UNIVERSITE SAAD DAHLAB DE BLIDA

Faculté des Sciences de l'ingénieur

Département d'Electronique

MEMOIRE DE MAGISTER

Spécialité : Image et Parole

MODIFICATIONS DE LA FREQUENCE FONDAMENTALE EN VUE DE LA SYNTHESE DE LA PAROLE A PARTIR DU TEXTE DE L'ARABE STANDARD

Par

Fayçal YKHLEF

Devant le jury composé de:

H. SALHI	Maître de Conférence, U. de Blida	Président
A. GUESSOUM	Professeur, U. de Blida	Examinateur
M. AREZKI	Chargé de Cours, U. de Blida	Examinateur
M. GUERTI	Maître de Conférence, E.N.P., Alger	Rapporteur
A. CHENTIR	Chargé de Cours, U. de Blida	Co-rapporteur

Blida, Septembre 2005

RESUME

Le but de notre travail est l'étude des techniques de modifications des paramètres prosodiques utilisées pour des signaux de parole en vue d'une synthèse de la parole à partir du texte (TTS : Text-To-Speech) de haute qualité. Nous nous sommes intéressés dans notre cas à la modification de la fréquence fondamentale. Pour ce faire, nous avons implémenté deux techniques de modifications, TD-PSOLA et la modélisation source conduit vocal par prédiction linéaire (LPC). Nous avons utilisé un corpus constitué de signaux de parole en Arabe Standard prononcés par un locuteur masculin, ensuite, nous avons fait une évaluation des techniques utilisées, globale et analytique, afin de tirer les meilleurs facteurs de modifications offrant une bonne qualité synthétique.

Mots clés: Prosodie, Fréquence fondamentale, TAP, LPC, TD-PSOLA, TTS (Text-To-Speech)

ABSTRACT

The aim of our work is to make a study of the prosodic modifications techniques used for Text To Speech Synthesis with high quality. In our case, we are interested by the PITCH modifications. With this intention, we implemented two techniques of modifications, TD-PSOLA and the source vocal tract modelling by the linear predictive Coding (LPC). We have used Standard Arabic speech signals pronounced by a masculine speaker. Then we made an evaluation of the techniques used, global and analytical, in order to get the best modifications factors offering a good synthetic quality.

Key words: Prosody, Fundamental frequency, ASP, LPC, TD-PSOLA, Text-To-Speech

ملخص

الهد ف من مشروعنا هو دراسة تقنيات تغيير العناصر العروضية المستعملة في إشارات الكلام لهدف التركيب الاصطناعي عن طريق النص و الحصول على كلام ذو جودة عالية.

اهتممنا في مشروعنا بتغيير التردد الابتدائي, ولذالك قمنا بتصميم طريقتين تغيير, الأولى تسمى TD-PSOLA و الثانية عبارة عن نموذج مصدر قناة – صوتية عن طريق تقنية LPC.

استعملنا إشارات كلام ملفوظة عن طريق متكلم ذكر باللغة العربية الأصيلة, ثم قمنا بعملية تقييم للطرق المستعملة, إجمالية و تحليلية لهدف اختيار أحسن عوامل تغيير لكلام ذو جودة حسنة.

كلمات مفاتيح : العروض , التردد الابتدائي, المعالجة الاوتوماتيكية للكلام, التنبؤ الخطي (LCP), TD-PSOLA, الكلام الاصطناعي عن طريق النص.

REMERCIEMENTS

Mes remerciements et ma gratitude se portent vers M^me^ M. GUERTI Maître de Conférence au département d'Electronique à l'Ecole Nationale Polytechnique ENP Alger, qui m'a encadré et guidé pendant ce travail et qui a su m'orienter vers les axes les plus pertinents du Traitement Automatique de la Parole. Je la remercie pour ses compétences, son ouverture d'esprit et sa grande disponibilité.

Je tiens également à exprimer toute ma gratitude et ma reconnaissance à ma Co-Directrice M^me^ A. CHENTIR chargée de cours au département d'Electronique à l'USD de Blida pour m'avoir assisté tout le long de ce travail. Je la remercie également pour ses nombreux conseils et remarques qui m'ont été d'une grande utilité.

Je remercie chaleureusement M^r^ H. SALHI, Maître de Conférence au département d'Electronique à l'USD de Blida de me faire un grand honneur en acceptant de présider le jury de ce mémoire.

Je remercie M^r^ A. GUESSOUM, Professeur au département d'Electronique à l'USD de Blida de l'intérêt qu'il ma clairement manifesté pour ce travail, et des remarques et corrections qu'il ma apporté à ce document.

Je tiens à exprimer mes remerciements à M^r^ M. AREZKI chargé de cours au département d'Electronique à l'USD de Blida de m'avoir accueilli au sein de son équipe, et de m'avoir apporté un bon environnement de travail. Je souhaite lui apporter mes plus vifs remerciements pour l'attention et remarques qu'il ma apporté à ce travail.

Je désire remercier également les collègues du Laboratoire de Traitement de la Parole du département d'Electronique USD de Blida pour l'aide précieuse qu'ils m'ont apportée pendant toute la durée de ce mémoire.

TABLE DES MATIERES

LISTES DES ILLUSTRATIONS, GRAPHIQUES ET TABLEAUX

INTRODUCTION

La parole constitue le moyen le plus naturel de communiquer entre les êtres humains. Les études se rapportant à elle ont retrouvé leur piédestal grâce à l'avènement des télécommunications et du traitement numérique du signal.

L'importance particulière du Traitement Automatique de la Parole (TAP) d'une manière générale s'explique par la position privilégiée de la parole comme vecteur d'informations dans notre société humaine.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée. Aucun signal, pourtant fort complexe, n'est cependant à la fois appelé à être *produit* et *perçu* instantanément par le cerveau, comme c'est le cas pour la parole. La parole est en effet produite par le conduit vocal, contrôlé en permanence par le cortex moteur.

L'étude des mécanismes de phonation permet donc de déterminer, dans une certaine mesure, ce qui est parole et ce qui n'en est pas. De même, l'étude des mécanismes d'audition et des propriétés perceptuelles qui s'y rattachent permet de dire ce qui, dans le signal de parole, est réellement perçu. Mais l'essence même du signal vocal ne peut être cernée de façon réaliste que dans la mesure où l'on imagine, bien au-delà de la simple mise en commun des propriétés de production et de perception de la parole, les propriétés du signal dues à la mise en boucle de ces deux fonctions.

Mieux encore, c'est non seulement la perception de la parole qui vient influer sur sa production par le biais de ce bouclage, mais aussi et surtout sa compréhension. On ne parle que dans la mesure où l'on s'entend et où l'on se comprend soi-même ; la complexité du signal qui en résulte s'en ressent forcément. S'il n'est pas en principe de parole sans cerveau humain pour la produire, l'entendre, et la comprendre, les techniques modernes de TAP

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

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

On comprend aisément que, pour obtenir de bons résultats dans chacune de ces tâches, il faut tenir compte des caractéristiques du signal étudié. Lorsque nous parlons, nous ne sommes en général pas conscients des mouvements complexes des muscles de la phonation, et il en va de même en particulier pour le contrôle de la hauteur et de l'intensité de la voix lors des vibrations des cordes vocales. Ces deux paramètres auxquels on joint habituellement les durées successives des segments syllabiques constituent en leur évolution la prosodie de la phrase. Ces paramètres prennent une importance particulière dans le TAP. En synthèse, ils améliorent le naturel et l'intelligibilité du signal synthétique en signalant les grandes articulations de la phrase ; en reconnaissance, ils peuvent servir d'indices pour l'identification d'éléments segmentaux déterminés et également signaler le type syntaxique de la phrase.

Notre travail s'insère dans le domaine de la synthèse de la parole à partir du texte (Text -To- Speech : TTS). 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. Les synthétiseurs reproduisent une voix tout à fait intelligible, mais les intonations et l'expressivité ne sont pas encore au point. De ce fait, une méthode de synthèse doit garantir une grande flexibilité de modifications au niveau suprasegmental et offrir ainsi la simplicité des variations prosodiques (la fréquence fondamentale ou F_0, l'énergie et la durée).

L'objectif principal de ce mémoire est d'élaborer un système de synthèse de la parole capable d'effectuer des modifications prosodiques et plus précisément des modifications de la fréquence fondamentale.

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

- le premier chapitre introduit les principes du TAP et une étude sur les principaux paramètres spécifiques à la langue Arabe Standard. Grâce à cette étude préalable, nous pouvons faire une base enrichissante sur ce domaine ;
- le second chapitre est consacré aux techniques de synthèse de la parole à partir du texte et à une étude sur les principales fonctions de la prosodie, suivie par une description de la manière d'évaluation des systèmes de synthèse de la parole à partir du texte en citant quelques méthodes d'évaluations ;
- dans le troisième chapitre, deux techniques de modifications du paramètre F_0 vont être développées, l'une est basée sur la modélisation physique par la prédiction linéaire (*Linear Predictive Coding*) et l'autre sur une nouvelle technique de synthèse appelée TD-PSOLA ;
- le dernier chapitre expose la mise en œuvre de notre application, l'interprétation des résultats obtenus ainsi que leurs évaluations ;

Nous terminons notre travail par des conclusions générales et des perspectives interprétant les différents résultats obtenus.

CHAPITRE 1
NOTIONS SUR LE TRAITEMENT AUTOMATIQUE DE LA PAROLE ET L'ARABE STANDARD

1.1. Introduction

Le traitement du signal de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et celui du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. Ce premier chapitre porte sur une analyse descriptive des différents niveaux non exclusifs des sons du langage (Français et Arabe), tout en exposant les principaux paramètres spécifiques à la langue Arabe Standard.

1.2. Généralités sur le Traitement Automatique de la Parole

La parole est la faculté de communiquer la pensée par la voix, c'est le moyen le plus privilégié entre les humains. L'information d'un message parlé réside dans les fluctuations de la pression de l'air engendrées puis émises, par l'appareil phonatoire. Ces fluctuations constituent le signal vocal. Elles sont détectées par l'oreille, laquelle procède à une certaine analyse. L'information portée par le signal de parole peut être analysée selon plusieurs techniques. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique, et pragmatique [1].

1.2.1. Niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur). Le signal résultant est le plus souvent numérisé. L'opération de numérisation, schématisée à la figure (1.1), requiert successivement : un filtrage de garde, un échantillonnage, et une quantification.

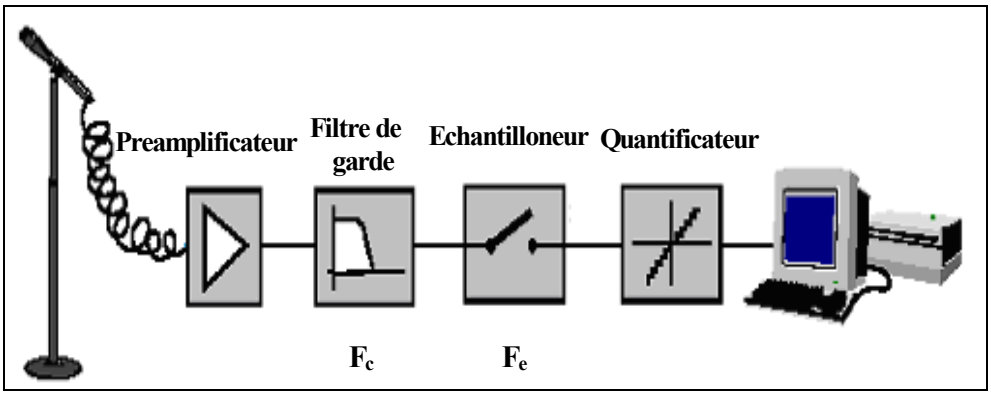


Figure 1.1 : Enregistrement numérique d'un signal acoustique.

1.2.1.1. Audiogramme

L'échantillonnage transforme le signal continu x(t) en signal discret x(nTe) défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage T_e; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage F_e. Pour ce qui concerne le signal vocal, le choix de F_e résulte d'un compromis. Son spectre peut s'étendre jusqu'à 12 kHz ; il faut donc en principe choisir une fréquence F_e égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon. Cependant, le coût d'un traitement numérique, filtrage, transmission, ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure F_c est choisie en fonction de la F_e retenue. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit Fe égale à 8000 Hz.

Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 8000 à 16000 Hz. Par contre pour le signal audio (parole et musique), on exige une bonne représentation du signal jusque 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées : 22.5 kHz, 11.25 kHz.

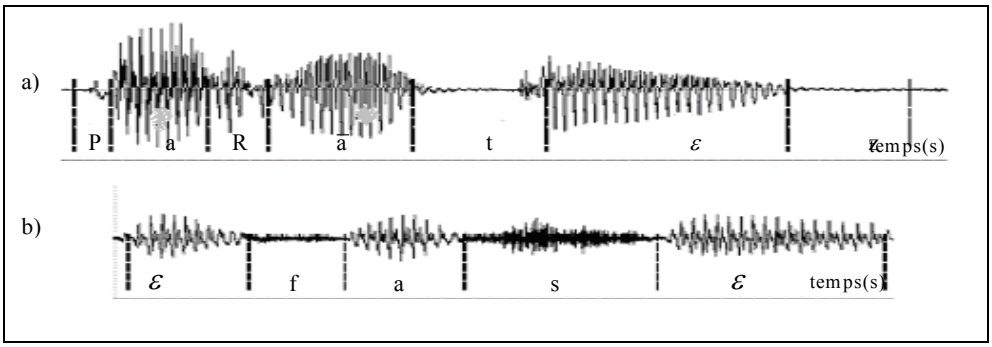


Figure 1.2 : Audiogrammes de signaux de parole des mots.
a) : parenthèse
b) : effacer

Nous constatons sur la représentation temporelle d'un signal vocal une alternance de zones assez périodiques et de zones bruitées, appelées successivement zones Voisées et Non Voisées (Figure 1.2). La figure (1.3) donne une représentation plus fine de tranches de signaux Voisés et Non Voisés.

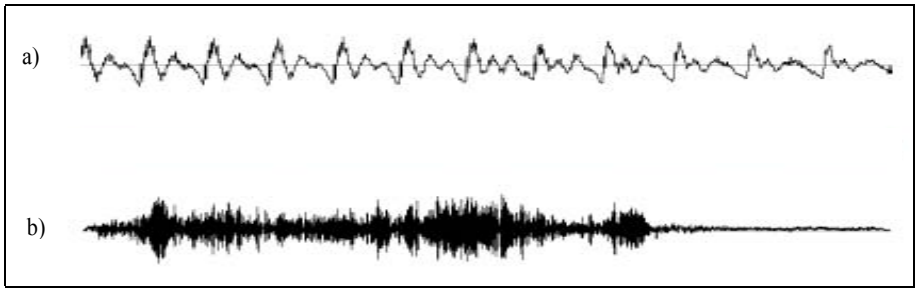


Figure 1.3 : Exemples de son Voisé (a) et Non-Voisé (b).

1.2.1.2. Paramètres prosodiques d'un signal de parole

Du point de vue acoustique, la prosodie désigne les phénomènes liés à la variation dans le temps des paramètres de hauteur, d'intensité et de durée. La hauteur est essentiellement liée à la fréquence fondamentale F_0 qui correspond, au niveau physiologique de la production de la parole, à la fréquence de vibrations des cordes vocales. L'intensité est essentiellement liée à l'amplitude et à l'énergie du son, mais dépend aussi partiellement de sa durée. La durée, correspond à son temps d'émission, sa durée acoustique. Si nous nous plaçons sur le plan perceptuel, la variation dans le temps de ces derniers correspond respectivement à la perception de la mélodie des phrases, de leur accentuation et de leur rythme.

La mélodie de la phrase correspond à l'évolution dans le temps de la hauteur. L'accentuation est un phénomène de plus haut niveau, qui consiste à mettre en relief une syllabe ou une more[1] par rapport à son environnement immédiat. Le rythme des phrases est perçu grâce à l'enchaînement des durées des segments. L'étude acoustique d'un signal de parole correspond à l'évaluation de ses paramètres prosodiques. Les modifications apportées à l'un d'eux peuvent altérer indéniablement les autres paramètres. Cependant, si nous voulons étudier ces paramètres d'un point de vue acoustique, nous pouvons les considérer comme étant parfaitement indépendantes.

1.2.1.2.1. Fréquence fondamentale

La fréquence la plus basse dans le signal de parole est la fréquence fondamentale F_0 de vibrations des cordes vocales (pitch), elle évolue lentement dans le temps. Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants. Les variations de 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 fréquence fondamentale. La figure (1.4) donne l'évolution temporelle de la fréquence fondamentale de la phrase "les techniques de traitement numérique de la parole". On constate qu'à l'intérieur des zones Voisées la F_0 évolue lentement dans le temps. La fréquence est donnée sur une échelle logarithmique; les sons Non-Voisés sont associés à une fréquence nulle.

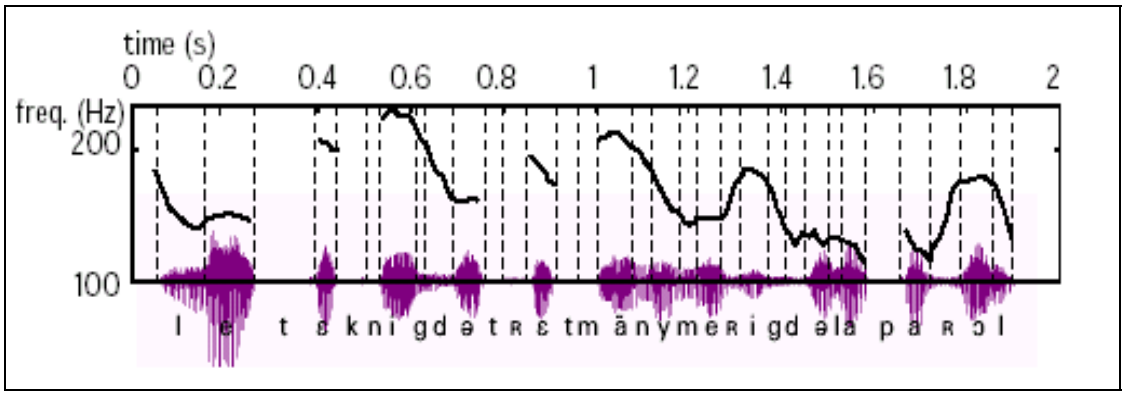


Figure 1.4 : Evolution de la fréquence de vibrations des cordes de la phrase : "les techniques de traitement numérique de la parole". [1]

[1] Une more est une unité inférieure à la syllabe et qui peut porter l'accent

1.2.1.2.2. Energie

L'énergie (intensité) E d'un son correspond à l'amplitude de la vibration acoustique. En d'autres termes, elle caractérise le volume sonore qui nous permet de distinguer un son fort d'un son faible. C'est le paramètre prosodique le plus facile à calculer, il peut être exprimé pour un signal échantillonné $x_n|_{n=1...T}$ et à support fini T par :

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

Etant donné sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

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

Pour un signal échantillonné quelconque, on calcule l'énergie à court terme en prenant des portions de signal convoluées avec une fenêtre glissante (généralement assez étroite, de l'ordre de 5 à 10ms). Pour éliminer la variabilité du gain (dû à des conditions d'enregistrements différentes, par exemple à la distance entre le locuteur et le microphone), l'énergie peut être normalisée par rapport a l'échantillon maximal sur la phrase.

1.2.1.2.3. Durée

La durée d'un signal correspond à son temps d'émission, sa durée acoustique. C'est le paramètre le plus difficile à préciser car rien n'indique comment le système de contrôle, de production ou de perception de parole mesure le temps. Les indices de durée classiques supposent généralement la donnée d'une segmentation, des frontières des unités dont on désire mesurer la durée : la durée d'une unité est alors mesurée par le nombre de trames qui séparent ses frontières de début et de fin. La plupart des systèmes utilisent une segmentation basée sur le phonème.

1.2.1.3. Représentation spectrale

La forme générale du spectre d'un signal vocale, appelée enveloppe spectrale, présente des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal (cavité orale et cavité nasale) et sont appelés formants et anti-formants.

L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son [2]. Il apparaît en pratique que l'enveloppe spectrale des sons Voisés est de type passe bas, avec environ un formant par kHz de bande passante, et dont

seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons Non Voisés présentent souvent une accentuation vers les hautes fréquences.

1.2.1.3.2. Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme à deux dimensions temps-fréquence.

On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération (Figure.1.5).

Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes Voisées apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont aussi utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones Voisées apparaissent sous la forme de bandes horizontales [3].

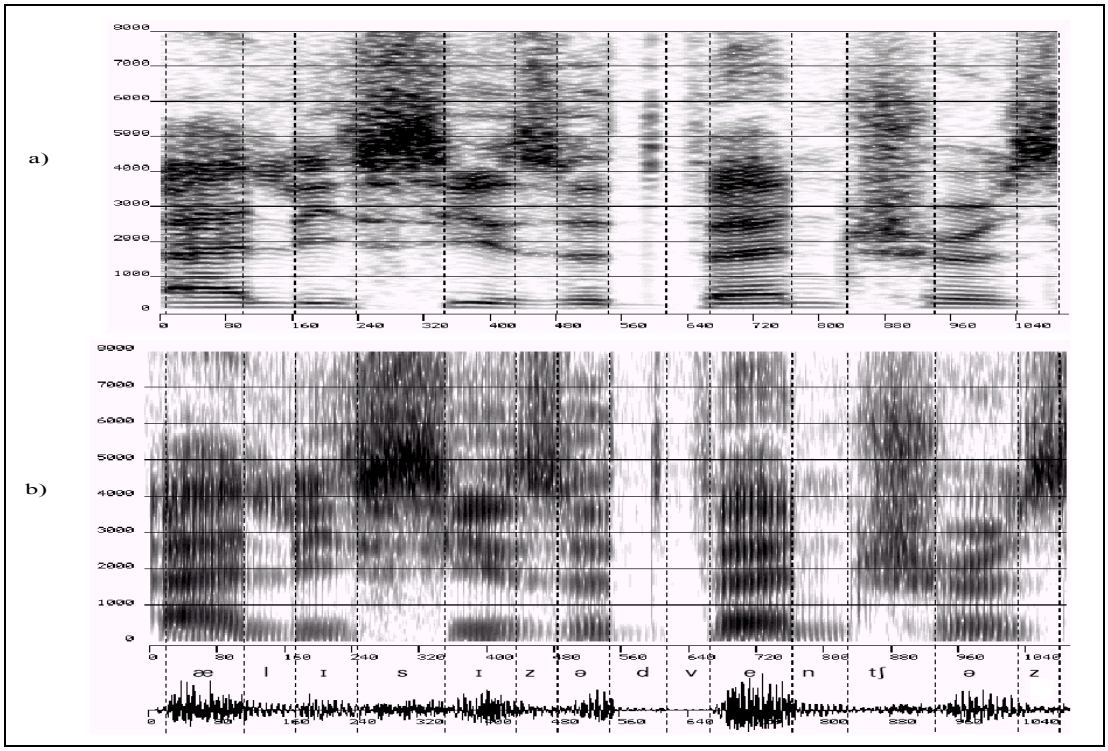


Figure 1.5 : Spectrogrammes et évolution temporelle de la phrase 'Alice's adventures'
a) : Spectrogramme a bande étroite.
b) : Spectrogramme a bande large.

1.2.2. Le niveau phonétique

Au contraire des acousticiens, ce n'est pas tant le signal qui intéresse les phonéticiens que la façon dont il est produit par le système articulatoire.

1.2.2.1. Phonation

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques. L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx ou la pression de l'air est modulée avant d'être appliquée au conduit vocal.

Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. 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 passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons Non-Voisés (ou sourds). Les sons Voisés (ou sonores) résultent au contraire de vibrations périodiques 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; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons.

Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le lieu d'articulation, région de rétrécissement maximal du canal buccal, ainsi que l'aperture, écartement des mâchoires au point d'articulation [4].

1.2.2.2. Transcription Orthographique Phonétique

La recherche en TAP et notamment en synthèse dans une langue donnée doit nécessairement passer par l'étude de la composante phonétique de cette langue. Cette étude nous permettra de dégager les principales caractéristiques relatives aux différents

phonèmes et aussi de cerner l'ensemble des paramètres acoustiques et physiologiques, en vue de les exploiter dans l'élaboration d'un système de synthèse de la parole.

La TOP, ou phonétisation, est une étape essentielle pour la synthèse à partir du texte. Elle consiste à produire la prononciation correspondant au texte en entrée sous la forme d'une liste de phonèmes. La complexité de cette tâche varie selon la langue traitée et selon l'application à laquelle est destinée. Par exemple, la langue Française est difficile à transcrire en raison de sa forme orthographique qui est différente de sa forme phonétique, contrairement à l'Arabe où la correspondance entre les graphèmes et les phonèmes est quasi-biunivoque.

Les phonéticiens symbolisent les sons du langage au moyen de signes divers auxquels on attribue une valeur conventionnelle. Diverses transcriptions phonétiques sont utilisées selon les auteurs, celle qu'on a choisie est IPA l'alphabet Phonétique International des sons Français (Tableau 1.1).

Tableau 1.1 : Transcription Orthographique Phonétique des sons Français.

Phonèmes	Exemples TOP	Phonèmes	Exemples TOP
[i]	l**i**t [li]	**[f]**	**f**ou [fu]
[y]	l**u** [ly]	**[s]**	**s**ous [su]
[u]	l**ou**p [lu]	**[∫]**	**ch**ou [∫ u]
[e]	le**s** [le]	**[m]**	**m**ou [mu]
[ø]	bl**eu** [bl ø]	**[n]**	**n**ous [nu]
[o]	l'**eau** [lo]	**[ŋ]**	**agn**eau [a ŋ o]
[ɛ]	l**ai**t [l ɛ]	**[v]**	**v**ous [vu]
[œ]	p**eu** [p **œ**]	**[z]**	bi**s**ou [bizu]
[ɔ]	l'**o**r [lɔ R]	**[3]**	**j**oue [3u]
[ɑ]	la [l **ɑ**]	**[l]**	**l**ou [lu]
[ə]	Petit**e** [petit **ə**]	**[R]**	**r**ou [Ru]
[$\tilde{\varepsilon}$]	br**ui** [br $\tilde{\varepsilon}$]	**[w]**	**oi**seau [wazo]
[$\tilde{\text{œ}}$]	br**un** [br$\tilde{\text{œ}}$]	**[Ч]**	n**ui**re [nЧR]
[$\tilde{\text{ɑ}}$]	mam**an** [mam $\tilde{\text{ɑ}}$]	**[j]**	trav**ail** [tRav j]
[$\tilde{\text{ɔ}}$]	b**on**jour [b$\tilde{\text{ɔ}}$juR]	**[b]**	**b**eau [bo]
[p]	**p**ou [pu]	**[d]**	**d**oux [du]
[t]	**t**out [tu]	**[g]**	**g**oût [gu]
[k]	**c**oup [ku]		

1.2.2.3. Phonétique articulatoire

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur mode articulatoire [1]. On distingue généralement deux classes principales : les voyelles (orales [i, y, u, e, ø, o, ɛ , œ, ɔ, ɑ] et nasales [$\tilde{\varepsilon}$, $\tilde{\text{ɑ}}$, $\tilde{\text{œ}}$, $\tilde{\text{ɔ}}$]) et les consonnes.

1.2.2.3.1. Voyelles

Diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal (et non, comme on l'entend souvent dire, par le degré d'activité des cordes vocales, déjà mentionné sous le terme de Voisement).

Si le conduit vocal est suffisamment ouvert pour que l'air issu des 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. La cavité buccale est dans ce cas un organe de production à part entière.

Les voyelles se différencient principalement les unes des autres par leur lieu d'articulation, leur aperture (ouverture), et leur nasalisation. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles antérieures, les voyelles moyennes, et les voyelles postérieures, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles fermées et ouvertes.

Les voyelles nasales diffèrent des voyelles orales selon que le voile du palais est abaissé pour leur prononciation, ce qui met en parallèle les cavités nasale et buccale.

Notons que, dans un contexte plus général que celui de la seule langue Française, d'autres critères peuvent être nécessaires pour différencier les voyelles, comme leur labialisation, durée, tension, stabilité, leur glottalisation, voire même la direction du mouvement de l'air [1].

Les paramètres prosodiques du signal de parole sont évidemment liés à sa production. L'intensité du son est liée à la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent.

Son spectre résulte du filtrage dynamique du signal glottique (impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses. Ainsi, par exemple, on peut approximativement représenter les voyelles dans le plan des deux premiers formants (Figure.1.6).

On observe en pratique un certain recouvrement dans les zones formantiques correspondant à chaque voyelle (dû à la variabilité du signal vocal).

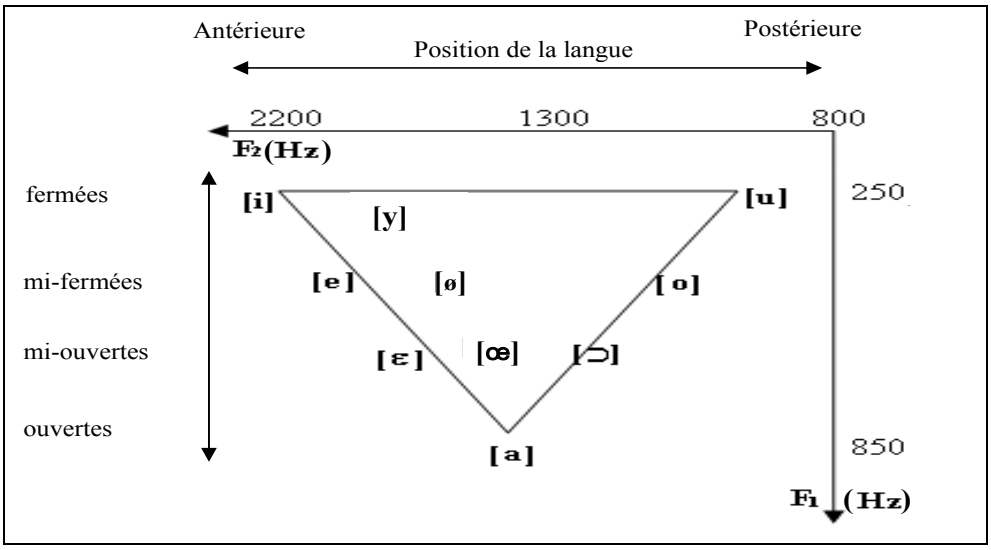


Figure 1.6 : Représentation acoustico-articulatoire des voyelles orales françaises [Triangle vocalique de Pierre Delattre, 1948].

1.2.2.3.2. Consonnes

On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation. Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général : l'organe articulatoire, la source sonore, l'intensité, l'aspiration, la palatalisation, et la direction du mouvement de l'air.

- Mode d'articulation

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré [5] :

- intervention des cordes vocales ou mise en vibration: articulation sonore ;
- fermeture momentanée du passage de l'air suivie d'une ouverture brusque (explosion): articulation occlusive ;
- rétrécissement du passage de l'air qui produit un bruit de friction ou de frôlement (articulation fricative) ;
- position abaissée du voile du palais: articulation nasale ;
- contact de la langue au milieu du canal buccal; l'air sort des deux côtés: articulation latérale ;
- une série d'occlusions brèves et séparées de la luette: articulation vibrante.

- Lieu d'articulation

Le lieu d'articulation est la zone du conduit vocale qui participe à la formation du son. Il présente la position de la constriction totale (cas des occlusives) ou partielle (cas des fricatives) d'une zone spécifique du conduit vocal lors du passage de l'airs provenant des poumons [6]. Le lieu d'articulation peut être bilabiale, glottale, labiodentale, etc (Tableau 1.2). En français, la distinction de mode d'articulation conduit à deux classes : les fricatives (ou constrictives) et les occlusives (ou plosives) [1].

- Les fricatives

Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais [∫ , 3], les dents [s, z], ou les lèvres [f, v]. Les fricatives Non Voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives Voisées combinent des composantes d'excitation périodique et turbulente : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète.

- Les occlusives

Les occlusives correspondent quant à elles, à des sons essentiellement dynamiques. Une forte pression est crée en amont d'une occlusion maintenue en un certain point du conduit vocal (qui peut ici aussi être le palais [k, g], les dents [t, d], ou les lèvres [p, d]), puis relâchée brusquement. La période d'occlusion est appelée la phase de tenue.

Les occlusives Voisées [b, d, g] sont tout d'abord émise par un son base fréquence grâce à la vibration des cordes vocales pendant la phase de tenue qui sera ensuite suivie par une phase d'explosion. La phase de tenue pour les occlusives Non Voisées [p, t, k] est un silence.

- Les consonnes nasales

Les consonnes nasales [m, n, ŋ] font intervenir les cavités nasales par abaissement du voile du palais.

- Les semi-voyelles

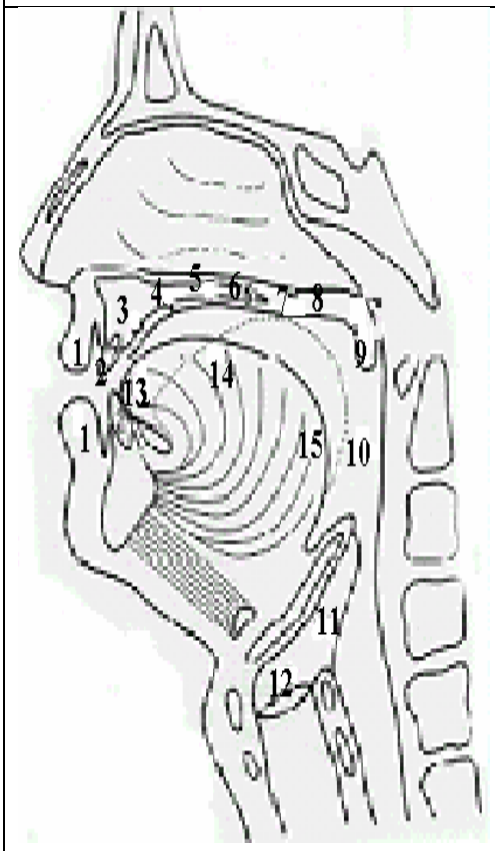
[j] [w] [Ч] ce sont des phonèmes intermédiaires entre les voyelles et les consonnes. Quand on les prononce, on entend le timbre d'une voyelle auquel s'ajoute le frottement

d'une consonne spirante (Constrictive). Leur fréquence d'emploi est liée à la vitesse du débit de la parole, plus celui-ci est rapide, plus il y aura de semi-voyelles.

- Les liquides

[l, R] sont assez difficiles à classer. L'articulation de [l] ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal. Le son [R], quant à lui, admet plusieurs réalisations fort différentes.

Tableau 1.2 : Principaux lieux d'articulation phonémique.

		Organe anatomique		Nomenclature phonétique correspondante	
	1	lèvres		Labiales	
	2	dents		Dentales	
	3	alvéole		Alvéolaire	
	4	palais dur		pré-palatales	
	5			médio-palatales	
	6			post-palatales	
	7	voile du palais		pré-vélaires	
	8			post-vélaires	
	9	luette(uvula)		Uvulaires	
	10	pharynx		Pharyngales	
	11	larynx		Laryngales	
	12	glotte		Glottales	
	13	apex	de la langue	apicales (pre- dorsales)	dorsales
	14	Dos		médeo-dorsales	
	15	Racine		radicales (post- dorsales)	

1.2.2.4. Timbre de la voix

Les courbes donnant par bandes de fréquences les niveaux sonores des sons Voisés mettent en évidence pour chacune d'elles, l'existence des formants. Deux d'entre eux sont toujours prépondérants, mais il est possible d'en définir un troisième de niveau beaucoup moins élevé. Il est fondamental de remarquer que, d'un individu à l'autre, l'intensité

relative et la fréquence d'un formant peuvent changer beaucoup, sans pourtant altérer la reconnaissance du son.

Le timbre de la voix est la couleur du son à partir de la quelle on peut identifier une personne à la simple écoute de sa voix. Il permet aussi de différencier deux sons de même hauteur et de même amplitude. C'est ainsi que l'on reconnaîtra, à l'oreille, deux instruments de musique jouant une même note ou deux personnes qui parlent. Le timbre d'un son dépend du nombre et de l'intensité des harmoniques contenues dans ce son.

Le timbre de la voix est fonction de trois critères :

- des conditions d'accolement des cordes vocales ;
- de leur épaisseur ;
- et enfin, des caractéristiques anatomiques des cavités de résonance (pharynx, cavités buccales et nasales) ;

L'accolement des cordes vocales peut être plus au moins ferme. Plus il est ferme, plus que le timbre vocal est riche en harmonique. Par ailleurs, selon que les ouvertures glottiques se font plus au moins rapidement, le spectre sonore est plus riche et inversement.

Les cavités de résonance contribuent aussi à la couleur de la voix. En modifiant le volume des cavités en obtient telle ou telle voyelle.

1.2.3. Niveau phonologique

Phonétique et phonologie sont deux sciences qui ont le même but : l'étude du langage. Ce qui les différencie, c'est le point de vue descriptif. La phonétique s'intéresse à la manière dont les sons du langage sont produits, transmis, perçus par des locuteurs [4], alors que la phonologie étudie les sons du point de vue des différences et des ressemblances phoniques fonctionnelles dans la langue c'est-à-dire du point de vue des différences et des ressemblances pertinentes pour la communication. A titre d'exemple il existe une différence fonctionnelle entre les sons [p] et [b] dans pain / bain : une différence de sens mais pas de différence pertinente entre les deux réalisations vibrante.

1.2.4. Niveau syntaxique

Du point de vue de la langue, la syntaxe est l'ensemble des règles écrites ou orales contraignant l'ordre des mots dans la phrase qui sont souvent appelés règles grammaticales. Les grammaires ne servent pas qu'à dresser la frontière entre les phrases régulièrement constituées ou pas : elles permettent également de décrire l'organisation hiérarchique des phrases, leur structure syntaxique.

