République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

ECOLE NATIONALE POLYTECHNIQUE

DÉPARTEMENT D'ÉLECTRONIQUE

Laboratoire de Signal et Communications

En vue de l'obtention du diplôme de Master en Electronique

Présentée par :

Mr BOUNABI Moussaab

Thème :

Synthèse de la Parole par méthode TDI-PSOLA

Soutenue le : 26 Juin 2013

Devant le Jury :

C. LARBES	Professeur	ENP	Président
M. GUERTI	Professeur	ENP	Rapporteur
M. MAAMRI	CC	ENP	Examinateur

Promotion: Juin 2013


بسم الله
الرحمن الرحيم

ملخص:

إن هذا العمل يهدف إلى دراسة تقنيات تغيير العناصر العروضية المستعملة في إشارات الكلام لهدف إجراء تغييرات عروضية للإشارات الصوتية والتركيب الاصطناعي للكلام اعتمدنا في مشروعنا على تقنيات التعديل بتغيير التردد الابتدائي وتغيير الزمن وعليه قمنا بتصميم طريقة تغيير تسمى TDI PSOLA حيث أخذنا عينة من الكلام وقمنا بتحليلها وتقييمها لهدف اختيار أحسن العوامل لتغيير الكلام بنوعية الجيدة.

كلمات المفاتيح: تركيب الكلام، التردد الابتدائي، TDI-PSOLA

Résumé

Le but de notre travail est d'étudier 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 **TDI-PSOLA** (**T**ime **D**omain **I**nterpolation **P**itch **S**ynchronous **O**verlap and **A**dd), 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, Fréquence fondamentale, TDI-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 TDI-PSOLA (Time Domain Interpolation Pitch Synchronous Overlap and Add). We made an evaluation of the technique used to get the best factors changes with good quality of synthetic speech.

Keywords: Speech synthesis, fundamental frequency, TDI-PSOLA.

Dédicaces

Je dédie cette humble travaille à ma très chère et tendre mère qui ces occupé de moi avec une grande affection et amour. A mon père qui a été et sera toujours derrière moi en m'encourageant a perfectionné mon travail.

Je le dédie aussi à :

- Mon frère Aymen et a mes deux sœurs Zineb et Fatima el Zahraa qui me soutiennent toujours dans le meilleure et dans le pire
- mon binôme Mohamed El-Amine qui a sue m'épaulé tous le long de ce travail
- Mes amis qui ont été toujours à mes coté dans les moments les plus rigoureux : Aymen, Billel, Yazid, Minou, Zinou, El bouz, amine, Khali, les « Samir », Oussama, Salah, djawade, Saad, Houdayfa, Ismail, Ben Taleb et surtout à celui que j'admire et que je respecte le plus qui est l'honorable « Cheheb Lotfi »
- et je n'oublie pas le reste de ma famille et tous ce qui m'aimes de m'avoir aidé commensurablement et indirectement à devenir l'être que je suis
- et je n'oublie pas aussi l'association scientifique « ELMAARIFA » et le GROUP « MAZAL WAKFIN » qui ont pris soin de moi.

MOUSSAAB

Remerciements

Tout d'abord je remercie Dieu de m'avoir donné la force et le courage d'accomplir ce travail.

Je remercie vivement ma promotrice 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 mémoire.

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

Je remercie également Monsieur ***MAMMERI Mohamed****, Maître de conférences à l'Ecole Nationale Polytechnique, d'avoir accepté de faire partie de mon jury.*

Je tiens à remercier également l'ensemble des enseignants qui ont contribué à ma formation.

Je remercie tous ceux, qui de près ou de loin, m'ont apportés leur contribution pour la réalisation de ce travail.

LISTE DES ABREVIATIONS

TAP : **T**raitement **A**utomatique de la **P**arole
RAP : **R**econnaissance **A**utomatique de la **Pa**role
API : **A**lphabet **P**honétique **I**nternational
F_0 : **F**réquence **f**ondamentale
F_1 F_5 : Formants
TTS : **T**ext-**T**o-**S**peech (Un Système de Synthèse à Partir du Texte)
OCR : **O**ptical **C**haracter **R**ecognition (un système de reconnaissance optique des caractères)
LPC : **L**inear **P**rédictive **C**oding (Codage Linéaire Prédictive)
TFD : **T**ransformée de **F**ourier **D**iscrète
TFR : **T**ransformée de **F**ourier **R**apide (**FFT**)
AR : **A**uto **R**égressif
ARMA : **A**uto **R**égressif à **M**oyenne **A**justée
MA : **M**oyenne **A**justée
SPR : **S**ynthèse **P**ar **R**ègles
TFI : **T**ransformée de **F**ourier **I**nverse
PSOLA : **P**itch **S**ynchronous **O**ver **L**ap and **A**dd
CT : **C**ourt **T**erme
OLA : **O**ver **L**app **A**dd
SI : **S**ystème **I**nternational
TDI-PSOLA : **T**ime **D**omain **I**nterpolation-Pitch Synchronous Over Lap and Add
FDI-PSOLA : **F**requency **D**omain **I**nterpolation-Pitch Synchronous Over Lap and Add
TDHS : **T**ime **D**omain **H**armonic **S**caling

Liste des Figures

page

Table des matières

Introduction générale

INTRODUCTION GENERALE

La parole étant le moyen de communication le plus naturel chez l'Homme, celui-ci a très vite cherché à l'intégrer dans les interfaces Homme-Machine. Cela a été rendu possible grâce aux efforts consentis en reconnaissance et en synthèse de la parole, alors que la première vise à reconnaitre les messages de l'utilisateur pour les traduire en action, la seconde a pour objectif de doter l'ordinateur de la capacité à lire des textes à haute voix. Cela rend le **T**raitement **A**utomatique de la **P**arole (**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 **R**econnaissance **A**utomatique et la **S**ynthèse de la **P**arole (**RAP ; SP)** ; Un thème important de la recherche actuelle dans le domaine du TAP, est la réalisation de véritables systèmes de dialogue oral entre l'Homme et la Machine

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 TDI-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.

