

Figure 3.4 : Signal synthétisé avec PSOLA [10].

3.3 ALGORITHME DE SYNTHÈSE DE LA TECHNIQUE PSOLA

Après avoir présenté le principe de la technique de synthèse PSOLA, nous pouvons décrire l'algorithme de synthèse –PSOLA (voir figure 3.5). 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.

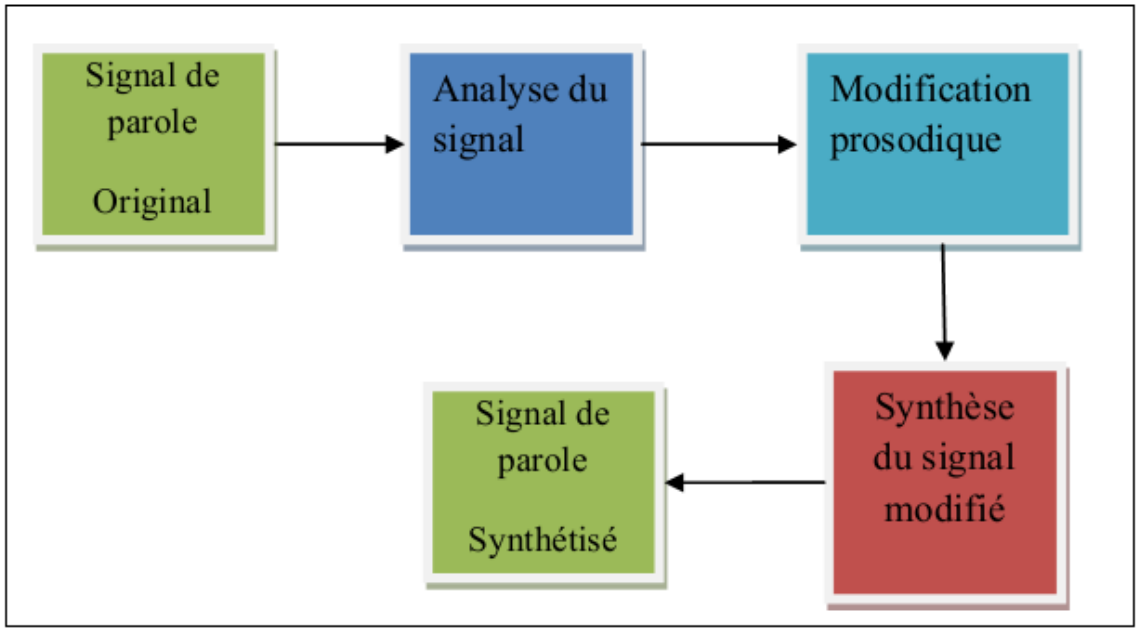

Figure 3.5 : Algorithme de synthèse –PSOLA

3.3.1 Analyse du signal de parole

Les opérations d'analyse à effectuer pour le marquage du signal dans la Figure 3.7. **S**ont les suivantes :

- ✓ **prétraitement :** séparation des composantes périodiques du signal (dites Voisées dans le cas de la voix) et des composantes aléatoires (dites Non Voisées). 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
- ✓ **Détection et détermination de Pitch :** la détection et détermination de la F0 avec une méthode appropriée. La phase de traitement est réservée à l'extraction de la fréquence fondamentale et dépend donc de l'algorithme utilisé.
- ✓ **Détermination des signaux à CT d'analyse.**