Dans un système de compréhension, le but de la syntaxe est de réduire le nombre de phrases autorisées à partir du vocabulaire choisi. Par exemple, on peut construire 250^8 ($\sim 10^{19}$) phrases de 8 mots à partir d'un vocabulaire de 250 mots, mais seules 10^7 d'entres elles environs ont un sens. On a donc divisé l'espace de recherche par 10^{12}.

1.2.5. Niveau sémantique

Si la syntaxe restreint l'ensemble des phrases acceptables pour une langue donnée, elle ne constitue cependant pas une limite exhaustive d'acceptabilité. En effet, bon nombre de phrases syntaxiquement correctes restent inadmissibles (ex : 'la politesse jaune pleure du pain'). Cette imprécision tient à la confusion qui est faite, par les grammaires, des mots appartenants à une même liste d'éléments du discours. L'étude des significations des mots, de la façon dont elles sont liées les unes aux autres, et des bases du choix lexical fait l'objet de la sémantique lexicale. Parmi les principales questions qu'il lui appartient d'examiner, les problèmes d'ambiguïté de portée prennent une part importante. Une phrase aussi simple que:

'Jean-François n'est pas parti à New York en avion'.

Peut en effet être comprise comme :

Quelqu'un d'autre est parti à New York en avion.

Jean-François est parti de New York en avion.

Jean-François est parti ailleurs.

Jean-François est parti à New York par un autre moyen de transport.

Notons que la différence entre sémantique et syntaxe reste relativement floue. Ainsi, une description syntaxique est souvent porteuse de sens (voir l'exemple de "les invités entendaient le bruit de leur fenêtre"). D'une façon générale, toute analyse syntaxique basée sur un nombre important de classes d'éléments du discours possède inévitablement un caractère sémantique.

L'étude de grammaires et d'analyseurs sémantiques est un sujet de recherche actuel en linguistique informatique. Seules des solutions partielles ont pu être obtenues jusqu'à présent (analyse dans un domaine sémantique restreint).

1.2.6. Niveau morphologique

Lorsqu'on étudie les formes écrites et phonétiques d'une langue, il est frappant de constater que les mots qui la composent, bien que très nombreux, sont eux mêmes

constitués d'unités plus petites (comme dans image, images, imagine, imagination, imagerie, inimaginable, etc.).

La morphologie est la branche de la linguistique qui étudie comment les formes lexicales sont obtenues à partir d'un ensemble réduit d'unités porteuses de sens, appelées morphèmes. On distingue les morphèmes lexicaux des morphèmes grammaticaux, qui apportent aux premiers des nuances de genre, nombre, mode, temps, personne, etc. Tout comme le phonème, le morphème est une unité abstraite. Elle peut être réalisée en pratique sous diverses formes appelées allomorphes, fonction de leur contexte morphémique. Ainsi le morphème grammatical du pluriel se manifeste t- il sous la forme d'un 's' dans 'pommes', d'un 'x' dans 'jeux' et d'un 'nt' dans 'jouent'.

1.2.7. Niveau pragmatique

Contrairement à la sémantique, que l'on qualifie souvent d'indépendant du contexte, le sens pragmatique est défini comme dépendant du contexte. Tout ce qui se réfère au contexte, souvent implicite, dans lequel une phrase s'inscrit et à la relation entre le locuteur et son auditoire, à quelque chose à voir avec la pragmatique. Son étendue couvre l'étude de sujets tels que les présuppositions, les implications de dialogue, les actes de parole indirects, etc. Elle est malheureusement bien moins développée encore que la sémantique [1].

1.3. Propriétés spécifiques du signal vocal

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

1.3.2. 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) [8].

1.3.2.1. Variabilité intra-locuteur

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

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

1.3.2.2. Variabilité inter-locuteur

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

1.3.2.3. Variabilité contextuelle

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

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

La réduction est due au fait que les cibles articulatoires sont moins atteintes dans le parler rapide. Ces phénomènes sont en grande partie responsable de la complexité des traitements réalisés sur les signaux de parole [8].

1.4. Segmentation du signal vocal

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

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

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

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

1.5. Notions fondamentales sur les sons de l'Arabe Standard

La recherche en traitement automatique de la parole dans une langue donnée doit nécessairement passer par l'étude de la composante phonétique. Cette étude nous permet de dégager les principales caractéristiques relatives aux différents phonèmes et ainsi de cerner l'ensemble des paramètres acoustiques, en vue de l'exploiter dans l'élaboration d'un système de synthèse de la parole.

1.5.1. Système phonétique de l'Arabe Standard

Le système phonétique de l'Arabe Standard (AS) comprend six voyelles et vingt-neuf consonnes *harūf* (en incluant la hamza) produit par seize lieux d'articulation [6]. Ces consonnes peuvent aussi être classées sur le plan acoustico-physiologique, selon leurs modes de production (*sifa*). Ils sont classés en sonores/sourdes, occlusives/spirantes, emphatiques/non emphatiques, etc.

Les voyelles de la langue arabe sont classées, en trois brèves ou courtes (harakāt) et trois longues (madd). Les voyelles brèves sont représentés par des signes diacritiques placés au-dessus des consonnes : (fatha) [a], (damma) [u] et (kasra) [i]. La harakāt est un mouvement aérien et organique dont a besoin un harf pour se produire dans un continuum sonore. L'absence de la harakāt dans la langue arabe s'appelle sukūn . Les voyelles longues : [ā], [ū] et [ī] sont des allongements temporels des voyelles brèves.

La classification des phonèmes de la langue arabe est basée, comme toutes autres langues, sur le lieu et le mode d'articulation (Tableau 1.3). Le phonème affriquée est considéré comme une semi-occlusive, cette consonne particulière se comporte à la fois comme une occlusive et une fricative.

Lors de sa prononciation, la langue ne s'écarte pas brusquement du palais, comme étant le cas des occlusives pures, mais plutôt d'une manière douce.

L'air libéré sera sous forme de friction. Dans l'Arabe il existe une seule affriquée c'est [ǧ].

1.5.2. Particularité de l'Arabe Standard

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

1.5.2.1. Voyelles longues

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

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

Sur le plan acoustique, les niveaux des formants entre chaque voyelle brève et son opposée longue sont assez rapprochés. L'allongement temporel effectué par les voyelles longues n'influe pas sensiblement sur les niveaux formantiques de ces derniers. La partie stable de la voyelle longue est beaucoup plus allongée par rapport à son opposée brèves [6].

Tableau 1.3 : Transcription Orthographique Phonétique de l'AS [7].

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

1.5.2.2. Phonèmes arrières

Le système phonétique de l'AS possède quatre phonèmes arrières spécifiques à cette langue, et n'ont leurs équivalents exacts dans aucune autre langue européenne [6] :

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

1.5.2.3. Emphase

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

Basé sur de nombreuses mesures spectrographiques, S. Al Ani définit ce phénomène comme étant produit dans la région vélaire et non la pharyngale. Cependant R. Jackbson proclame que les emphatiques sont réalisées par la pharyngalisation qui se produit lors de la contraction de la partie supérieure du pharynx.

Les phonèmes emphatiques sont caractérisés par une tonalité plus pleine et grave car ils exigent la dépense d'un volume d'air important et une tension organique supérieure par rapport aux autres consonnes. L'intérêt porté par ce phénomène remonte jusqu'aux grammairiens arabes du moyen-âge. Attirés par le système phonétique de leur langue, ils ont pu déterminer par de simples constatations ciblées, le fonctionnement et les positions des principaux organes entrant dans la production d'un son emphatique que se soit sur le plan auditif ou physiologique [7].

1.5.2.4. Gémination

Appelée aussi redoublement, la gémination correspond au phénomène de renforcement d'une articulation consonantique qui tend à prolonger la durée de la consonne tout en augmentant son intensité.

En abordant la description du système phonétique de l'AS, Sibawyh le célèbre grammairien arabe s'est particulièrement intéressé aux phénomènes qui affectent la Langue Arabe notamment la gémination. L'absence d'une voyelle intermédiaire dans une séquence

constituée de deux consonnes identiques permet l'insertion du phonème dans l'autre engendrant ainsi, le renforcement de l'articulation et le prolongement de la consonne. En effet certaines expériences ont permis de constater que la consonne géminée ne présente pas deux mouvements articulatoires distincts, mais un mouvement unique qui diffère de la consonne simple par sa stabilité articulatoire et par sa durée importante. Son spectre présent en générale une énergie uniformément répartie et sans discontinuité.

Selon J.F Bonnot, la durée phonémique n'est le critère principal de la discrimination entre une consonne géminée et son homologue simple. Il est indispensable suggère-t-il. « D'aborder une très grande attention aux autres indices ». Il existe deux sortes de gémination :

- Une gémination succédant toujours une voyelle. Elle permet la différenciation entre deux mots. L'exemple de [hamām] et [hammām] (Figure.1.7).
- La gémination euphonique est utilisée pour pallier les groupes de sons qui paraissent durs à l'oreille [7].

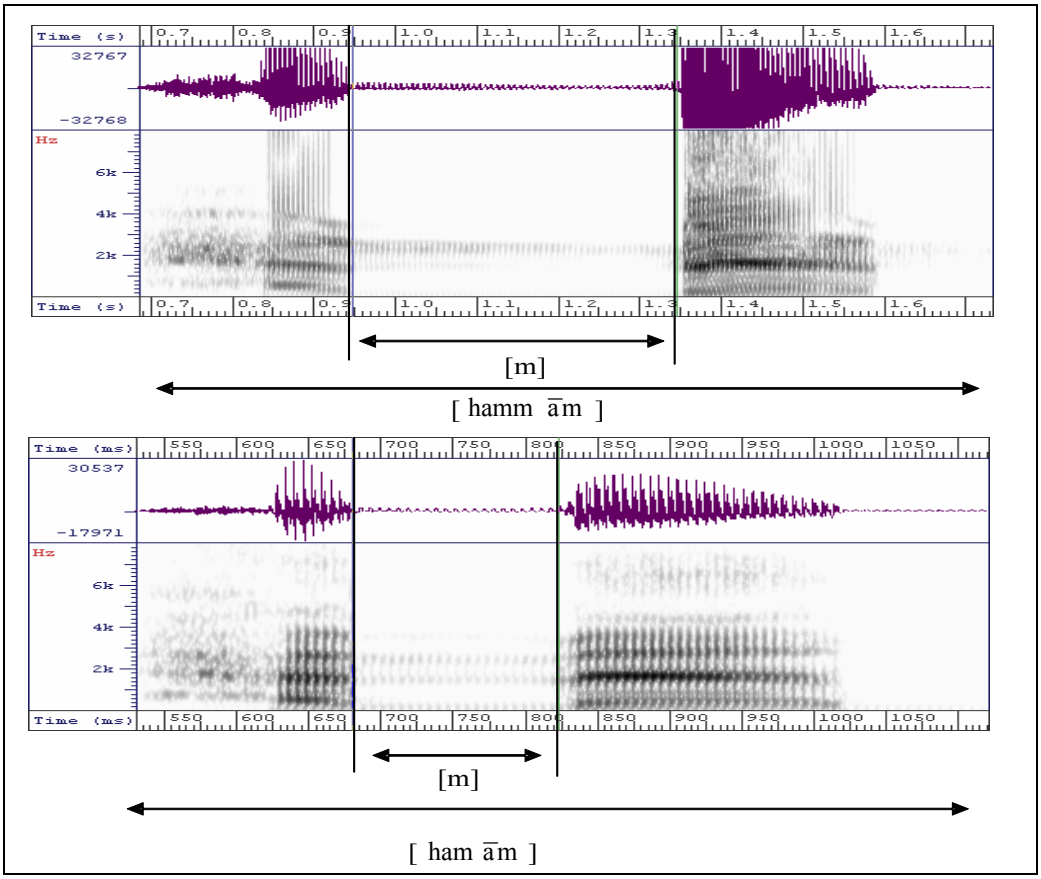


Figure 1.7 : Spectrogramme de la consonne [m] représentant une gémination

1.6. Conclusion

Dans ce chapitre nous avons exposé sommairement une description des différents niveaux des sons du langage (Français et Arabe), et nous avons fait une étude brève sur le système phonétique de l'Arabe Standard en présentant quelques généralités phonétiques et certaines propriétés spécifiques de l'AS en donnant une description très simplifiée, probablement erronée aux yeux d'un phonéticien mais suffisante pour comprendre les particularités du traitement de la parole.

CHAPITRE 2
SYSTEMES DE SYNTHESE DE LA PAROLE

2.1. Introduction

Ce deuxième chapitre va être consacré à l'étude des machines parlantes, et plus particulièrement à la conception des systèmes de synthèse à partir du texte (Text-To-Speech : TTS). Nous allons dans un premier temps décrire le fonctionnement général d'un tel système pour ses différentes composantes de traitement de signal. Nous étudions les principales fonctions de la prosodie des langues utilisées dans les systèmes de synthèse de la parole. La deuxième phase est consacrée à une étude des différentes techniques d'extraction de la F_0, et nous terminons ce chapitre par une description de la manière d'évaluation des systèmes de synthèse de la parole a partir du texte en citant quelques méthodes d'évaluations.

2.2. Historique de la synthèse de la parole

À plusieurs reprises au cours de l'histoire, on a tenté de reproduire la voix humaine. Au XVIII^{e} siècle, on met au point à cet effet des dispositifs mécaniques équipés de soufflets et d'anches vibrantes. Au XX^{e} siècle, l'apparition de l'électricité et de l'électronique autorisent des tentatives plus ambitieuses : en 1922, J.C. Stewart fabrique une machine capable de reproduire des voyelles, des diphtongues et quelques mots simples ; plusieurs années plus tard en 1939, H. Dudley présente, à l'occasion de l'exposition universelle de New York, le VODer (Voice Operation Demonstrator), appareil mis au point par les laboratoires Bell [9].

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.).

2.3. Principe de la 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 ait jamais songé à l'appeler « synthétiseur » !

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

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

« La synthèse de la parole permet de produire des sons de la parole à partir d'une représentation phonétique du message » [10].

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.

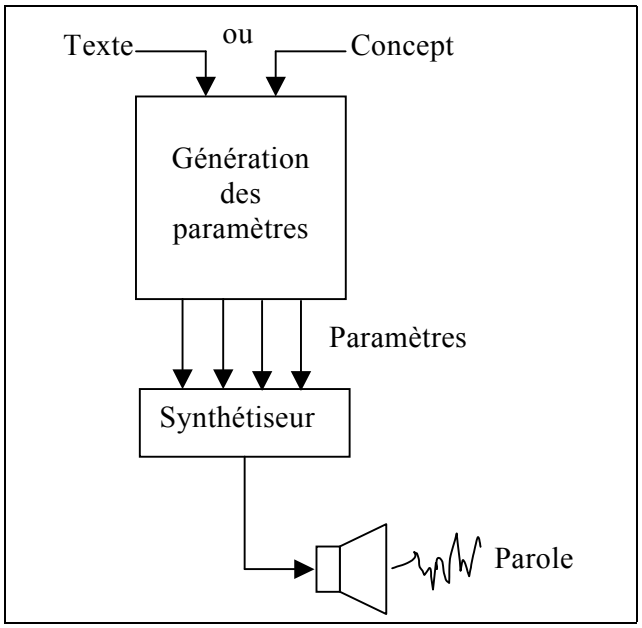


Figure 2.1 : Système de synthèse de la parole [10].

On peut identifier trois composantes fonctionnelles dans un système T T S [11] :

- l'analyse du texte, pour passer du texte écrit à une suite de symboles phonétiques, qui représentent la prononciation du texte, et des symboles représentant l'accentuation et l'intonation du texte. Cette première composante relève essentiellement du traitement automatique de la langue;
- la description symbolique du texte doit être convertie en paramètres numériques (courbes mélodiques, intensité de la voix), à l'aide de procédures phonétiques. Cette seconde composante utilise divers types d'algorithmes comme des systèmes de règles, des automates stochastiques, ou des réseaux de neurones;
- la dernière composante, ou synthétiseur acoustique, calcule le signal de parole à l'aide des paramètres. Cette composante relève du traitement numérique du signal.

2.4. Synthèse de la parole

Comme il n'est pas envisageable de considérer toutes les configurations acoustiques d'un même phonème et que le phonème reste tout de même le seul lien avec le texte à prononcer, deux méthodes assurent un pivot entre une description phonétique abstraite et une réalité acoustique concrète. Soient, la synthèse par règles et la synthèse par concaténation d'unités acoustiques.

2.4.1. Synthèse par règles

Indépendamment du modèle de signal considéré, une méthode de synthèse par règles consiste à modéliser dynamiquement les transitions entre phonèmes sous la forme de règles [12].

Par opposition avec l'approche par concaténation d'unités acoustiques, il s'agit d'une méthode qui apporte un pouvoir explicatif aux processus de phonation. Une fois le modèle paramétrique représentant l'évolution temporelle du signal de parole imposé, des règles décrivant l'évolution des paramètres du modèle sont inférées, le plus souvent par un expert linguistique, à partir de données d'exemples. Pour des raisons historiques et de facilité d'interprétation, le modèle du signal de parole le plus souvent associé à une synthèse par règles est un synthétiseur à formants. Un ensemble de règle de contrôle constituant quelques dizaines de paramètres situés au niveau d'un modèle acoustique. Les principales difficultés de ces approches sont la mise au point des règles et le choix de leur formalisation. Nous donnons un aperçu simple et succinct pour ce type de synthèse.

Dans un premier temps, on enregistre sous forme numérique un grand nombre de mots (généralement de type Consonne-Voyelle-Consonne) prononcés par un locuteur professionnel.

Les mots sont choisis de façon à constituer un corpus représentatif des transitions phonétiques et des phénomènes de coarticulation dont on veut se rendre compte. On modélise alors ces données à l'aide d'un modèle paramétrique de parole (séparer les contributions respectivement de la source glottique et le conduit vocal) plus adéquat à l'établissement des règles. On commence par inspecter globalement l'ensemble des données, de façon à établir la forme générale des règles produites. On précise les valeurs numériques des paramètres intervenant dans ces règles (par exemples, les fréquences des formants ou les durées des

transitions) par un examen minutieux du corpus. La mise au point du synthétiseur s'achève par un long processus d'essais erreurs, afin d'optimiser la qualité de la synthèse.

Lorsqu'un nombre suffisant de règles a été établi, la synthèse proprement dite peut commencer. Les entrées phonétiques du synthétiseur déclenchent l'application des règles, qui produisent elles mêmes un flux de paramètres liés au modèle de parole utilisé. Cette séquence temporelle de paramètres est alors transformée en parole. La conception de tels systèmes requiert une part importante d'expérimentation par essais-erreurs, ce qui allonge le temps de développement et alourdi le coût [13].

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

2.4.2. Synthèse par concaténation d'unités acoustiques

Puisque le contenu acoustique d'un signal de parole peut être décrit à l'aide d'un ensemble fini d'unités linguistiques « les phonèmes », il suffirait de juxtaposer 'ou de concaténer' un représentant acoustique pour chaque phonème à synthétiser.

Un modèle trivial peut simplement consister à stocker un seul exemplaire acoustique du phonème (ou encore un phone) qui sera utilisé pour tous les messages de synthèse.

La principale difficulté dans cette méthode reste de calculer un signal final dans lequel les transitions entre unités ne sont pas perceptibles. Le choix des unités joue un rôle primordial pour ce type de synthèse. Le dialectique est simple : Les unités courtes sont économiques, mais ne permettent pas d'obtenir une bonne qualité. Les unités longues sont plus coûteuses mais permettent généralement d'obtenir une meilleure qualité de synthèse. On distingue plusieurs méthodes dont les plus utilisées sont la synthèse par phrases, mots, phonème, syllabes, et diphones etc.

La synthèse par phrases n'est pas réellement une véritable synthèse, le principe de base est de stocker et de restituer de la parole continue, en vue d'une application bien définie. A titre d'exemples on peut citer le cas de l'horloge parlante où on stocke dans la mémoire et selon la langue voulue, une phrase porteuse fixe de type :

« Au troisième top, il sera exactement '. . . heures' '. . . minutes' '. . . secondes' ».

On ajoute trois parties variables constituées par des nombres préalablement enregistrés dans un dictionnaire de sons. Un répondeur automatique se déclenche lors d'un appel, en intégrant dans la phrase porteuse les nombres correspondants aux heures, minutes et secondes.

Le principe de la synthèse par mots est de préenregistrer les éléments de message sous forme de mots, et de les juxtaposer pour former une phrase. Des règles prosodiques peuvent améliorer le naturel en tenant compte des types de phrases.

En ce qui concerne la synthèse par phonèmes, le principe est le même avec des éléments préenregistrés sous forme phonémique, son inconvénient est que les effets de discontinuité qui se produisent aux jonctions inter-phonémiques ne donnent pas une parole synthétique intelligible.

La synthèse par syllabes est peu utilisée du fait que la frontière d'une syllabe reste difficilement détectable à cause du phénomène de la coarticulation [cv, vc ou cvc, etc].

Avant de procéder à la synthèse par diphones (Figure.2.2), il faut donner sa définition : « Un élément sonore caractéristique de la transition entre deux phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant » [10].

Dans cette méthode, la synthèse est alors réalisée par simple concaténation de ces éléments, les liaisons se font aux niveaux des parties spectralement identiques. Les éléments mémorisés, sous forme d'un dictionnaire, sont donc essentiellement les transitions spectrales entre deux phonèmes. La taille est, par conséquent inférieure à un mot et supérieure à un phonème.

Cette méthode présente un inconvénient qui se résume en la difficulté de dégager une zone stable pour segmenter, dans un contexte de liquides ou semi-voyelles très sensibles aux effets de coarticulation. La figure 2.2 présente la mise en œuvre du principe d'une synthèse par diphone.

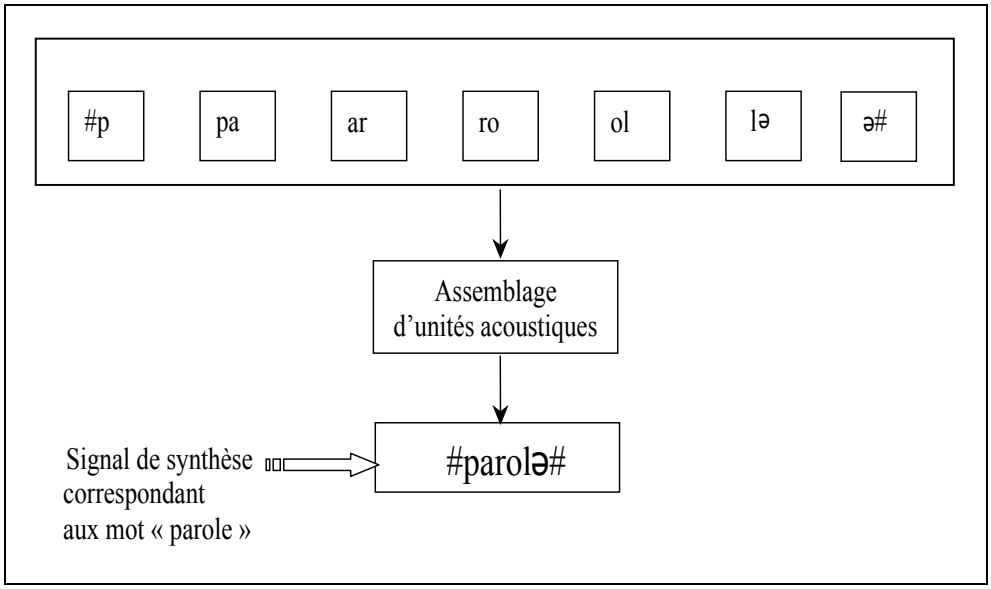


Figure 2.2 : Synthèse par diphones du mot « parole » prononcé de manière isolée. Le signe # représente une indication de silence.

2.4.3. Modélisation du signal de parole

Que l'on adopte une méthode de synthèse fondée sur un pouvoir explicatif (approche par règles) ou seulement descriptive (approche juxtaposition d'unités préenregistrées), l'évolution temporelle du signal de parole peut être, ou non, représentée par un modèle.

L'intérêt d'un tel modèle réside dans sa capacité à réduire la redondance du signal acoustique et à définir des paramètres mieux appropriés aux traitements acoustiques.

Un système de synthèse par règles ne peut faire l'économie d'un tel modèle, car il est indispensable de réduire la quantité d'information acceptable pour une solution algorithmique fondée sur un ensemble fini et maîtrisable de règles de traitement.

Par contre, pour une méthode de synthèse par concaténation d'unités acoustiques préenregistrées, le signal est, pour la plupart des systèmes, conservé sous formes temporelle – c'est à dire sans modèle-, ou bien à l'aide de codeurs, comme les méthodes de prédiction linéaire. L'intérêt d'un modèle paramétrique est double. En permettant de réduire la redondance présente dans le signal, il permet une compression des données là où les capacités de stockage pour un système de synthèse sont limités –ce qui est le cas pour les applications qui sont embarquées. Il permet, en outre, d'offrir des paramètres qu'il est facile de relier avec ses informations suprasegmentales.

De nombreux modèles décrivant l'évolution temporelle du signal de parole ont été proposés au cours de ces quarante dernières années. Même si le signal à modéliser est considéré comme la simple variation d'une grandeur physique en fonction du temps, la majorité d'entre eux font une hypothèse issue des mécanismes de production de la parole : l'hypothèse de décomposition Source-Filtre.

Les modèles de représentation du signal de la parole peuvent être classés en deux catégories selon l'hypothèse méthodologique considérée. On distingue les modèles :

- s'appuyant sur la production ; et qui décrivent la génération d'un signal de parole à partir de paramètres physiologiques. Il s'agit de reproduire le comportement de l'appareil phonatoire humain ;
- placés sous une hypothèse de perception ; et qui décrivent comment générer un signal de parole qui pourra être perçu comme un signal de parole naturelle.

L'hypothèse de perception est celle qui rencontre le plus de suffrage dans la réalisation concrète des systèmes de synthèse de la parole car elle est propice à une formalisation simple.

Cependant, celle de la production apporte un grand fondement à l'explication des mécanismes de production de la parole.

La séparation Source-Filtre considère que le signal de parole résulte de la combinaison d'une énergie aérienne couplée à une Fonction de Transfert (FT) déterminée par la forme supra glottique. Cette description trouve sa justification à partir des travaux J.Muller [13] menés sur la mécanique phonatoire. Remplacé dans un contexte de traitement du signal par G.Fant [13], le modèle Source-Filtre présente alors le signal de parole x(n) comme la convolution d'un signal d'excitation e(n) par un filtre h(n), on obtient alors :

$$x(n)=e(n)*h(n) \tag{2.1}$$

e(n) : excitation glottique.

h(n) : Réponse Impulsionnelle du filtre vocal.

Il est en pratique impossible d'avoir accès simultanément et sans trop perturber les mesures de x(n), e(n) et h(n). Seule la mesure de x(n) est simple à mettre en œuvre, un microphone permet de capter un signal de parole à la sortie des lèvres. Pour lever l'ambiguïté d'une équation à deux inconnues, les différents modèles que nous allons décrire feront tour à

tour des hypothèses sur l'une des variables permettant l'estimation de la seconde à partir du signal connu x(n).

Après une rapide explication de l'hypothèse Source-Filtre, nous aborderons la présentation des modèles articulatoires, à formants, vocodeur à canaux, Auto-Régressifs (AR) et enfin le modèle de compositions harmoniques et bruit, du signal de parole.

2.4.3.1. Modélisation articulatoire

Un modèle articulatoire permet de produire un signal de parole à partir d'une description physiologique. Historiquement, les premières études du comportement dynamique des articulateurs se sont surtout attachées à leur description externe : les lèvres, les mâchoires, les joues. Le développement de la photographie par rayons X a permis de disposer d'une information plus précise sur le comportement interne du conduit vocal : section du conduit, position des articulateurs internes comme la langue ou le voile du palais. De nos jours, des techniques de mesure moins dangereuses telles que la résonance magnétique offrent des outils de mesure efficaces. La méthodologie de l'approche articulatoire consiste à déterminer une fonction de passage entre un plan articulatoire et un plan acoustique (inversion articulatori-acoustique) sous la contrainte d'hypothèse linguistique [13].

Une fois la configuration du conduit vocal déterminée, plusieurs méthodes de calcul du signal de parole sont possibles :

- la moins coûteuse au niveau des calculs consiste à revenir à un plan acoustique par des techniques d'acoustiques linéaires;
- des systèmes calculant la variation de la pression acoustique au niveau des lèvres par résolution des équations de propagation d'une onde de pression dans un milieu dynamique.

2.4.3.2. Modélisation formantique

Dans l'approche formantique, le procédé de synthèse se trouve lié à une hypothèse d'interprétation à la fois phonétique et acoustique. Un formant caractérise une résonance du conduit vocal. L'analyse d'un signal acoustique naturel consiste, à partir de son spectre, à caractériser ses résonances par leurs positions fréquentielles et bandes passantes. La synthèse,

part d'une séquence finie de représentation de formants –position, amplitudes, et bandes passantes- pour générer un signal de parole à partir d'un spectre artificiel.

L'établissement d'une FT artificielle est assuré par l'association en série ou en parallèle de cellules résonnantes du second degré. Pour les sons Voisés, ce système est excité par une onde quasi périodique dont la forme est aussi proche que possible de l'onde glottique. Pour les sons Non Voisés, l'excitation est un bruit blanc.

Si cette approche semble séduisante puisqu'elle assure un lien entre une interprétation phonétique et une réalité acoustique, l'acquisition des données reste relativement complexe.

La détection des formants d'un signal de parole est en effet délicate notamment lorsque ceux-ci sont très proches (Exemple : le F_2 des voyelles [i] et [e]).

2.4.3.3. Vocodeur à canaux

Le vocodeur à canaux est un appareil destiné à transmettre la parole à un faible débit d'information. Il se compose de deux parties l'analyseur et le synthétiseur. La fonction d'analyse de l'enveloppe spectrale est effectuée à l'aide de canaux dont le nombre peut varier, suivant les réalisations, entre 10 et 20. Chaque canal traite une bande de fréquence déterminée.

Le signal de parole issue d'un microphone est analysé au moyen d'un banc de filtres passe bande contigus, couvrant l'étendue de la bande téléphonique 300 à 3400 Hz [14].

Le signal délivrer par chacun des filtres subit une détection puis traverse des filtres passe bas, dont les fréquences de coupures sont de l'ordre de 20 à 50 Hz, parce que les variations énergétiques dans les canaux sont lentes (à l'image de variation lente de l'articulation).

L'analyse du vocodeur comporte également un détecteur de voisement. Ce dernier permet de différencier les sons Voisés et de donner la valeur de la F_0.

La synthèse est effectuée à l'aide d'un banc de filtres passe-bande analogue à celui de l'étage d'analyse. Pour chaque canal, le signal basse fréquence issue du filtre d'analyse est multiplié par le signal d'excitation dans un modulateur. Selon que la parole à reproduire est détectée comme Voisée ou Non Voisée, le signal d'excitation attaque les modulateurs provient soit d'un générateur périodique soit d'un générateur de bruit. Le signal de sortie est obtenu par addition des sorties des filtres de synthèse. L'intelligibilité est assez bonne, bien que

l'agrément et le naturel de la voix soient dégradés par le traitement. La partie délicate est constituée par le détecteur de pitch.

2.4.3.4. Modélisation Auto-Régressive

G.Fant [13] introduisit en 1960 une modélisation linéaire de l'évolution temporelle du signal de parole. Il ajoute à l'hypothèse Source-Filtre une hypothèse d'indépendance de l'onde glottique et du conduit vocal. Ce dernier est modélisé par un filtre tout pôles que l'on appelle filtre Auto-Régressif (AR).

L'onde glottique, quant à elle, est modélisée approximativement par un train d'impulsions ayant pour période la période fondamentale pour un son voisé et par un bruit blanc de moyenne nulle et de variance unité pour un son Non Voisé [4].

Ce type de modélisation est largement répondu en TAP et notamment en codage. Cependant, le nombre de paramètres du modèle est relativement faible et la modification des paramètres prosodiques en contexte de synthèse de la parole est relativement aisée. En effet, pour modifier la durée d'un son, il suffit d'allonger ou de raccourcir le signal d'excitation et la modification du fondamental est opérée par une simple modification de la période fondamentale du train des impulsions. C'est une modélisation qui a été largement utilisée à la fin des années 1980 et qui a cédé la place à des techniques plus complexes qui offrent une meilleure qualité du signal comme par exemple le modèle harmonique et bruit et la technique PSOLA (Pitch Synchronous Overlap and Add).

2.4.3.5. Modélisation Harmonique et Bruit

Le principe original du Modèle Harmonique et Bruit (MHB) consiste à décrire formellement le signal de parole comme la superposition d'une composante harmonique (Voisement) et d'une composante bruit (Non-Voisement). Ce modèle permet de regrouper sous un même formalisme aussi bien la description d'un signal vocal comme une voyelle, qu'un signal bruité comme une consonne sourde.

S'il est justifié de poser une telle hypothèse d'analyse à partir d'une description des principales caractéristiques des signaux de parole, il reste que la définition d'un critère objectif séparant la partie harmonique de la partie bruitée est le plus souvent arbitraire. La plupart des travaux traitant ce sujet considèrent le bruit comme un résidu de calcul qu'il est difficile d'interpréter à un niveau acoustique. Cependant, certains travaux ont apporté des solutions à

l'estimation de la composante de bruit tout en conservant une justification perceptive à ce bruit [13].

La modification des paramètres prosodiques de la partie harmonique est relativement facile dans la mesure ou la variable temporelle et la période de distribution des harmoniques sont explicites dans le modèle [3]. Le MHB offre une modélisation du signal de la parole avec une bonne qualité ; la phase d'estimation des paramètres reste cependant relativement délicate.

Il faut noter que ce modèle offre un cadre formel unique pour des modifications de haute qualité des paramètres prosodiques de la voix ainsi que des modifications sur l'enveloppe spectrale caractérisant le timbre du locuteur [13].

2.4.3.6. Définition de la technique PSOLA

PSOLA (Pitch Synchronous Overlap and Add) n'est pas à proprement parler un modèle du signal de la parole. Il s'agit d'une technique de traitement du signal de parole qu'il soit naturel ou de synthèse dont l'objectif est de modifier ses paramètres prosodiques.

L 'originalité et l'efficacité d'une telle technique consiste à modifier le fondamentale et la durée du signal de manière indépendante et locale sans altérer le timbre de la voix d'origine.

La mise en œuvre de cette technique peut se faire dans un cadre temporel TD PSOLA-Time Domain PSOLA ou fréquentiel FD-PSOLA, (Fréquency Domain-PSOLA*)*. Dans tous les cas, la qualité du signal de parole modifié est transparente[1] pour des excursions de paramètres prosodiques qui restent compatibles avec ce que prédisent les modèles prosodiques d'un système de synthèse. Toute l'efficacité de la technique PSOLA réside dans la manipulation de signaux de parole à court terme dont la relation de phase est explicite. Un traitement indispensable dont la précision influe sur la qualité du signal modifié par PSOLA consiste donc à déterminer les instants de synchronisme pour lesquels les signaux élémentaires seront considérés en phase [13]. Le couplage d'une méthode de synthèse par concaténation d'unités acoustiques et de la technique PSOLA est aujourd'hui le meilleur compromis qualité/efficacité algorithmique dans le domaine du développement des systèmes de synthèse de la parole. On peut noter quelques variations autour de l'approche PSOLA comme par exemples le prétraitement à pitch constant du répertoire des unités acoustiques en utilisant une technique d'analyse /synthèse du signal de parole.

[1] C'est-à-dire que la voix ne doit pas être dégradée lorsque l'on ne fait pas de modifications prosodiques.

2.4.4. Applications d'un système de synthèse de la parole

L'énumération des produits industriels s'appuyant sur une technologie de synthèse de la parole serait une tâche fastidieuse. Cependant, il est intéressant de dégager une taxonomie du domaine applicatif de cette technologie.

L'intelligibilité et la qualité de la parole de synthèse sont encore de nos jours inférieurs de celles de la parole naturelle. Il serait donc absurde de tenter de remplacer un service rentable diffusant des messages de parole naturelle par des messages en voix de synthèse. Un auditeur, usagé d'un service d'informations vocales, supporte la qualité d'une voix de synthèse s'il ne peut faire autrement. La synthèse de la parole est couramment utilisée dans plusieurs domaines d'applications on peut citer :

Tableau 2.1 : Applications de la synthèse de la parole.

APPLICATIONS	DOMAINES
Aide aux personnes handicapées	- machine à lire pour aveugles ; - annonces et renseignement parlés.
Services de télécommunications	Rendre toute information écrite disponible via le téléphone - horaires de cinéma ; - horaires de train ; - informations routières ; - état d'un compte en banque ; - dernière facture téléphonique.
Applications en bureautique	- terminaux parlants ; - lecture des E-mails par la voix.
Applications dans les transports	- informations dans les automobiles ; - aide à l'exploitation des trains ; - lecture de cadrans dans les avions.
Apprentissage des langues étrangères	- dictionnaires électroniques avec prononciation intégrée ; - logiciels d'apprentissage des langues étrangères - traduction automatique.
Livres et jouets parlants	- pour les enfants, la voix est le seul moyen de communication avec la machine
Dialogue Homme-Machine	- la communication avec la machine de manière plus naturelle

2.5. Principales fonctions de la prosodie

La prosodie intervient à tous les niveaux dans les phénomènes de parole et remplit des rôles variés dans la communication parlée. En voici quelques unes :

2.5.1. Distinction entre homonymes

On distingue généralement deux catégories de langues, des langues à accent :

- fixe pour lesquelles la position accentuable du mot est toujours placée sur une syllabe déterminée (en français la dernière syllabe du mot).
- libre pour lesquelles la position accentuable dépend de la fonction lexicale et de la structure morphologique du mot.