Pour atteindre notre objectif, nous avons structuré notre travail en trois parties :

- dans la première, nous allons décrire d'une manière générale des notions sur le traitement de la parole ainsi que sa production, l'appareil phonatoire humain, des spécifications du signal vocal et finissons par la classification des sons ;
- la deuxième donne une brève définition de la synthèse de la parole, 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 la dernière partie, nous étudions la technique qui permet de faire la synthèse d'un signal de parole, soit TDI-PSOLA et FDI-PSOLA en tenant compte de la comparaison entre la technique précédente et la TD-PSOLA.
- Nous terminons notre travail par des conclusions et perspectives

Chapitre I

Notions sur la parole

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 automatique de la parole et de sa production, ensuite nous présentons l'appareil phonatoire humain et finissons par complexité du signal vocal et classifications des sons.

1.2 QU'EST-CE-QUE LE TRAITEMENT AUTOMATIQUE DE LA PAROLE (TAP) ?

Le **T**raitement **A**utomatique de la **P**arole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage.

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 reconnaisseurs** ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse.
- **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 deux types de synthétiseurs : les synthétiseurs de parole à partir d'une représentation numérique, et les synthétiseurs de parole à partir d'une représentation symbolique
- enfin, le rôle des **codeur**s 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 [1].

1.3 L'APPAREIL PHONATOIRE

L'appareil phonatoire est l'ensemble des organes qui permettent de produire les sons constituant la voix. Elle est décomposé en trois parties correspondant à trois entités fonctionnelles différentes : les voies aériennes inférieures composées des poumons et de la trachée artère, le larynx, et le conduit vocal (Fig.1.1).

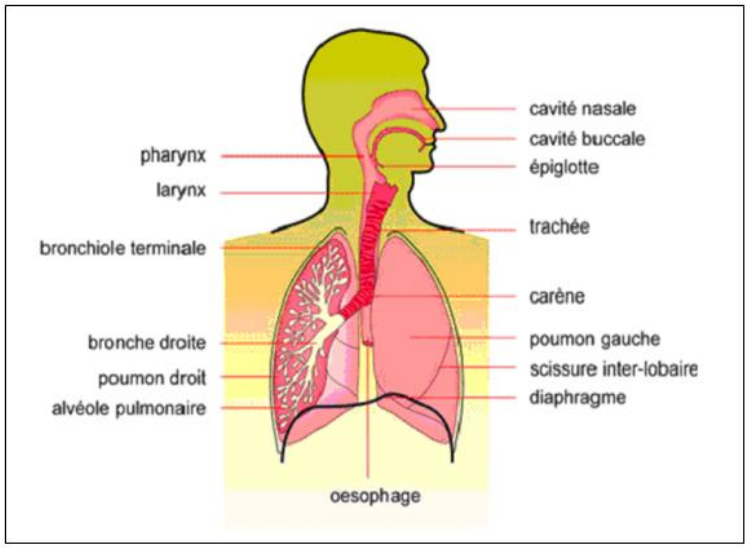

Figure 1.1 : Modèle simplifié de l'appareil phonatoire [2]

1.3.1 Les voies aériennes inférieures

Les voies aériennes inférieures correspondent à la partie de l'appareil phonatoire située dans le thorax et sont composées de deux poumons reliés à la trachée qui elle-même remonte jusqu'aux voies aériennes supérieures, les poumons jouent le rôle de réservoir de pression et permettent de générer l'écoulement d'air à l'origine de la production de sons et notamment des vibrations des cordes vocales.

1.3.2 Le larynx

Le larynx est l'organe qui fait la jonction entre la trachée et le pharynx. Il se situe dans la gorge et est donc le siège de la production des sons voisés qui implique la vibration des cordes vocales(Fig.1.2).

Lors de la production des sons voisés de la parole (comme les voyelles par exemple), c'est la vibration des cordes vocales qui constitue la source des ondes acoustiques.

1.3.3 Le conduit vocal

Le conduit vocal est la partie des voies aériennes supérieures située au-dessus du larynx, il est localisé dans la tête et constitué du pharynx et de deux cavités résonnantes séparées par le palais : la cavité orale (ou buccale) et la cavité nasale. Lorsque les cordes vocales vibrent, les ondes acoustiques générées se propagent dans ces cavités qui agissent comme un résonateur acoustique (Fig.1.2) [3].