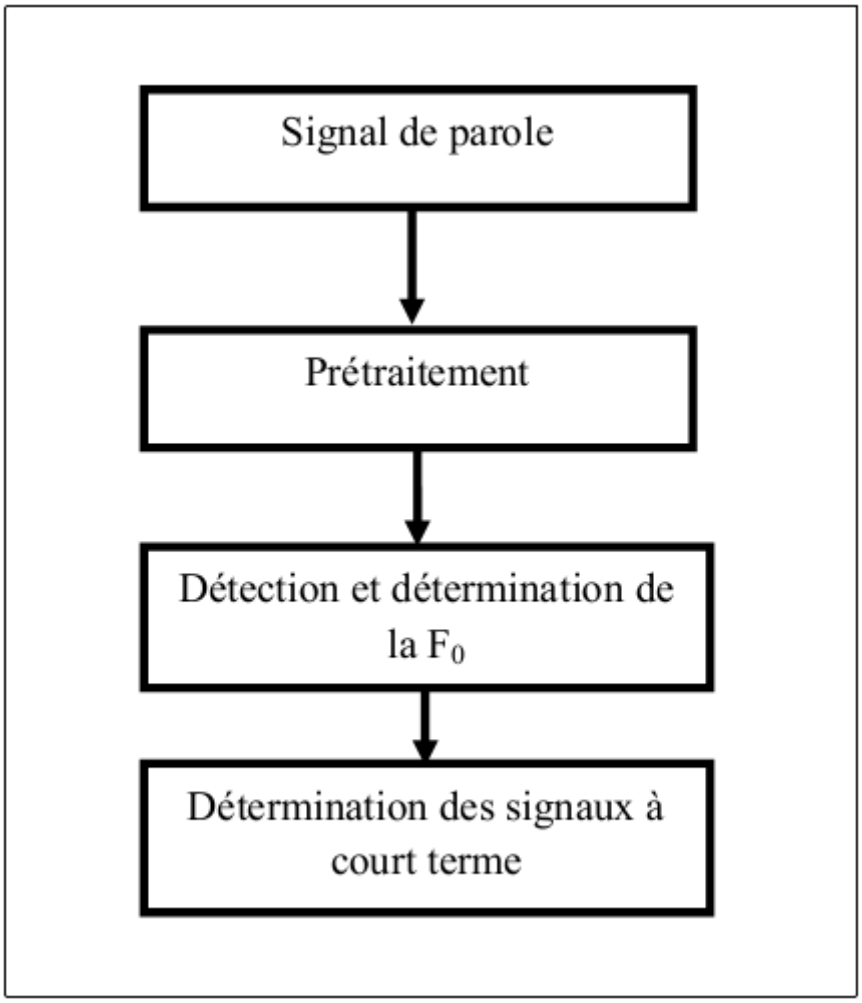

Figure 3.6 : Analyse du signal de parole.[13]

3.3.2 Les Méthodes de détection du Pitch

Les méthodes de détection de pitch, sont souvent classées en trois catégories principales :

- ✓ **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.
- ✓ **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.

- ✓ **Méthodes combinatoires :** Il existe un très grand nombre de méthodes pour extraire la F0, 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 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. [14]

3.3.3 Détermination des signaux à CT (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 :

$$S_i(n) = X(n)\, W_i(n - iT_0) \quad \text{.....(3.1)}$$

$$S(n) = \sum s_i(n - i(T - T_0)) \quad \text{.....(3.2)}$$

Ces fenêtres successives sont obtenues par placement de marques appelées marques de lecture de manière synchrone au pitch du signal (la différence entre deux marques de lecture successive est égale à T0 locale). Voir Figure.3.7. Le signal est alors découpé à l'aide de fenêtres d'analyse centrées sur ces marques de lecture.

$$S_i(n) = X(n)\, W(n - t_r^i) \quad \text{.....(3.3)}$$

Le signal de synthèse S(n) sera donc obtenu par Superposition/Addition des signaux élémentaires centrés en de nouvelles positions que nous appelons marques d'écriture. Ce sont ces positions qui déterminent le pitch et la durée du signal de synthèse. [15]

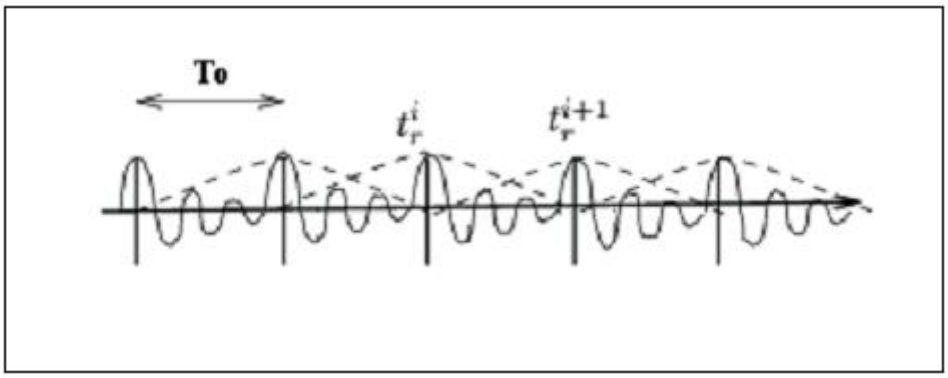

Figure 3.7: Placement des marques de lecture. [15]