Dans les langues à accent libre et dans les langues à tons, la prosodie est associée aux choix lexicaux. Elle permet en particulier de discriminer des homonymes. En Anglais, la position de l'accent tonique peut distinguer un nom d'un verbe :

Exemples :

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

En espagnol l'accent tonique permet de différencier certains mots :

- **pla**tano (une banane)
- pla**ta**no (un platane)

2.5.2. Structuration de l'énoncé

Généralement, dans toutes les langues, la prosodie participe à la segmentation de l'énoncé en segments minimaux. Elle indique aussi le type de relations que ces groupes entretiennent. Elle permet aussi la hiérarchisation au sein de la phrase. Cette structure n'est pas identique à la structure syntaxique, mais elle est en relation avec elle.

Exemples illustrant comment la prosodie participe au regroupement de mots :

- L'instituteur, dit le directeur, est un incapable.
- L'instituteur dit : "le directeur est un incapable".

- ذلك الكتاب لا ريبِ# فيه هدى للمتقين
- ذلك الكتاب لا ريب فيهِ# هدى للمتقين

Les situations où il y a risque d'ambiguïté entre deux phrases ou expressions, appelées oronymes, sont aussi courantes en Anglais :

Exemples :

- That reflects the secretariat's fear of competence.
- That reflects the secretariat's sphere of competence.

Dans cette situation la prosodie peut jouer un rôle, en conjonction avec d'autres sources d'informations comme le contexte sémantique de la phrase.

Au-delà de la phrase, la prosodie joue un rôle dans la structuration du discours. Elle peut par exemple, marquer un changement de sujet dans la conversation. L'indice prosodique le plus souvent étudié est la dynamique locale ("pitch range"). Le rythme et l'intensité sont deux autres indices importants.

2.5.3. Emphase et Focalisation

L'accentuation est un moyen d'insister sur tel ou tel mot :

- **Je** vais terminer (par opposition à quelqu'un d'autre).
- Je **vais** terminer (par opposition à une action déjà accomplie).
- Je vais **terminer** (par opposition à une autre action).

L'accentuation peut aussi renforcer une opposition :

- Non, pas le six... le **dix** août.

2.5.4. Modalité

La prosodie, et plus particulièrement la mélodie, est liée au mode de la phrase : affirmatif, interrogatif, impératif, ou exclamatif. Il est très courant de donner une intonation interrogative ou exclamative à une phrase tout en conservant sa structure grammaticale d'affirmation.

Exemples :

- Il va venir?
- Il va venir.
- Il va venir !

2.5.5. Attitude et Interaction

La prosodie véhicule l'attitude du locuteur vis-à-vis de l'énoncé ou ses intentions vis-à-vis de l'interlocuteur. En indiquant son adhésion plus ou moins forte envers l'énoncé, le locuteur exprime selon le contexte la conviction ou le doute, l'accord ou le désaccord, l'approbation ou la désapprobation, ou encore une invitation, une incitation. La dynamique de la F_0 est l'une des manifestations acoustiques. Si l'on inclue les phénomènes d'hésitation, la prosodie permet aussi de gérer les tours de parole dans une conversation. Un ton final, un silence laissant le champ libre à l'interlocuteur pour intervenir. Un ton continuatif, une pause verbale ("filled pause"), servant à conserver son tour pour parler.

2.5.6. Fonctions non linguistiques

La prosodie au sens large traduit l'état psychologique du locuteur : calme ou énervé, triste ou gai, enthousiaste, surpris, etc. Elle caractérise aussi le locuteur en tant qu'individu ou membre d'un groupe (exemple : l'accent régional).

2.6. Différentes phases d'un algorithme de détection de Pitch

La fréquence fondamentale est un paramètre très important pour la synthèse de la parole ; l'oreille est en effet très sensible à ces variations qui constituent un élément essentiel de la prosodie. En plus le nombre élevé de méthodes de détection de pitch existant dans la littérature, souligne son importance. L'estimation du pitch est une tâche difficile pour de nombreuses raisons telles que la non-stationnarité du signal, certaines irrégularités dans l'excitation glottique ou encore une interaction avec F_1. En effet, une des premières phases des algorithmes de détection de pitch est la décision du Voisement. Il est bon de rappeler, qu'il n'existe jusqu'a présent aucun dispositif infaillible dans ce domaine.

Un algorithme d'extraction de la F_0 peut en général se décomposer en trois phases successives (Figure. 2.3) :

- un prétraitement et un changement de représentation ;
- l'extraction du fondamental ou phase de traitement ;
- un post-traitement visant à corriger les erreurs.

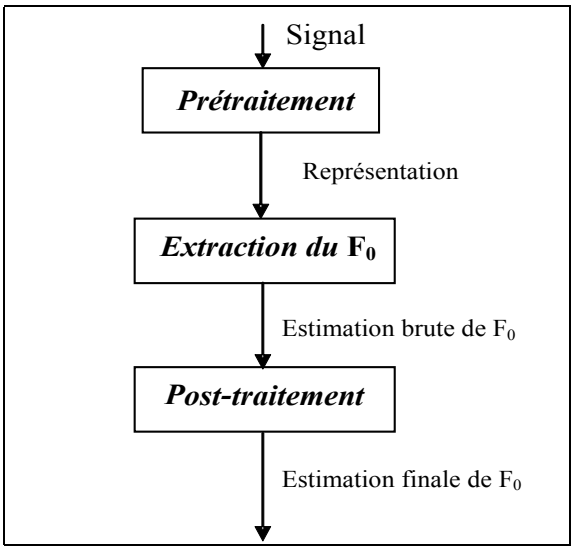


Figure.2.3 : Schéma global d'un algorithme d'extraction du pitch

2.6.1. Phase de prétraitement

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

- la condition de stationnarité souvent exigée par les algorithmes de traitement ;
- l'effet de bord lié aux fenêtres de pondération appliquées, et d'assurer ainsi la présence d'au moins une période du fondamental. La durée de la trame est généralement choisie entre 20 et 30ms avec un recouvrement de 30 à 50% ;
- a la détection de silence garantissant la présence du signal utile;
- a la détection du voisement puisque la F_0 n'est présent que sur une séquence Voisée;
- a une préaccentuation afin de rehausser l'énergie des hautes fréquences.

Dans cette phase de prétraitement, nous trouvons souvent d'autres techniques permettant d'améliorer la rapidité d'extraction. Il s'agit de toutes sortes de techniques de filtrage permettant d'atténuer ou même d'éliminer les formants d'ordres supérieur ou égal à 2, et de minimiser l'effet du bruit sur la détection. La décimation dans le rapport de 5 à 1 est souvent utilisée par des détecteurs à temps réels [1]. Cependant la décimation ne peut être utilisée dans les systèmes où une grande précision est requise (telle que la reconnaissance, l'identification, la modification de la F_0 etc.).

2.6.2. Phase de traitement

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

2.6.3. Phase de post-traitement

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

- les erreurs de voisement : lorsqu'une valeur F_0 a été trouvée sur une zone non voisée, ou lorsque aucune n'a été trouvée sur une zone voisée (phénomène de coarticulation);
- les erreurs grossières : lorsque la F_0 correspond à une harmonique ou une sous-harmonique. Ce type d'erreur peut facilement être corrigé en tenant compte du voisinage ou en effectuant un lissage;
- les erreurs fines : la valeur trouvée est située à plus ou moins 10% de la valeur réelle.

Rappelons qu'une désaccentuation est souvent nécessaire si toutefois il y a eu une pré-accentuation. Nous voyons donc, par le biais de ces trois phases, que d'innombrables techniques sont introduites dans le souci d'améliorer le temps de traitement et d'augmenter la précision, et qui malheureusement sont souvent sources d'erreurs supplémentaires.

2.7. Méthodes de détection du Pitch

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

2.7.1. Méthodes temporelles

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

Lors de la mise en œuvre de ces méthodes, le signal est découpé 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.

2.7.1.1. Algorithmes de type corrélation

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

2.7.1.1.1. Fonction d'auto-corrélation

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

$$FP_{AutoC}(\tau) = \frac{1}{N}\sum_{i=1}^{N-\tau} x_i x_{i+\tau} \qquad (2.2)$$

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

Les maxima de la FP correspondent à des multiples de la période fondamentale (Figure. 2.4). Le premier pic (le décalage donnant la meilleure corrélation) indique la valeur de la F_0. Cette méthode suppose que le signal soit stationnaire, tout au moins dans la trame utilisée or, cette hypothèse est rarement valide sur les signaux étudiés. Avec un signal de parole, l'extraction de la période est donc moins simple car la moindre irrégularité du signal peut provoquer l´apparition de pics dont le décalage est inférieur à T_0. Dans ces conditions, pour améliorer la recherche du pic correspondant à la période, il est possible de choisir d'autres heuristiques, par exemple, en ne retenant que les pics d'amplitude supérieure à un certain seuil (par exemple 50% du maximum de la fonction d'autocorrélation) [15].

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

- le choix de la fenêtre adéquate pour le calcul à court terme, afin d'atténuer son influence sur la fonction d'autocorrélation. Cependant, la fenêtre idéale doit contenir 2 à 3 périodes de pitch. Sa durée doit être située entre 5 et 20 ms pour des valeurs élevées de la F_0 et entre 20 et 50 ms pour des valeurs plus faibles.
- le premier maximum peut être lié à la structure des formants surtout pour les voix féminines ou enfantines.

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

L'avantage de cette méthode est qu'elle est très simple, ne nécessite pas un temps de calcul trop coûteux et donne des résultats relativement satisfaisants [1].

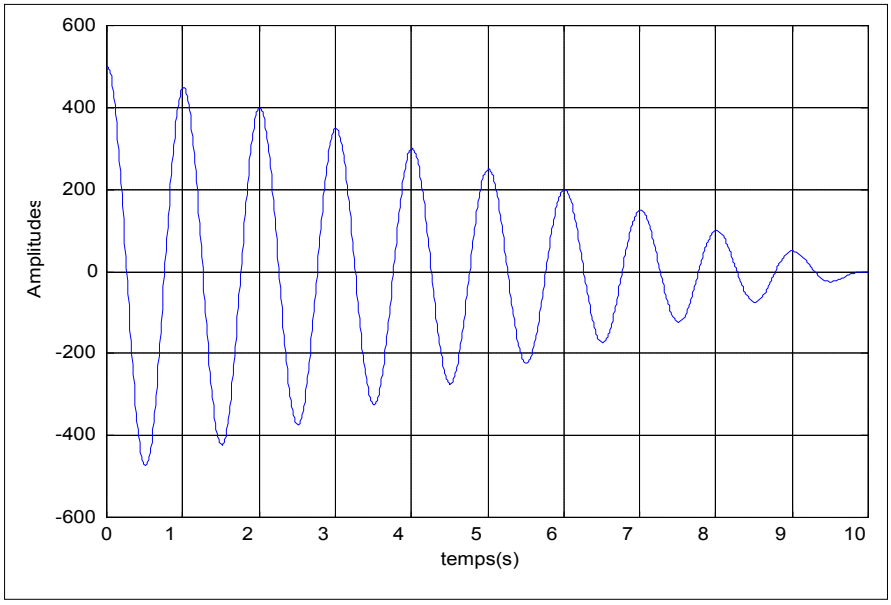


Figure 2.4 : Fonction d'Autocorrelation d'un signal périodique (sinus)

2.7.1.1.2. Fonction de distance (AMDF)

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

$$FP_{AMDF}(\tau) = \sum_{i=1}^{N} \left| x_i - x_{i+\tau} \right| \qquad (2.3)$$

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

La FP présente un minimum au niveau des multiples de la période. Cette méthode n'utilise pas l'hypothèse de stationnarité du signal. D'ailleurs, l'ambiguïté entre les pics T_0, $2T_0$, $3T_0$, . . . nT_0 est souvent atténuée par la non-stationnarité du signal analysé : plus le décalage est grand, plus le signal, de part sa non-stationnarité, intègre des différences par rapport à la trame de départ; le signal présente alors plus de différences pour un décalage de T_0 que pour un décalage de $2T_0$ [15]. Une conséquence de la non utilisation de l'hypothèse de stationnarité est que le choix de la taille des fenêtres de signal et celui décalé est libre : l'AMDF utilise des fenêtres de taille fixe, mais il est possible de concevoir des algorithmes avec des tailles de fenêtre variable, par exemple égale au décalage testé. Cette résistance au problème des erreurs grossières et la rapidité de calcul font de l'AMDF une méthode couramment employée.

2.7.1.1.3. Super résolution (SRPD)

Le SRPD (Super Resolution Pitch Determination) est un algorithme proposé par Medan, Yair et Chazan [15], avec comme objectif initial de réduire le plus possible les "erreurs fines" d'estimation de la F_0. L'idée de l'algorithme est de comparer selon une mesure de ressemblance deux fenêtres de signal décalées de la valeur de la période test. Cela ressemble fort aux algorithmes du type AMDF, la différence essentielle étant que la taille des fenêtres est ici variable, et plus exactement égale à la période de test. Ainsi l'algorithme vise à positionner au mieux deux fenêtres successives représentant deux périodes successives du signal. La FP de l'algorithme SRPD s'écrit alors :

$$FP_{SRPD}(\tau) = \frac{\sum_{i=1}^{\tau} x_i x_{i+\tau}}{\sum_{i=1}^{\tau} x_i^{\,2} \sum_{i=1}^{\tau} x^{2}_{i+\tau}} \qquad (2.4)$$

Soquet [15] a trouvé que le SRPD, malgré sa grande simplicité donne des taux d'erreurs très acceptables, seules les erreurs de sous-harmoniques présentent un score relativement élevé.

2.7.1.2. Algorithmes basé sur le filtre inverse

La modélisation LPC (Linear Predictive Coding), est en effet aussi applicable en détection de pitch. Markel en 1972 a proposé une méthode de détection qui pourrait être considérée comme temporelle [4]. Elle est basée sur l'examen de la fonction d'autocorrélation du résidu LPC. Cette particularité de la méthode permet en fait de travailler directement sur la source évitant ainsi l'interaction source-conduit vocal, cette méthode est connue sous le nom de méthode SIFT (Simplified Inverse Filtrer Tracking).

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

Le signal microphonique capté à la sortie des lèvres est, en fait le résultat de différents filtres mis en cascade. Chacun de ces filtres apporte une certaine déformation au signal de parole. Ainsi le signal vocal peut être exprimé par l'équation (2.1). Il suffit d'appliquer à x(n) un autre filtre h'(n) qui est l'inverse de h(n), c'est- à- dire :
H'(f)=1/H(f), pour obtenir le signal d'excitation e(n) (déconvolusion de la sortie) (Figure.2.5)

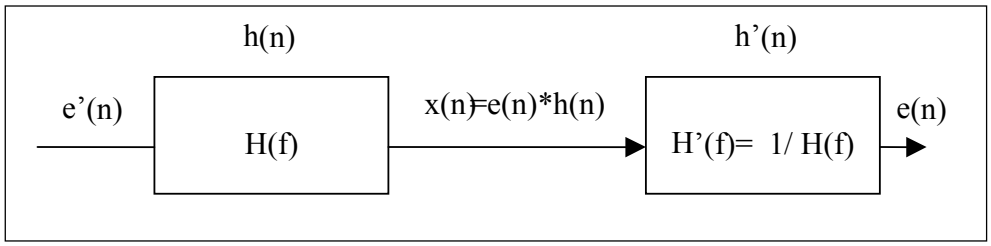


Figure.2.5 : Modélisation du filtre inverse

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

Afin de faciliter le calcul en temps réel, on opère une décimation dans le rapport 5 à 1 après passage par un filtre passe-bas. On procède après à une analyse LPC (P=4) pour définir

le filtre inverse. L'ordre 4 a été choisi comme étant l'ordre optimal suffisant pour la gamme de fréquence utilisée, soit 0-900 Hz et il permet une bonne vitesse de traitement [16]. Les résultats pratiques ont montré que cette méthode est bien adaptée aux applications en temps réel. Cependant la précision atteinte avec ce détecteur est assez médiocre.

2.7.2. Méthodes spectrales

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

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

2.7.2.1. Méthode du cepstre

Avec la méthode du cepstre, on arrive à séparer la source du conduit vocal, en prenant logarithme du spectre. On passe ainsi d'un produit à une somme. On calcule ensuite la Transformée de Fourier Inverse (TFI), et on obtient le cepstre dans le domaine des quéfrences. La figure (2.6) représente le schéma de la méthode du cepstre.

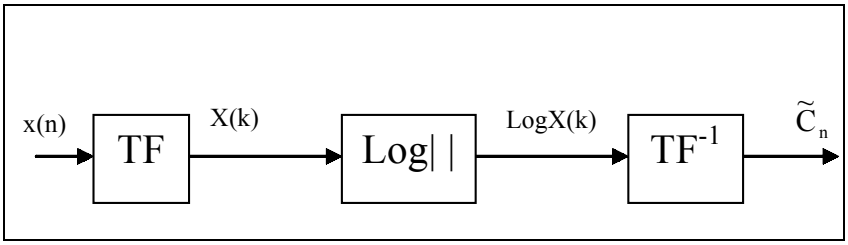


Figure 2.6 : Méthode du cepstre

$$\text{Cepstre}(x(n)) = TF^{-1}(\text{Log}\,|X(f)|) \qquad (2.5)$$

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

Le signal à la sortie du microphone est donné par équation 2.1.Par TF on obtient

$$X(f) = E(f).H(f) \tag{2.6}$$

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

$$\mathrm{Log}(X(f)) = \mathrm{Log}(E(f)) + \mathrm{Log}(H(f)) \tag{2.7}$$

Par TFI on obtient :

$$TF^{-1}(\mathrm{Log}(X(f))) = TF^{-1}(\mathrm{Log}(E(f))) + TF^{-1}(\mathrm{Log}(H(f))) \tag{2.8}$$

Enfin

$$\mathrm{Cepstre}(x(n)) = \mathrm{Cepstre}(e(n)) + \mathrm{Cepstre}(h(n)) \tag{2.9}$$

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

2.7.2.2. Méthode d'intercorrelation avec la fonction peigne

La méthode d'intercorrelation avec une fonction peigne proposée par Martin [15] utilise une représentation fréquentielle à court terme. Les algorithmes utilisant cette représentation exploitent généralement la structure harmonique des signaux périodiques.

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

$$FP_{\text{peigne}}(f) = \sum_{i=1}^{n(f)} \alpha_i \left| S(i * f) \right| \tag{2.10}$$

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

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

Dans ces conditions, la FP présente un maximum pour la F_0 même si cette fréquence est la seule composante du spectre dépourvu de toute harmonique, ou si seules deux harmoniques consécutives sont présentes. Le volume de calcul de cette méthode est du même ordre que celui du cepstre.

2.7.3. Méthodes combinatoires

Il existe un très grand nombre de méthodes pour extraire la F_0, chacune présentant des avantages et des inconvénients, mais aucune ne permet d'évaluer la fréquence fondamentale avec une précision absolue. Ces observations ont conduit Hess [15] 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.

2.8. Evaluation de la synthèse

L'évaluation de la synthèse de la parole est un domaine de recherche et d'application en prise directe avec le développement de la synthèse elle-même. En effet, si l'on savait évaluer précisément et diagnostiquer les défauts de qualité des synthétiseurs, on saurait aussi comment y remédier, ou au moins comment chercher les solutions.

Donc les progrès de l'évolution, de l'analyse de la qualité de la parole synthétique, sont intimement liés à ceux des algorithmes de synthèse.

2.8.1. Pourquoi évaluer la parole synthétique ?

L'évaluation de la synthèse est une nécessité, et elle le restera tant que la parole synthétique sera d'une qualité inférieure à la parole naturelle (ce qui sera évidemment toujours

le cas, puisque la parole fait appel à toutes les facultés, physiques, psychiques et spirituelles de l'homme).

Au stade de la conception des systèmes, il faut être capable de mesurer objectivement les progrès (dont le concepteur n'a qu'une idée très biaisée à cause de son implication). Au niveau de l'utilisation des systèmes, il faut pouvoir décider et choisir. Concepteurs et utilisateurs ont finalement des besoins communs, pour le diagnostic, la comparaison et la normalisation des tests :

- tests diagnostiques des systèmes ou de leurs composantes, afin de vérifier le niveau de qualité des systèmes ou de leurs composantes;
- comparatifs pour évaluer les mérites et classer les systèmes ou leurs composantes;
- tests normatifs, pour choisir un système pour un usage donné.

En général, la référence de l'évolution des systèmes est la parole naturelle, ou de la parole naturelle avec un certain niveau de dégradation de signal.

La grande majorité des tests utilise le jugement d'un groupe d'auditeurs. Cela implique d'une part des précautions méthodologiques (par exemple : effets d'accoutumance et d'apprentissage, compétences et habitudes linguistiques ou auditives des sujets, types de tâches, motivation des sujets). Cela implique aussi qu'en synthèse, fort peu de tâches d'évaluation peuvent être entièrement automatisées. Ainsi l'évaluation de la synthèse, dépendante d'un facteur humain, sera un processus lent et coûteux, dès qu'on exige des garanties d'objectivité plutôt que des impressions.

La synthèse à partir du texte est en quelque sorte une chaîne de traitements, depuis le texte jusqu'au signal acoustique. C'est le maillon le plus faible de la chaîne qui va en limiter la qualité, donc il est important dans ce contexte d'évaluer chaque maillon, en plus de l'évaluation globale.

On distingue deux types d'évaluation des systèmes de synthèse de la parole à partir du texte : l'évaluation analytique, ou interne, « boîte verre » et l'évaluation globale ou externe « boîte noire ».

2.8.2. Evaluation globale

Par évaluation globale, on entend évaluation de la sortie du système de synthèse, sans se préoccuper de son fonctionnement interne et sans chercher la source des défauts éventuels.

Ce type d'évaluation sert à la fois au concepteur du système et aux utilisateurs. On peut recenser dans la littérature deux grandes familles de tests consacrés à l'évaluation externe de la parole de synthèse :

- la première famille concerne l'évaluation quantitative et qualitative de l'intelligibilité de la synthèse ;
- la deuxième famille vise à donner une note d'appréciation plus globale sur la qualité subjective de la parole synthétique suivant des analyses multi échelles ou des analyses sémantiques.

Parmi les méthodes utilisées pour l'évaluation globale, nous pouvons citer :

- la méthode catégorielle de jugement des dégradations qui permet une comparaison directe de plusieurs systèmes, appelée DCR : (Degradation Cathegory Rating) ;
- la méthode catégorielle de jugement absolu qui est utilisée pour évaluer et comparer la qualité des systèmes par rapport à une référence, fait entendre les différents systèmes séparément, elle est appelée ACR : (Absolute Cathegory Rating);

Exemples :

Verbmobil : Une procédure multidimensionnelle d'évaluation de la qualité globale de la langue allemande en 1995 sous le projet Verbmobil.

JEIDA (Japan Electronic Industry Devlopment Association) : Il s'agit d'échelles sémantiques, utilisant des paires de mots opposés pour une évaluation multidimensionnelle de la langue Japonaise.

2.8.3. Evaluation analytique

L'évaluation globale renseigne sur la qualité atteinte par les systèmes, en référence par exemple à la parole naturelle plus au moins dégradée, ou à de la parole codée.

Pour le concepteur de systèmes, ou pour des applications spécifiques, il est nécessaire de tester également les composantes des systèmes séparément.

2.8.3.1. Transcription Graphème Phonème

Même si la conversion graphème phonème semble plutôt plus simple à évaluer que d'autres composantes d'un système de synthèse, de nombreuses questions se posent au préalable :

- Quel alphabet phonétique adopter ?
- Evaluer sur un texte ou sur une lexique ?
- Quelle est le format d'entrée, vierge ou non de toute imperfection comme les fautes d'orthographes, etc.….

Des réponses pratiques à ces questions ont été apportées pour le français dans [16].

2.8.3.2. Module prosodique

Les recherches sur l'évaluation de la synthèse ont montré que les sujets testés sont généralement plus confiants dans leur propre jugement lorsqu'il s'agit de comparer des différences acoustiques globales plutôt que des différences plus spécifiques au niveau suprasegmentale ou prosodique. Il s'agit d'évaluer si la prosodie synthétique permet de percevoir correctement le contenu linguistique désiré par exemple, affirmation, question, doute, etc.

En effet, les scores obtenus pour des tests concernant la prosodie d'un système vont étroitement dépendre de la qualité segmentale et inversement. L'évaluation de la prosodie en soit est donc délicate, et aucun test standard n'existe.

2.8.3.3. Synthétiseur acoustique

Le synthétiseur acoustique intervient dans un système de synthèse pour l'analyse/modification/synthèse de signaux de parole utilisé en synthèse de signaux de parole utilisés en synthèse par concaténations, ou pour le calcul du signal à travers un synthétiseur à formants. Le problème de l'évolution est de mesurer l'aptitude des systèmes d'analyse / synthèse ou de synthèse à effectuer des déformations diverses de signaux naturels, ou à produire des signaux de synthèse réaliste, pour diverses tâches prosodiques, au sens large (changement de la F_0 de la durée, d'intensité, de qualité de la voix).

Dans la liste potentielle des synthétiseurs, on trouve des systèmes opérants :

- directement sur le signal temporel ;
- sur une séparation source - conduit vocal.

La modification prosodique consiste à déformer un signal synthétique source de manière à approcher au mieux les caractéristiques extraites d'un autre signal, la cible prosodique.

Le signal original peut être soit naturel, soit constitué de segments (diphones, phonèmes) extraits d'une base de données.

De nombreuses comparaisons existent déjà sur le jugement de qualité de telle ou telle méthode de concaténations, mais une approche plus ambitieuse est décrite dans [13]. Il s'agit de constituer des couples de phrases prononcées avec divers débits, diverses intonations, diverses forces d'articulation, divers styles, de manière à évaluer la capacité du système à transférer la prosodie de l'une sur l'autre. L'avantage est de fixer une tâche que l'on peut étalonner et dont on connaît la référence absolue. Ce banc d'essai peut inclure divers niveaux d'évaluation (transfert de la fréquence fondamentale, de durée, mais aussi du type de voix : soufflée, rauque, pressée, de la réduction vocalique, etc).

2.9. Conclusion

La synthèse de la parole est loin d'être un problème résolu. Si les synthétiseurs offrent aujourd'hui une bonne qualité segmentale, et donc une bonne intelligibilité, il n'en va pas de même de leur naturel. C'est sans doute ce qui freine encore leur apparition dans un bon nombre de produit grand public comme le montre en effet de façon flagrante la synthèse par concaténations avec transplantation prosodique (c'est-à-dire la production par un synthétiseur de phrase dont l'intonation et la durée des phonèmes sont directement copiées sur celles mesurées sur une prononciation humaine préalable du même texte). Dès qu'il s'agit de laisser à la machine le choix de ces partions intonatifs et rythmiques la faiblesse de génération prosodique ce fait cruellement sentir. On peut dans certains cas obtenir une parole de synthèse qu'il est difficile de distinguer d'une voix humaine. L'obtention d'un synthétiseur de très haute qualité ne se fera donc qu'aux prix d'une étude approfondie des problèmes qui subsistent en sélection totale, de façon à trouver un optimum intelligibilité/naturel.

La fréquence fondamentale est un paramètre très important pour la synthèse de la parole ; l'oreille est en effet très sensible à ses variations, lesquelles constituent un élément essentiel de la prosodie. La mesure de ce paramètre prosodique peut se faire à partir d'un signal de parole dans le domaine temporel ou fréquentiel. Les méthodes spectrales résistant au bruit, conviennent pour l'étude des macro variations de la F_0. Les dispositifs opérant dans le domaine temporel seront par contre souhaitables pour l'analyse de la micro mélodie (les variations cycle par cycle) [10]. Malgré leur complexité, tous les dispositifs proposés à ce jour présentent des défaillances dans des conditions spécifiques.

CHAPITRE 3
TECHNIQUES DE MODIFICATIONS DE LA FREQUENCE FONDAMENTALE

3.1. Introduction

Effectuer une synthèse de haute qualité demande un contrôle précis des paramètres de qualité vocale, qui dépendent principalement de la source de voisement.

Les paramètres acoustiques qui portent les informations prosodiques (F_0, intensité, durée des unités phonétiques et des pauses, énergie du signal) doivent être ajustés contextuellement et évalués par rapport aux capacités de la perception. Par ailleurs, même si les commandes peuvent être activées indépendamment, certains liens fonctionnels entraînent des modifications corrélées des paramètres. L'évolution de la fréquence laryngienne porte les traces de l'ensemble des variations des commandes; cette particularité justifie la prééminence des travaux sur la mélodie. Les phénomènes prosodiques sont complexes et se manifestent conjointement sur plusieurs paramètres.

Dans ce chapitre nous allons faire une étude des différentes techniques qui permettent la modification de la fréquence laryngienne, en faisant une comparaison de point de vue qualité sonore et complexité d'élaboration des systèmes. Nous nous intéressons principalement aux techniques de modifications de la F_0, et plus précisément à celles utilisées pour un signal de parole. En effet, il existe trois méthodes pour la modification de la F_0, deux d'entres elles sont applicables pour un tel signal et une essentiellement applicable pour des signaux musicaux. Soient la technique PSOLA, la modélisation physique et la technique IPS (Instrumental Pitch Shifting).

3.2. Techniques de modifications de la fréquence fondamentale

La modification de la F_0 et le lissage spectral de signaux de parole sont des tâches difficiles à réaliser sans qu'il en résulte des dégradations dans la qualité du signal. Une bonne partie des recherches menées en synthèse vocale durant les années 80 et 90 ont précisément porté sur la mise au point des méthodes d'analyse-synthèse de parole permettant de résoudre ces problèmes avec plus ou moins de succès.

3.2.1. Modification instrumentale du pitch IPS

La modification instrumentale (IPS) permet la modification du pitch d'un son. Elle est utilisée dans le cas des sons musicaux pour l'obtention des différentes fréquences. Cet algorithme a approximativement le même effet sur le spectre du son que le rééchantillonnage, il procède par un changement d'échelle de l'axe des fréquences du spectre (Figure.3.1).

La différence entre l' IPS et le rééchantillonnage est que ce dernier compresse ou élargit l'échelle des temps. Suréchantillonner un son donnera un pitch plus élevé mais le signal résultant sera également plus court, souséchantillonner un son donnera un pitch inférieur mais le signal résultant aura également un temps plus long, tandis que la technique IPS rééchantillonne le spectre, mais n'affecte pas l'échelle des temps.

Nous pouvons voir sur la figure (3.1) que le spectre est élargi dans le cas où on veut augmenter la F_0.

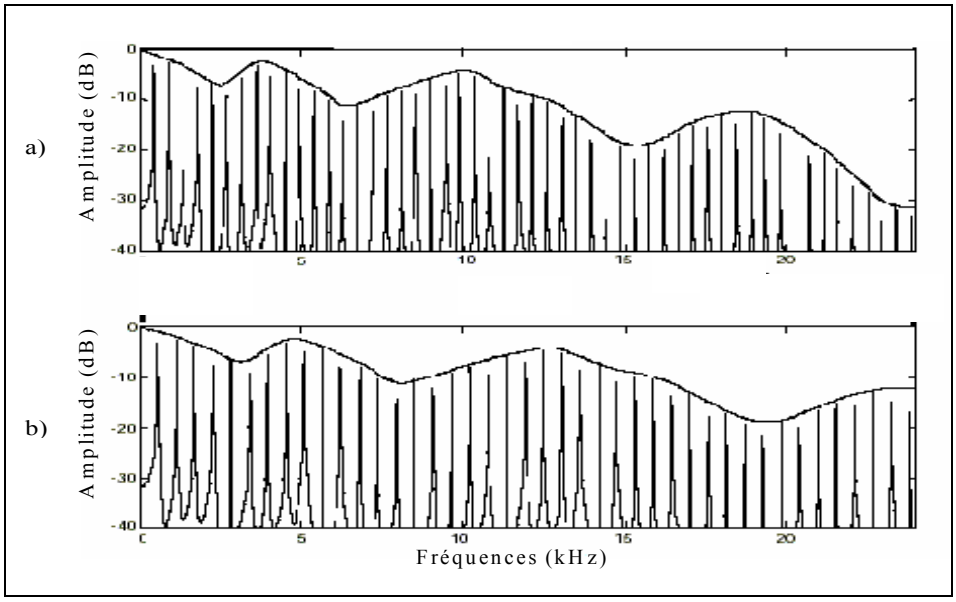


Figure 3.1 : Opération d'augmentation du pitch par la technique IPS [17].
a) : Enveloppe spectrale du signal réel.
b) : Enveloppe spectrale du signal transposé.

La technique IPS consiste à rééchantillonner le signal original uniquement dans les parties Voisées. Le signal synthétique résultant est alors constitué des parties Non Voisées, qui n'ont pas été modifiées mais simplement recopiées sur le signal de synthèse, et des parties Voisées du signal qui ont été modifiées sans changer la durée totale du signal original.

Avant de pouvoir rééchantillonner le signal, il faut que les zones Voisées et Non Voisées du signal soient déterminées ainsi que la valeur du pitch pour chaque trame du signal de manière à réaliser la modification du pitch uniquement sur les zones Voisées du signal original.

Pour ne pas changer la durée totale du signal original, il faut que la durée d'une trame reste constante soit pour l'augmentation du pitch ou soit pour sa diminution. On est amené donc à dupliquer ou éliminer autant de fois que nécessaire les périodes modifiées de manière à occuper toute la trame.

- Pourquoi IPS ne fonctionne t-elle pas pour la voix ?

Comme nous avons vue au chapitre 1, la voix humaine est faite d'un signal produit par les cordes vocales (le passage de l'air expiré à travers les cordes vocales) et filtré par le conduit vocal.

Le conduit vocal a quelques fréquences de résonance (formants) qui dépendent de la position de la mâchoire, la langue…, ces fréquences de résonance peuvent être vues en tant que maximums locaux de l'enveloppe spectrale.

Comme nous pouvons indépendamment contrôler les cordes vocales (la source du voisement) et le conduit (le résonateur), nous pouvons changer la fréquence fondamentale et les fréquences formantiques séparément.

Si nous changeons le pitch d'un signal vocal à l'aide d'un IPS, nous transposons les formants aussi bien que le pitch, et modifions ainsi le timbre de la voix [17].

Par exemple, la transposition à pitch plus élevé augmente les fréquences de résonance. De même, la transposition à pitch inférieur diminue les fréquences de résonance, dans les deux cas, la voix semble également fortement artificielle.

Afin d'être conformes aux caractéristiques humaines de voix, nous devons changer le pitch sans modifier les fréquences formantiques (Figure.3.2).

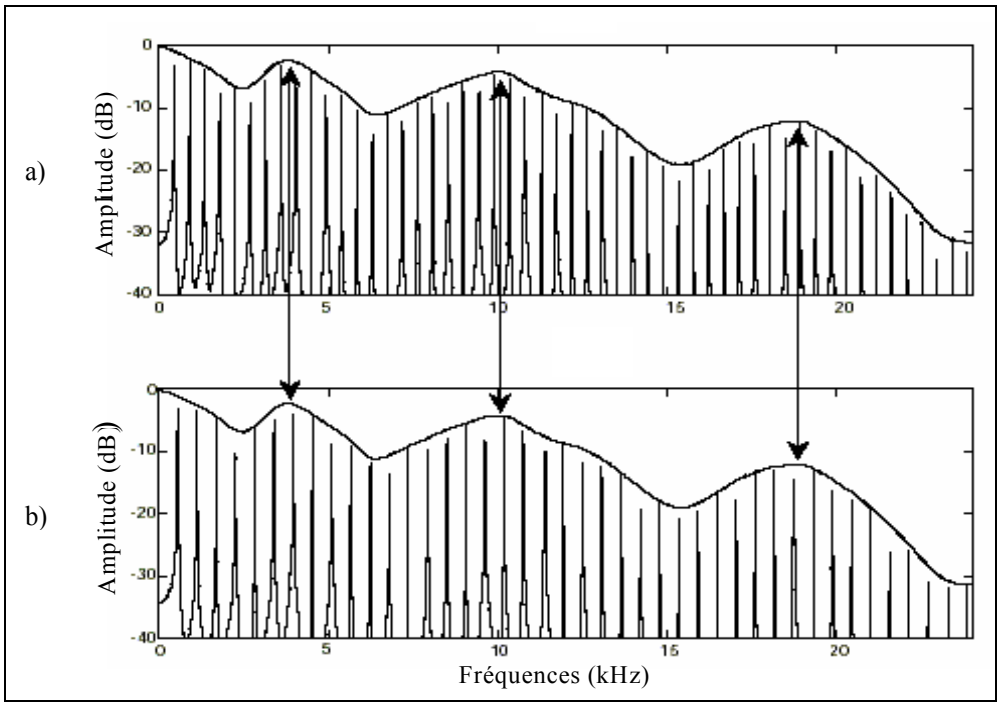


Figure 3.2 : Opération d'augmentation du pitch sans changement d'enveloppe spectrale [17].
a) : Enveloppe spectrale du signal réel.
b) : Enveloppe spectrale du signal transposé.

Nous pouvons voir dans la figure (3.2) que le changement de pitch est effectué (la distance entre les harmoniques a augmenté) sans altérer l'enveloppe spectrale [17].

A l'aide de cette transformation nous pouvons contrôler les divers paramètres qui définissent le caractère (timbre) d'une voix qui est malheureusement non réalisable à l'aide de la technique IPS. Nous allons voir par la suite que les méthodes subséquentes tentent à réaliser ces buts.

3.2.2. Modélisation physique Source - Filtre

Le signal vocal peut être décrit comme source + modèle de filtre, la source étant les cordes vocales (le passage de l'air expiré à travers les cordes vocales) et le filtre étant le conduit vocal. L'idée de la modélisation physique est d'analyser le signal d'entrée pour séparer l'information glottale de l'information conduit. Le conduit vocal peut se subdiviser en trois cavités, le pharynx (du larynx au voile du palais et à l'arrière de la langue), la cavité orale (du pharynx au lèvres) et la cavité nasale. La géométrie du conduit vocal dépend des organes articulatoires : mâchoires, lèvres, langue.

Nous commençons d'abord par établir un modèle du conduit vocal d'un locuteur, connaissant la sortie de ce modèle (la voix du locuteur), nous pouvons estimer le signal

d'entrée (l'excitation glottique). Comme nous pouvons manipuler le signal glottal et le modèle du conduit séparément, nous pouvons contrôler le pitch (l'information glottale) et les formants (l'information du conduit vocal) indépendamment.