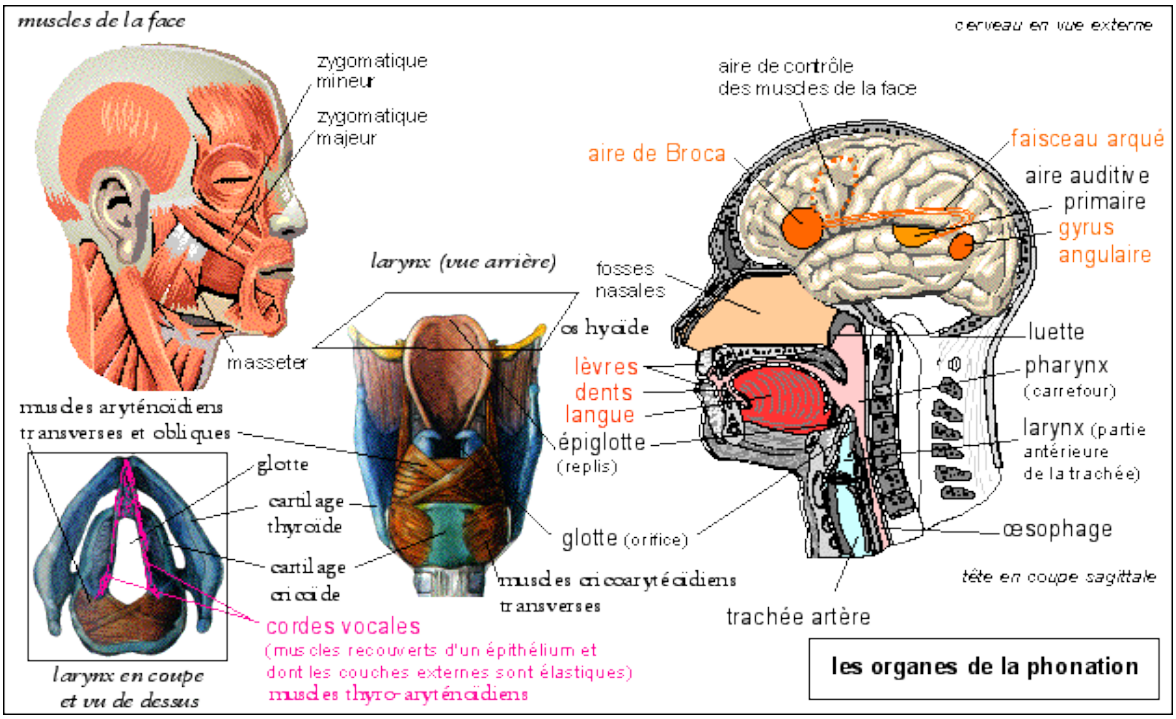

Figure 1.2 : les organes de la phonation [4]

1.4 LA PRODUCTION DE PAROLE

La production de la parole est l'opération la plus complexe de l'activité biologique humaine, et du monde vivant connu. Elle met en jeu un très grand nombre de muscles aux mouvements particulièrement précis, caractérisés par de très nombreuses unités motrices, et dont la synchronisation doit être parfaitement contrôlée pour créer l'objet sonore porteur de sens.

La production de la parole est un système dynamique, dont le comportement à un moment donné dépend de ses états antérieurs. Le système est donc dépendant d'une variable paramétrable fonction du temps qui dans ce cas est un geste articulatoire.

La phonologie articulatoire est basée sur la définition des phonèmes en termes de gestes, qui sont les unités d'action et les bases de contraste des items linguistiques, les atomes de la description phonologique. Les phonèmes sont donc définis par des groupes de gestes [5].

1.5 LES PARAMETRES PROSODIQUES ET ACOUSTIQUES 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 caractéristiques prosodiques influencent directement sur l'intelligibilité de la parole synthétique.

1.5.1 La Fréquence Fondamentale

La Fréquence Fondamentale ou $\mathbf{F_0}$ 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.

Elle permet de diviser l'ensemble des sons de parole en trois grandes macros classes [6]:

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

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.

1.5.2 La durée

La durée est une mesure très variable. Elle représente le temps de la prononciation d'un phonème. 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.

1.5.3 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 :

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

$$E_{db} = 10*\log_{10}\left(\frac{1}{T}\sum_{n=1}^{T} x_n^{\,2}\right) \qquad (1.2)$$

1.5.4 Les Formants

Lorsqu'un excitateur entre en vibrations et fournis un signal, ce dernier passe à travers une cavité de résonnance (le résonateur) qui a amplifier certains composantes. On obtient alors ce qu'on appelle les formants qui sont un facteur essentiel dans la caractérisation du timbre.

L'appareil phonatoire étant constitué de différente cavités Lors du passage de l'air à travers ces cavité il est amplifié et subit différentes transformations due aux degrés d'ouverture et de fermeture au niveau de chaque cavité à la position de la langue des lèvres etc. Ces cavités possèdent des fréquences de résonance qui renforcent certaines régions du spectre de sources excitatrices. Les maxima de la courbe de réponse en fréquences du conduit vocale sont appelés Formants. Chaque son à ses formants caractéristique. Sur un spectrogramme, les formants sont représentés par des bandes noires (le degré de noirceur correspondant à l'énergie) (Fig.1.3).

La fréquence fondamentale est responsable de la hauteur perçue d'un son. Les fréquences d'harmonique renforcées, responsables du timbre d'un son, sont elle aussi numérotées. F_1 correspond à la 1 ère zone d'harmoniques renforcées, F_2 à la 2 éme et ainsi de suite jusqu'à F_5.

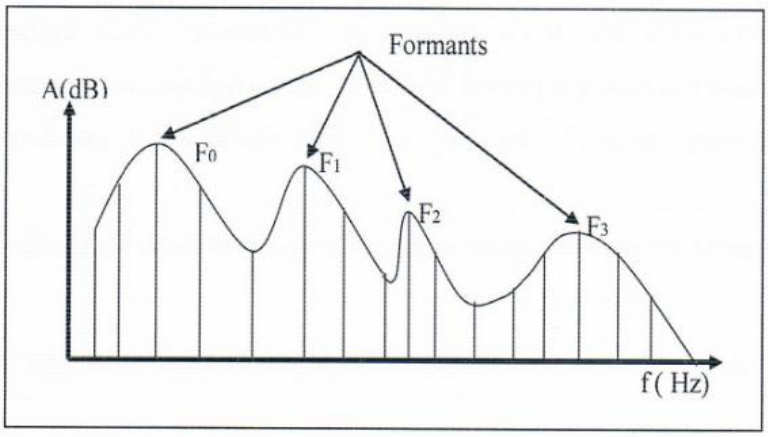

Figure 1.3 : Représentation des Formants d'un son voisé [7]

Généralement, nous pouvons aller jusqu'à cinq ou six formants pour produire une parole de très haute qualité. Les formants nous permettent de décrire aussi les cibles vocalique correspondant aux zones stables ainsi que les zones de transitions (passage entre deux son consécutifs) ce qui montre leur très grande importances pour l'analyse acoustique en phonétiques au moins trois formants sont exigés pour produire les différentes voyelles généralement, on peut aller jusqu'à cinq formants pour produire une parole de haute qualité [7].