3.4 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

Résultats et Discisions

Chapitre 4:

Résultats et Discisions

4.1 INTRODUCTION

Après avoir présenté, au chapitre précédent, la technique de synthèse et de Modification utilisée 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 ce projet, nous avons utilisé un corpus des phrases affirmatives en AS, 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 «الشباك رقم واحد» énoncée en langue arabe et (Figure 4.1), où toutes les algorithmes seront appliqués sur la dite phrase.

Phrase **: الشباك رقم واحد**

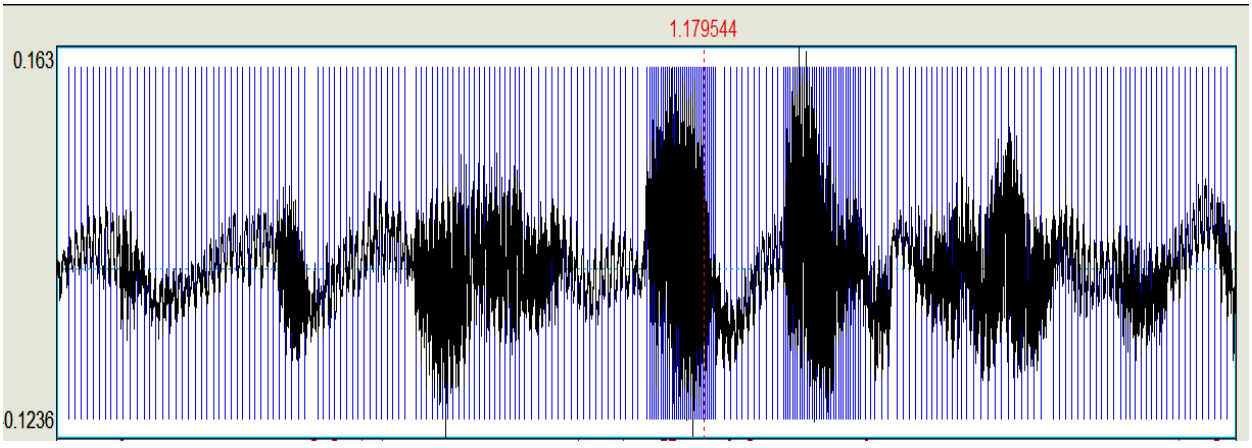

Figure 4.1 : Représentation temporelle de la phrase (الشباك رقم واحد)

4.2.1. Chargement de son en mémoire

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

4.3. La méthodologie suivie dans ce travail

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 et la frèquence fondamentale F0. **(الشباك رقم واحد)**

TESTE1 :

Dans la première partie nous avons 10 concaténations par phonème , les rèsultat donnèe beaucoup de silence ;