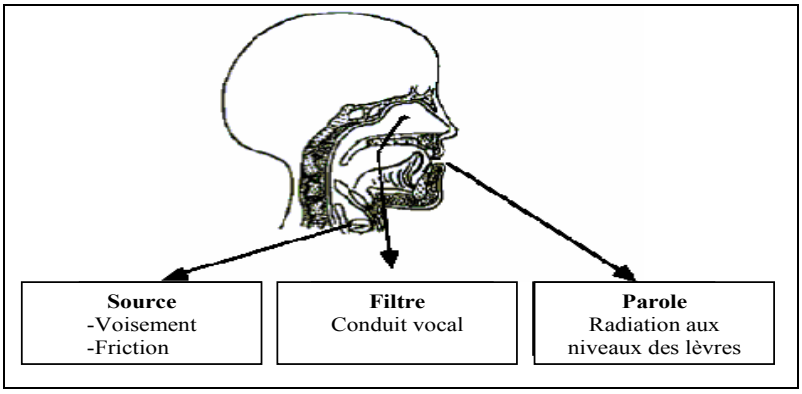


Figure 3.3 : Modélisation physique Source -Filtre

3.2.2.1. Principe de fonctionnement de la modélisation physique Source - Filtre

Nous supposons que le signal glottal est approximativement un train d'impulsions pour les signaux Voisés, et un bruit blanc pour les signaux Non Voisés, ainsi l'enveloppe spectrale du signal glottal est plat (Figure 3.4).

Comme le signal d'entrée du conduit vocal est censé avoir une enveloppe spectrale plate, la forme du spectre de la voix dépend seulement du conduit vocal. Nous pouvons dériver de cette figure que le conduit vocal agit en tant qu'un filtre qui a comme réponse en fréquence l'enveloppe spectrale du signal de sortie. En conséquence, si nous pouvons concevoir un filtre équivalent, nous aurons un modèle du conduit vocal.

Le modèle physique décrit le signal vocal comme un signal glottal (avec une enveloppe spectrale plate) plus un filtre du conduit vocal contenant les informations sur les formants.

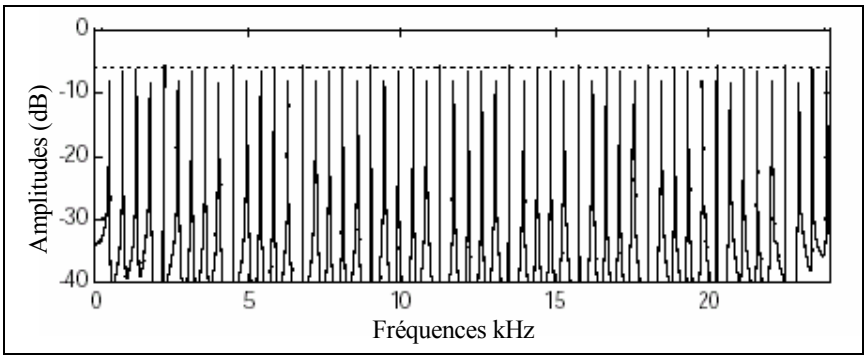


Figure 3.4 : Enveloppe spectrale du signal glottique.

3.2.2.2. Modélisation Source- Filtre par prédiction linéaire

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage, ou de reconnaissance. Elle repose en général sur un modèle. Celui-ci possède un ensemble de paramètres numériques, dont les plages de variations définissent l'ensemble des signaux couverts par le modèle. Pour un signal et un modèle donné, l'analyse consiste en l'estimation des paramètres du modèle dans le but de lui faire correspondre le signal analysé.

Pour ce faire, on met en oeuvre un algorithme d'analyse, qui cherche généralement à minimiser la différence, appelée erreur de modélisation, entre le signal original et celui qui serait produit par le modèle (signal approché).

S'il existe de nombreux modèles de parole, il en est un que l'on retrouve partout, et dans un nombre croissant d'appareils « grand-public » : le modèle prédictif linéaire (LPC : Linear Predictive Coding). Particulièrement adapté à la modélisation Source Filtre des signaux de parole, nous en étudions dans ce paragraphe les principes théoriques avant de nous intéresser à ses avantages et inconvénients dans le cadre d'analyse synthèse des signaux de parole.

3.2.2.3. Principe de la prédiction linéaire

Sur le système d'entrée/sortie modélisé (Figure 3.5), le signal de sortie x(n) s'écrit comme une combinaison linéaire des échantillons du signal de sortie observés aux p instants précédents, et des échantillons du signal d'entrée u(n*)* observés à l'instant présent et aux q instants précédents (Modèle ARMA, AutoRegressive Moving Average).[18]

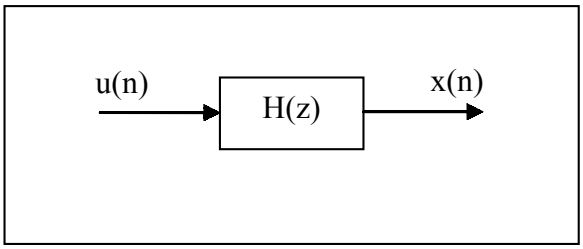


Figure 3.5 : Système d'entrée/sortie

Dans le domaine temporel on a :

$$x(n) + \sum_{i=1}^{p} a(i)x(n-i) = u(n) + \sum_{i=1}^{q} b(i)u(n-i) \qquad (3.1)$$

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

Dans le domaine fréquentiel, en désignant par H (z) la FT du système, l'équation (3.1) s'écrit :

$$H(z) = \frac{X(z)}{U(z)} = \frac{1 + \sum_{i=1}^{q} b(i) z^{-i}}{1 + \sum_{i=1}^{p} a(i) z^{-i}} \tag{3.2}$$

Où

$$X(z) = \sum_{i=0}^{\infty} x(n) z^{-i} \tag{3.3}$$

Les racines du numérateur et du dénominateur sont respectivement les zéros et les pôles du modèle. Caractériser le signal x(n) revient donc à estimer les coefficients {a(i), b(i)} pour une source connue u(n) (séquences d'impulsions ou bruit blanc).

Souvent pour simplifier la résolution de ce problème, on suppose que b(i)=0, $i \geq 1$ ce qui rend le modèle AR. Différentes variantes sont décrites pour déterminer les coefficients a(i) [10].

3.2.2.4 Modèle Auto-Régressif

Un signal x(n) sera dit Auto-Régressif s'il obéit au modèle suivant, lui aussi dit autorégressif :

$$x(n) + a(1)x(n-1) + ... + a(p)x(n-p) = e(n) \tag{3.4}$$

Où e(n) est un bruit blanc centré, gaussien, de variance σ^2. Les coefficients a(i) seront dits prédicteurs, car la quantité :

$$\hat{x}(n) = -\sum_{i=1}^{p} a(i)x(n-i) \tag{3.5}$$

est la prédiction de x(n) conditionnellement au passé {x(n-1), ….} infini. La relation (3.4) peut s'interpréter par la transformée en z :

$$X(z) = \frac{1}{A(z)} E(z) \tag{3.6}$$

$$\text{avec} \quad A(z) = 1 + a(1)z^{-1} + ... + a(p)z^{-p}$$

Comme la description de {x(n)} en tant que sortie du filtre de FT $\frac{1}{A(z)}$, c'est-à-dire d'un filtre récursif ou d'un système tout pôle.

L'entrée de ce système est le bruit blanc e(n). En combinant (3.4) et (3.5), on interprète aussi e(n) comme une erreur de prédiction [10] :

$$e(n) = x(n) - \hat{x}(n) \tag{3.7}$$

ce qui donne :

$$e(n) = \sum_{i=0}^{p} a(i)x(n-i); \quad \text{avec } a(0) = 1 \tag{3.8}$$

Ce qui justifie de rechercher les coefficients a(i) optimaux en minimisant cette erreur ou plus exactement en minimisant sa variance $\sigma_e^{\,2}$.

3.2.2.5. Variance de l'erreur de prédiction

Nous supposons dans cette section que le signal x(n) est aléatoire et stationnaire ; les coefficients a(i) sont donc indépendants du temps et peuvent être estimés une fois pour toute en prenant tous le temps qui convient [3].

L'estimation des coefficients de prédiction est basée sur la minimisation de la variance de l'erreur de prédiction :

$$\sigma_e^{\,2} = E\left[e(n)^2\right] = E\left[\sum_{i=0}^{p} a(i)x(n-i)\sum_{j=0}^{p} a(j)x(n-j)\right]$$

$$= E\left[\sum_{i,j=0}^{p} a(i)a(j)x(n-i)x(n-j)\right]$$

$$\sigma_e^{\,2} = \sum_{i,j=0}^{p} a(i)a(j)R_x(i-j) \tag{3.9}$$

Dans cette dernière expression, $R_x(k)$ représente la fonction d'autocorrélation du signal x

$$R_x(k) = E\left[x(n)x(n+k)\right] = \frac{1}{N}\sum_{i=0}^{N-k-1} x(i)x(i+k) \tag{3.10}$$

avec

$$R_x(0) = \sigma_x^{\,2} \tag{3.11}$$

La moyenne de x est supposée nulle; une composante continue ne porte en général aucune information utile et l'on peut aussi si nécessaire l'extraire à l'aide d'un filtre très simple [1].

Le vecteur des coefficients de prédiction d'ordre p sera noté :

$$a = [1, a(1), a(2), ..., a(p)]^T = [1, \underline{a}^T]^T \ . \tag{3.12}$$

et la matrice d'autocorrection du signal x(n) s'écrit :

$$R_x = \begin{bmatrix} R_x(0) & R_x(1) & \cdots & R_x(p-1) & R_x(p) \\ R_x(1) & R_x(0) & \cdots & \cdots & R_x(p-1) \\ \vdots & & & & \vdots \\ \vdots & & & & R_x(1) \\ R_x(p) & \cdots & \cdots & R_x(1) & R_x(0) \end{bmatrix} \tag{3.13}$$

Selon (3.9) et (3.12), la variance σ^2 de l'erreur de prédiction peut s'écrire :

$$\sigma_e{}^2 = [1, \underline{a}^T].R_x.[1, \underline{a}^T]^T \tag{3.14}$$

C'est une forme définie positive en a(i), ce qui assure l'unicité de son minimum[1].

On remarque que la matrice d'autocorrélation possède une structure très particulière, les éléments situés le long de chaque diagonale parallèle à la diagonale principale sont égaux ; une telle matrice est appelée matrice de Toeplitz. De plus, en l'occurrence, il s'agit d'une matrice de Toeplitz symétrique.

Si l'on dérive l'expression (3.14) par rapport aux coefficients a(i), (i=1,2,…,p) on obtient un système d'équations linéaires en ces a(i). Comme nous le verrons, la matrice de ce système est la matrice R_x dont la structure particulière conduit à un algorithme de résolution rapide et efficace.

En fait, cette résolution rapide est basée sur une méthode récurrente sur l'ordre de la prédiction : un vecteur $\underline{a}$ d'ordre m va être calculé à partir de celui d'ordre m-1, et ce pour m=1,2,…,p.

Ceci nous oblige à préciser les notations utilisées jusqu'à présent. L'indice x est omis lorsque cela est possible sans nuire à la bonne compréhension ; ainsi la matrice R_x de (3.13) sera notée R_p. Le vecteur $[1, \underline{a}]^T$ sera noté $[1, \underline{a}_p]^T$ et nous posons :

$$r(p) = \left[R_x(1), R_x(2) .... R_x(p)\right]^T \tag{3.15}$$

$$R_p = \begin{bmatrix} \sigma_x^{\,2} & r_p^{\,T} \\ r_p & R_{p-1} \end{bmatrix} \tag{3.16}$$

Observons que la matrice d'autocorrélation R_p (3.13) est constituée par la matrice R_{p-1} bordée par le vecteur r_p, par ce même vecteur transposé et par la variance du signal $R_x(0) = \sigma_x^{\,2}$.

La forme quadratique de l'équation (3.14) peut s'écrire :

$$\begin{aligned} \sigma_e^{\,2} &= [1, \underline{a}_p^{\,T}] . \begin{bmatrix} \sigma_x^{\,2} & r_p^{\,T} \\ r_p & R_{p-1} \end{bmatrix} \begin{bmatrix} 1 \\ \underline{a}_p \end{bmatrix} \\ &= \sigma_x^{\,2} + 2.r_p^{\,T}.\underline{a}_p + \underline{a}_p^{\,T}.R_{p-1}.\underline{a}_p \end{aligned} \tag{3.17}$$

On a :

$$\frac{\partial \sigma_e^{\,2}}{\partial \underline{a}_p} = 2.r_p + 2.R_{p-1}\underline{a}_p = 0 \tag{3.18}$$

De sorte que le système à résoudre soit :

$$R_{p-1}.\underline{a}_p = -r_p \tag{3.19}$$

Soit sous la forme développée :

$$R_x(k) = -\sum_{i=1}^{p} a_p(i).R_x(k-i) \quad \text{pour } k = 1,2,....,p \tag{3.20}$$

$$\text{Avec } R_x(0) = \sigma_x^{\,2}$$

La valeur minimisée de la variance de l'erreur de prédiction $\sigma_e^{\,2}$, qui sera notée $\sigma_p = \sigma_{e,min}^2$ vaut :

$$\sigma_{e,min}^2 = \sigma_x^{\,2} + r_p^{\,T}.\underline{a}_p = \sum_{i=0}^{p} a_p(i).R_x(i) = \alpha_p \tag{3.21}$$

D'autre part si l'on réunit (3.19) et (3.21), on obtient avec la contrainte $a(0) = 1$ les équations de Yule-Walker [1].

$$\begin{bmatrix} \sigma_x^{\,2} & r_p^{\,T} \\ r_p & R_{p-1} \end{bmatrix} . \begin{bmatrix} 1 \\ \underline{a}_p \end{bmatrix} = \begin{bmatrix} \alpha_p \\ \vdots \\ 0 \end{bmatrix} \tag{3.22}$$

D'où pour un processus AR-p causal la relation entre les paramètres du modèle et les autocorrection R(k), est donnée par :

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p) \\ R(1) & R(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & R(1) \\ R(p) & \cdots & R(1) & R(0) \end{bmatrix} \begin{bmatrix} 1 \\ a(1) \\ \vdots \\ a(p) \end{bmatrix} = \begin{bmatrix} \sigma_{e,min}{}^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.23}$$

Les équations (3.23) sont dites équations normales ou équation de Yule-Walker.

3.2.2.6. Estimation des paramètres d'un modèle AR

3.2.2.6.1. Calcul direct par la résolution du système

Les équations de Yule-Walker fournissent une relation entre les paramètres du modèle AR et ses coefficients d'autocorrélation. Elles fournissent donc un moyen d'estimer les paramètres d'un modèle AR en substituant aux instants d'autocorrection leurs valeurs estimées. Nous pouvons alors calculer les paramètres a(i) du modèle ainsi que la densité spectrale du signal synthétique par la résolution d'un système de (p+1) équations à (p+1) inconnues sans l'utilisation d'un algorithme de résolution.

3.2.2.6.2. Algorithme de Levinson

Nous avons vu que la matrice d'autocorrélation était une matrice de Toeplitz. Nous allons donner à présent un algorithme rapide dû, à l'origine, à Levinson et qui permet de résoudre les équations de Yule-Walker. Si k désigne la dimension de la matrice d'autocorrélation, cet algorithme est d'une complexité de k^2 alors qu'un algorithme général est en k^3. L'algorithme de Levinson est récursif : il calcule les coefficients de prédiction au rang m à partir de ceux obtenus au rang (m-1) [19]. Pour établir cette récursion nous abordons la notation suivante :

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(m-1) \\ R(1) & R(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & R(1) \\ R(m-1) & \cdots & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_{m-1}(0) \\ a_{m-1}(1) \\ \vdots \\ a_{m-1}(m-1) \end{bmatrix} = \begin{bmatrix} v_{m-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.24}$$

où l'indice (m-1) indique la solution au rang (m-1), avec $a_{m-1}(0)=1$ et v_{m-1} désigne l'erreur quadratique qui, dans le cas d'un AR d'ordre p, représente aussi pour $m \geq p$ la variance du processus d'excitation.

De façon plus compacte on peut écrire :

$$\left.\begin{aligned} R_{m-1}a^{F}_{m-1} &= \begin{bmatrix} v_{m-1} \\ 0_{m-2} \end{bmatrix} \\ R_{m-1}a^{B}_{m-1} &= \begin{bmatrix} 0_{m-2} \\ v_{m-1} \end{bmatrix} \end{aligned}\right\} \tag{3.25}$$

Où nous avons posé :

$$\left.\begin{aligned} a^{F}_{m-1} &= [1 \;\; a_{m-1}(1) \;\; \cdots \;\; a_{m-1}(m-1)]^{T} \\ a^{B}_{m-1} &= [a_{m-1}(m-1) \;\; \cdots \;\; a_{m-1}(1) \;\; 1]^{T} \end{aligned}\right\} \tag{3.26}$$

Les exposants F (comme Forward) et B (comme Backward) indiquent que les vecteurs d'autocorrélation sont pris respectivement dans le sens direct et dans le sens rétrograde.

En passant au rang m, la matrice d'autocorrélation s'écrit :

$$R_m = \begin{bmatrix} R_{m-1} & r^{B}_{m} \\ r^{BT}_{m} & R(o) \end{bmatrix} = \begin{bmatrix} R(0) & r^{FT}_{m} \\ r^{F}_{m} & R_{m-1} \end{bmatrix} \tag{3.27}$$

Où

$$\left.\begin{aligned} r^{B}_{m} &= [R(m) \;\; \cdots \;\; R(1)]^{T} \\ r^{F}_{m} &= [R(1) \;\; \cdots \;\; R(m)]^{T} \end{aligned}\right\} \tag{3.28}$$

En utilisant l'expression de la solution au rang (m-1), on a alors :

$$R_m \begin{bmatrix} a^{F}_{m-1} \\ 0 \end{bmatrix} = \begin{bmatrix} R_{m-1} & r^{B}_{m} \\ r^{BT}_{m} & R(0) \end{bmatrix} \begin{bmatrix} a^{F}_{m-1} \\ 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} v_{m-1} \\ 0_{m-2} \end{bmatrix} \\ r^{BT}_{m} a^{F}_{m-1} \end{bmatrix} \tag{3.29}$$

et

$$R_m \begin{bmatrix} 0 \\ a^{B}_{m-1} \end{bmatrix} = \begin{bmatrix} R(0) & r^{FT}_{m} \\ r^{F}_{m} & R_{m-1} \end{bmatrix} \begin{bmatrix} 0 \\ a^{B}_{m-1} \end{bmatrix} = \begin{bmatrix} r^{FT}_{m} a^{B}_{m-1} \\ \begin{bmatrix} 0_{m-2} \\ v_{m-1} \end{bmatrix} \end{bmatrix} \tag{3.30}$$

Par combinaison linéaire :

$$R_m \left[\begin{bmatrix} a^{F}_{m-1} \\ 0 \end{bmatrix} + k_m \begin{bmatrix} 0 \\ a^{B}_{m-1} \end{bmatrix} \right] = \begin{bmatrix} v_{m-1} + k_m r^{FT}_{m} a^{B}_{m-1} \\ 0_{m-2} \\ r^{BT}_{m} a^{F}_{m-1} + k_m v_{m-1} \end{bmatrix} \tag{3.31}$$

En choisissant :

$$k_m = -\frac{r^{BT}_{m} a^{F}_{m-1}}{v_{m-1}} \tag{3.32}$$

On annule le dernier terme et on obtient par identification la solution au rang m qui donne

$$v_m = v_{m-1} + k_m r_m^{FT} a_{m-1}^B \tag{3.33}$$

Sachant que par définition

$$r_m^{FT} a_{m-1}^B = r_m^{BT} a_{m-1}^F \tag{3.34}$$

On déduit d'après l'équation 3.32 que :

$$r_m^{FT} a_{m-1}^B = -k_m v_{m-1} \tag{3.35}$$

Et donc que :

$$v_m = v_{m-1}(1 - |k_m|^2) \tag{3.36}$$

Notons que $0 \le v_m \le v_{m-1}$ et que, $|k_m| \le 1$

Dans la littérature du traitement du signal, les coefficients k_m portent le nom de coefficients de réflexion. Par identification, on déduit à partir de l'équation 3.31 les coefficients au rang m en fonction de ceux obtenus au rang (m-1) :

$$a_m^F = \begin{bmatrix} a_{m-1}^F \\ 0 \end{bmatrix} + k_m \begin{bmatrix} 0 \\ a_{m-1}^B \end{bmatrix} \tag{3.37}$$

En particulier $a_m(0) = 1$ et $a_m(m) = k_m$. En résumé, partant de la suite des autocorrélations R(k) ou en pratique de leurs estimées, l'algorithme de Levinson est :

Tableau 3.1 : Algorithme de Levinson.

Valeurs initiales : $a_0(0) = 1$ et $v_0(0) = R(0)$
Pour $m = 1, \ldots\ldots, K$, répéter :
1. $k_m = -\dfrac{R(m)a_{m-1}(0) + \cdots + R(1)a_{m-1}(m-1)}{v_{m-1}}$
2. $a_m(0) = 1, a_m(m) = k_m$
3. Pour $j \in \{1, \cdots, m-1\}$:
$a_m(j) = a_{m-1}(j) + k_m a_{m-1}(m-j)$
4. $v_m = v_{m-1}(1 - \|k_m\|^2)$

Dans le cas où le processus est AR d'ordre p, on montrera pratiquement que les coefficients $a_m(m)=0$ pour $m \geq p+1$, ce qui permet d'arrêter la boucle précédente [19].

3.2.2.6.3. Algorithme de Burg

L'idée de J.P. Burg est d'estimer directement à partir des données les coefficients de réflexion et ce sans passer par le calcul préalable des autocorrélations.

Une fois les coefficients de réflexion calculés, on déduit les paramètres du modèle par l'équation récurrente 3.36. L'algorithme proposé par J.P.Burg est donné par [20].

3.2.2.7. Stabilité du modèle AR

Il est clair qu'un modèle doit être stable sous peine d'être inutilisable, un modèle AR a pour FT $\frac{1}{A(z)}$. Il est stable lorsque toutes les racines du A(z) sont situées à l'intérieur du cercle unitaire. A(z) est à phase minimale ce qui garanti la stabilité du filtre H(z) puisque les zéros de A(z) deviennent les pôles de H(z) [18].

Le critère de stabilité s'exprime très simplement en fonction des coefficients de réflexions k_m qui sont souvent appelés coefficients de corrélation partielle : par la condition suivante [1] : Un modèle AR est stable si seulement si $|k_m| \leq 1$; qui est vérifié pratiquement dans l'algorithme de Levinson.

3.2.2.8. Interprétation de la prédiction linéaire

Dans ce qui précède, le signal e(n) a été considéré comme une erreur de prédiction dont on a minimisé la variance pour calculer les coefficients de prédiction linéaire a(k). Sa définition était comme le montre l'équation 3.6 interprété par la transformée en z.

La première représentation montre que l'on peut reconstruire le résidu e(n) de l'estimation à partir du signal x(n) à l'aide d'un filtre non récursif représenté par la FT A(z).

Inversement et c'est la deuxième représentation, on notera que le résidu e(n) peut être considéré comme un signal d'excitation servant à créer le signal x(n) avec l'aide d'un filtre récursif tous pôles H(z) = $\frac{1}{A(z)}$. Dans le cas de la parole, ce signal d'excitation peut être périodique (sons Voisés) ou aléatoire (sons Non Voisés).

3.2.2.9. Modèle du conduit vocal

La production des sons met en oeuvre un certain nombre de muscles modifiant la forme du conduit vocal dans lequel circule un flux d'air. On y trouve les cordes vocales qui vibrent pour les sons Voisés et restent au repos pour les sons Non Voisés. Viennent ensuite le pharynx et la cavité buccale en parallèle avec la cavité nasale. La forme de ces parties est constamment modifiée pour créer le message sonore.

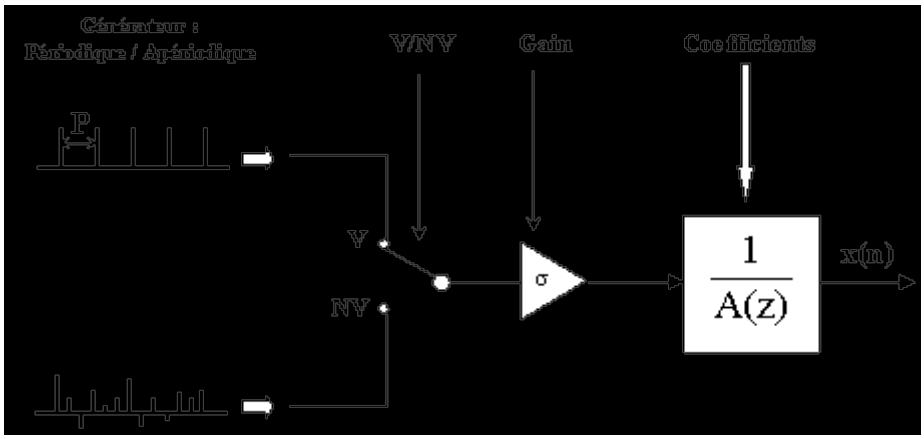


Figure 3.6 : Modèle du conduit vocal

Le modèle généralement adopté pour créer artificiellement des sons est grossi par rapport à la complexité du système phonatoire mais il est tout à fait satisfaisant pour les besoins de la téléphonie. Ce modèle comprend (Figure 3.6) :

- un générateur périodique d'impulsions unités ;
- un générateur de nombres aléatoires à valeur moyenne nulle et à variance unité ;
- un commutateur servant à choisir les sons Voisés ou Non Voisés ;
- un facteur σ appelé gain du modèle, il est choisi égal à $\sigma_{e,min} = \sqrt{\alpha_p}$, donner par l'équation 3.21.
- un filtre tous pôles $H(z) = \frac{1}{A(z)}$.

3.2.2.10. Modifications de la fréquence fondamentale

Les modifications de l'intonation des segments stockés sous forme LPC est une opération élémentaire. La période de pitch T_0 du signal étant un paramètre du modèle, il suffit d'en imposer directement la valeur de synthèse T (ce qui a pour effet de modifier la période du train périodique d'impulsions utilisé en excitation des zones Voisés).

Si l'on veut que cette opération n'affecte pas la puissance du signal, il faut modifier simultanément la valeur de σ à partir de sa valeur initiale σ_0 [1].

$$\sigma = \sigma_0 \sqrt{\frac{T_0}{T}} \tag{3.38}$$

3.2.2.11. Qualité segmentale

La modélisation Auto-Régressive souffre d'erreurs intrinsèques et extrinsèques qui en limitent fortement la qualité. L'expérience montre d'ailleurs qu'il est difficile d'améliorer cette qualité en augmentant l'ordre du modèle, où la F_e, ou encore la fréquence de rafraîchissement des paramètres.

Ainsi, les sons nasalisés, dont le spectre présente des antiformants autour de 1kHz, sont intrinsèquement mal modélisés. Par ailleurs, cette erreur intrinsèque induit une erreur d'estimation des formants eux-mêmes. Souvent, on met également en cause la forme même du signal d'excitation pour les sons Voisés.

On obtient une meilleure qualité de parole lorsqu'on utilise un signal dont l'amplitude spectrale est maintenue identique à celle de l'excitation de base, mais dont le spectre de phase est différent.

L'interrupteur Voisé/Non Voisé ne permet pas de rendre compte de façon réaliste des sons mixtes (comme les fricatives voisées) ; pour lesquels la glotte n'est jamais complètement fermée. L'utilisation d'une excitation mixte rend à la parole synthétique une certaine rondeur qui était absente dans le modèle binaire V/NV.

Enfin, la configuration du conduit vocal évalue parfois très vite, comme c'est le cas pour les plosives. Il est clairement impossible de rendre compte avec précision de phénomènes transitoires avec un modèle qui suppose le signal stationnaire sur une durée de 20 ou 30 ms.

Malgré tous ces problèmes de modélisation, le modèle AR a été et reste fort utilisé en synthèse vocale, vu sa grande simplicité et les rapports de compression exceptionnels qu'il permet d'obtenir.

3.2.3. Technique PSOLA (Pitch Synchronous Overlap and Add)

On a vu apparaître ces dernières années des méthodes basées sur une modification temporelle directe de la forme du signal.

L'idée sous-jacente est qu'il est possible de modifier l'intonation et la durée d'un signal sans l'usage d'aucun modèle paramétrique, en évitant ainsi toute possibilité d'erreur de modélisation.

La mise au point d'une nouvelle technique de synthèse PSOLA, applicable à tous les systèmes de synthèse dit «par concaténations », y attribue de façon essentielle dés les années 1990, en permettant d'avoir un timbre de voix notablement plus naturel que celui fourni par les systèmes antérieurs.

Sa variante dans le domaine temporel TD-PSOLA (Time Domain Pitch Synchronous Overlap Add) est relativement flexible, moins complexe et a donné de meilleurs résultats [1].

TD-PSOLA est un algorithme purement du domaine temporel. Contrairement à L'IPS, les opérations "Lecture-Ecriture" sont exécutées à la même F_e. Il n'y a aucun rééchantillonnage de la forme d'onde, qui évite la compression ou l'élargissement de l'enveloppe spectrale, ce qui explique son avantage à maintenir le timbre de la voix le plus naturel que possible.

3.2.3.1. Principe de fonctionnement

Si x(n) est un signal purement périodique, en effet il est possible d'obtenir un signal s(n) de même enveloppe spectrale que x(n) mais de F_0 différente en additionnant des fenêtre d'OLA (Overlap and Addition) $s_i(n)$, extraites par multiplication de x(n) par une fenêtre de pondération w(n) synchronisée sur le pitch T_0 de x(n).

La modification de la F_0 se fait en changeant l'écartement temporel entre les fenêtres d'OLA successives (de sa valeur de départ T_0 à une valeur T quelconque), et en réadditionnant les unes aux autres les fenêtres d'OLA ainsi écartées.

L'abaissement de la F_0 (qui correspond à une augmentation de la période pitch) est exprimé par un écartement entre les fenêtres (signaux à cout-terme) ; l'élévation consistera à rétrécir la distance entres les fenêtres (Figure. 3.7).

Pour augmenter la durée d'un signal, il convient donc de dupliquer un certain nombre de signaux à cout-terme, pour réduire la durée d'origine, il convient dans ce cas de supprimer certains signaux à court-terme. Ainsi le signal de synthèse s(n) est alors obtenu par addition des signaux à cout-terme.

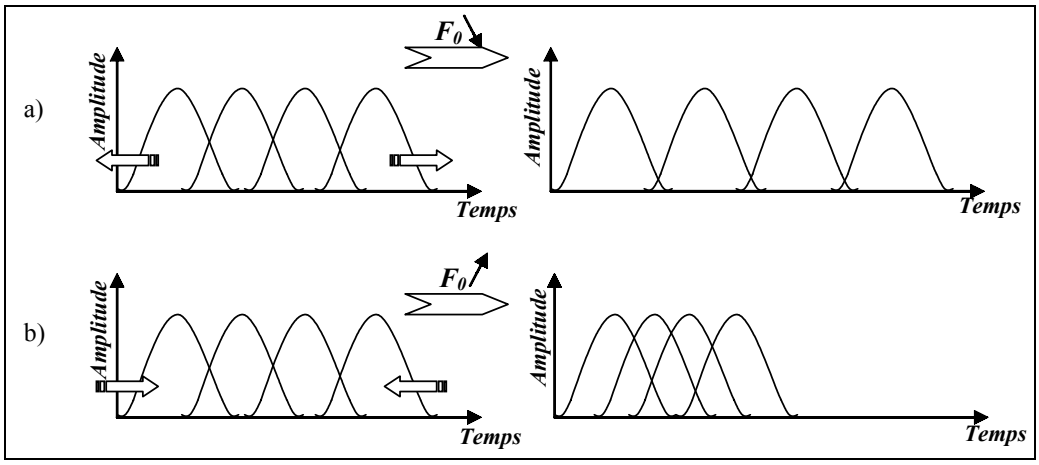


Figure 3.7 : Procédure de modifications de la fréquence fondamentale.
a) : Diminution de la F_0.
b) : Augmentation de la F_0.

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

$$s_i(n) = x(n)w_i(n - iT_0) \tag{3.39}$$

$$s(n) = \sum_i s_i(n - i(T - T_0)) \tag{3.40}$$

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

$$s_i(n) \overset{\Im}{\longleftrightarrow} S_i(\omega) \quad \text{alors} \quad s(n) \overset{\Im}{\longleftrightarrow} \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} S_i(n\frac{2\pi}{T})\,\delta(\omega - n\frac{2\pi}{T}) \tag{3.41}$$

Il s'ensuit que si la fenêtre de pondération w(n) est choisie de façon à ce que le spectre de $s_i(n)$ approxime l'enveloppe spectrale de x(n), l'équation 3.40 fournit un moyen très simple de modifier la F_0 d'un signal périodique.

Rappelons le *Théorème de* ***POISSON***

Suivant la formule de Poisson, la somme d'une infinité de versions décalées d'un même signal f(t) conduit à un signal périodique dont les raies spectrales viennent se positionner exactement sur le spectre du signal du départ [1] *:*

$$\textit{Si} \quad f(t) \overset{\Im}{\longleftrightarrow} F(\omega), \quad \text{alors} \quad \sum_{n=-\infty}^{\infty} f(t - nT_0) \overset{\Im}{\longleftrightarrow} \frac{2\pi}{T_0} \sum_{n=-\infty}^{\infty} F(n\frac{2\pi}{T_0})\ \delta(\omega - n\frac{2\pi}{T_0}) \tag{3.42}$$

Les méthodes de synthèse utilisant ce principe mathématique comme TD-PSOLA, MBROLA[1], LP-PSOLA[2] sont maintenant très utilisées en synthèse par concaténations. Elles requièrent en effet une très faible charge de calcul et leur qualité segmentale est excellente.

3.2.3.2. TD-PSOLA

La technique dite d'**addition recouvrement de fenêtres temporelles synchrones avec le pitch TD-PSOLA** ; (Time Domain Pitch Synchronous Overlap Add); applique le principe de la réharmonisation spectrale directement sur le signal de parole. Appliquée à la synthèse par concaténation, elle conduit à une base de données « paramétriques » où les seuls paramètres stockés sont les marqueurs du pitch indiquant le milieu des fenêtres d'OLA dans les signaux de base de données des segments. Ces marqueurs sont positionnés en synchronisme avec le pitch à l'aide d'un algorithme d'extraction de pitch, et régulièrement espacés sur les zones Non Voisées où aucune modification de pitch ne devra de toute façon être effectuée. La fenêtre utilisée doit garantir l'atténuation des lobes secondaires, car elles seront candidates à une sommation ultérieure, et elles portent des informations sur l'identité des fenêtres voisines du signal (Figure.3.8).

On choisit souvent une fenêtre de Hamming ou une fenêtre Triangulaire, avec une longueur égale à deux fois la période du pitch du signal. Une fenêtre plus large fait apparaître des harmoniques dans le spectre de $s_i(n)$; une fenêtre plus courte n'approxime que très grossièrement l'enveloppe spectrale de x(n) (pour une discussion du choix, du type et de la taille des fenêtres [21]).

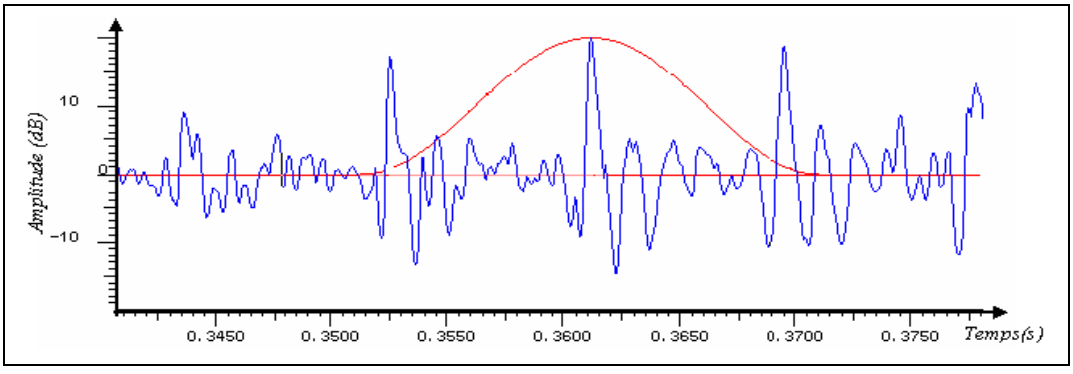


Figure 3.8 : Fenêtrage du signal de parole

[1] MBROLA : MultiBand Overlap and Add est une technique qui réalise l'addition recouvrement après resynthèse multibandes.
[2] LP-PSOLA : Linear prediction PSOLA est une technique qui réalise l'addition recouvrement sur le résiduel du signal.

Les formes d'onde des segments sont stockées telles quelles dans la base de données « paramétriques ».

La fenêtre Triangulaire n'est pas trop compliquée, elle est la plus simple des fenêtres qui amplifie les valeurs centrales, pénalisant les valeurs extrêmes (side lobes) [22].

$$w(k) = \begin{cases} 1 - 2|k|/N & \text{pour } |k| \le \dfrac{N}{2} \\ 0 & \text{partout ailleurs} \end{cases} \qquad (3.43)$$

Elle n'est pas trop utilisée parce que des résultats meilleurs peuvent être obtenus sans beaucoup plus de complexité.

La fenêtre de Hamming est couramment utilisée, elle est codifiée en tables pour éviter le calcul du cosinus.

$$w_H(k) = \begin{cases} \alpha + (1-\alpha)\cos(2\pi k/N) & \text{pour } |k| \le \dfrac{N}{2} \\ 0 & \text{partout ailleurs} \end{cases} \qquad (3.44)$$

La synthèse se résume dans ce cas à une simple modification des paramètres prosodiques. Le système est basé sur la concaténation prosodique des segments à utiliser, qui figure dans des fichiers prosodiques et doit disposer de segments d'analyse sur lesquels les contours de synthèse doivent être ajustés.