1.6 LA COMPLEXITE DU SIGNAL VOCAL

La grande difficulté du TAP et en particulier celui de la **R**econnaissance **A**utomatique de la **P**arole **(RAP)** provient du caractère du processus de la communication parlée et des caractéristiques intrinsèques du signal vocal. La parole est un signal continu d'énergie finie, non stationnaire. Sa structure est complexe et variable dans le temps ; périodique ou plus exactement pseudo périodique pour les sons voisés, aléatoire pour les sons fricatifs et impulsionnel pour les sons occlusifs

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 silence 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.

1.6.2 Variabilités

La parole présente une très grande variabilité qui résulte de plusieurs facteurs et ceci que ce soit pour un même ou plusieurs locuteurs (la fatigue, l'état émotionnel, l'âge, le sexe, l'origine géographique et le milieu social). Parmi ces facteurs, les perturbations apportées par le microphone (selon le type, la distance et l'orientation) et l'environnement (bruit et réverbération). De telles variations ne donnent pas naissance à de nouveaux phonèmes, puisqu'elles ne portent aucune information sémantique [8]

1.6.2.1 Variabilité intra-locuteur

La variabilité intra-locuteur 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 qui affecte le timbre et le rythme de la voix ;
- les maladies affectant les organes de la voix.

1.6.2.2 Variabilité interlocuteur

Des différences acoustiques 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 milieu social.

1.6.2.3 Variabilité contextuelle

En effet les mouvements articulatoires peuvent 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.*

1.6.3 Coarticulation

Le signal de parole est constitué d'une succession d'unités différentes. Cependant, contrairement à ce qu'on pourrait croire, ces unités ne sont pas indépendantes les unes des autres mais s'influencent mutuellement : c'est le phénomène de coarticulation [9].

1.6.4 Redondance

Le signal de la parole est très redondant. Son traitement automatique nécessite, de réduire au maximum cette redondance afin de diminuer l'encombrement en mémoire et de limiter les durées du traitement, lequel doit se faire en temps réel. A l'inverse, le débit ne doit pas être trop faible pour conserver un bon rapport signal/bruit [8].

1.7 CLASSIFICATION DES SONS

D'un point de vue linguistique, 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. Le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air (Fig.1.4).

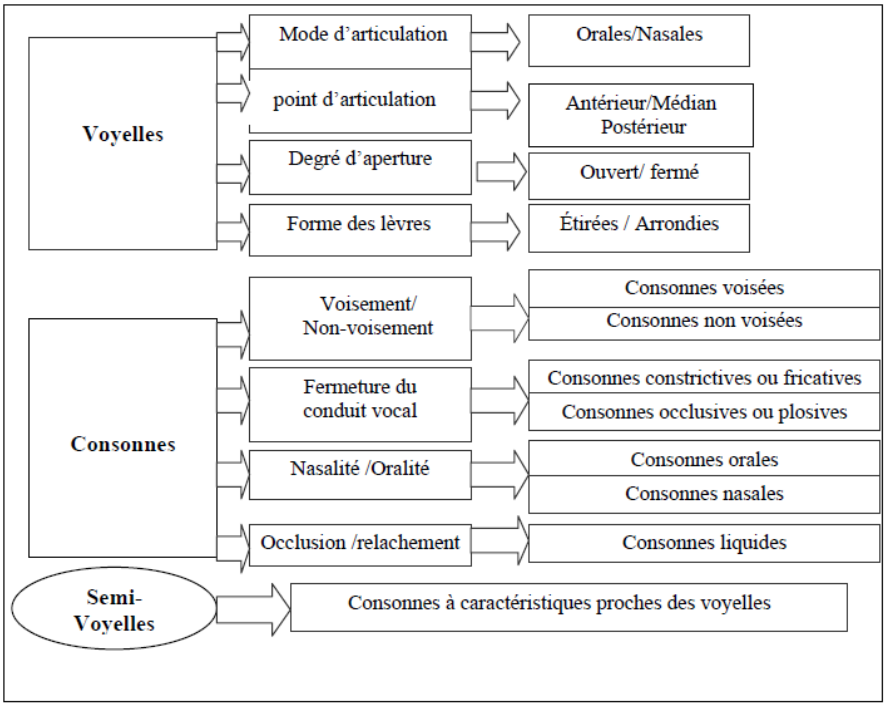

Figure 1.4 : Classification des sons du langage

1.7.1 Les sons voisés ou sonores

Les vibrations des cordes vocales produisent les sons voisés (voyelles, semi-voyelles, consonnes nasales…etc.). Les cartilages sur lesquels s'accrochent les cordes vocales régularisent la tension des cordes, donc la fréquence des vibrations, au moyen des muscles du larynx s'appelle la fréquence fondamentale ou $\mathbf{F_0}$.

1.7.2 Les sons non voisés ou sourds

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 ou de plosion, provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais. Dans cette catégorie de sons les cordes vocales ne vibrent pas.