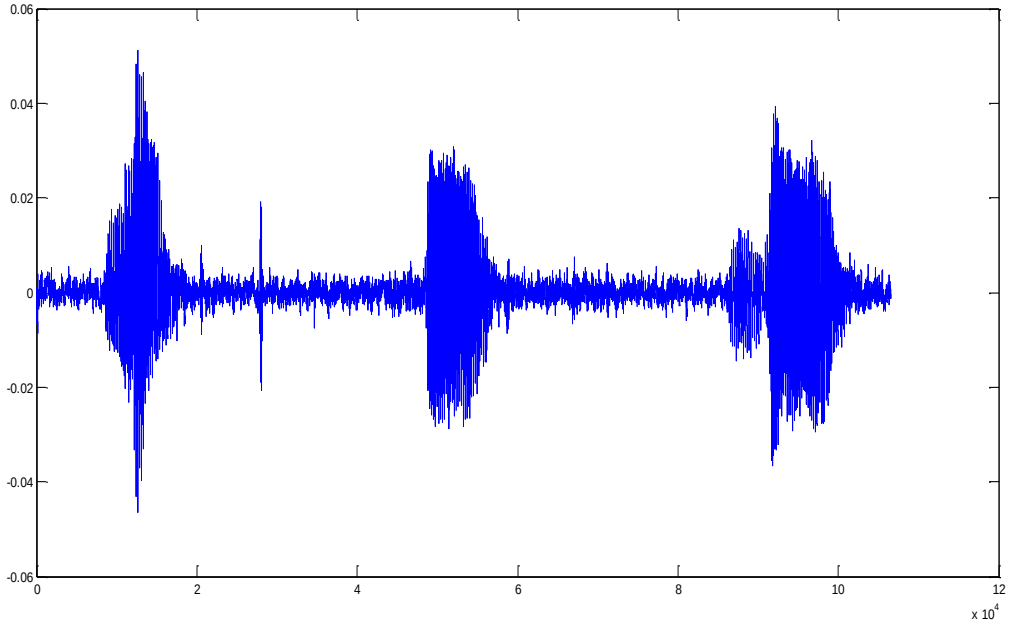

Figure 4.2 : Teste avec silence

TESTE2 :

Dans la 2eme partie en a minimisèe le silence

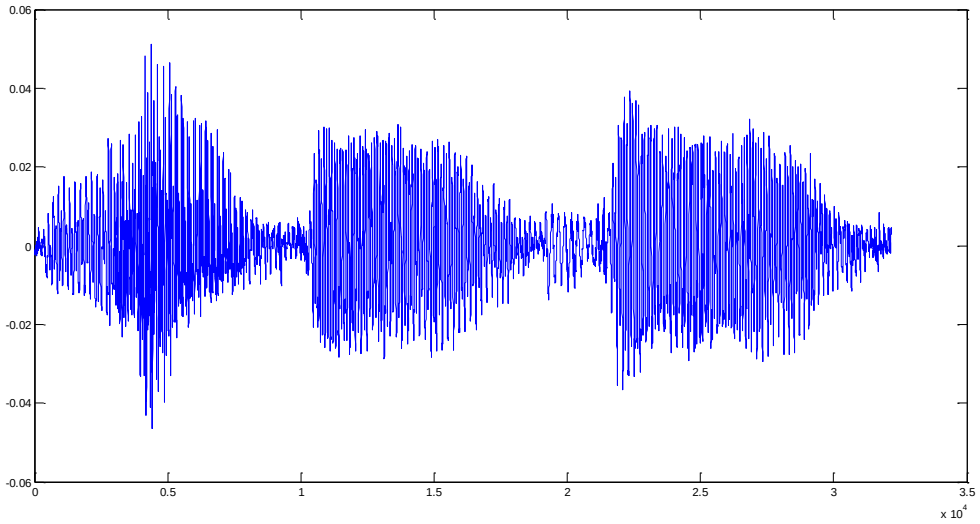

Figure 4.3 : Teste avec minimisation du silence

TESTE3 :

Dans le 3eme teste encoure minimisèe le silence avec le lissage et emplifie le volume de son