La synthèse passe par plusieurs étapes qui sont :

- la préparation du segment d'analyse ;
- la création du segment d'analyse par :
  - la définition du contour de la F_0 de synthèse ;
  - le remplissage du segment de la synthèse avec des éléments du segment d'analyse déformé en fonction du nouveau contour prosodique ;
  - l'ajustement de l'énergie.

La notion de segment représente une portion du signal déterminée à l'aide de marqueurs de frontières positionnés par un programme de segmentation, de durée connue, et divisé en sous éléments qu'on appelle période. Ces sous éléments sont caractérisées par leur type (Voisé ou Non Voisé), leur durée en nombre d'échantillons (qui représente la F_0 en cas de voisement) et leur énergie à court terme.

La création du segment de synthèse consiste à recopier certaines périodes du segment d'analyse et à leur donner de nouvelles caractéristiques (durée et énergie) en

fonction du contour prosodique imposé. Le choix des périodes à recopier pendant l'étape de préparation du segment de synthèse est déterminant pour la qualité de la parole synthétique.

3.2.3.3. Algorithme de synthèse TD-PSOLA

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

- Analyse du signal d'origine qui peut se résumer en :
  - séparation des composantes périodiques du signal (dites Voisées dans le cas de la voix) et des composantes aléatoires (dites Non/Voisées) ;
  - détermination de la F_0 par une méthode appropriée ;
  - détermination des signaux à court terme d'analyse.
- modification prosodique apportée à ces signaux à court terme.
- synthèse du signal modifié par recouvrement addition des signaux à court terme.

3.2.3.3.1. Analyse du signal vocal d'origine

L'existence d'une composante périodique claire dans les sons Voisés rend cette tâche un peu plus fiable, mais elle nécessite une quantité considérable de calculs. Généralement, cette détection est inclue dans la détection du pitch.

La séparation des composantes périodiques et aléatoires d'un signal de parole se fait par différents algorithmes de segmentation qui se basent sur le calcul de la fonction d'autocorrelation et le Taux de Passage par Zéros (TPZ). Les composantes aléatoires dans un signal de parole présentent un TPZ élevé par rapport aux composantes périodiques.

Dans le chapitre suivant on va voir des techniques pratiques des décisions Voisées/Non Voisées qui nous ont permises d'avoir de bons résultats en terme de vitesse et précision de calcul, plus de détails sur les techniques de segmentation sont donnés en réf. [23].

L'estimation de la F_0 se base essentiellement sur la méthode utilisée, ainsi que sur la fiabilité de l'algorithme de segmentation utilisé. A partir des informations issues de la caractérisation du signal, nous pouvons calculer les signaux à court-terme qui vont nous donner par la suite la possibilité de faire des modifications prosodiques.

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.

Ces fenêtres successives sont obtenues par placement de marques appelées marques de lecture t_r^i de manière synchrone au pitch du signal (la différence entre deux marques de lecture successive est égale à T_0 locale) (Figure.3.9). Le signal est alors découpé à l'aide de fenêtres d'analyse centrées sur ces marques de lecture.

$$s_i(n) = x(n)w(n - t_r^i) \tag{3.45}$$

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

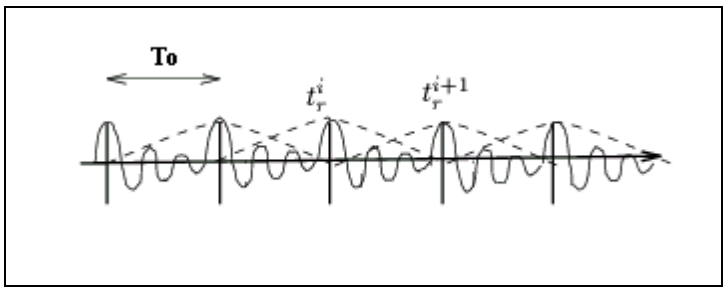


Figure 3.9: Placement des marques de lecture

- Opération de marquage de la fréquence fondamentale

La précision du placement des marques d'écriture détermine en grande partie la qualité du signal de synthèse obtenu. Les marques doivent être placées non seulement de manière synchrone au pitch du signal, mais également de manière à ce que le fenêtrage préserve au maximum les caractéristiques temporelles du signal. Elles doivent aussi se trouver prés des maxima locaux d'énergie [24].

Un algorithme de marquage vise à optimiser les contraintes suivantes :

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

La position des marques est optimisée sur l'ensemble des marques appartenant à une région Voisée. Pratiquement l'optimisation des différentes contraintes est difficile à réaliser, ainsi il existe plusieurs façons d'élaborer un algorithme de marquage. Le principe est d'essayer d'avoir une méthode moins complexe avec des résultats acceptables [25].

3.2.3.3.2. Modifications prosodiques des signaux à court terme

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

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

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

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

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

La figure (3.10**)** montre le principe d'adition recouvrement dans le cas d'une modification simple de la fréquence fondamentale (abaissement par un facteur constant). Dans le prochain chapitre nous allons voir une solution sous optimale des règles de duplication élimination qui nous a donnée des résultats satisfaisants.

3.2.3.3.3. Synthèse du signal vocal

La synthèse est effectuée par Superposition/Addition des signaux élémentaires (obtenue à partir des $s_i(n)$ placés en de nouvelles positions t_w^i). Ces positions sont déterminées selon la hauteur voulue (Figure.3.10).

$$s(n) = \sum_i s_i (n - t_w^i) \tag{3.46}$$

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

$$s(n) = \frac{\sum_i s_i(n - t_w^i)}{\sum_i w_i(n - t_w^i)} \tag{3.47}$$

Nous utilisons souvent la méthode de Giffin et Lim [26] des moindres carrées donnée par.

$$s(n) = \frac{\sum_i r_i(n) w_i(n - t_w^i)}{\sum_i w_i^{\,2}(n - t_w^i)} \tag{3.48}$$

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

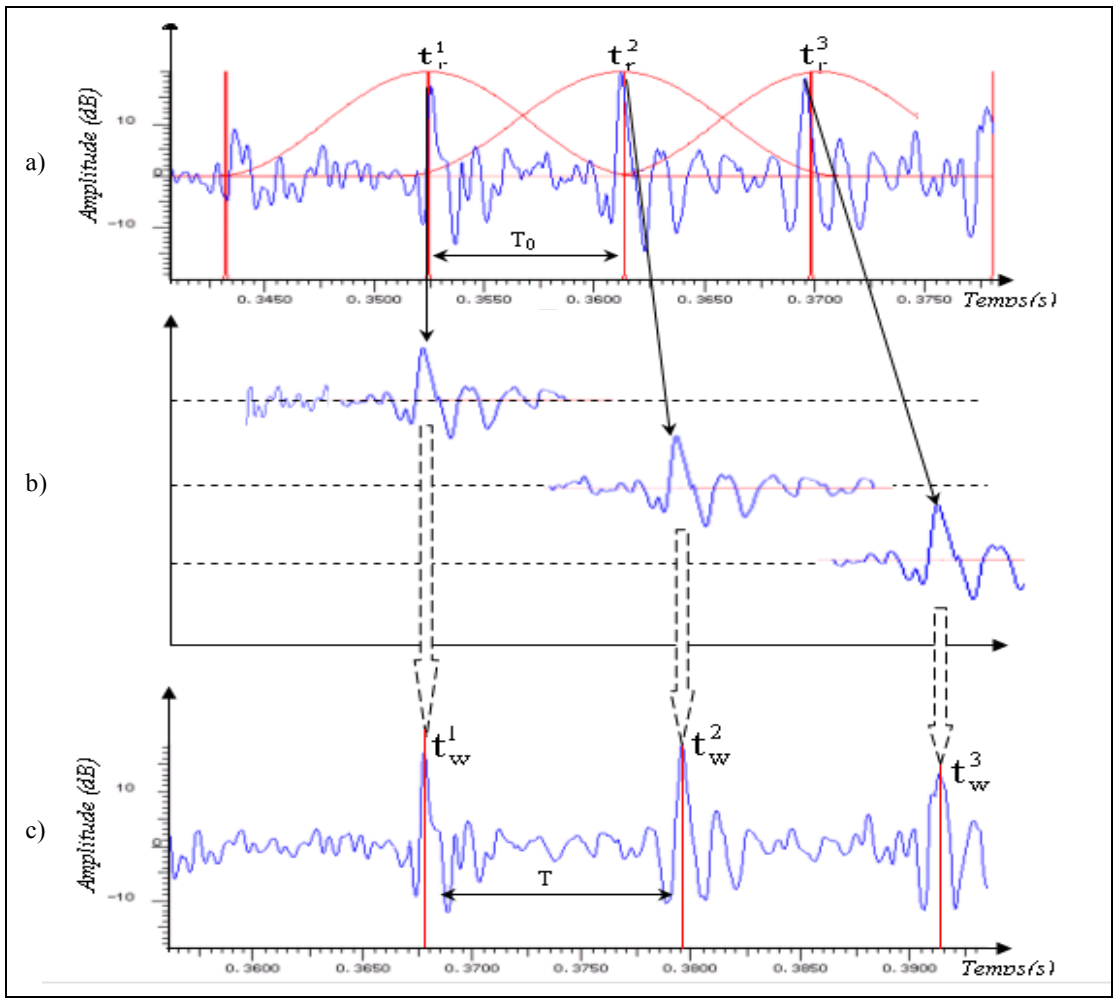


Figure 3.10 : Modification de la F_0 par un facteur 1.2 avec TD – PSOLA
a) : Signal de parole original ainsi que les positions centrales des signaux à court terme.
b) : Signaux à court terme décalés.
c) : Signal modifié obtenu par addition des signaux à court terme décalés.

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

D'après l'équation 3.39 et 3.40. Si nous posons

$$r_i(n) = s_i(n - i(T - T_0)) \tag{3.49}$$

Le signal de synthèse s'écrit :

$$s(n) = \sum_i r_i(n) \tag{3.50}$$

s(n) peut s'écrire :

$$s(n) = \sum_i x(n - i(T - T_0))\ w_i(n - iT) \tag{3.51}$$

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

$$s(n) = \sum_i x(n)\ w_i(n - iT) \tag{3.52}$$

Donc

$$s(n) = x(n) \sum_i w_i(n - iT) \tag{3.53}$$

$$s(n) = cst \times x(n) \tag{3.54}$$

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

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

$$s(n) = \frac{\sum_i r_i(n)\, w_i(n - iT)}{\sum_i w_i{}^2(n - iT)} \tag{3.56}$$

On utilisant les équations 3.39, 3.40 et 3.49, nous obtenons :

$$s(n) = \frac{x(n) \sum_i w^2{}_i(n - iT)}{\sum_i w_i{}^2(n - iT)} = x(n) \tag{3.57}$$

On retrouve le signal d'origine

3.2.3.4. Qualité segmentale avec TD-PSOLA

La qualité segmentale offerte par TD-PSOLA dans le cadre d'une modification prosodique du signal continu est excellente. Lorsqu'on ne modifie ni l'intonation, ni la durée, PSOLA produit même un signal de synthèse identique à celui de l'analyse [27].

Les problèmes se posent plutôt lorsqu'on utilise cette technique pour la synthèse par concaténation de segments. On est en effet alors loin du cas théorique du signal stationnaire.

Trois cas de discontinuités peuvent se présenter : discontinuité de phase, de pitch et d'enveloppe spectrale.

La discontinuité de pitch est causée par une différence de la F_0 dans les segments gauches et droits.

La discontinuité de phase provient d'un positionnement relatif des marqueurs de pitch (par rapport à la période qu'ils marquent) différents dans les segments gauches et droits. Il est très difficile de l'éviter lorsqu'on positionne les marqueurs de pitch sur toute la base de données de segments.

La discontinuité de l'enveloppe spectrale est due à la coarticulation, qui affecte différemment les segments gauches et droits, vu qu'ils proviennent en général de contextes phonétiques différents.

Une fois encore, ne mettant pas en œuvre de modélisation paramétrique des signaux qu'ils modifient, TD-PSOLA ne dispose pas de moyen simple de lisser ces discontinuités. On en est souvent conduit à tester les diphones en concaténations, à détecter les problèmes flagrants, à éliminer les segments qui posent problèmes et à les remplacer par d'autres. Cette optimisation de la base de données est cependant longue et coûteuse [1].

Le coût de calcul associé à la méthode PSOLA est très raisonnable (typiquement, moins de 10 multiplications-additions par échantillon de signal) [9]. D'ailleurs il faut noter que, préalablement à toute modification prosodique, il est nécessaire de déterminer la période du signal de parole (ainsi que son caractère Voisé ou Non Voisé), opération qui est en général plus coûteuse que l'algorithme PSOLA lui-même ; en synthèse par concaténations d'unités, ceci n'est pas très gênant, cette opération étant effectuée une fois pour toute lors de l'enregistrement des unités.

3.3. Conclusion

La transformation vocale est une méthode qui consiste à modifier une voix naturelle ou synthétique pour la transformer en une nouvelle voix semblant la plus naturelle que possible. Cela consiste à usurper la voix de quelqu'un en utilisant le synthétiseur. Elle est souvent utilisée par les développeurs de voix de synthèse pour créer plusieurs voix à partir d'une seule. Pour y parvenir, il faut opérer de deux formes de transformations : soit par transformation prosodique, soit par transformation spectrale avec changement de fréquence d'échantillonnage. La transformation prosodique est censée modifier les paramètres tels que la hauteur, le rythme, le débit et l'intensité. La technique PSOLA et la modélisation physique Source-Filtre sont deux approches pour la manipulation de la parole, ils représentent une étape importante dans le développement des techniques du traitement de la parole.

Le signal synthétique produit par prédiction linéaire souffre d'une qualité segmentale fortement limitée aux prix d'une charge de calcul importante. La synthèse dans le domaine temporel (TD-PSOLA) permet d'avoir une parole de synthèse de très bonne qualité pour un temps de calcul dix fois moindre. Cet avantage est cependant contrebalancé par le fait que les synthétiseurs paramétriques offrent des possibilités de compression nettement supérieures à celles des synthétiseurs dans le domaine temporel.

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
SOLUTIONS POUR LA MODIFICATION DE LA FREQUENCE FONDAMENTALE

4.1. Introduction

Après avoir présenté, au chapitre précédent, les deux techniques de modifications de la F_0 utilisées pour les signaux de parole, nous développons dans ce chapitre deux solutions complètes qui nous permettent de détailler toutes les étapes utilisées pour simuler ces deux techniques : soient la modélisation physique Source Filtre par prédiction linéaire et la technique de synthèse dans le domaine temporel TD-PSOLA. Nous présentons par la suite une évaluation globale et analytique des deux techniques employées.

4.2. Description du corpus utilisé

Nous allons mettre en œuvre la réalisation de synthétiseurs de parole en utilisant les deux techniques décrites auparavant afin de pouvoir effectuer des modifications prosodiques du signal de parole; il s'agit de la modification de la F_0 dans notre cas.

Le signal d'entrée est obtenu à partir d'une phrase énoncée en langue arabe et prononcée par un locuteur masculin. La phrase est échantillonnée à 16 kHz (Figure.4.1).

Phrase : [addarso assabiɛ]. الدرس السابع

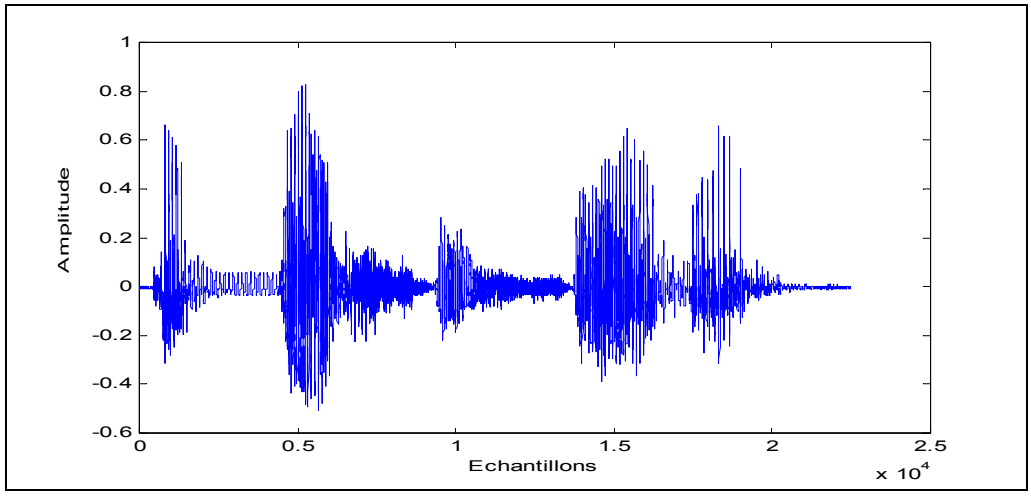


Figure 4.1 : Représentation temporelle de la phrase [addarso assabiɛ]

Les moyens informatiques dont nous disposons sont constitués d'un Micro Ordinateur type P3 de RAM 128Mo dont la fréquence d'horloge est de 900 Mhz. Le logiciel utilisé est le MATLAB (Version 5.30.1 ou plus) dont le but est de faciliter les calculs intermédiaires.

4.3. Détection de la fréquence fondamentale

4.3.1. Filtrage du signal vocal

Le problème lié à la structure formantique et aux effets de bruit peut être atténué en proposant un filtrage préalable selectif de type passe bas. Le domaine spectral qui nous préoccupe ici est donc inférieur à 800 Hz. Il est ainsi préférable, avant de poursuivre l'analyse, de commencer par éliminer les fréquences supérieures à 800 Hz à l'aide d'un filtre passe-bas de Butterworth, généralement d'ordre 8 et de fréquence de coupure d'environ 800 Hz (Figure.4.2).

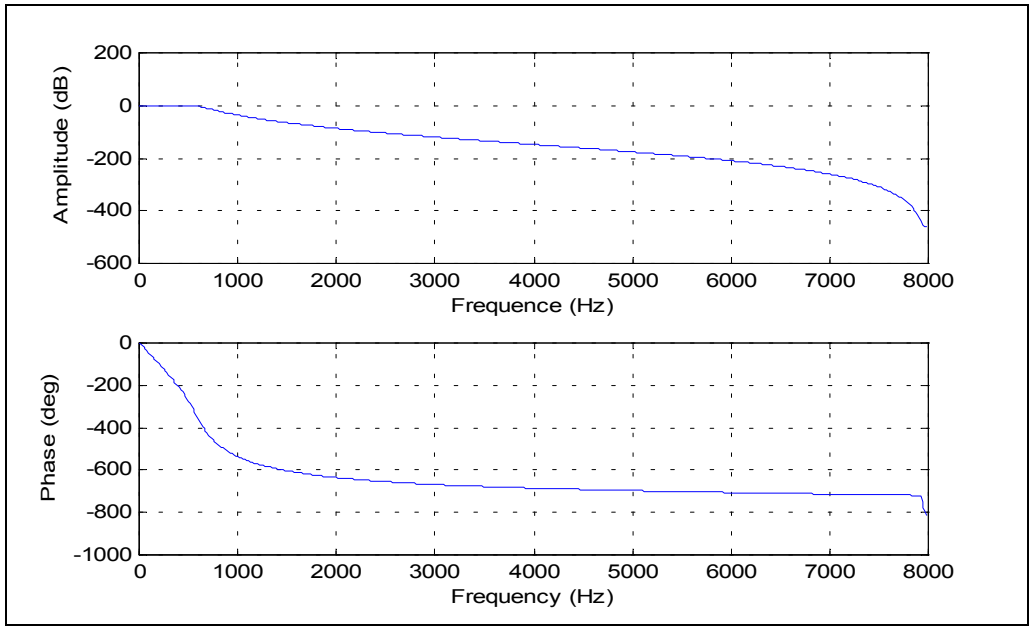


Figure 4.2 : Réponse en fréquence du filtre passe-bas utilisé

4.3.2. Critères de décision

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

Le but de la segmentation est de fournir une résolution avec le plus que possible de précision, dans la mesure où tous les traitements qui vont suivre reposent sur les résultats de cette segmentation.

Il existe plusieurs méthodes pour segmenter le signal vocal qui ont en commun l'analyse d'un intervalle de parole précis pour assurer la stationnarité de ses segments, et de calculer l'énergie moyenne et le TPZ.

4.3.2.1. Décision basée sur la fonction d'autocorrélation

L'existence d'une composante périodique claire dans les sons Voisés rend cette tâche un peu plus fiable, mais elle nécessite une quantité considérable de calculs.

Généralement, cette détection est inclue dans la détection du pitch. Par exemple, dans la méthode de corrélation par produit, le signal est considéré Voisé si l'amplitude du maximum recherché est supérieure à 50% de la valeur maximale initiale [22] (Figure.4.3).

$$R_{max} > 0.5\ R(0) \tag{4.1}$$

D'autres niveaux peuvent être choisis pour les autres méthodes. Pour pouvoir évaluer cette technique, la décision Voisée Non Voisée (V/NV) prendra comme valeurs :

$$V/NV = \begin{cases} 1 & \text{si la trame est voisée.} \\ 0 & \text{si non.} \end{cases} \tag{4.2}$$

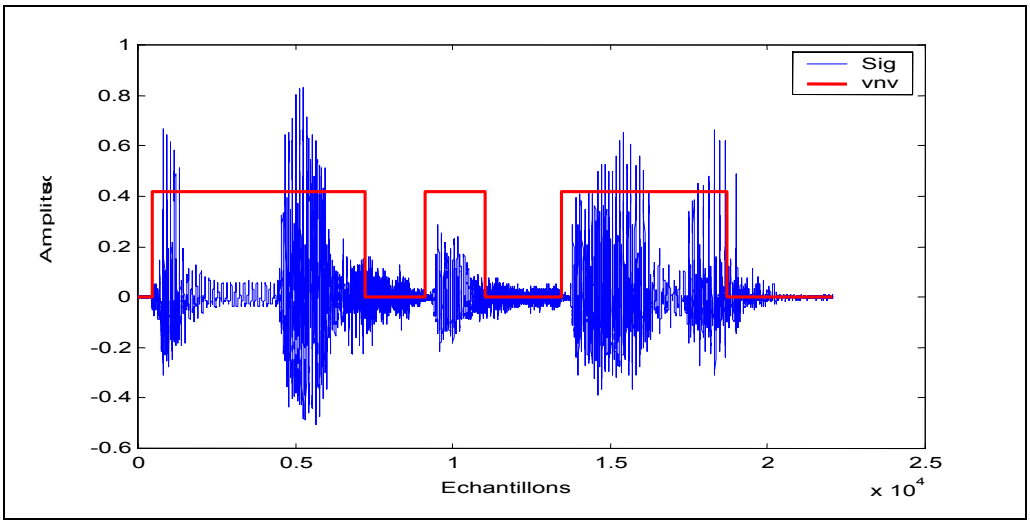


Figure 4.3 : Décision basée sur la fonction d'autocorrélation

4.3.2.2. Décision basée sur le calcul du Taux de Passage par Zéros et l'énergie du signal

En cas d'incertitude, on peut également utiliser le TPZ et l'énergie de la tranche considérée pour aider à la décision.

Pour ce faire, on estime en permanence l'énergie du signal de parole donnée par la l'équation 1.1, ainsi que l'état de passage par zéros EPZ qui prend comme valeurs :

$$EPZ = \begin{cases} 1 & \text{si } x_{n-1} \neq 0 \text{ et } x_n = 0 \text{ ou } (x_{n-1} x_n) < 0. \\ 0 & \text{ailleurs} \end{cases} \quad (4.3)$$

Où x_n est le signal de parole

Nous sommes amenés à faire une décision V/NV afin de pouvoir estimer la F_0 sur toutes les trames Voisées du signal d'entrée. Réellement le calcul d'énergie est essentiellement utilisé pour détecter les zones de silence présentes sur un tel signal (qui se caractérisent par une faible énergie).

Dans notre cas, les zones de silences sont considérées comme des zones Non Voisées où l'absence de la F_0 est déterminée.

L'idée de base est de comparer successivement l'énergie et le TPZ des trames d'analyse à des seuils bien déterminés afin de réaliser une décision finale.

Le principe de cette technique est résumé dans l'organigramme suivant où V/NV représente la décision donnée par l'équation 4.2.

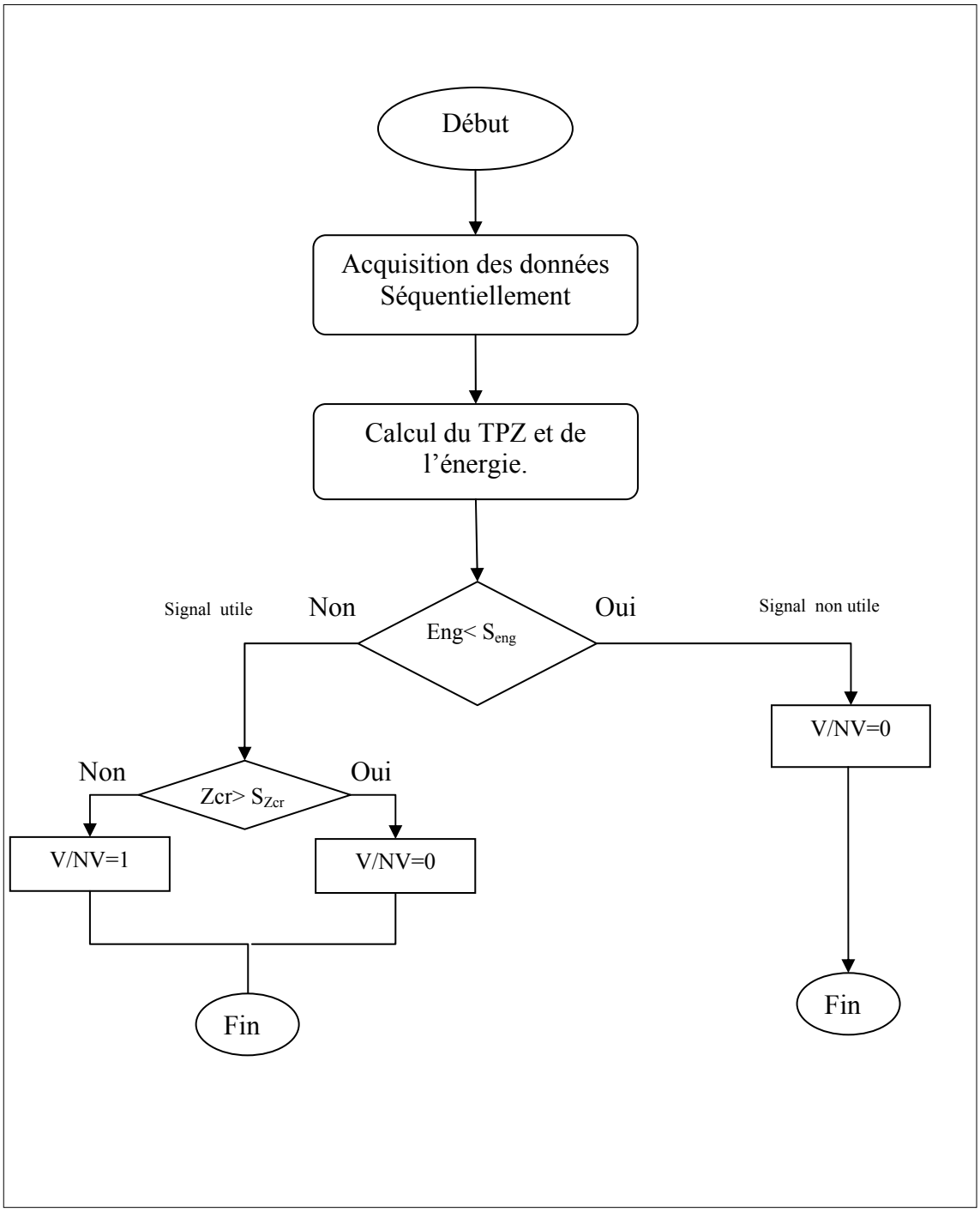


Figure 4.4 : Organigramme représentant la procédure de la décision basée sur le TPZ et l'énergie du signal.

Eng représente l'énergie de la trame actuelle qui est comparée à un seuil S_{eng} et Zcr est le Taux de Passage Par Zéro (zero crossing rate) qui est comparé à un seuil S_{Zcr}. Le TPZ (noté Zcr) peut être exprimée par la relation suivante :

$$Zcr_n = Zcr_{n-1} + \frac{1}{2}\left|sign(x_n) - sign(x_{n-1})\right| \qquad (4.4)$$

Où Zcr est calculée sur une trame de longueur L avec une valeur initiale nulle et sign représente la fonction signe qui détermine le signe d'échantillons.

Les seuils de l'énergie et le TPZ sont déterminés pour chaque signal d'entrée à partir des tests pratiques qui donnent de meilleurs résultats. Afin d'automatiser le système ; nous proposons une technique de détermination des seuils.

Nous commençons tout d'abord par établir un vecteur qui est constitué des différentes énergies calculées de toutes les trames du signal d'entrée, et cela par un ordre croissant.

Ensuite, parmi ces différentes successions d'énergie, nous déterminons celle qui représente le 25^{eme}% de l'ensemble d'indices (positions) du vecteur d'énergie, qui dessine enfin le seuil d'énergie voulu.

Le seuil du TPZ est calculé de la même manière, en déterminant le taux qui représente cette fois le 75^{eme}% de l'ensemble d'indices du vecteur du TPZ.

Les valeurs des seuils pour notre cas, et pour le signal d'entrée de la phrase donnée au paragraphe § 4.2 sont : $S_{eng=}$-18.70dB et $S_{Zcr=}$120.

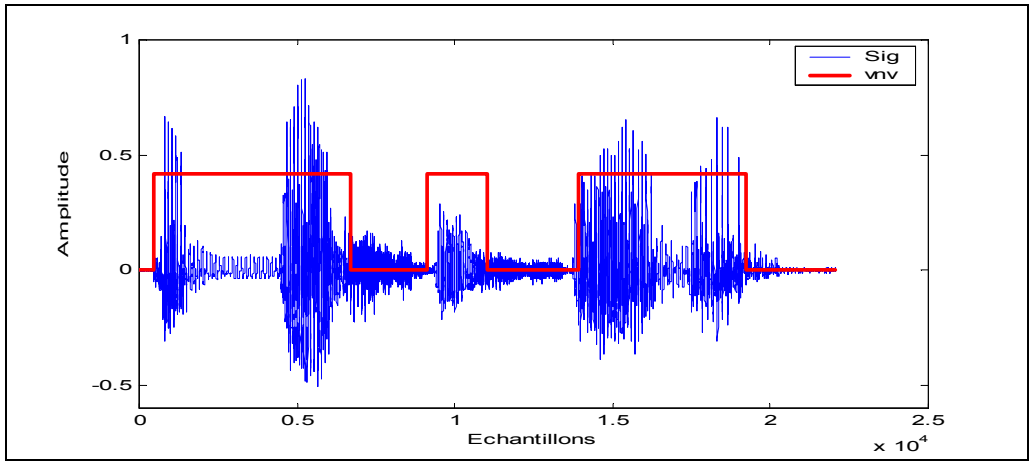


Figure 4.5 : Décision basée sur le calcul TPZ et l'énergie du signal d'entrée.

Les seuils comparatifs peuvent être déterminés automatiquement pour les deux techniques ; malheureusement les critères utilisés ne sont pas toujours garantis de manière automatique surtout dans le cas des sons transitoires. Alors, pour avoir une grande précision, nous préférons souvent déterminer ces dernières manuellement par des tests pratiques qui donnant les meilleurs résultats afin d'assurer le bon fonctionnement de la technique de segmentation qui représente une étape préliminaire pour tout traitement surtout en synthèse de la parole.

4.3.3. Evaluation de la technique

La période du fondamental (appelé communément le pitch) est un paramètre très important pour la synthèse de la parole ; l'oreille est en effet très sensible à ses variations, qui constituent la prosodie ou timbre du locuteur.

L'extraction du pitch a été une tâche particulièrement difficile pour trois raisons : premièrement, les vibrations des cordes vocales n'ont pas nécessairement une périodicité complète, particulièrement au commencement des sons Voisés. Deuxièmement, il est difficile de séparer le pitch des effets des paramètres vocal. Troisièmement, la plage de dynamique de la F_0 est très grande.

La plupart des procédures étudiées comportent une de ces deux erreurs : soit elles dupliquent le pitch, soit elles le divisent par deux. Normalement ces cas ne sont pas très ennuyeux mais il existe des applications où la robustesse du détecteur est très importante.

Dans notre cas, nous avons utilisé une technique temporelle qui est souhaitable pour l'analyse de la micro mélodie, elle est basée sur le calcul de la fonction d'autocorrelation.

C'est une technique de base pour la plupart des techniques utilisées à présent et qui donne des résultats satisfaisants avec moins de complexité.

Avant de procéder au calcul du pitch, le signal d'entrée doit être divisé à des trames de même longueur « L » variant entre 20 et 30 ms pour respecter la condition de stationnarité d'un tel signal. Nous sommes ensuite amenés à calculer la valeur du pitch pour chaque trame étudiée. Si nous évaluons la fonction d'autocorrelation décrite par l'équation 2.2 pour les différentes trames obtenues et sur une longueur L. Le résultat est un vecteur de longueur L avec un maximum en sa valeur initiale R(0). Si le signal est périodique, d'autres pics distants de la valeur du pitch seront présents. Pour trouver ce dernier, il suffit de mesurer cette distance (Figure. 4.6).

La F_0 se situe entre 50 Hz et 250 Hz pour les voix masculines, entre 150 Hz et 400 Hz pour les voix féminines et entre 200 et 600 chez les enfants. Pour faciliter la tâche et limiter la plage dynamique de la F_0 et pour enlever toute ambiguïté qui peut surgir, le calcul de la fonction d'autocorrelation est effectué entre 2 et 20 ms (F_0 entre 50 et 500 Hz, pour limiter la plage dynamique du pitch), de cette manière nous pouvons garantir l'obtention des valeurs de la F_0 qui ne sont pas erronées.

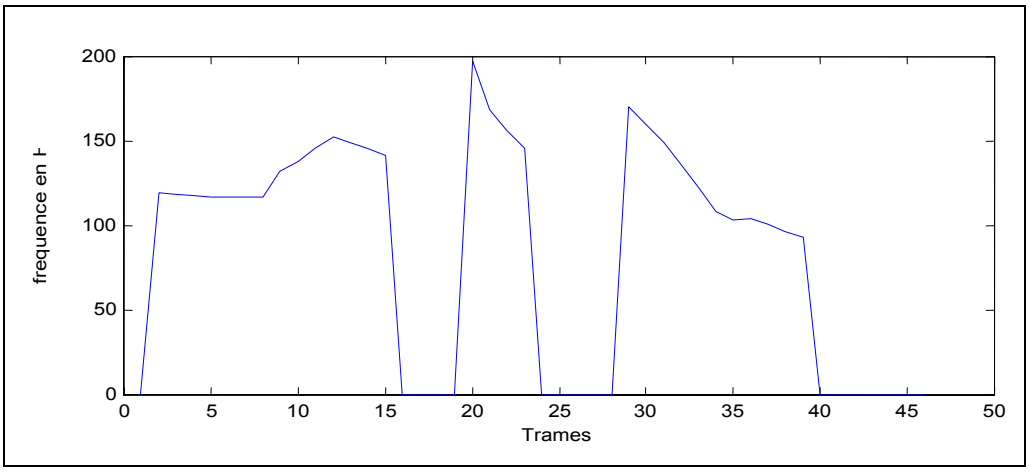


Figure 4.6 : Evaluation de la F_0 du signal d'entrée en fonction des blocs d'analyse par la méthode d'autocorrélation.

Remarque : Comme le signal est passablement bruité, la recherche de la période est grandement facilitée si on l'effectue sur la fonction d'autocorrélation du résidu d'excitation plutôt que sur le signal lui-même. Cette technique peut être employée dans le cas d'un synthétiseur LPC où le calcul du résiduel peut être effectué après la modélisation AR.

4.4. Technique TD-PSOLA

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

Ces méthodes nécessitent au préalable un marquage des périodes du fondamental. Nos travaux se situent dans le cadre de l'apprentissage des langues, sur la modification de la fréquence fondamentale. C'est ainsi que nous avons choisi d'utiliser la méthode TD-PSOLA.

Elle est rapide et permet de modifier le pitch avec une bonne qualité segmentale du signal synthétique résultant.

TD-PSOLA est basée sur la décomposition du signal de parole en fenêtres recouvrantes synchronisées sur les périodes du fondamental. Les marques de synchronisation (ou du fondamental) indiquent le centre de ces fenêtres. Les modifications consistent à manipuler les marques d'analyse pour générer de nouvelles marques de synthèse. Cela correspond à la duplication ou l'élimination des fenêtres dont l'écartement peut être modifié.

La principale exigence, commune aux techniques de décomposition du signal en courtes fenêtres, est de garder la cohérence mutuelle des emplacements des marques pour préserver la structure temporelle originale du signal étudié. Par conséquent, il est crucial d'obtenir un marquage précis des périodes du fondamental, car il influe directement sur la qualité du signal.

De nombreuses méthodes de marquage ont été décrites dans la littérature. Elles sont généralement basées sur la recherche d'évènements précis dans le signal de parole : Instants de fermeture glottale, extrema du signal, instants d'excitation des modèles LPC . . . etc [24] ; le principe est de respecter les conditions de marquage décrites au chapitre 3.

Comme le souligne P.Veldhuis [26]**,** ces techniques souffrent d'une certaine rigidité face au critère numérique exploité. En particulier, le critère numérique peut forcer le marquage d'échantillons qui satisfont ce critère mais dont la distance avec les marques voisines est éloignée de la période du fondamental. Dans le contexte de la synthèse de la parole, il est concevable de corriger quelques erreurs à la main, ce qui n'est plus possible si nous souhaitons modifier des phrases de manière automatique pour l'apprentissage des langues.