1.7.3 Les voyelles

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Quand ce dernier est suffisamment ouvert pour que l'air expiré par les poumons, le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la cavité buccale se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il

se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. Une voyelle se caractérise par un passage libre de l'air dans le conduit vocal et par les vibrations des cordes vocales.

Elles se différencient principalement les unes des autres par leur lieu d'articulation (position de la langue), leur degré d'ouverture (espace compris entre la pointe de la langue et le palais), et leur nasalisation. Nous distinguons ainsi, selon la localisation de la masse de la langue, les voyelles antérieures ou avant, les moyennes, et les voyelles postérieures (ou arrières), et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles fermées et ouvertes (Fig1.5).

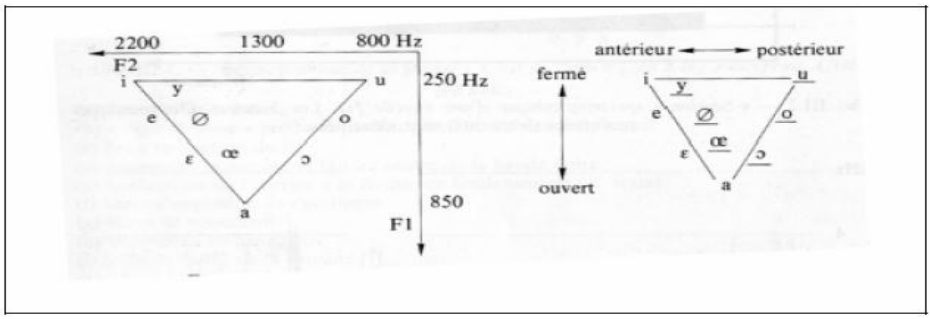

Figure 1.5 : Relation acoustico-articulatoire des voyelles orales du Français [10]

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.

1.7.5 Les semi-voyelles

Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement

soudain de cette position produit une friction qui est typique des consonnes. Enfin, elles sont assez difficiles à classer [8].

1.8 CONCLUSION

Dans ce chapitre nous avons exposé des notions de base sur le traitement de la parole, des spécifications du signal vocal.

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 sur systèmes de synthèse de la parole et ses variantes.

Chapitre II

Techniques et méthodes de la synthèse de la parole

2.1 INTRODUCTION

Ce chapitre nous permet de présenter les principales techniques de la synthèse de la parole, en premier lieu, nous allons donner une brève définition de la synthèse de la parole, et le, 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.

2.2 DEFINITION DE LA SYNTHESE 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 [9].

2.3 LE SYSTEME TEXT-TO-SPEECH (TTS)

Un Système de Synthèse à Partir du Texte (**TTS** : **T**ext-**T**o-**S**peech) 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** : **O**ptical **C**haracter **R**ecognition), 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 [1].

2.4 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.

2.4.1 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 [8].

2.4.2 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.

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.

2.4.2.1 Codage Prédictif Linéaire (LPC)

Cette méthode connus de la production sous le sigle LPC (Linear Predictive Coding) se base sur les connaissances de la production de la parole et suppose que le modèle de la production est linéaire (Fig.2.1).

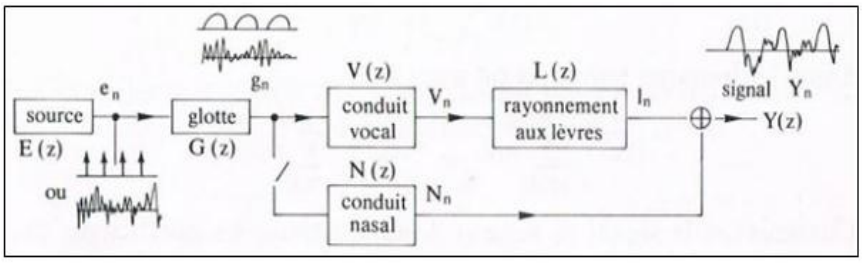

Figure 2.1 : Modèle général de production de la parole [10]

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 suivant : 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 $\mathbf{e_n}$ de ce filtre est un train d'impulsions de période T_0 pour les sons voisés ou un bruit blanc pour les sons non voisés (source).

Le modèle du conduit vocale est un filtre tout pôle (AR : autorégressif) d'ordre 2M décomposable en une cascade de résonateur à 2 pôles en série (tuyaux résonants). Le model du conduit nasal est un filtre pole zéro ARMA (auto régressif a moyen ajusté) et le rayonnement aux lèvres peux se modélisé par un filtre tout zéro (MA moyenne ajusté).

L'ensemble des conduits se comporte donc comme un système linéaire ARMA.

Modèle glottale :

$$G(z) = \frac{1}{(1-e^{-2\pi f_g T} z^{-1})^2} \qquad f_g = 100\ MHz \tag{2.3}$$

Modèle du conduit vocal :

$$V(z) = \prod_{i=1}^{M} \left(\frac{1}{1-2e^{-2\pi B_i T}.\cos(2\pi F_i T) z^{-1} + e^{-4\pi B_i T} z^{-2}}\right) \tag{2.4}$$

F_i : Fréquence du formant, B_i sa bande passante

Modèle du conduit nasal :

$$N(z) = \frac{1-2e^{-2\pi B'_N T}.\cos(2\pi F'_N T) z^{-1} + e^{-4\pi B_N T} z^{-2}}{1-2e^{-2\pi B_N T}.\cos(2\pi F_N T) z^{-1} + e^{-4\pi B_N T} z^{-2}} \tag{2.5}$$

Avec F_N et F'_N formant nasal et anti-formant nasal et respectivement, B_N et B'_N leurs bandes passantes [10].

2.4.2.2 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 F_0 et des formants.