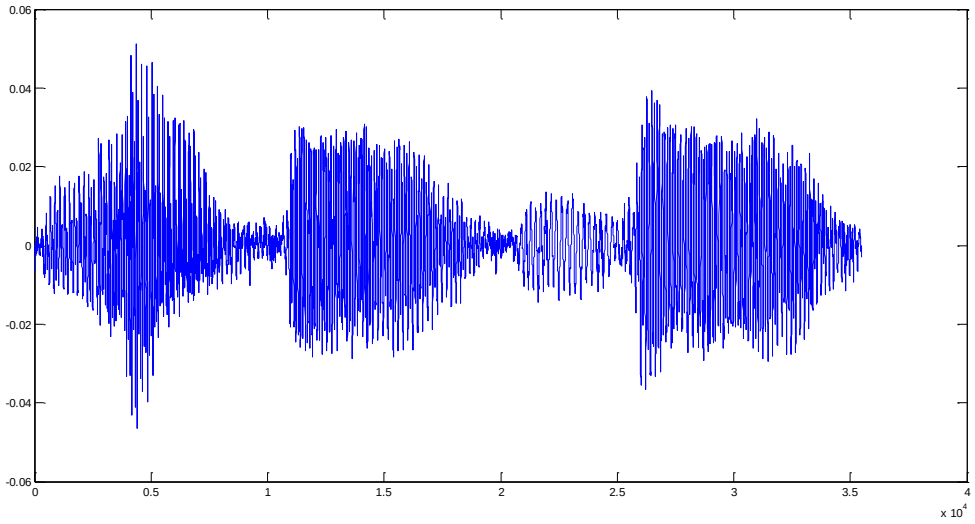

Figure 4.4 : Minimisation du silence avec lissage

TESTE4 :

Dans cette dernière partie du programme on applique la méthode **PSOLA** en modifiées le durèe et la frèquence et le pat de signal et en a rèpète le teste avec la modification des paramètres jusqu'à l'obtenir bonne rèsultat.

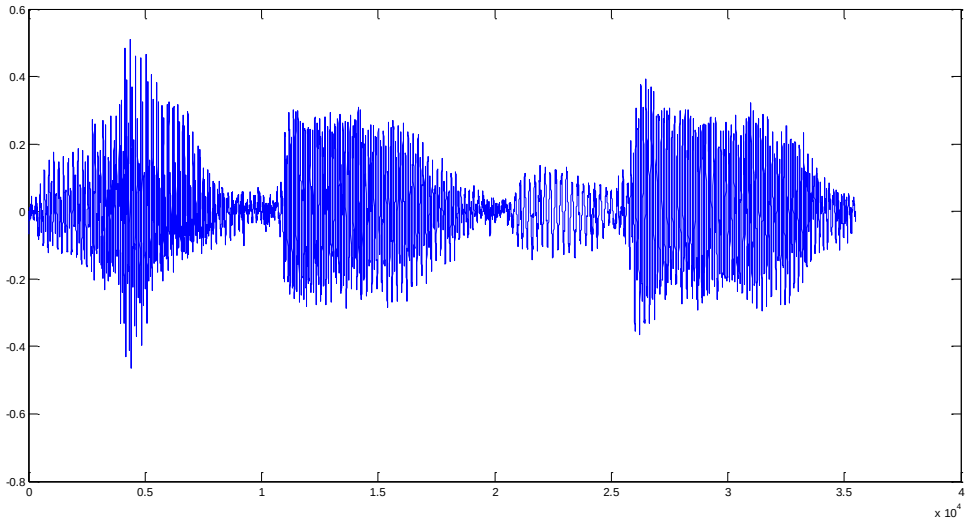

Figure 4.5 : avec le technique PSOLA

En la figure 4.6, les verticals lines sont pitch marks. Et le horizontal line est pitch contour. Et les ètoiles **(*)** sont pitch mark candidates of search regions.