Il est donc important de développer des algorithmes offrant une meilleure précision du marquage du fondamental et de la localisation des marques de synthèse. Nous proposons un algorithme de marquage qui exploite les résultats de l'extraction de la F_0 et assure la cohérence des marques sur l'ensemble de la phrase utilisée.

La procédure à suivre pour cette technique peut se résumer à :

- acquisition du signal x(n) ;
- constitution de blocs de 20ms de durée ;
- décision V/NV et détermination de la F_0 de chaque bloc voisé par l'algorithme proposé ;
- détermination des marques de lecture synchrone au pitch et constitution des fenêtres OLA ;
- détermination des marques d'écriture en fonction du facteur de modification et positionnement des fenêtres analytiques à de nouvelles cadences en utilisant des règles de duplication/élimination ;
- constitution du signal synthétique par addition/recouvrement.

4.4.1. Marquage du fondamental

4.4.1.1. Marquage dans les zones Voisées

Le marquage du pitch constitue une contrainte fondamentale pour l'application de l'algorithme TD-PSOLA. On se basant sur le principe de la décomposition du signal en des trames de 20ms, l'idée la plus simple est d'effectuer un marquage à pitch constant, calculé pour chaque trame du signal de départ.

Le principe de détermination des marques est donc de parcourir le signal et de déterminer un ensemble de points distants d'une période de pitch pour chaque trame étudiée.

Il faut tout d'abord déterminer un marquage primaire sur une zone Voisée afin de pouvoir déterminer toutes les marques restantes à partir de celle ci.

Le marquage initial sera donc exprimé par la localisation du premier maximum sur une zone Voisée, à partir de ce dernier nous pouvons déterminer les marques suivantes en ayant une connaissance a priori sur la valeur du pitch calculé pour chaque trame. L'efficacité de cette technique de marquage se base essentiellement sur la fiabilité de la technique de détection du fondamentale employée. L'organigramme suivant illustre le principe de cette idée sur une zone Voisée (Figure. 4.7).

Sur cet organigramme la variable N représente le nombre de trames Voisées localisé par la technique de décision afin de pouvoir effectuer un marquage qui est fonction d'une marque fondamentale M_1 comme il est décrit ci-dessous, où F_j représente la fréquence de la j^{eme} trame. i et j sont respectivement un compteur de trames localisés sur une zone Voisée et un indicateur d'indice (position) des trames du signal de départ. A l'aide de cette technique la synchronisation avec le fondamental est assurée.

4.4.1.2. Marquage dans les zones Non Voisées

Dans le cas d'une zone Non Voisée, les temps de correspondance (les marques) sont placées de manière équidistante (absence de la F_0) ; cependant les transitoires d'attaques (Non Voisées) ne doivent subir aucune modification, leur dilatation étant perspectivement désagréable [25]. L'ensemble des fenêtres OLA correspondant à une transitoire d'attaque sont donc recopiées sans espacement temporel et seront reproduits de la même cadence du départ.

La difficulté qui surgit dans ce cas est de pouvoir effectuer une segmentation automatique pouvant assurer la séparation des transitoires d'attaques et les autres types de sons.

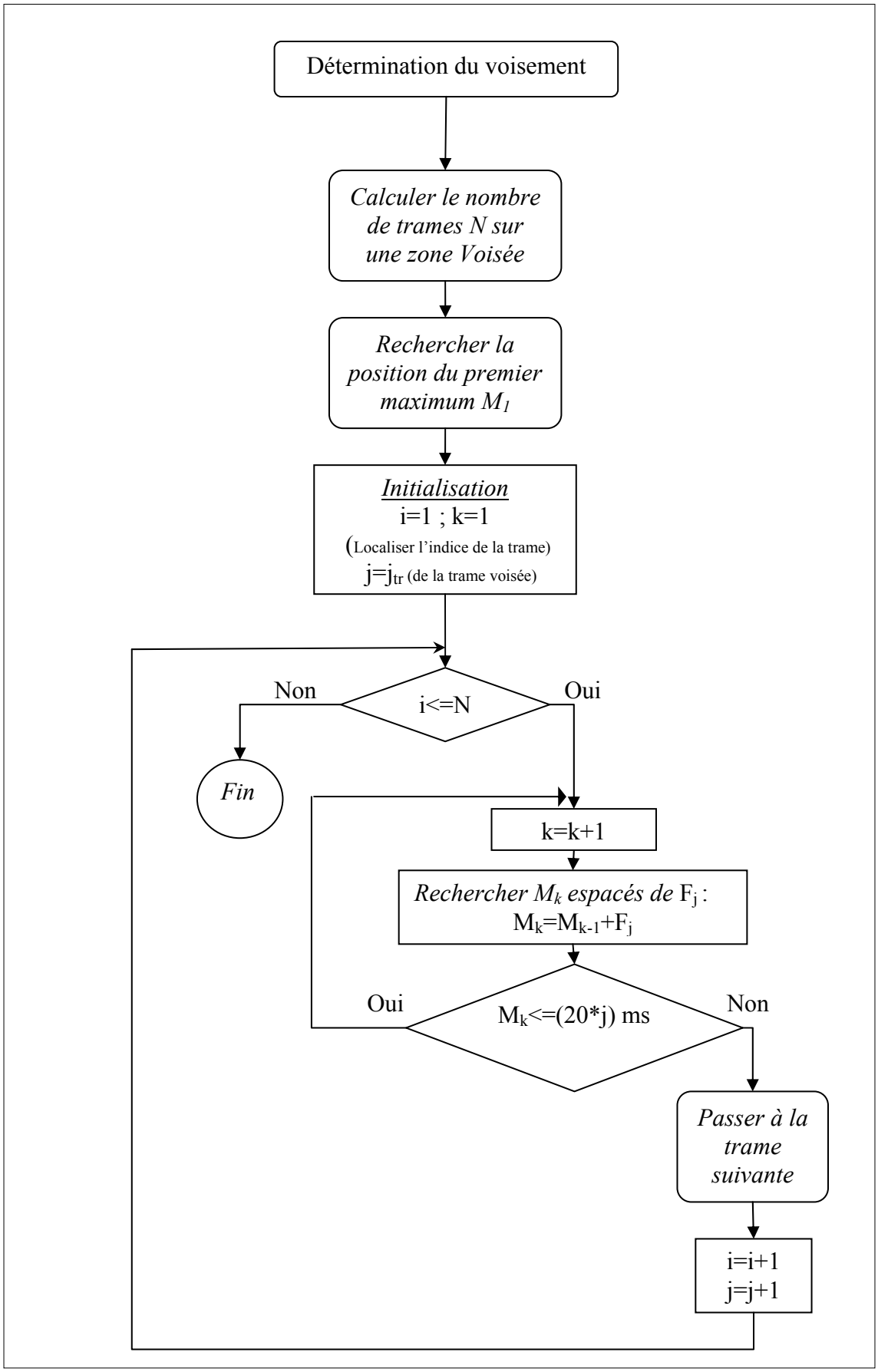


Figure 4.7 : Organigramme représentant la procédure du marquage du fondamental.

Pratiquement il n'existe aucun dispositif automatique fiable pouvant effectuer cette opération sans trop perturber le signal résultant. La décision du Voisement utilisée dans notre cas ne permet pas de rendre compte des sons mixtes (comme les fricatives Voisées) ; pour lesquels la glotte n'est jamais complètement fermée, et non pas la séparation des transitoires d'attaques.

La solution est soit d'effectuer un marquage équidistant pour toute catégorie de sons, ou tout simplement de recopier l'ensemble des signaux élémentaires sans modifications d'espacement temporel alors sans marquage. Les tests pratiques favorisent la première solution qui assure des résultats raisonnables avec une qualité sonore acceptable puisque une telle technique temporelle de modifications prosodiques nécessite une grande précision de localisation des différentes zones qui sont objets à un traitement définissant la qualité du signal résultant.

Le marquage dans ces zones est effectué de manière équidistante le long de la trame décidée Non Voisée. Cette distance est repérée par le calcul de l'espacement séparant la valeur initiale de la fonction d'autocorrelation et le premier pic détecté entre 2 et 20 ms définissant l'amplitude de la fonction d'autocorrelation du point recherché et qui est recalculée pour chaque trame. L'avantage d'une telle technique est de pouvoir effectuer une modification de la F_0 sans le biais d'une segmentation préalable qui nécessite une précision remarquable et qui requiert à son tour une complexité d'implémentation.

Dans les deux cas (cas des zones Voisées et Non Voisées), le marquage est effectué à partir de la prévision d'une distance qui se base sur le calcul de la fonction d'autocorrelation assurant la synchronisation des marques.

Dans le cas des zones Voisées, cette distance mesure la période fondamentale, dans l'autre cas (cas des zones non voisées) elle mesure une distance à partir de laquelle le positionnement des marques est effectué de manière équidistante à cette dernière. Alors l'organigramme 4.7 peut être utilisé pour effectuer le marquage du fondamentale pour toutes les trames du signal sans le biais d'une décision V/NV.

4.4.2. Règles de duplication et élimination

Le principe de la modification du paramètre temporel du signal de parole est simple : il s'agit de dupliquer – ou de soustraire un segment relativement court du signal. L'astuce va donc consister à trouver un procédé optimal pour positionner chaque nouveau segment sur l'axe temporel par rapport au signal existant de manière à minimiser l'effet perceptif du raccord, effet portant sur trois discontinuités possibles : le pitch, la phase et l'enveloppe

spectrale. Basé sur la supposition que deux fenêtres (OLA) successives sont pratiquement identiques tant que le pitch ne fluctue pas beaucoup au cours du temps ; nous pouvons déterminer plusieurs règles de duplication/élimination conservant la durée du signal analytique ; le but est d'avoir une bonne qualité segmentale sans perte d'informations.

Puisque dans notre cas nous supposons que le pitch reste constant sur une trame donnée, éliminer où dupliquer une fenêtre ne fera pas altérer le signal de parole. Les règles suivantes sont élaborées pratiquement afin d'avoir de meilleurs résultats sans altérer le signal et d'avoir une bonne qualité segmentale pour obtenir un signal de sortie d'une fréquence différente.

Une fois les marques de lecture sont déterminées, nous sommes amenés à positionner les fenêtres OLA à une nouvelle cadence pour effectuer le changement de pitch ; alors il faut déterminer les marques d'écriture correspondantes au nouveau pitch voulu. L'idée de base est de faire une correspondance entre marques d'écriture et lecture sur une trame déterminée; ce qui veut dire, si nous désignons Na le nombre de marques de lecture et Ns le nombre de marques d'écriture, nous sommes amenés à les correspondre une à une ; la première marque d'écriture correspond à la première marque de lecture, la deuxième marque d'écriture correspond à la deuxième marque de lecture et ainsi de suite jusqu'au $Ns^{éme}$ marque d'écriture qui correspondra à la $Na^{éme}$ marque de lecture.

Malheureusement, il n'y a pas vraiment une correspondance bien définie entre marques d'écriture et lecture ; augmenter la valeur de la période du pitch donne moins de marques d'écriture, diminuer la période du pitch donne plus de marques, c'est pour cette raison que nous dupliquons ou nous éliminons certaines marques pour pouvoir assurer ainsi la conservation de la durée de départ.

4.4.2.1. Règle d'élimination

La règle d'élimination est utilisée dans le cas où nous voulons augmenter la valeur de la période du pitch. Dans ce cas, il est possible d'avoir moins de marques synthétiques (d'écriture) qu'analytiques (de lecture) puisque la période du pitch est augmentée sur une trame donnée le long de toute la zone voisée.

Le principe de cette approche est de calculer tout d'abord le nombre de marques (écriture et lecture) et de faire une correspondance entre marques d'écriture/lecture sur une trame donnée ensuite d'éliminer les (Na-Ns) marques de lecture restantes.

L'organigramme suivant résumera le principe de cette technique sur une trame du signal de départ. Et on suivra le même principe pour toutes les trames analysées le long d'une zone Voisée (Figure 4.8).

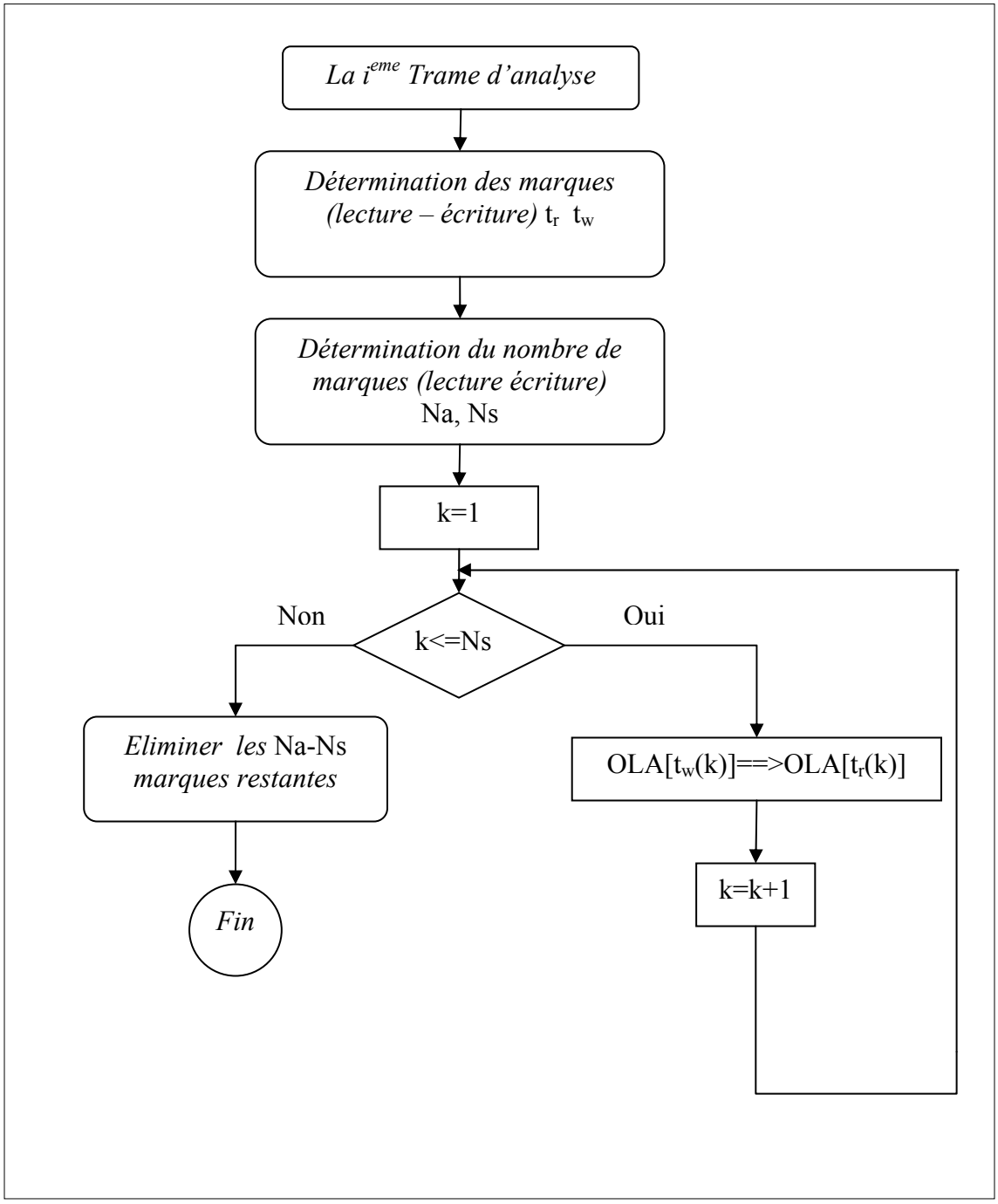


Figure 4.8 : Organigramme représentant la règle d'élimination des fenêtres OLA.

Il est possible qu'il y'ait une correspondance totale entre marques d'écriture et lecture sur une trame donnée. Cette correspondance peut être obtenue dans le cas d'une

légère augmentation de la période synthétique ; donc dans ce cas cette règle n'est pas appliquée (Na=Ns).

La règle d'élimination étudiée dans ce paragraphe est théoriquement valable pour des valeurs du facteur de modifications de la période du pitch compris entre 1 et 2 puisque la longueur de la fenêtre de pondération utilisée dans cette application est de longueur correspondante à deux périodes de pitch.

Pratiquement dés que nous dépassons une valeur de 1.6 du facteur de modification, la qualité du signal obtenu devient de plus en plus médiocre, ceci est dû essentiellement à l'élimination des (Na-Ns) fenêtres de pondération qui peut être utile pour l'obtention du signal synthétique et en grande partie à l'approximation faite sur la valeur du pitch par la fonction d'autocorrélation et au positionnement approximative des marques de pitch.

4.4.2.2. Règle de duplication

La règle de duplication est utilisée dans le cas d'une diminution de la T_0 pour un facteur de modification inférieur à 1. Dans ce cas, il va y avoir plus de marques d'écriture que de lecture, ce qui veut dire que nous devons répéter certaines marques de lecture pour pouvoir occuper une trame synthétique étudiée. Dés que nous diminuons suffisamment le facteur de modification, nous ne parlons plus de dupliquer une fenêtre ou une marque, nous parlons plutôt d'une reproduction des fenêtres et cela est en fonction des marques d'écriture obtenues dans une trame donnée.

Le principe de la règle adopté dans notre cas est de calculer le nombre de marques (écriture/lecture) dans une trame donnée, et de dupliquer Ns-Na marques afin d'occuper toute la trame. La règle utilisée est un peu grossière à cause de sa simplicité, cependant elle donne de bons résultats de point de vue qualité segmentale et compréhension du message vocal. Dans l'organigramme suivant (Figure.4.9), la variable Qos représente l'entier supérieur du quotient calculé à partir du nombre de marques d'écriture et lecture.

Dans le cas où celui-ci est inférieur ou égal à deux (supérieur à 1 bien sûr), nous dupliquons Ns-Na marques de lecture de manière à occuper toute la trame. Dans le cas contraire (Qos>2), on reproduit Qos fois les marques d'écriture, cette reproduction peut être au nombre de 3 ou 4 ou même plus en fonction du facteur de modifications utilisé et du nombre de marques occupant la trame. C'est une solution importante pour des petites valeurs du facteur de modification qui sont inférieurs à 0.5. Pour éclaircir, des exemples explicatifs pour ces deux règles seront donnés en annexes. Dans le cas où Ns=Na, cette règle n'est pas appliquée.

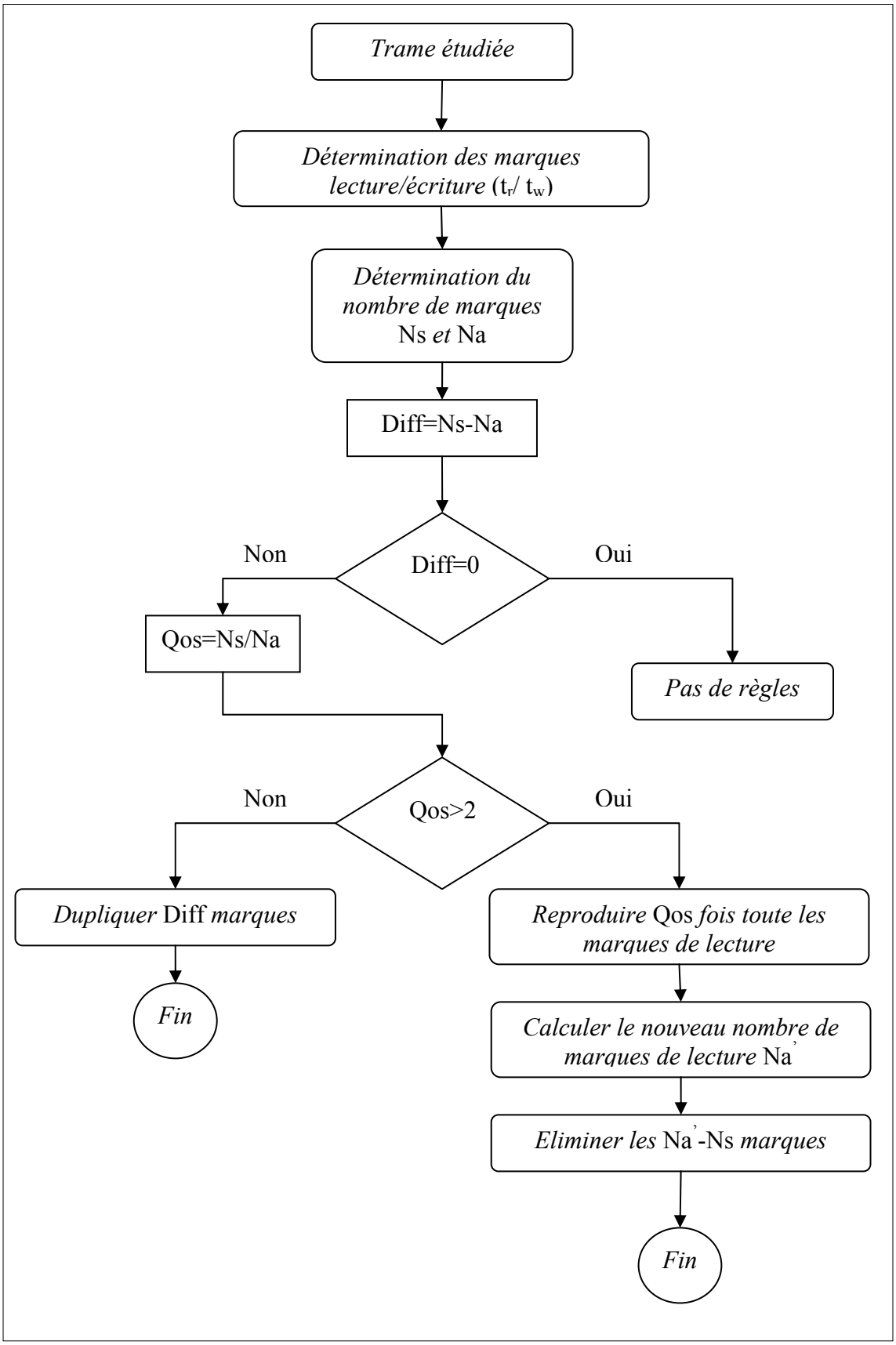


Figure 4.9 : Organigramme représentant la règle de duplication des fenêtres OLA.

4.4.3. Synthèse du signal vocal

La synthèse est effectuée par addition recouvrement des fenêtres OLA prises à des cadences dites marques de lecture et reproduites à des nouvelles cadences dites marques d'écriture au moyen de l'équation 3.47. Les résultats obtenus à l'aide de cette technique de synthèse sont satisfaisants et seront présentés au paragraphe suivant.

4.4.4. Résultats de la simulation TD PSOLA

Nous présentons dans ce paragraphe, des tests réalisés sur le signal de parole de notre corpus pour l'obtention d'un signal synthétique de fréquence fondamentale différente pour trois facteurs de modifications différents, le facteur 1.3 (facteur supérieur à 1), le facteur 0.8 (facteur inférieur à 1) et enfin pour un facteur égal à1 qui normalement donne le même signal de départ.

Le facteur de modification égal à 0.8 correspond à un abaissement de la période pitch, c'est à dire à une augmentation de la F_0 et le facteur égal à 1.3 correspond à une diminution de la F_0, c'est à dire une augmentation de la période du pitch.

La figure 4.10 donne une représentation temporelle du signal décrivant un intervalle temporel du phonème [a] extrait du signal original à partir du mot « **addarso** », qui possède une F_0 égale à 144.45 Hz. Ce dernier est superposé à un autre signal obtenu par l'algorithme TD PSOLA au même intervalle temporel du même phonème mais possédant une période fondamentale multipliée par un facteur égal à 1.3 par rapport à la période du signal de départ. La F_0 du signal résultant est égale à 111.11 Hz, qui vérifiera exactement l'hypothèse de départ.

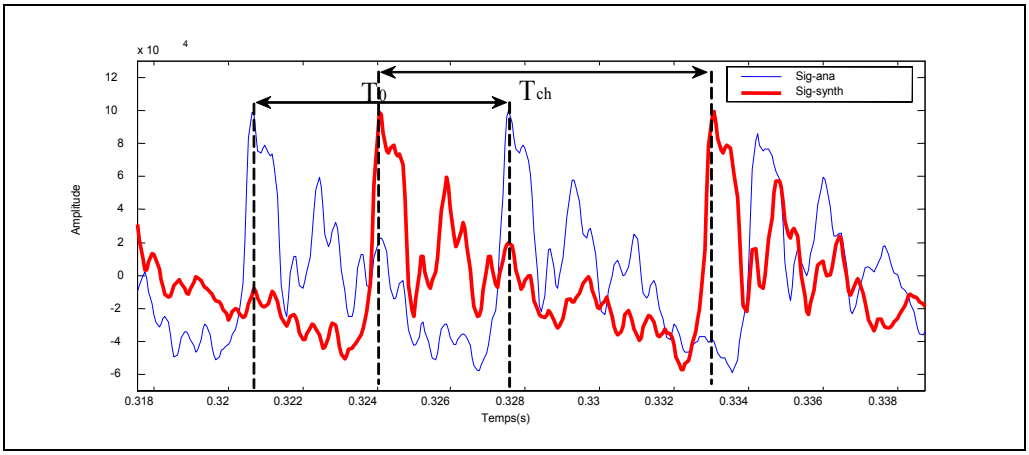


Figure 4.10 : Opération d'augmentation de la T (facteur =1.3) par TD-PSOLA sur un intervalle de 20 ms du phonème [a].

De même, si nous voulons obtenir un signal synthétique d'une période fondamentale inférieure à celle d'origine, il suffit de préciser la valeur du facteur de modification et d'appliquer l'algorithme TD-PSOLA. La figure 4.11 donne la représentation temporelle du signal analytique et synthétique pour le même intervalle temporel et avec un facteur de modification égal à 0.8. La fréquence fondamentale du signal obtenue est égale cette fois à 181.7356 Hz. Dans ces deux figures T_0 représente la période du signal analytique et T_{ch} celle du signal synthétique reconstitué et sig-ana, sig-synth représentent respectivement le signal analytique et synthétique d'un intervalle de 20 ms du phonéme [a].

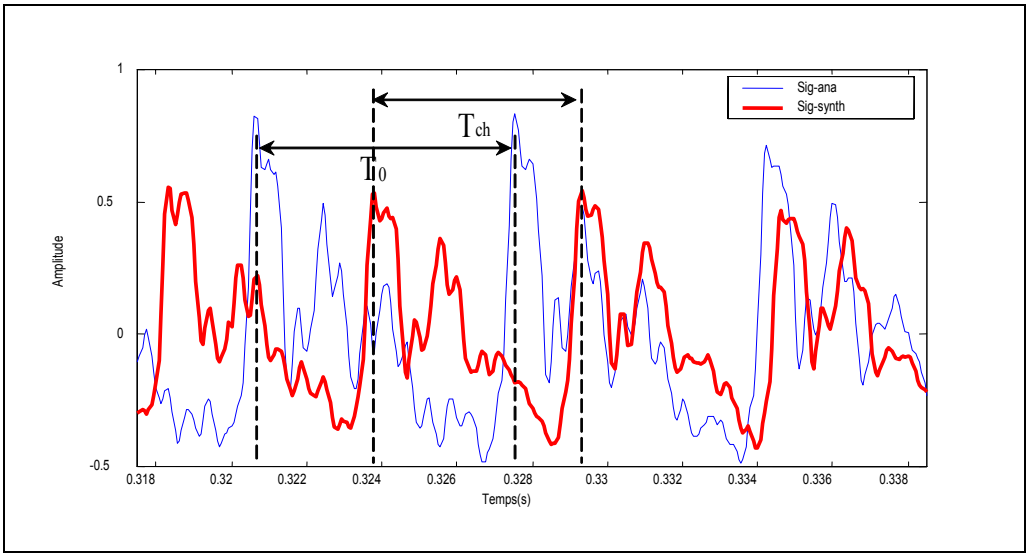


Figure 4.11 : Opération de diminution de la T_0 (facteur =0.8) par la TD-PSOLA sur un intervalle de 20 ms du phonème [a].

Nous remarquons que le signal reconstitué a subit une distorsion d'amplitude. Cela est dû essentiellement au positionnement des marques proposées sur le signal original qui ne sont pas très précises pour pouvoir déterminer rentablement les périodes de pitch et de vérifier toutes les conditions de marquage décrites au chapitre précèdent. De plus, sur l'approximation faite sur la valeur du pitch par la fonction d'autocorrélation ainsi que sur l'efficacité de précision des règles de duplication/élimination. Malgré l'inconvénient, la distorsion en amplitude n'est pas vraiment un problème et pourra être remédiée en assurant le rapport d'amplitudes.

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

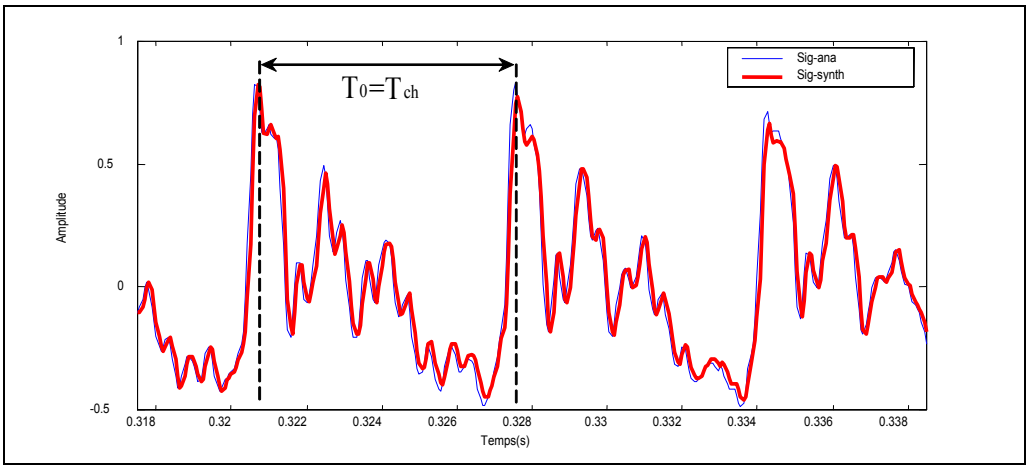


Figure 4.12 : Signal synthétique et analytique du phonème [a] obtenus par la TD-PSOLA à un facteur égal à 1 sur un intervalle de 20 ms.

L'avantage d'une telle technique est qu'elle réalise la modification de la F_0 d'un signal de parole sans changement de l'enveloppe spectrale qui à son rôle préserve le timbre de la parole aussi bien que possible.

L'observation de l'enveloppe spectrale du signal original et des signaux interpolés (pour un facteur 1.3 et 0.8) nous montrent dans les deux cas que les résultats obtenus sont quasi équivalents (Figure.4.13), assurant ainsi le maintien de l'enveloppe spectrale ; ou DSP0p8 et DSP1p3 représentent respectivement l'enveloppe spectrale des signaux interpolés pour des facteurs égaux à 0.8 et 1.3 et DSPorg est l'enveloppe spectrale du signal original (ou avec un facteur de modification égal à 1).

En plus, on remarque que la Densité Spectrale de Puissance du signal synthétique obtenu par un facteur égal à 0.8 présente une légère diminution d'énergie par rapport à la densité spectrale du signal original, cela est dû au problème de distorsion d'amplitude qui a causé à son tour une dégradation de l'énergie du signal (Figure 4.13).

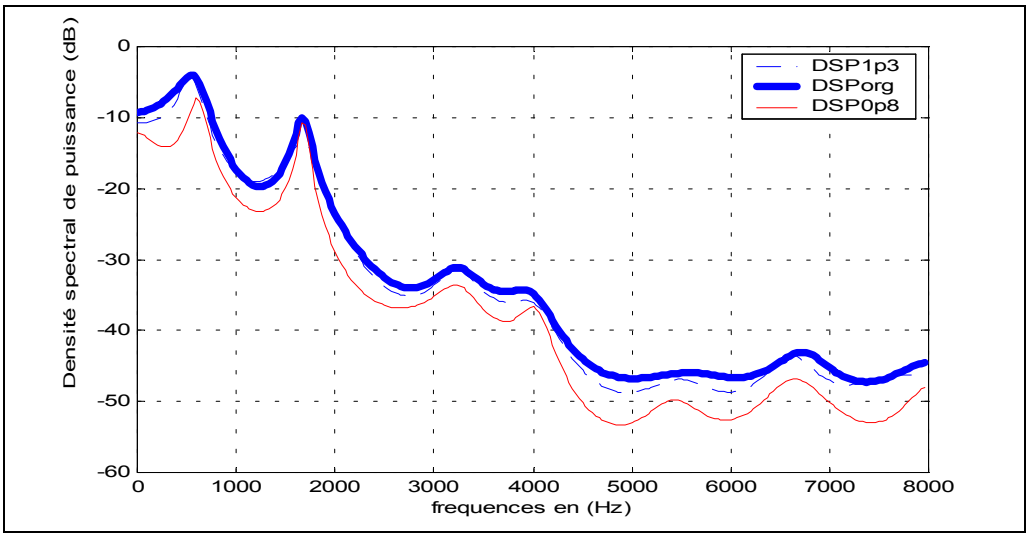


Figure 4.13 : Enveloppes spectrales des signaux interpolés et du signal original du phonème [a] sur un intervalle de 30 ms.

La figure 4.14 effectue une comparaison entre les spectrogrammes des deux signaux modifiés soit par un facteur de 0.8 ou 1.3 et le spectrogramme du signal original.

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

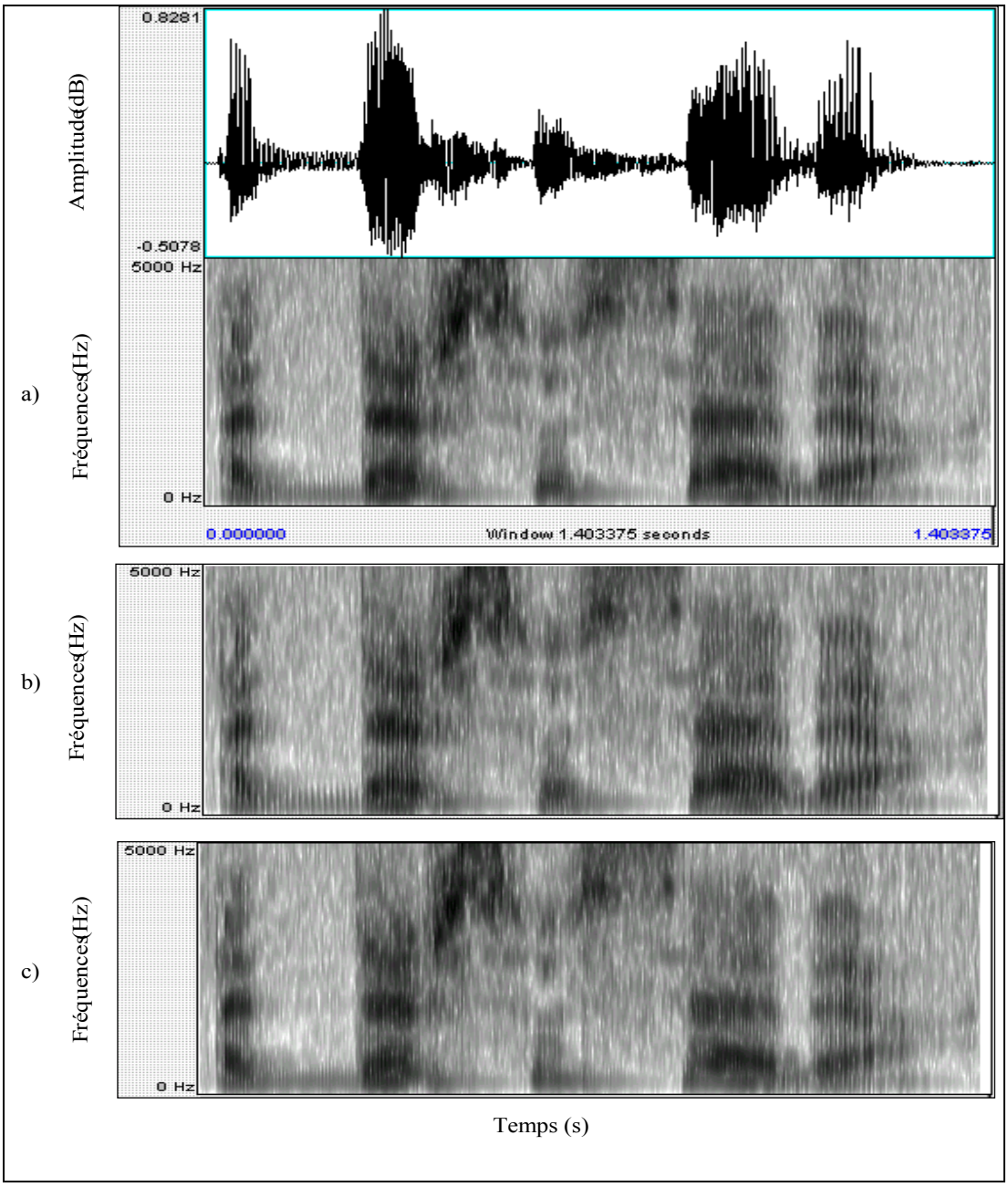


Figure 4.14 : Spectrogrammes des signaux interpolés et du signal original de la phrase « addarso assabiɛ » obtenue par la TD-PSOLA.
a) : Signal original.
b) : Signal reconstitué avec un facteur de 1.3.
c) : Signal reconstitué avec un facteur de 0.8.

4.5. Modélisation Source - Filtre par prédiction linéaire

La prédiction linaire ou LPC (Linaire Prédictive Coding) est un modèle paramétrique, dans notre cas, du signal de parole. A partir du modèle Source-Filtre, nous allons établir un modèle d'analyse et de synthèse du signal qui dépend d'un nombre réduit de paramètres et des méthodes d'estimation de ces paramètres.