Pour cela, nous faisons l'hypothèse que le signal vocal y_n est produit par le signal excitateur $u_n$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 que un 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.

$$Y(z) = B(z).U(z) \quad (2.9)$$

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

$$\log|Y(z)| = \log|U(z)| + \log|B(z)| \quad (2.10)$$

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 :

$$C(n) = FT^{-1}\{\log(FT\{y(n)\})\} \quad (2.11)$$

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 formantiques) est obtenue par une transformation supplémentaire

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) [8].

2.5 LES METHODES DE SYNTHESE DE LA PAROLE

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.

2.5.1 Synthèse Par Règles (SPR)

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].

2.5.2 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.

2.5.2.1 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 pré-enregistrées 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 [11].

2.5.2.2 Synthèse fondée sur l'algorithme PSOLA

L'algorithme **PSOLA** (**P**itch **S**ynchronous **O**ver**L**ap and **A**dd) 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 [7].

2.6 LES APPLICATIONS DE LA SYNTHESE DE 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 muettes, laryngectomisées ou 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 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.)[6].

2.7 INTERETS DE LA REPRESENTATION FREQUENTIELLE DU SIGNAL DE PAROLE

La représentation fréquentielle de la parole est d'une très grande importance dans le domaine de la Communication Parlée. Elle a permis l'extraction des paramètres pertinents du signal de parole comme la fréquence fondamentale, l'intensité et les formants. Ces paramètres sont d'une importance capitale dans de nombreux domaines comme :

- les différentes méthodes de synthèse ;
- la Reconnaissance Automatique de la Parole et du Locuteur ;
- l'Identification Automatique des Langues ;
- et bien d'autres domaines.

2.8 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 III

La méthode TDI-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 TDI-PSOLA et les méthodes de détection de pitch.

3.2 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 (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 (3.1)

3.2.1 Cas d'un signal voisé

Nous désirons obtenir un signal de synthèse de fréquence f(t). Pour ce faire nous plaçons les signaux élémentaires (représentant chacun une période fondamentale) distants de $\frac{1}{f(t)}$ Ceci determine alors le placement des marques d'écriture t_w^j.

Supposons connu t_w^{j-1}, on en déduit t_w^j de la manière suivante :

$$\forall\, j \quad , \qquad t_w^j = t_w^{j-1} + \frac{1}{t_w^i - t_w^{i-1}} \int_{t_w^{i-1}}^{t_w^i} \frac{1}{f(d)}\, dt \tag{3.1}$$

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 espacées d'une durée de l'ordre de 10ms.

Afin de connaitre la correspondance entre marques de lecture et d'écriture nous introduisons le temps dit de : "correspondance". t_{co}^{j} est la position correspondant a t_w^j sur le signal original (fig.1).

Il dépend non seulement de t_w^j mais également du facteur de dilatation d(t)

Supposons connus t_{co}^{j-1}, t_w^{i-1} et t_w^i , supposons de plus d(t) constant sur la largeur d'une période, on en déduit t_{co}^{j} de la manière suivante :

$$\forall\, j \quad , \qquad t_{co}^{j} = t_{co}^{j-1} + \frac{t_w^i - t_w^{i-1}}{d(t)} \tag{3.2}$$

3.2.2 Cas d'un signal non-voise

Dans le cas d'une zone non-voisée, les marques d'écriture et 1 es temps de correspondance sont places de manière équidistante. Leur position s'obtient à l'aide de (eq.3.2) et (eq.3.3) avec

$f(t) = \frac{1}{f_O(t)}$ .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.

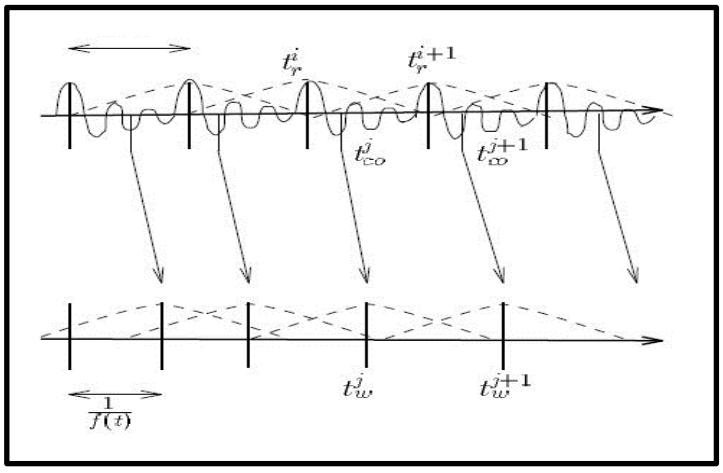

Figure 3.1 Placement les marques de lecture t_r^i et d'écriture t_w^i et les temps de correspondance t_{co}^i[12]

3.3 DETERMINATION DES SIGNAUX ELEMENTAIRES UTILISES

Les signaux $s_j(t)$ utilises dans (eg.2) sont déterminés à partir des temps de correspondance t_{co}^j calcules en (eq. 4). Nous distinguons ici les méthodes de TD-PSOLA, TDT-PSOLA et FDT-PSOLA selon l'association qui est faite entre marques d'écriture et temps de correspondance.

• Dans une approche simple (TD-PSOLA) nous prenons le $s_i(t)$ associe a la marque de lecture t_r^j la plus poche de t_{co}^j. Le signal $s(t)$ est alors forme de la somme des $s_j(t) = s_i(t_r^j - t_w^j)$

Cette méthode présente le désavantage d'introduire des discontinuités dans le signal de synthèse. En effet, pour des facteurs d(t) recopiée plusieurs fois avant passage à la forme d'onde suivante.