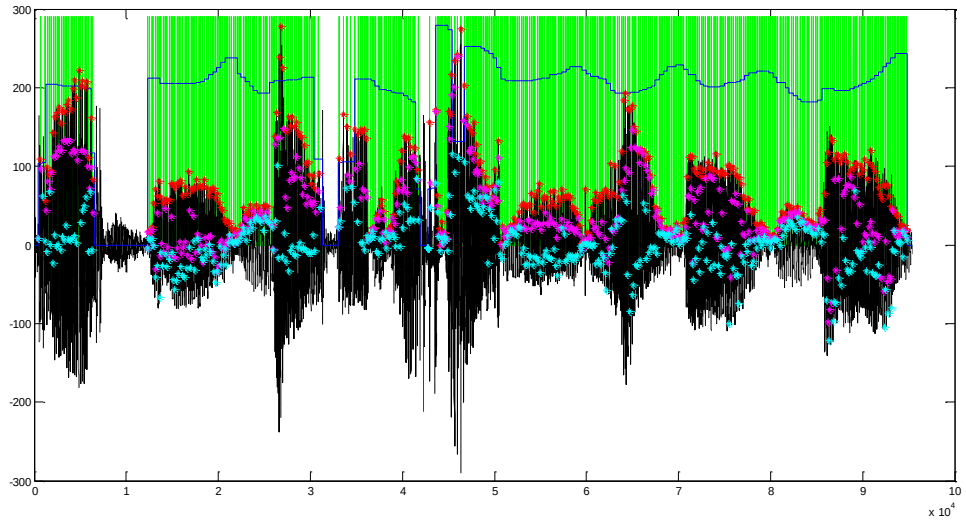

Figure 4.6: pitch marking illustration

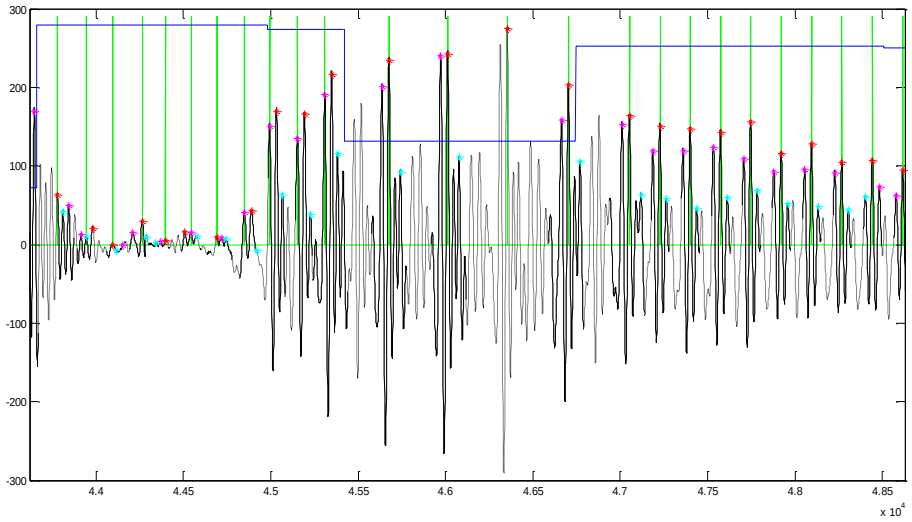

Figure 4.7: pitch marking illustration (ZOOM)

4.5CONCLUSION

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 PSOLA et qui appartient aux familles des synthétiseurs acoustiques dans le domaine temporel. L'Algorithme 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 avons effectivement préservées l'identité vocale.

Conclusion Générale

Conclusion Générale

L’objectif de notre travail tout au long de ce mémoire était la réalisation d’un système de synthèse de parole par unités variables (phrases et mots combinés), Ils nous ontpermis 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 nonparamétriques. Parmi ces techniques de modifications étudiées jusqu’à présent, la plus efficace est l’algorithme PSOLA. C’est une technique de traitement du signal de paroledont l’objectif est de modifier les paramètres. Si maintenant on la considère comme une technique de synthèse de parole 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 PSOLA est qu’elle opère directement sur la forme d’onde du signal de parole. L’idée de base est d’extraire du signaldes grains de sons élémentaires, représentant les caractéristiques locales du signal, et dejouer avec ces grains élémentaires pour réaliser les modifications désirées. L’analyse par PSOLA pour changer le pitch est identique à l'analyse pour étirer le temps, la différenceest 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 latonalité, 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 par plusieurs facteurs appliquées sur la langue Arabe Standard du corpus choisi. Ainsi, nous avons déduitque les dits résultats sont acceptables et satisfaisants.

Références

References:

1. Pouget, M.l., Synthèse incrémentale de la parole à partir du texte. 2017, Grenoble Alpes.
2. P. Taylor and R. Caley, Book, "The architecture of the Festival speech synthesis system", Edition 1, Jean1998.
3. N. Schnell et all. "Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)". in ICMC. 2000.
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. 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.
6. M. Aissiou, Application des Algorithmes Génétiques au Décodage Acoustico-Phonétique de la parole en Arabe Standard, Thèse de Doctorat, ENP, Alger, 2008
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. 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.
9. V A.Dubesset, La Langue française Parlée Complétée (LPC) : Production et Perception, Thèse de Doctorat, Institut National Polytechnique De Grenoble, France, 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, Faculté Polytechnique de Mons, LCTS Lab, France, 2000
13. R. Boite, H. Bourlard, T. Dutoit, Traitement de la parole, Collection électronique, Presses Polytechniques et Université Romandes, 1999.

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