La procédure à suivre pour cette méthode peut être résumée comme suit :

L'analyse du signal consiste en une :

- acquisition du signal ;
- préactuation éventuelle par passage dans un filtre de transmittance $(1-\mu\ z^{-1})$, $\mu \cong 0.95$. Cette opération vise à accentuer la partie haute fréquence du spectre qui entraîne une réduction de la dynamique du signal ;
- constitution de blocs de « N » échantillons avec décalages de « L » échantillons ;
- décision Voisée/Non Voisée et détermination de la F_0 de chaque bloc Voisé par l'algorithme proposé ;
- détermination des paramètres du modèle pour chaque bloc.

La synthèse consiste à reconstituer chaque tranche du signal sonore à partir des paramètres du générateur et du filtre.

4.5.1. Choix des conditions d'analyse

Il faut au premier lieu fixer les conditions suivantes :

- la fréquence d'échantillonnage Fe ;
- l'ordre de la prédiction p ;
- le nombre d'échantillons par tranche d'analyse ;
- le décalage entre les tranches successives ;
- la préaccentuation éventuelle.

Le choix de F_e est indépendant de la méthode d'analyse. Pour une représentation complète des sons fricatifs, il faudrait choisir pour F_e une valeur au moins égale à 20 kHz, par contre pour une transmission téléphonique, on se contentera de reproduire la bande 300-3400 Hz et on choisit F_e=8 kHz. Pour une synthèse de haute qualité, la fréquence d'échantillonnage est choisie entre 12 et 20 kHz. Dans ce qui suit nous utilisons pour fixer les idées une valeur moyenne de 16 kHz.

En général, nous estimons que la FT du modèle doit comporter une paire de pôles par kHz de bande passante ; l'excitation glottique et la radiation des lèvres exigent ensemble 3 à 4 pôles ; alors dans le cas où F_e=10 kHz, nous choisissons p=13 ou 14 [1]. Pour notre cas F_e=16kHz l'ordre p est choisi égal à 19 ou 20 (16 pour la FT et 3 ou 4 pour l'excitation glottique et la radiation aux lèvres). Une base objective peut être trouvée dans l'évolution de l'énergie résiduelle de prédiction avec l'ordre p d'un modèle AR dans la référence [1].

La durée des tranches d'analyse dépend beaucoup de la méthode choisie et les conditions dans lesquelles elle est appliquée. La pratique montre que la fenêtre doit empiéter sur plusieurs périodes du fondamental pour les sons Voisés ; on utilise couramment des fenêtres de 30 ms décalées de 20 ms. Pour F_e = 16kHz, l'analyse porte sur des tranches de N= 480 échantillons décalées de L = 320 échantillons ; on garde les même valeurs pour les sons Non Voisés.

La préaccentuation du signal consiste à le faire passer dans un filtre de transmittance $(1-\mu\ z^{-1})$ (μ compris entre 0.9 et 1), ce qui a pour effet d'accentuer la partie haute fréquence du spectre. Ce prétraitement assure un bon conditionnement des algorithmes de résolutions, au moins pour les sons Voisés. La valeur de μ n'est pas critique et on choisit une valeur fixe par exemple μ =0.95.

4.5.2. Estimation des paramètres du modèle

L'estimation des paramètres du modèle peut être effectuée en utilisant plusieurs techniques ; soit par calcul direct par résolution d'un système d'équations ou en utilisant l'algorithme de Levinson, qui sont tous les deux basés sur le calcul d'autocorrelation ; soit par l'utilisation de l'algorithme de Burg qui est basé sur l'estimation directe des coefficients de réflexion à partir des données sans passer par le calcul préalable des autocorrélations.

Puisque dans notre approche nous nous sommes basés sur le calcul de la fonction d'autocorrelation pour estimer le pitch, alors nous essayons d'estimer les paramètres du modèle en utilisant des techniques qui se basent sur le calcul préalable des autocorrélations. Pour fixer les idées, on considère le processus AR-3 défini par :

$$x(n)-0.9x(n-1)+0.49x(n-2)-0.441x(n-3)=w(n) \qquad (4.5)$$

où w(n) est un bruit blanc centré de variance 1.

Le modèle AR-3 peut être vu comme la sortie d'un filtre H(z) excité par w(n), où $H(z)=\frac{1}{A(z)}$.

Le filtre H (z) peut alors s'écrire :

$$H(z) = \frac{\sigma}{1 - 0.9z^{-1} + 0.49z^{-2} - 0.441z^{-3}} \qquad (4.6)$$

Où σ représente le gain du modèle calculé à partir de l'équation 3.21. Par la suite, nous allons estimer les paramètres de ce modèle en les comparant avec les valeurs réelles définies précédemment.

4.5.2.1. Estimation des paramètres par la résolution du système

La solution donnée estime la suite à des P coefficients d'un modèle AR ainsi que le gain du modèle à partir d'un échantillon x et de l'ordre P supposé du modèle.

L'estimation des paramètres a(i) nécessite la résolution d'un système de (p+1) équations à (p+1) inconnues. Posons comme conditions initiales que a(0)=1 et que le signal d'entrée soit stationnaire sur la durée d'analyse. Cette solution est vérifiée pratiquement pour différents ordres P du modèle.

Si nous considérons le modèle décrit en § 4.4.2 excité par un bruit blanc centré de variance unitaire et de longueur « N » échantillons, nous obtenons à la sortie un signal dont le modèle AR est celui du filtre H(z). Nous effectuons M = 100 tirages à l'identique et nous prenons comme indice d'erreur d'estimation la moyenne sur M tirages de l'écart quadratique entre un paramètre estimé et sa vraie valeur. Nous refaisons l'expérience pour plusieurs valeurs de N de la taille d'échantillons. Nous constatons dans la courbe que l'écart diminue quand N augmente (Figure.4.15). On peut refaire l'expérience pour évaluer les écarts sur l'estimation d'un autre paramètre du modèle où changer l'ordre du modèle.

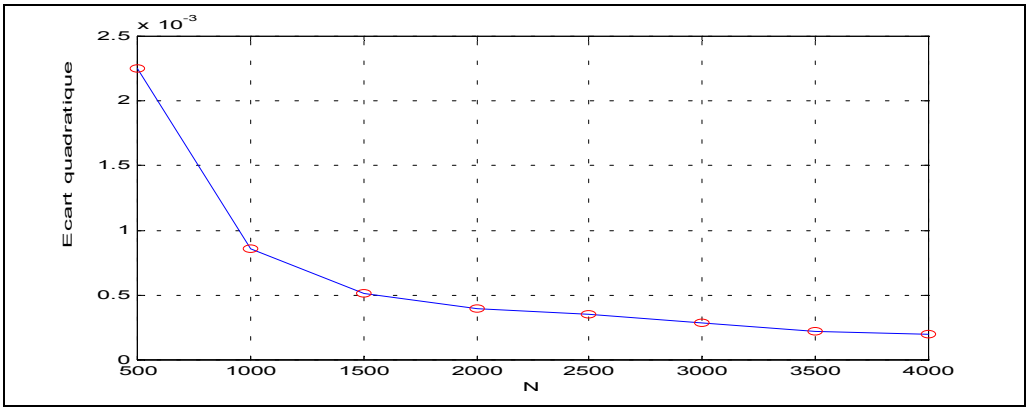


Figure 4.15 : Evaluation de l'écart quadratique entre un paramètre estimé et sa vraie valeur.

La matrice d'autocorrélation pour cette exemple est donnée par :

$$\begin{bmatrix} R(0) & R(1) & R(2) & R(3) \\ R(1) & R(0) & R(1) & R(2) \\ R(2) & R(1) & R(0) & R(1) \\ R(3) & R(2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} 1 \\ a(1) \\ a(2) \\ a(3) \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad (4.7)$$

Les paramètres réels du modèle sont alors :

a(0)=1; a(1)=-0.9; a(2)=0.49; a(3)=-0.441.

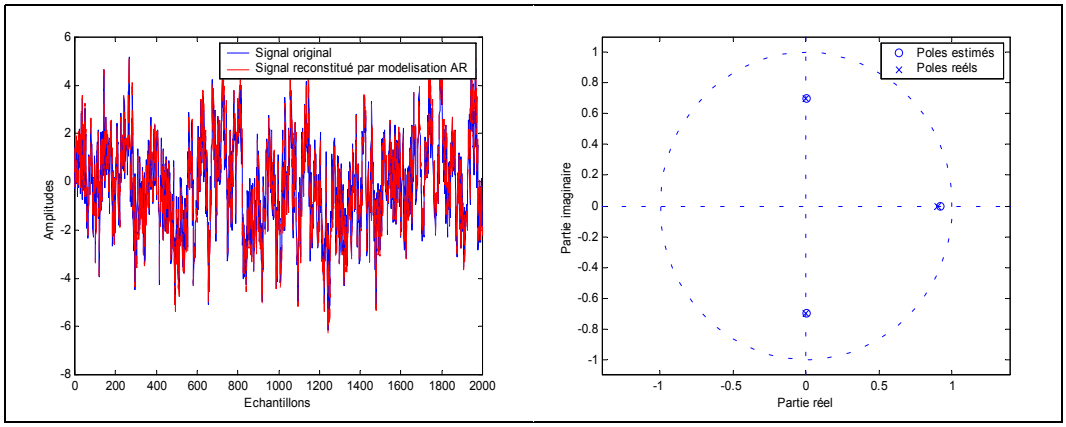
et

$\sigma^2 = R(0)a(0) + R(1)a(1) + R(2)a(2) + R(3)a(3)$ .Qui est égal à $\sigma^2 = 1.0273$

L'estimation des paramètres du modèle du signal de sortie calculée par cette technique et pour la FT donnée ci-dessus pour N=1000 (Figure. 4.16) est donnée par :

$a(0)_{estimé} = 1.0000$; $a(1)_{estimé} = -0.9011$; $a(2)_{estimé} = 0.5033$; $a(3)_{estimé} = -0.4510$

et $\sigma_{estimé}^2 = 1.05093$.



(a) (b)

Figure 4.16 : Signal original et reconstitué par modélisation AR (Calcul direct).
a) : Représentation temporelle.
b) : Pôles de la fonction de transfert.

4.5.2.2. Estimation des paramètres par l'algorithme de Levinson

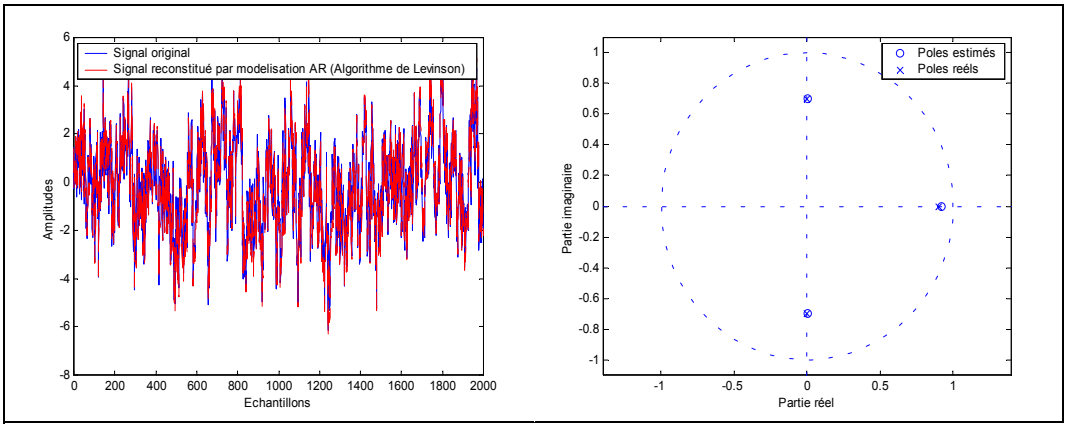
Nous allons utiliser un algorithme rapide, dû à l'origine à Levinson et qui permet de résoudre les équations de Yule-Walker sans passer par la matrice d'autocorrélation. (Voir chapitre 3). Si nous considèrons comme dans le premier cas le modèle décrit en § 4.4.2 excité par un bruit blanc centré de variance unitaire et de longueur N=1000 échantillons,

nous obtenons à la sortie un signal dont le modèle AR est celui du filtre H(z) (Figure 4.16), en posant les mêmes conditions initiales. Les paramètres du modèle sont donnés auparavant.

L'estimation des paramètres du modèle du signal de sortie calculé par cette technique est donnée par :

$a(0)_{estimé} = 1.0000$; $a(1)_{estimé} = -0.9106$; $a(2)_{estimé} = 0.4995$; $a(3)_{estimé} = -0.4534$;

$\sigma_{estimé}^{2} = 1.01876$;



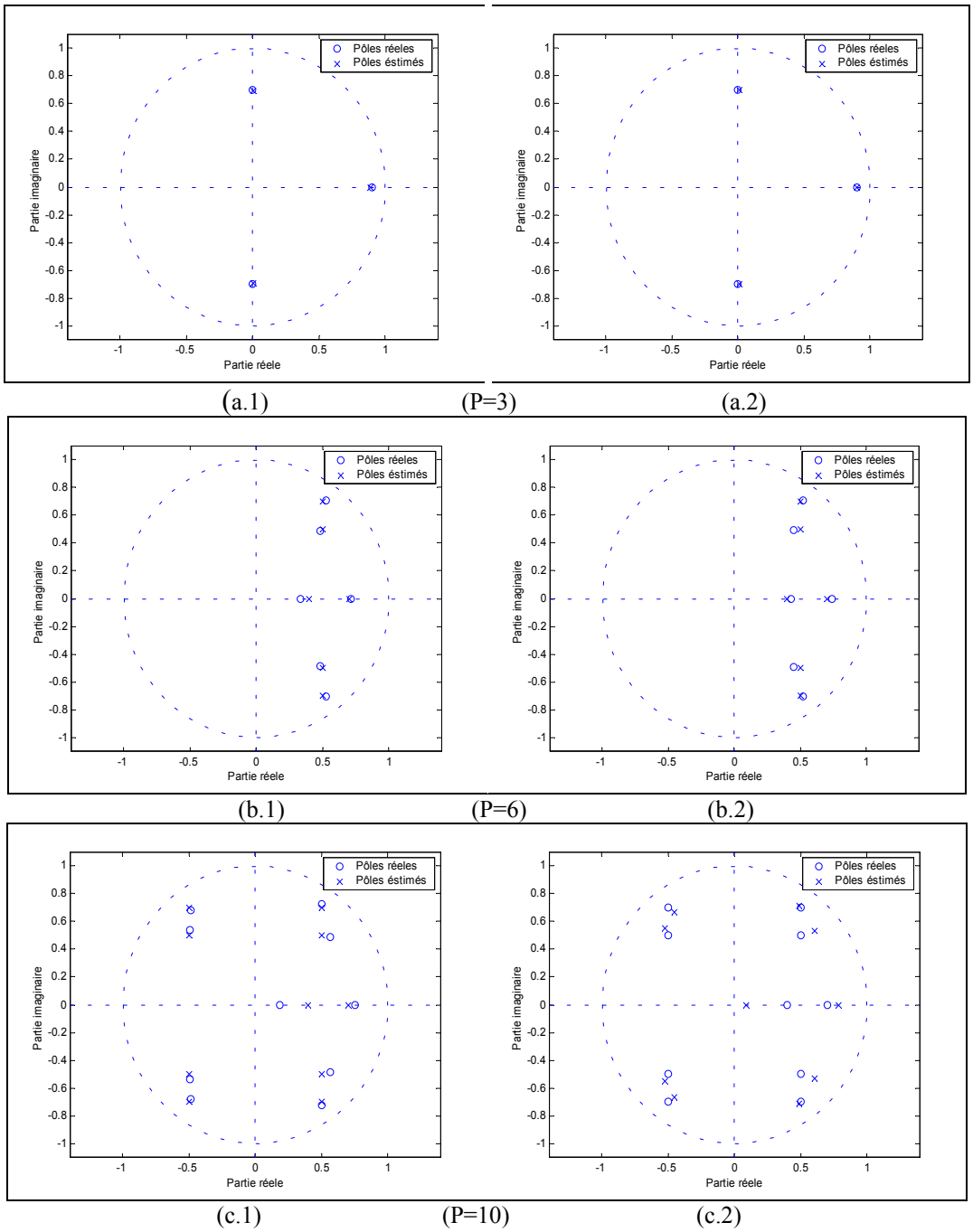
(a) (b)

Figure 4.17 : Signal original et reconstitué par modélisation AR (Algorithme de Levinson)
a) : Représentation temporelle.
b) : Pôles de la Fonction de Transfert.

Les figures 4.16 et 4.17 montrent la représentation temporelle et les pôles de la FT du 3^{eme} ordre, calculées à partir des deux techniques et comparées avec les pôles réels.

Nous allons essayer de résumer quelques tests pratiques effectués pour différents ordres du modèle qui donnent les meilleurs résultats des pôles de la FT en suivant les mêmes étapes précédentes.

Une première expérience consiste à varier l'ordre du modèle et de fixer la valeur de N (N=2000), où on remarque que l'écart quadratique (erreur d'estimation) augmente avec l'ordre du modèle pour les deux techniques (Figure 4.18).



(a.1) (P=3) (a.2)

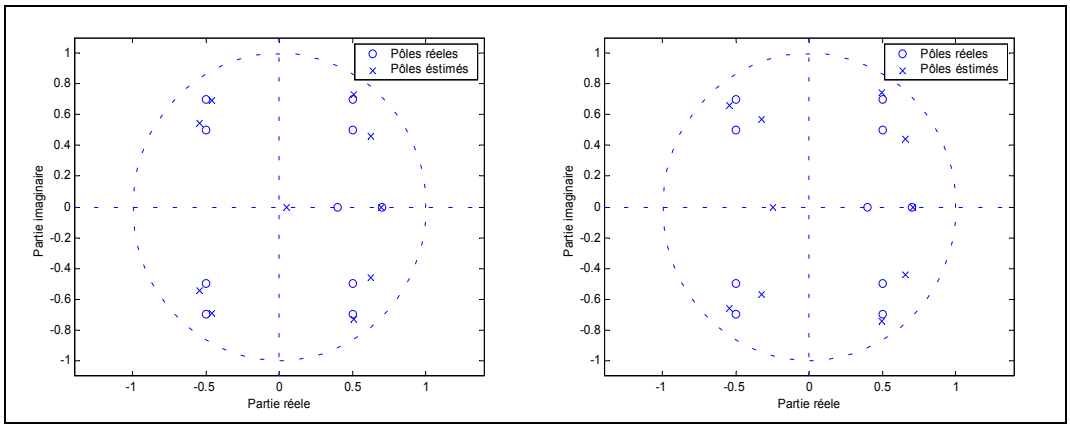
(b.1) (P=6) (b.2)

(c.1) (P=10) (c.2)

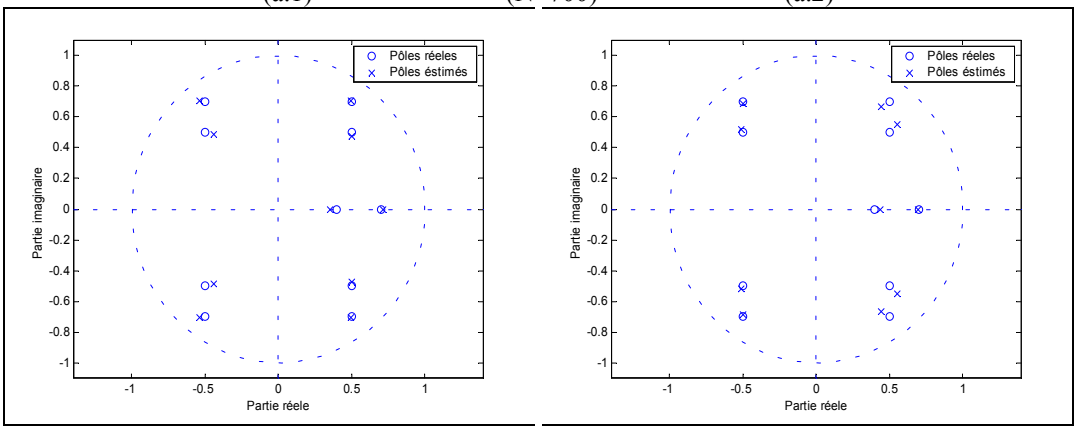
Figure 4.18 : Pôles de la Fonction de Transfert avec N=2000.
1) : Par Calcul direct.
2) : Par L'algorithme de Levinson.

Une erreur d'estimation est remarquable sur deux pôles situés sur l'axe des réels qui peut être résolue en augmentant la valeur de N (Figure 4.18.c).

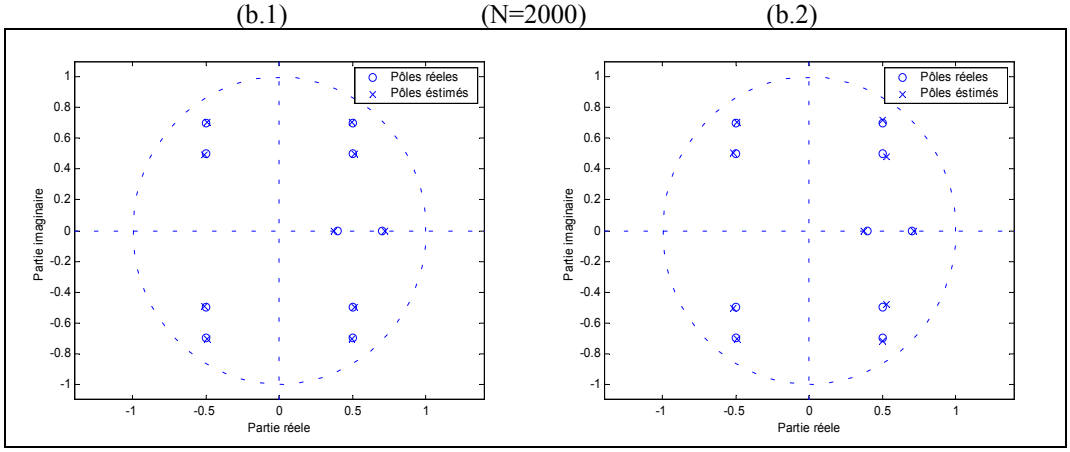
La deuxième expérience consiste à fixer l'ordre du modèle (P=10), et de varier la valeur de N pour diminuer l'écart quadratique et minimiser l'erreur d'estimation pour les deux techniques (Figure 4.19).



(a.1) (N=700) (a.2)



(b.1) (N=2000) (b.2)



(c.1) (N=3000) (c.2)

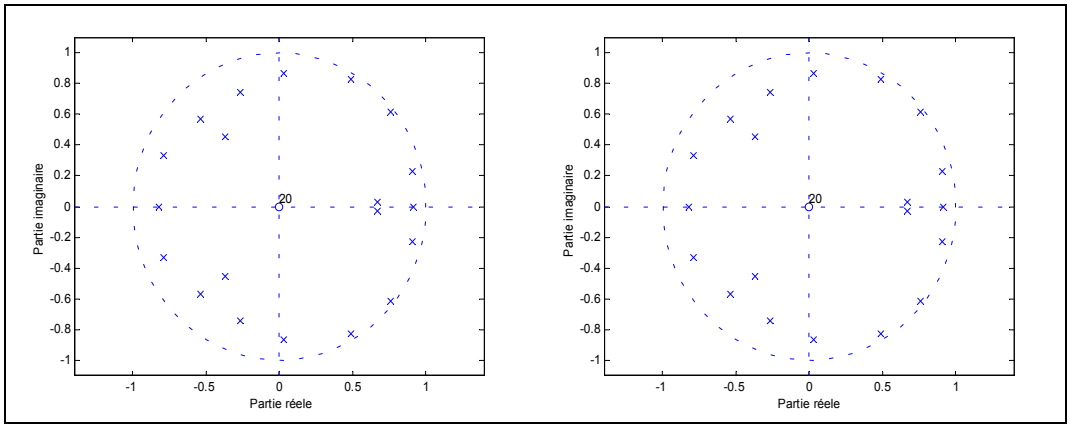
Figure 4.19 : Pôles de la Fonction de Transfert avec P=10.
1) : Par calcul direct.
2) : Par l'algorithme de Levinson.

Nous pouvons constater à partir de ces tests qu'une bonne estimation est atteinte pour N=2000 en comparant à celle où N=700. Pour obtenir une grande précision avec un écart quadratique minimal, il suffit d'augmenter la valeur de N (N=3000) où nous obtenons une estimation optimale des pôles de la FT (Figure 4.19). Comme conclusion, il est restrictif de supposer qu'un processus AR ne dépend que d'un nombre fini de paramètres.

A partir de ces tests, nous pouvons envisager la difficulté d'une modélisation AR surtout pour un signal de parole.

4.5.3. Fonction de Transfert du conduit vocal

Les coefficients LPC représentent un polynôme en z^{-1} qui n'est autre que le dénominateur de la FT du conduit vocal. Comme ce dernier est par essence stable, les pôles doivent se trouver à l'intérieur du cercle de rayon unité. La donnée de la FT sous la forme du tracé des pôles du conduit vocal dans le plan complexe avec un ordre de prédiction P=20 pour les deux techniques et pour une durée de 20 ms du phonème [a] du contexte **« darso »** de la phrase donnée en §4.2 est représentée dans la figure (4.20).



(a) (20ms) (b)

Figure 4.20 : Pôles de la fonction de transfert du phonème [a] sur 20 ms avec P=20.
a) : Par calcul direct.
b) : Par l'algorithme de Levinson.

La réponse fréquentielle du conduit vocal est donnée sur la figure 4.21 :

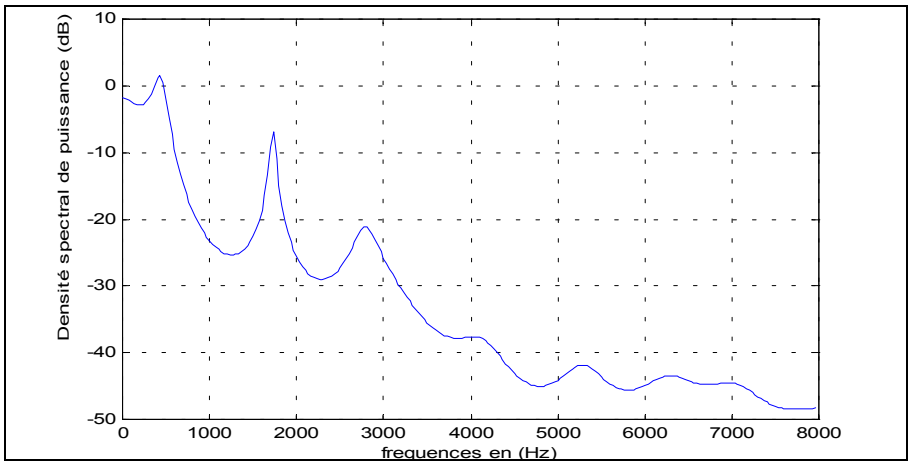


Figure 4.21 : Enveloppe spectrale du phonème [a] sur un intervalle de 30 ms

4.5.4. Synthèse du signal par prédiction linéaire

La reconstitution du signal consiste à utiliser les paramètres du modèle pour synthétiser une tranche après l'autre.

Pour atténuer l'effet des variations éventuelles et brutales des coefficients d'une tranche à la suivante, nous effectuons une pondération des tranches par une fenêtre de Hamming de longueur d'une trame et un chevauchement de 10 ms sur les sorties successives calculées sur des tranches de temps de 30 ms, la synthèse se fait par addition recouvrement des fenêtres qui se chevauchent.

Il faudra tout d'abord créer le signal d'excitation (périodique ou bruité selon que le signal est Voisé ou Non), le filtrer à travers la FT H(z) et adapté son amplitude.

Le signal bruité est un bruit blanc de moyenne nulle et de variance unité ; le signal périodique est un train impulsionnel, sa période étant la période fondamentale de chaque tranche d'analyse.

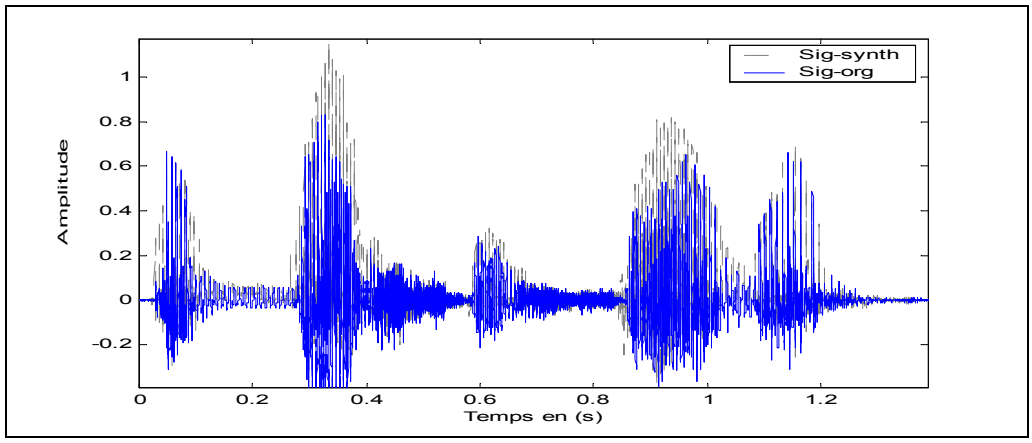
Comme les sons évoluent constamment, le générateur et le filtre doivent être modifiés en permanence. Notons que la forme de l'impulsion est un peu critique pour une telle méthode de synthèse qui traduit à son tour l'intelligibilité du signal obtenu. Une désaccentuation $(1/(1-\mu\ z^{-1}))$, avant la recomposition est effectuée si une préaccentuation a eu lieu en analyse du signal.

4.5.5. Résultats de la simulation LPC

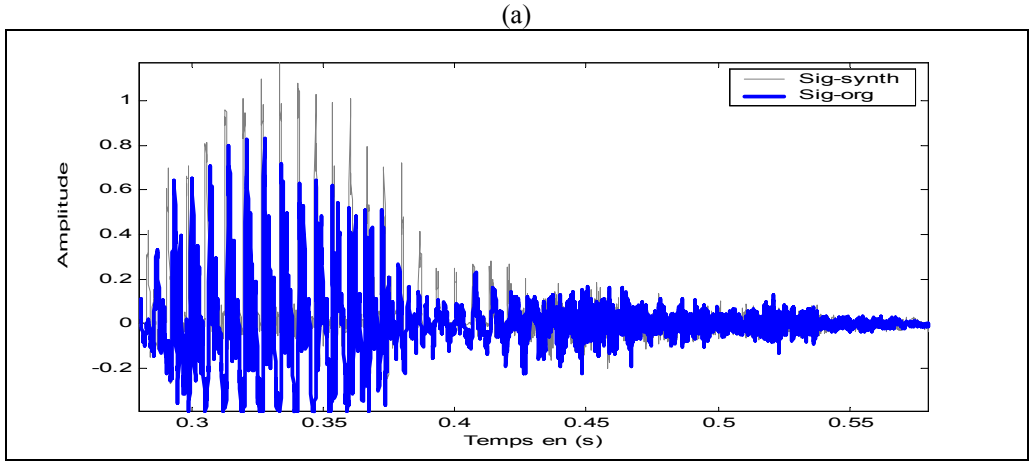
Nous présentons dans ce paragraphe les tests réalisés sur le signal de parole décrit au paragraphe §4.2 pour l'obtention d'un signal synthétique, puis nous effectuons des modifications de la F_0 à l'aide de cette technique.

La figure 4.22 montre les résultats de la synthèse de la phrase test. Ces différences visuelles nous paraissent difficilement acceptables.

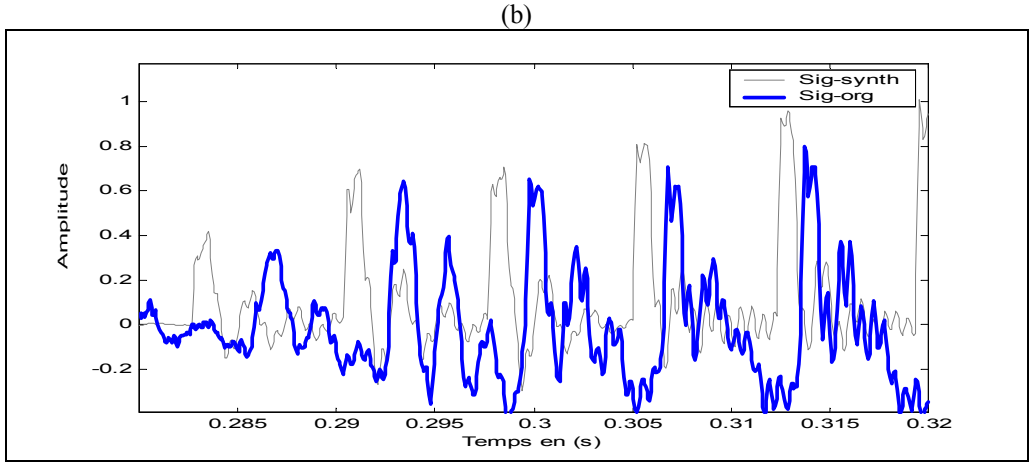
Il ne faut cependant pas oublier que la parole est un message très redondant et que seule l'écoute de la phrase synthétisée permet de juger de la qualité du codage LPC.



(a)



(b)



(c)

Figure 4.22 : Représentation temporelle du signal original et synthétique.
a) : phrase entière.
b) : 30 ms du phonème [a].
c) : 10 ms du phonème [a].

Pour l'obtention d'un signal synthétique de F_0 différente, comme pour le premier cas, trois facteurs de modifications différentes sont présentés ; le facteur 1.3 (facteur supérieur à 1), le facteur 0.8 (facteur inférieur à 1) et enfin un facteur égal à 1 qui est pour cette technique le signal synthétique représenté par la figure 4.22 où aucune modification n'est

appliquée. Sur ces figures les signaux originaux et synthétiques sont notés respectivement sig-org et sig-synth.

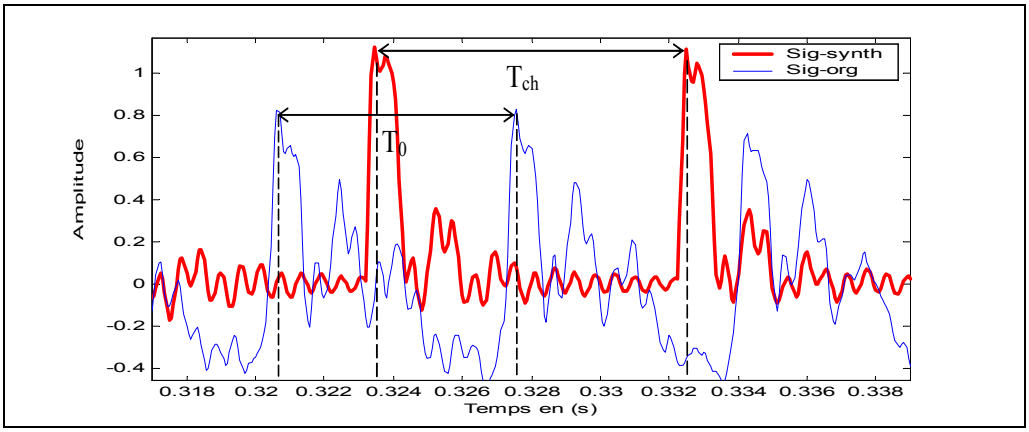


Figure 4.23 : Opération d'augmentation de T_0 (facteur =1.3) par la modélisation Source-Filtre par prédiction linéaire sur un intervalle de 20 ms du phonème [a].

La figure 4.23 donne une représentation temporelle du signal décrivant un intervalle temporel du phonème [a] extrait du signal original à partir du mot « **addarso** », qui possède une F_0 égale à 144.45 Hz. Ce dernier est superposé à un autre signal obtenu par la modélisation Source Filtre par prédiction linéaire du même intervalle temporel et du même phonème mais possédant une période fondamentale multipliée par un facteur égal à 1.3 par rapport à la période du signal de départ. La F_0 du signal résultant est égale à 111.16Hz, qui vérifie l'hypothèse de départ.

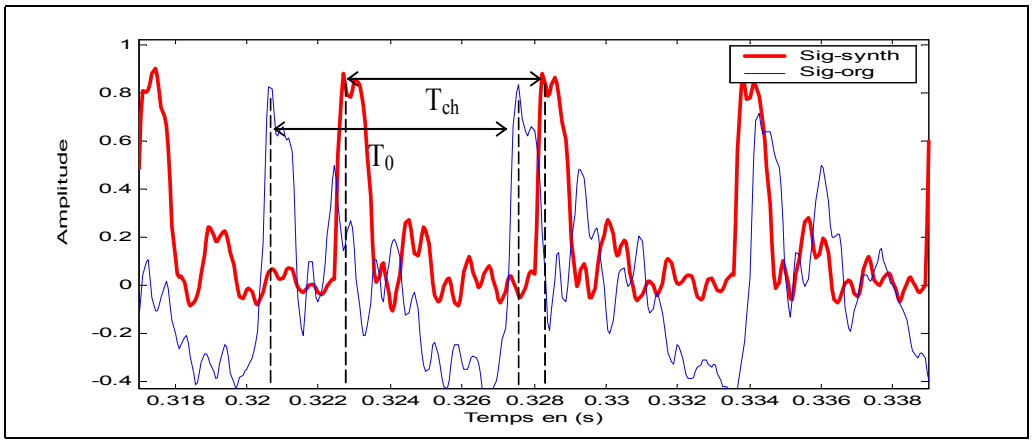


Figure 4.24 : Opération de diminution de la T_0 (facteur =0.8) par la modélisation Source-Filtre par prédiction linéaire sur un intervalle de 20 ms du phonème [a].

L'obtention d'un signal synthétique d'une T_0 inférieure à celle de l'originale est donnée par la figure 4.24 qui étale la représentation temporelle du signal analytique et synthétique pour le même intervalle temporel et avec un facteur de modification égal à 0.8. La F_0 du signal obtenu est égale cette fois à 181.810 Hz. Dans ces deux figures, T_0 représente la période du signal analytique et T_{ch} celle du signal synthétique reconstitué.