• Une solution plus sophistiquée, inspirée de la concaténation entre diphones [12], consiste à prendre les signaux élémentaires associes aux marques t_r^jet t_r^{j+1} encadrant t_{co}^j et à effectuer leur interpolation temporelle fréquentielle (FDI-PSOLA). (TDI-PSOLA) ou Le signal s(t) est alors forme de la somme des signaux interpoles. Lors d'une dilatation du son le signal passe alors continuellement d'une forme d'onde à une autre.

3.4.1 Interpolation temporelle TDI-PSOLA

L'interpolation temporelle s'effectue directement sur les formes d'onde et encadrant le temps de correspondance t_{co}^j

$$S_j(k) = (1-\alpha)S_i(k) + \alpha S_{i+1}(k) \tag{3.3}$$

$$\alpha = \frac{t_{co}^j - t_r^j}{t_r^{j+1} - t_r^j} \tag{3.4}$$

3.4.2 Interpolation fréquentielle FDI-PSOLA

L'interpolation s'effectue sur les spectres d'amplitude et de phase des deux signaux encadrant t_{c0}^j .

$\forall\, k \in [0, N]$:

$$\tilde{A}(k) = (1-\alpha)\tilde{A}_i(k) + \alpha\tilde{A}_{i+1}(k) \tag{3.5}$$

$$\varphi(k) = (1-\alpha)\varphi_i(k) + \alpha\varphi_{i+1}(k) \tag{3.6}$$

L'interpolation des spectres de phase pose cependant un problème. La phase étant estimée en valeurs principales, on ne peut interpoler les spectres directement (exemple de 1' interpolation de

$\varphi_i(k) = -\pi$, Avec $\varphi_{i+1}(k) = \pi$ donnant$\varphi(k) = 0$.

Il est nécessaire d'ajouter une procédure de "déroulement" de la phase.

Le déroulement de la phase s'avère une manipulation tout aussi risquée puisque n'ayant de pertinence que dans les zones fréquentielles voisées. Toute erreur dans la détection du voisement peut être fatale.

Notre solution consiste à dérouler seulement le spectre de phase du deuxième signal élémentaire $\varphi_{i+1}(k)$ De manière à ce que $\varphi_{i+1}(k) + 2n\pi - \varphi_i(k) < \pi$.

3.4.3 Dans les zones non-voisées

Afin de permettre la dilatation de ces zones tout en évitant Introduction d'une périodicité artificielle ("'phasy-effect"), le signal élémentaire $S_i(t)$ détermine par le t_r^i le plus proche de t_{c0}^i est décalé aléatoirement autour de sa position. Nous appliquons de plus une inversion alternative de l'axe de la fréquence de $S_i(t)$ lorsque celui-ci est répète successivement [12]. Cette méthode permet de conserver le spectre d'amplitude tout en brouillant le spectre de phase.

3.5 **OVERLAP-ADD**

La synthèse du signal est alors effectuée par Superposition/Addition des $S_i(t)$. Nous utilisons la méthode de Grifin et Lim des moindres carres [13]

$$s(t) = \frac{\sum_j s_j(t) h_j(t)(t - t_w(j))}{\sum_j h_j^2(t)(t - t_w^i)} \qquad (3.7)$$

3.6 COMPARAISON TD/TDI/FDI-PSOLA

La méthode TD-PSOLA quoique très simple d'application introduit dans le signal des discontinuités qui n'apparaissent pas avec les méthodes utilisant l'interpolation. Cela se traduit perceptivement par une certaine : "rugosité" du signal de synthèse. L'expérience suivante est réalisée afin de permettre la comparaison des signaux de synthèse issus des deux méthodes d'interpolation avec le signal original.

• La durée et la hauteur des signaux de synthèse sont gardées identiques à celles du signal original.

• Les formes d'ondes utilisées pour la synthèse sont le résultat de l'interpolation des signaux élémentaires en posant a = 0.5 (cas le plus critique).Les signaux sont comparés uniquement dans les zones voisées stables (la forme d'onde évolue très peu d'une période à l'autre).

D'un point de vue perceptif, l'interpolation fréquentielles préserve mieux la brillance du signal original comparée à l'interpolation temporelle. Ceci pondéré quelque peu [12]. L'analyse de l'enveloppe spectrale du signal (calcul de l'enveloppe par filtre LPC d'ordre 40) sur l'ensemble d'une zone stationnaire du signal corrobore notre impression (fig. 2). Ceci peut s'expliquer en considérant l'interpolation temporelle en termes de filtrage passe-bas (somme de deux signaux élémentaires quasi-identiques mais dont le marquage PT diffère par rapport au signal élémentaire respectif).

L'observation des spectres du signal original et des signaux interpoles (fig.3) nous montre que les deux interpolations donnent des résultats quasi-équivalents en basse fréquence. Aux fréquences élevées (> 4000MHz)) les spectres des signaux interpoles différent. Les raisons de cette divergence ne sont pas les mêmes pour les deux méthodes. En effet si l'influence d'une différence de marquage se traduit par un filtrage en TDI-PSOLA, en FDl-PSOLA celle-ci se répercute sur le spectre de phase. La rotation du spectre de phase entrainée rend leur interpolation difficile. Les interpolations fréquentielles et temporelles sont donc l'une comme l'autre extrêmement sensibles à la position des marques de lecture.

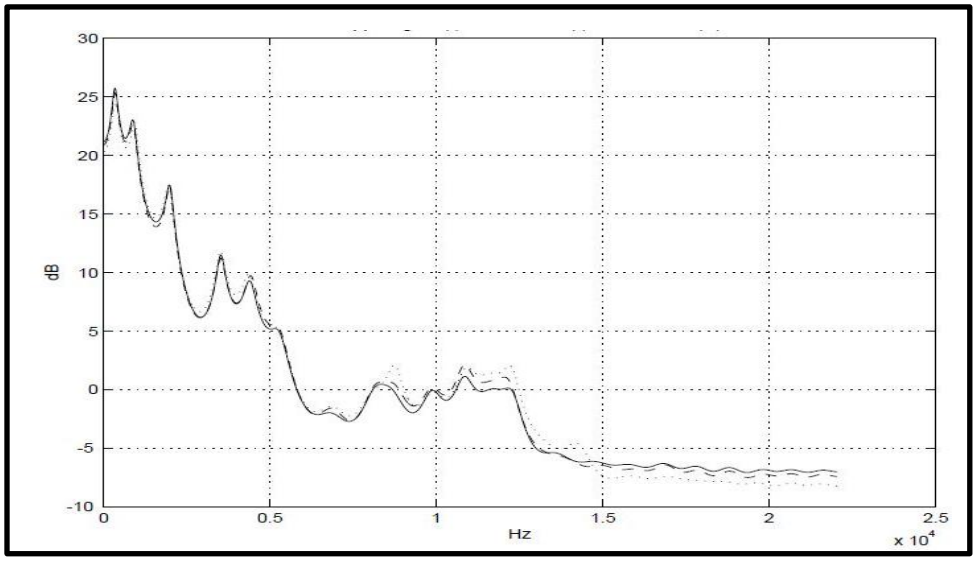

Figure 3.2 : Enveloppe signal : TDI-PSOLA /FDI-PSOLA [13]

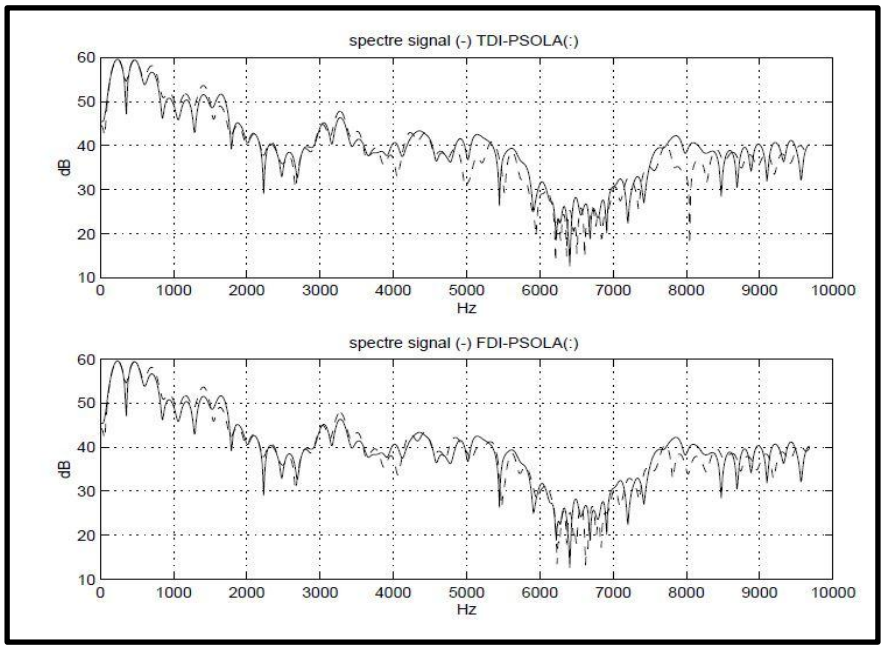

Figure 3.3 : Spectre signal : TDI-PSOLA /FDI-PSOLA [13]

3.7 CONCLUSION

La technique TDI-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.

Conclusions générales et perspectives

CONCLUSIONS GENERALES ET PERSPECTIVES

La caractéristique la plus remarquable de la technique TDI-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 TDI-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.

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 FDI-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

REFERENCES BIBLIOGRAPHIQUES

[1] T. Dutoit, Introduction au traitement automatique de la parole notes de cours /DEC2, Faculté Polytechnique de Mons, LCTS Lab, France, 2000.

[2] http : //www.infovisual.info.

[3] J. Cisonni, Modélisation et inversion d'un système complexe de production de signaux acoustiques Application à la voix et aux pathologies, Thèse de Doctorat, Institut Polytechnique De Grenoble, France, Novembre 2008.

[4] http://users.skynet.be/illusionsauditives/images/page802.gif

[5] L'étude instrumentale des gestes dans la production de la parole ; importance de l'aérophonométrie ; Manuscrit auteur, publié dans "Les Dysarthries, P. Auzou (Ed.) (2007) 115-117

[6] O. Amine, Synthèse de la parole en arabe standard, Mémoire de Magister, ENP, Alger, Algérie, 2011.

[7] G. Djeghiour, thèse magister : application des réseaux de neurones à la synthèse de la parole en arabe standard, Ecole nationale supérieure des sciences humaines ; 2011.

[8] A. Mohamed, Application des Algorithmes Génétiques au Décodage Acoustico-Phonétique de la parole en Arabe Standard, Thèse de Doctorat, ENP, Alger 2008

[9] V A.Dubesset, La Langue française Parlée Complétée (LPC) : Production et Perception, Thèse de Doctorat, Institut National Polytechnique De Grenoble, France, 2005

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

[11] E. Moulines, O. Cappé, synthèse de la parole à partir du texte Techniques de l'ingénieur, Vol. H1 960, p.6-7.

[12] F. Charpentier and E. Moulin Text-to-speech algorithms based on FFT synthesis , ICASSP ,1988.

[13] D. Griffin and J. Lim signal Estimation from modified short time Fourrier transform,IEEE Trans,ASSP vol ASSP-32 1984. p.236-243