Nous remarquons que le signal reconstitué subit une distorsion d'amplitude. Le maintien du rapport d'amplitude est assuré théoriquement par la modification simultanée de la valeur de gain σ à partir de sa valeur initiale σ_0 au moyen de l'équation 3.38**.**

Malheureusement la modélisation AR souffre d'erreurs qui en limitent fortement la qualité et ne permettent pas d'assurer le rapport d'amplitude provoquant ainsi des pertes d'énergie qui sont dans certains cas non négligeables et affectent la qualité segmentale du signal synthétique. Il est difficile d'améliorer cette qualité en augmentant l'ordre du modèle, ou même la F_e. L'avantage d'une telle technique est quelle réalise la modification de la fréquence fondamentale d'un signal de parole sans changement de l'enveloppe spectrale qui à son rôle préserve le timbre de la parole aussi bien que possible (Figure 4.25).

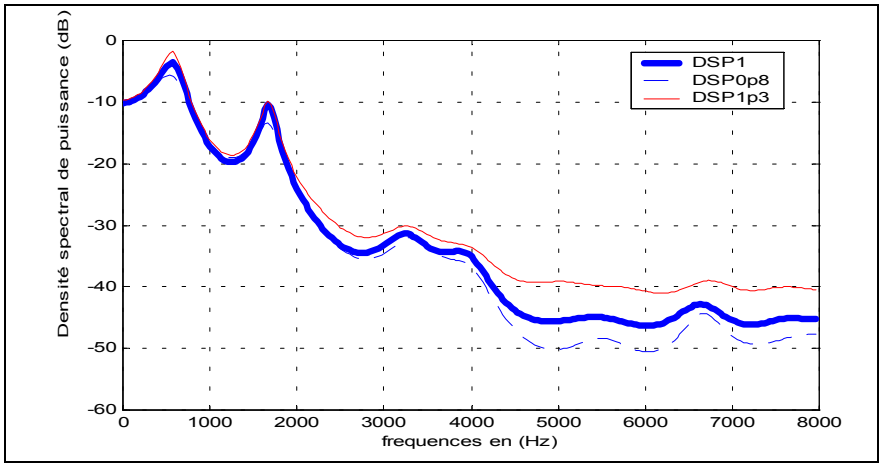


Figure 4.25 : Enveloppes spectrales des signaux interpolés et du signal synthétique du phonème [a] sur un intervalle de 30 ms.

L'observation de l'enveloppe spectrale du signal synthétique et des signaux interpolés (pour un facteur 1.3 et 0.8) nous montrent dans les deux cas que les résultats obtenus sont quasi équivalents (Figure 4.25), assurant ainsi le maintient de l'enveloppe spectrale ; ou DSP0p8 et DSP1p3 représentent respectivement l'enveloppe spectrale des signaux interpolés pour des facteurs égaux à 0.8 et 1.3 et DSP1 est l'enveloppe spectrale

du signal synthétique sans changement de la F_0. On remarque que la Densité Spectrale de Puissance du signal synthétique obtenu par un facteur égal à 0.8 présente une légère diminution d'énergie par rapport à la densité spectrale du signal original, cela est du au problème de distorsion d'amplitude qui a causé à son tour une dégradation de l'énergie du signal (Figure 4.25).

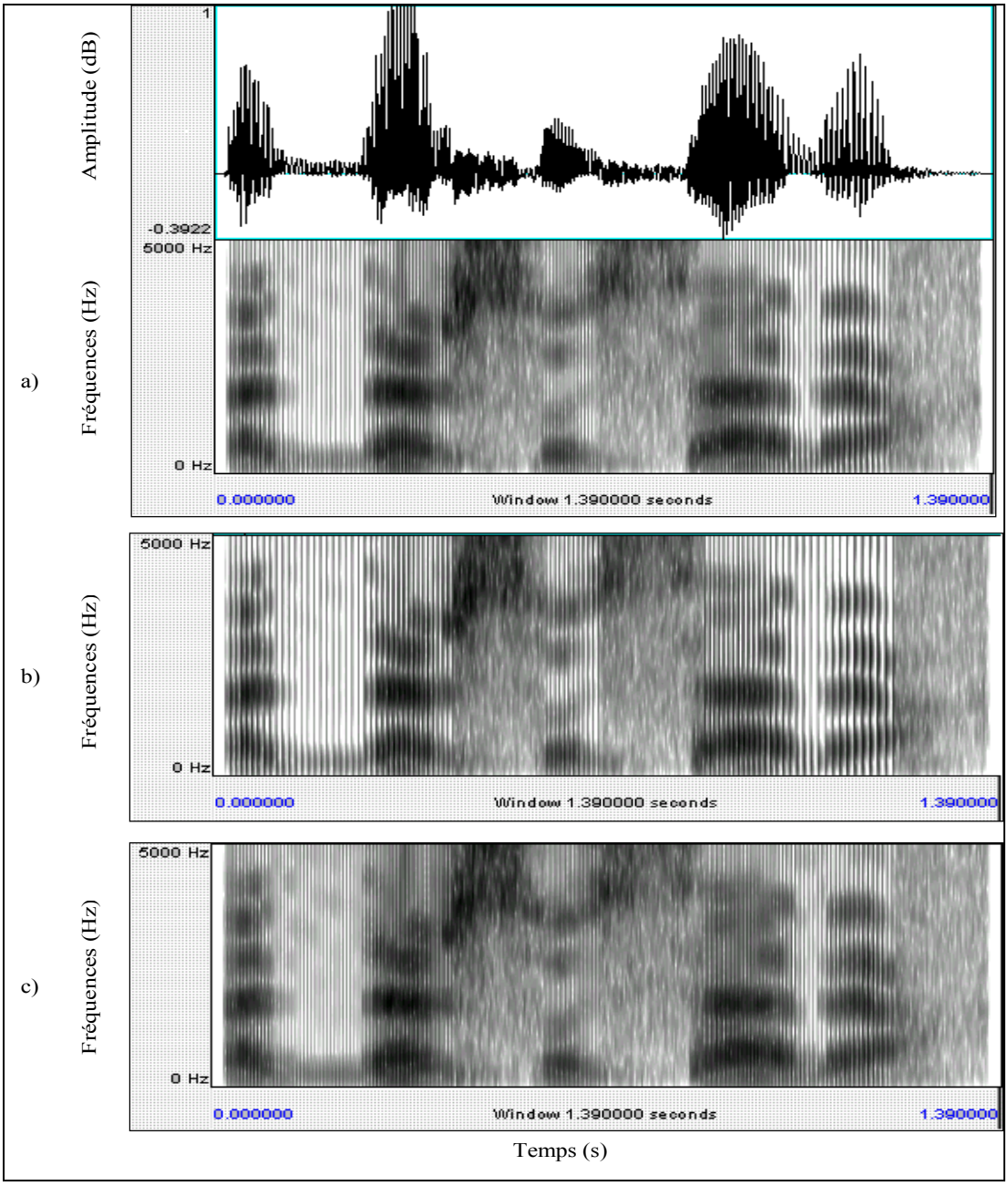


Figure 4.26 : Spectrogrammes des signaux interpolés et du signal synthétique de la phrase test obtenu par la modélisation source conduit vocal par prédiction linéaire.
a) : Signal synthétique.
b) : Signal synthétique avec un facteur de 1.3.
c) : Signal synthétique avec un facteur de 0.8.

La figure 4.26 effectue une comparaison entre les spectrogrammes des deux signaux modifiés soit par un facteur de 0.8 ou 1.3 et le spectrogramme du signal synthétique. Il est bien clair dans cette figure que la valeur et la trajectoire des formants sont maintenus le long du signal, ainsi on peut conclure qu'on a pu aboutir à des signaux de fréquences fondamentales différentes à partir d'un signal de référence en maintenant l'enveloppe spectrale inchangée et en préservant ainsi le timbre de la voix.

4.6. Evaluation des techniques

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

Dans un premier temps, nous allons donner une appréciation sur la qualité globale des deux techniques de synthèse employées en se basant sur des tests d'écoute qui visent à juger l'intelligibilité et la qualité du message parlé obtenu à la sortie de notre système sans se préoccuper de son fonctionnement interne et sans chercher la source des défauts éventuels.

Le second test vise à estimer l'aptitude des techniques à effectuer des modifications de la F_0 tout en conservant l'intelligibilité et la qualité de la parole synthétisée.

Réellement les tests d'écoute doivent être réalisés auprès des experts audio qui vont par suite donner leurs appréciations sur la qualité et l'intelligibilité du signal synthétique. Dans notre cas, nous allons essayer de donner une évaluation nous mêmes, en se basant toujours sur des tests d'écoute en utilisant plusieurs messages vocales prononcés en langue Arabe et appliqués à l'entrée de notre système pour les deux techniques (tableau 4.1).

Pour ce faire, nous avons réalisé une interface graphique à l'aide du logiciel MATLAB qui nous permet d'effectuer plusieurs opérations d'analyse / synthèse /modifications de la F_0 par le biais des deux techniques décrites auparavant.

Tableau 4.1 : Messages vocales utilisés.

1	الدرس السابع
2	وسكت الصوت مرة أخرى
3	الموقع الجغرافي
4	ظهور الإسلام و انتشاره
5	أهل الجنة
6	الصحافة العربية

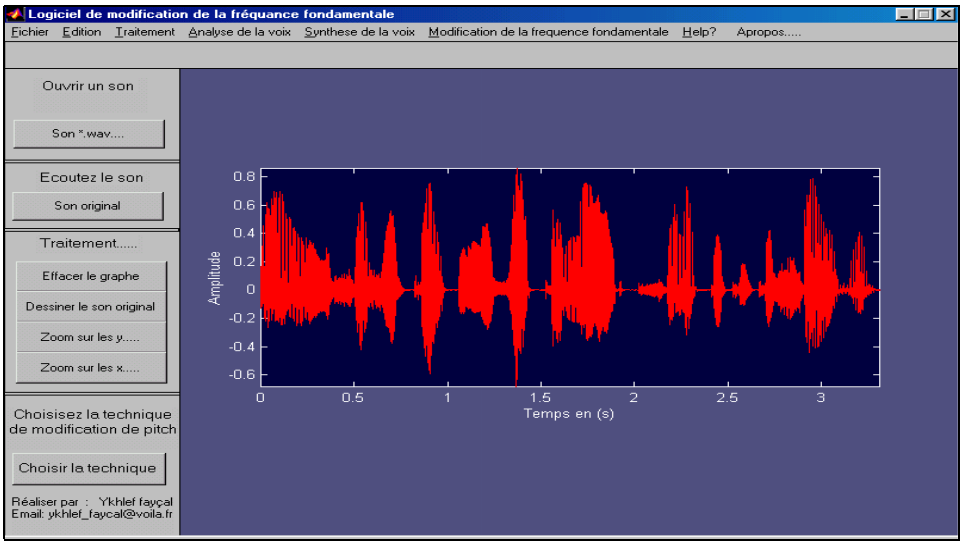
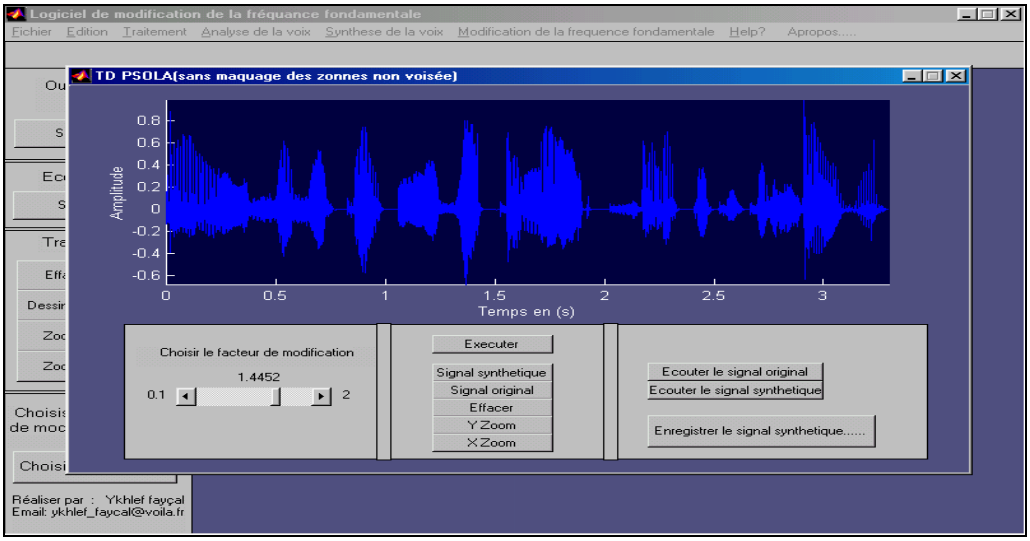


Figure 4.27 : Interface Graphique Réalisée.

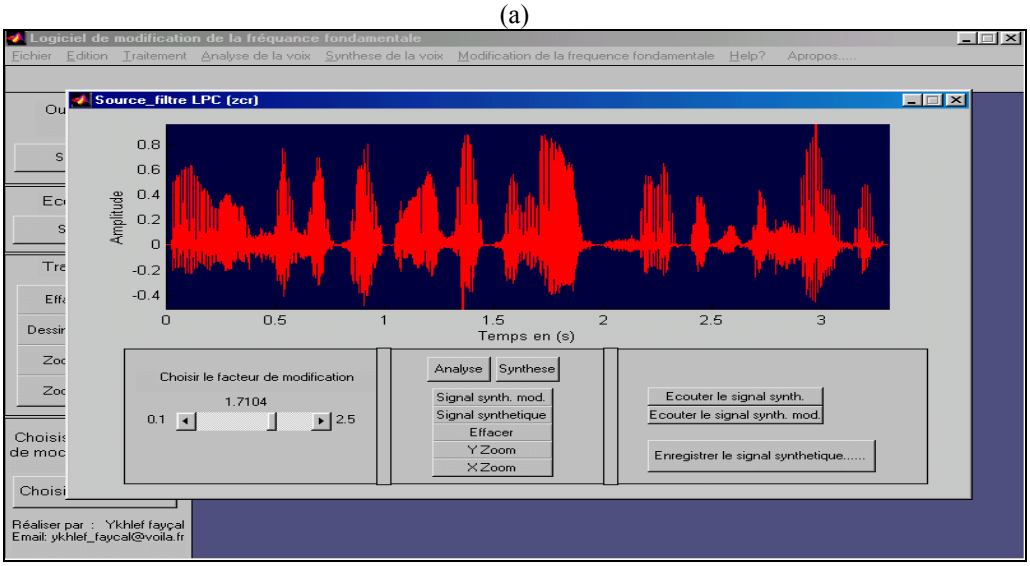
Ces opérations peuvent être résumées comme suit :

- ouverture d'un son à partir d'un fichier *.wav ;
- représentation temporelle du signal vocal ;
- reproduction sonore du message vocal en variant la fréquence d'échantillonnage ;
- l'analyse de la voix par trames de 30 ms de durée qui comprend :
  - l'enveloppe spectrale des trames du signal vocal de 30 ms de durée ;
  - une représentation des pôles estimés de la FT des trames de 30 ms de durée du signal vocal dans le plan des Z, par deux techniques différentes ; l'algorithme de Levinson et la résolution directe des équations du système ;
  - une représentation temporelle des deux techniques de segmentation V/NV étudiée ;
  - évolution de la F_0 en fonction des trames d'analyse par la technique d'autocorrélation ;
  - le spectrogramme du signal vocal ;
  - la Transformée de Fourrier du signal vocal ;

- la synthèse de la voix par :
  - la modélisation source - conduit vocal par prédiction linéaire ;
  - TD-PSOLA (pour un facteur de modification fixé égal à 1).
- la modification de la fréquence fondamentale par :
  - la modélisation source conduit vocal par la prédiction linéaire (suivant les deux techniques de segmentation) Figure 4.28.a ;
  - TD PSOLA (Avec les deux techniques de marquage proposées au paragraphe § 4.3.1 (Figure 4.28.b).



(a)



(b)

Figure 4.28 : Module de modification de la fréquence Fondamentale.
a) : TD-PSOLA.
b) : Modélisation Source-Filtre par prédiction linéaire.

Grâce à cette interface, nous pouvons effectuer tout test sonore d'une simple manière afin de pouvoir donner nos appréciations sur les deux techniques et de pouvoir déterminer les seuils des facteurs de modifications qui conservant l'intelligibilité et la qualité du message synthétiser pour les deux techniques étudiées.

4.6.1. Evaluation globale

Afin de pouvoir estimer la qualité globale du message parlé, notre choix porte sur une procédure multidimensionnelle d'évaluation de la qualité globale qui à été proposée en 1995 pour l'Allemand dans le cadre du projet Vebmobil que nous allons appliquer par la suite à la langue Arabe. Les huit dimensions d'analyse et les échelles associées sont :

Le naturel est :
1 : très naturel.
2 : naturel.
3 : plutôt naturel.
4 : plutôt peu naturel
5 : peu naturel
6 : très peu naturel.

L'intelligibilité est :
1 : très facile.
2 : facile
3 : plutôt facile.
4 : plutôt difficile.
5 : difficile
6 : très difficile

La compréhension est :
1 : très facile.
2 : facile
3 : plutôt facile.
4 : plutôt difficile.
5 : difficile
6 : très difficile

L'agrément est :
1 : très agréable.
2 : agréable
3 : plutôt agréable.
4 : plutôt désagréable.
5 : désagréable
6 : très désagréable

La netteté est :
1 : très claire.
2 : claire.

3 : plutôt claire.
4 : plutôt brouillée.
5 : brouillée.
6 : très brouillée.

Le débit est :
1 : beaucoup trop lent.
2 : trop lent.
3 : légèrement trop lent.
4 : légèrement trop rapide.
5 : trop rapide.
6 : beaucoup trop rapide.

Les erreurs de prononciation sont :
1 : pas gênantes.
2 : légèrement gênantes.
3 : plutôt gênantes.
4 : gênantes.
5 : très gênantes.

Les erreurs d'accentuation sont :
1 : pas gênantes.
2 : légèrement gênantes.
3 : plutôt gênantes.
4 : gênantes.
5 : très gênantes.

4.6.1.1. Evaluation de la technique TD-PSOLA

TD-PSOLA n'est pas à proprement parlée une technique de synthèse du signal de la parole. Il s'agit d'une technique de traitement du signal de parole- qu'il soit naturel ou de synthèse – dont l'objectif est de modifier les paramètres prosodiques de celui-ci. Si maintenant on la considère comme une technique de synthèse de parole (si le facteur de modification de F_0 est égal à 1) ; elle va sans aucun doute avoir les meilleurs échelles de Vebmobil puisque c'est une reproduction excellente de signal naturel (d'après des tests d'écoute). L'évaluation est résumée comme suite :

Le naturel est :
1 : très naturel.

L'intelligibilité est :
1 : très facile.

La compréhension est :
1 : très facile.

L'agrément est :
1 : très agréable.

La netteté est :
1 : très claire.

Le débit est :
Naturel

Les erreurs de prononciation sont :
1 : pas gênantes.

Les erreurs d'accentuation sont :
1 : pas gênantes.

4.6.1.2. Evaluation de la modélisation Source – Filtre par prédiction linéaire

C'est une modélisation qui a été largement utilisée à la fin des années 1980 et qui a cédé la place à des techniques plus complexes et offre une meilleure qualité du signal. Nous allons par suite donner une évaluation de cette technique en se basant toujours sur des tests d'écoute et sans se préoccuper de son fonctionnement interne.

L'évaluation est résumée comme suit :

Le naturel est :
5 : peu naturel.

L'intelligibilité est :
3 : plutôt facile.

La compréhension est :
3 : plutôt facile.

L'agrément est :
3 : plutôt agréable.

La netteté est :
3 : plutôt claire.

Le débit est :
4 : légèrement trop rapide.

Les erreurs de prononciation sont :
3 : plutôt gênantes.

Les erreurs d'accentuation sont :
4 : gênantes.

4.6.2 Evaluation analytique

Contrairement à l'évaluation globale ou externe qui traite le système comme étant une boite noire, et effectue une évaluation de la sortie du système de synthèse, cependant l'évaluation analytique se préoccupe du fonctionnement interne et cherche la source des défauts éventuels « boite verre ».

Pratiquement , si nous avons un système de synthèse de la parole , nous sommes amenés à faire une évaluation séparée des différents maillons du système que ça soit le module de la TOP ou le module prosodique ou même le synthétiseur acoustique. Dans notre cas, nous allons nous préoccuper du synthétiseur acoustique puisque le but de notre travail est d'effectuer des modifications prosodiques et plus précisément les modifications de la F_0.

En premier lieu nous allons citer quelques défauts des deux techniques utilisées qui limitent la qualité ou l'intelligibilité du signal synthétique obtenu à la sortie du système, puis nous allons déterminer les extrémités des facteurs de modifications de la F_0 que nous pouvons aboutir à l'aide de ces deux techniques sans trop perturber le signal de sortie.

D'une façon globale, une étape préliminaire pour tout système d'analyse - synthèse – modifications prosodiques est la détection du fondamental qui peut être dans le cas d'une erreur de détection, la source des différentes dégradations éventuelles qui peut surgir et affecter le signal synthétisé qui est dans notre cas une erreur commune entre les deux techniques employées pour réaliser la modification de la F_0. Cette erreur est appelée erreur d'estimation du pitch et détection du voisement.

- *Erreurs de détection du voisement* : la détermination du voisement est une étape nécessaire dans un algorithme de détection du fondamental. Il existe plusieurs méthodes de décision du voisement qui ont en commun d'analyser un intervalle de parole précis pour assurer la stationnarité de ces segments. C'est à partir de cette décision que nous allons donner une estimation de la période du fondamentale.
- *Erreurs dues à l'estimation de la fréquence fondamentale :* l'approximation faite sur la valeur du pitch par la fonction d'autocorrélation ne permet pas d'estimer les fréquences des trames du signal avec une grande précision et cela est dû essentiellement aux problèmes cités au § 4.3.3. La période du fondamental est un paramètre très important pour la synthèse de la parole, c'est à partir de cette détection que nous allons synthétiser et modifier la F_0 du signal étudié.

La propagation de l'erreur d'estimation va fortement diminuer la qualité du signal synthétique puisque cette détection est une tâche préliminaire pour tout système d'analyse/synthèse/ modifications prosodiques d'un signal vocal.

Comme pour le cas précèdent, TD-PSOLA n'est pas à proprement parler une technique de synthèse du signal de la parole alors on ne peut rien en juger sur la qualité ou l'intelligibilité du signal obtenu avec un facteur qui est égal à un puisque c'est une reproduction du naturel. Le problème se trouve dans le cas d'une modification de la F_0 où la dégradation de la qualité du message vocal obtenu lors d'une modification est remarquable et augmente en élevant ou en diminuant suffisamment le facteur de modifications. Les problèmes ou les raisons de cette dégradation peuvent être résumés comme suit :

- *Erreurs dues au marquage du fondamentale :* la précision du placement des marques détermine en grande partie la qualité du signal synthétisé. La technique de marquage proposée dans notre cas ne permet pas d'avoir une très bonne qualité du signal synthétique obtenu puisqu'elle ne vérifie pas toutes les conditions du marquage décrites au paragraphe § 3.2.3.3. Elle garantit une synchronisation à la période locale ; un espacement entre deux marques successives qui est égale à la période locale mais ne permet pas cependant l'éloignement minimal des marques aux maxima d'énergies locaux vu la difficulté d'optimisation des différentes contraintes surtout pour un signal de parole.

- *La précision des règles de duplication élimination : l*es règles de duplication/ élimination proposées pour garantir le maintien de la durée totale du signal synthétisé en modifiant sa F_0 sont élaborées par des tests pratiques donnant les meilleurs résultats perceptuels avec moins de complexités. Ces règles ne sont cependant pas très utiles pour des différences de facteurs de modifications assez élevées où nous obtenons des signaux synthétisés avec une dégradation remarquable de la qualité du message parlé. L'élaboration de règles plus sophistiquées se basant sur la précision des marques de pitch est souvent souhaitée.

Pour la modélisation physique source conduit vocal par prédiction linéaire, nous pouvons citer quelques problèmes qui peuvent être classés comme suit :

- *Erreurs dues à la modélisation AR :* la modélisation AR souffre d'erreurs intrinsèques et extrinsèques qui en limitent fortement la qualité et qu'il est difficile d'améliorer en augmentant l'ordre du modèle, ou même la fréquence d'échantillonnage. Les sons nasalisés sont intrinsèquement mal modélisés ce qui induit par ailleurs une erreur d'estimation des formants eux-mêmes. Avec un modèle qui suppose le signal stationnaire sur une durée de 20 ou 30 ms, il est impossible de rendre compte avec précision des phénomènes transitoires dont la configuration du conduit vocal évolue trop vite (ex. les sons plosifs).

- *Erreurs dues à la source d'excitation :* le modèle adopté pour créer artificiellement des sons est approximé par rapport à la complexité du système phonatoire. Notons que la forme de l'impulsion glottique utilisée est un peu critique pour une telle méthode de synthèse. Il existe des modèles d'excitations glottiques plus sophistiqués tels que ceux de Rosenberg ou de Liljencrants-Fant qui approximent plus la forme du signal d'excitation [28].

- *Erreurs dues à la détection Voisé/Non Voisé :* la détection V/NV ne permet pas de rendre compte de façon réaliste des sons mixtes (comme les fricatives Voisées). L'utilisation d'une excitation mixte rend à la parole de synthèse une certaine rondeur qui était absente dans le modèle binaire V/NV.

Malgré tous ces problèmes de modélisation, le modèle AR a été et reste fort utilisé en synthèse vocale, vu sa grande simplicité et les rapports de compression exceptionnels qu'il permet d'obtenir.

Les seuils des facteurs de modifications atteints par une technique de modification de la F_0 sont théoriquement compris entre deux extrémités l'une maximale et l'autre minimale. Pour la technique TD-PSOLA l'extrémité maximale est obligatoirement inférieure à deux périodes de pitch puisque nous utilisons une fenêtre de pondération de longueur égale a deux périodes et qui donne une différence maximale entre deux marques de pitch qui est obligatoirement inférieure à deux périodes locales.

L'autre extrémité n'est mathématiquement pas limitée, qui est de même pour les deux extrémités concernant la modélisation source conduit vocal par la prédiction linéaire

à condition que les deux extrémités restent raisonnables et donnent cependant un message vocal intelligible.

Il est pratiquement remarqué dés qu'on est en dehors d'une extrémité maximale d'un facteur de modification qui est égal à 2.5 et une extrémité minimale de 0.4 le message vocal obtenu par la modélisation source conduit vocal a tendance de devenir de plus en plus robotique d'où une perte du naturel. Les seuils des facteurs de modifications atteints par les deux techniques donnant un signal synthétique d'une F_0 différente avec une bonne qualité et intelligibilité sonore sont :

Pour TD-PSOLA ; une extrémité minimale de 0.6 et une extrémité maximale de 1.4. Pour la modélisation source conduit vocal par prédiction linéaire ; une extrémité minimale de 0.4 et une extrémité maximale de 2 en précisant que le signal synthétisé par TD-PSOLA est plus intelligible que celui synthétisé par la modélisation source conduit vocal. On remarque une grande marge de modifications du pitch de la modélisation Source-Filtre par rapport à TD-PSOLA qui est traduit par la simplicité de la réalisation de la modification prosodique de cette technique. D'une façon générale, il est difficile de favoriser une technique par rapport à une autre puisque chacune d'entres elle présente des avantages et des inconvénients.

4.7. Conclusion

Nous avons pu concevoir dans ce chapitre deux solutions complètes de modifications prosodiques qui se démarquent fortement des autres par le contrôle des divers paramètres qui définissent le timbre de la voix. La première est basée sur la modélisation source conduit vocal par la prédiction linéaire, et la deuxième sur le principe de la ré harmonisation spectrale ; nommé TD-PSOLA et qui appartient aux familles des synthétiseurs acoustiques dans le domaine temporel.

La modélisation par la prédiction linéaire, aussi appelée synthétiseur LPC, donne de bons résultats de point de vue capacité et simplicité à réaliser des modifications de la F_0, cependant souffre d'une qualité segmentale médiocre qui affecte l'intelligibilité et la compréhension du message parlé. Cette mauvaise qualité est essentiellement due aux erreurs de modélisation, de détection du voisement et du pitch et aux erreurs dues à la source d'excitation. La source du Voisement utilisée est un peu grossière, elle ne peut approximer l'excitation glottale naturelle.

Contrairement à la modélisation source conduit-vocal, la qualité segmentale offerte par TD-PSOLA est très bonne dans le cas où on ne modifier aucune donnée prosodique (Pitch). Les imperfections surgissent lorsque nous l'utilisons pour réaliser des modifications de la F_0. Ces problèmes sont essentiellement des discontinuités de phase, de pitch et d'enveloppe spectrale qui sont dûs essentiellement aux problèmes de détection et marquages du fondamental et aux imperfections des règles de duplication/élimination. La discontinué du pitch à l'intérieur des trames est amortie par le fait de considérer ce dernier comme constant le long de la trame d'analyse. La discontinuité de phase n'est évitée, vu les erreurs de marquage du pitch qui surgit, surtout l'éloignement minimal des maxima locaux qui n'est pas garanti par une telle technique de marquage. La discontinuité de l'enveloppe spectrale n'est jamais évitée ou amortie puisqu'elle est due aux effets de la coarticulation, la seule façon d'y remédier est d'identifier les diphones qui posent problème et de les remplacer. C'est donc une étape longue et coûteuse. La charge de calcul prix par un algorithme TD-PSOLA est dix fois moins à celle prise par un synthétiseur LPC [1]. Cette charge de calcul est en grande partie le temps nécessaire de l'analyse et l'extraction des différents paramètres d'un tel signal.

CONCLUSION

Notre objectif est la réalisation des techniques de modifications de la fréquence fondamentale utilisées dans les systèmes de synthèse de la parole à partir du texte.

Pour y parvenir, il faut opérer sur deux formes de transformations : soit par transformation prosodique, soit par transformation spectrale avec changement de fréquence d'échantillonnage.

La transformation prosodique contrôle les divers paramètres qui définissent le timbre d'une voix qui est malheureusement non réalisable à l'aide d'une transformation spectrale avec changement de fréquence d'échantillonnage.

Nous avons utilisé comme techniques de modifications de la F_0 la modélisation source conduit vocal par prédiction linéaire (*Linear Predictive Coding*) et la TD-PSOLA (*Time Domain Pitch Synchronous Overlap and Add*).

La première, décrit le signal de la parole comme un système source filtre ; la source étant le passage de l'air expirées à travers les cordes vocales et le filtre étant le conduit vocal. Dans notre cas, le conduit vocal est modélisé par un filtre Auto-Règressif.

La source est modélisée pour les sons Voisés par un train d'impulsions de même fréquence que la F_0 du signal de parole, et pour les sons Non Voisés, par un bruit blanc.

De ce fait, la modification de l'intonation est une opération élémentaire. La période de pitch du signal étant un paramètre du modèle, il suffit d'en imposer directement la valeur de synthèse en modifiant la période du train périodique d'impulsions utilisé en excitation des zones Voisés.

La voix obtenue est de mauvaise qualité à cause des erreurs éventuelles qui sont essentiellement des erreurs de modélisation, des erreurs dues à la source d'excitation et à la décision V/NV et des erreurs de détection du fondamental. L'intérêt principal vient du petit nombre de paramètres à faire évaluer (les coefficients du filtre AR). C'est donc très intéressant pour certaines applications et surtout pour la modification prosodique.

La deuxième technique, TD-PSOLA, repose sur le principe de réaharmonisation spectrale qui appartient à la famille des synthétiseurs acoustiques dans le domaine temporel qui peut être traduit en Français par recouvrement et addition des fenêtres temporelles

synchrones avec le pitch. Les points forts de cette technique sont la facilité, la flexibilité et la qualité de modifications de la F_0.

Les phases les plus délicates dans la conception de cette technique sont le marquage du fondamental et l'élaboration des règles duplication/élimination. Ces dernières jouent un rôle très important dans la qualité du signal synthétique. Nous nous sommes basés sur des tests d'écoute pour l'élaboration de meilleures règles et conditions de marquages avec moins de complexité et qui peuvent assurer un fonctionnement automatique.

La qualité segmentale est très bonne, si bien que si l'on ne modifie aucune donnée prosodique, le signal de sortie est identique au signal en entrée. Cette méthode présente néanmoins quelques imperfections lorsqu'elle est utilisée pour réaliser des modifications de la F_0. Ces problèmes sont essentiellement dûs aux des discontinuités de phase, de pitch et d'enveloppe spectrale.

Pour terminer cette étude il faut préciser qu'un synthétiseur LPC basé sur la modélisation source conduit vocal est peu utilisé puisque la qualité est largement meilleure avec d'autres techniques peu complexes. La technique qui est encore la plus courante et la plus fiable est la synthèse de la parole dans le domaine temporel basée sur le principe d'une réaharmonisation spectrale PSOLA.

Il existe d'autres variantes autour de la technique PSOLA qui combinent les deux techniques étudiées, et qui donnent des meilleurs résultats, soit la technique LP PSOLA qui réalise l'addition recouvrement sur le résiduel du signal.

Nous pouvons dire que les résultats obtenus sont satisfaisants, ce qui nous encourage à poursuivre le travail, notamment pour la modification de la durée segmentale d'un tel signal qui sera une opération élémentaire. Une fois la modification de la fréquence Fondamentale réalisée, le changement de la durée avec la technique TD-PSOLA peut se faire en dupliquant ou en éliminant autant de fois que nécessaire les fenêtres à courts termes. Pour le synthétiseur LPC le changement de la durée se fait en allongeant ou en rétrécissant le résiduel d'excitation pour aboutir à des trames de durées différentes pour l'obtention d'un signal synthétique de durée totale différente. Cette étude brève dans le monde de la synthèse de la parole à partir du texte nous a permis de maîtriser des techniques de traitement du signal, et de se lancer dans des domaines de recherche liés à la synthèse de la parole. Ce qui présente un gain considérable et une bonne base à la recherche scientifique dans différentes spécialités et domaines.

A la fin, il serait intéressant de souligner le fait que malgré l'existence de certains produits sur le marché, les synthétiseurs vocaux ne sont pas complètement prêts. Ces outils ne s'expriment pas avec tout le naturel dont l'être humain est capable. Actuellement l'adaptation de ces produits au grand public ne semble pas être une bonne idée car les synthétiseurs sont encore trop précoces et le grand public risque de refuser le produit. Il faut signaler que la synthèse vocale, bien qu'intelligible, demande plus de concentration de la part de l'auditeur que lorsqu'il s'agit d'une discussion entre humains. Cela provient justement du fait qu'elle ne s'exprime pas avec naturel. Son application ne peut donc pas s'étendre sur des services nécessitant une longue interaction avec celui-ci. Néanmoins, si on estime que la plupart des appels téléphoniques durent en moyenne 3 minutes, cette technologie garde tout son intérêt, notamment pour la consultation de service rapide (e-mail, agenda, etc…).

APPENDICE A

LISTE DES SYMBOLES ET DES ABREVIATIONS

ACR	Absolute Cathegory Rating
AMDF	Average Magnitude Difference Function
API	Alphabet Phonétique International
AR	Auto-Regressive
ARMA	Auto-Regressive Moving Average
AS	Arabe Standard
DCR	Degradation Cathegory Rating
EPZ	Etat de Passage par Zéros
F_0	Fréquence Fondamentale
FD-PSOLA	Fréquency Domain-PSOLA
F_e	Fréquence d'échantillonnage
F_i	$I^{éme}$ Formant
FP	Fonction de Périodicité
FT	Fonction de Transfert
IPS	Instrumental Pitch Shifting
JEIDA	Japan Electronic Industry Development Association
LPC	Linear Predictive Coding
LP-PSOLA	Linear prediction PSOLA
MBROLA	MultiBand Overlap and Add
MHB	Modèle Harmonique et Bruit
OLA	OverLap and Addition
PSOLA	Pitch synchronous Overlap and Add
SIFT	Simplified Inverse Filter Tracking
SRPD	Super Resolution Pitch Determination
T_0	Période Fondamentale
TAP	Traitement Automatique de la Parole
TD-PSOLA	Time Domain PSOLA
TF	Transformer de Fourier
TFI	Transformée de Fourier Inverse
TOP	Transcription Orthographique Phonétique
TPZ	Taux de Passage par Zéros
TTS	Text-To-Speech
V/NV	Voisés /Non Voisés
ZCR	Zero Crossing Rate

APPENDICE B

EXEMPLES DE REGLES DUPLICATION ELIMINATION

B.1) Règle d'élimination

Soit aléatoirement une trame j du signal de départ dont l'ensemble des marques de lectures est représentée sur la figure 1.

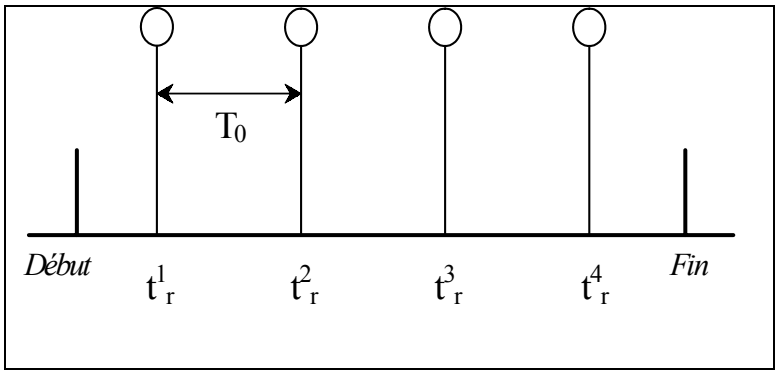


Figure 1 : Marques de Lecture.

Pour cet exemple, le nombre de marques de lecture est égal à 4 (Na=4). Si, par exemple on veut augmenter la valeur de la période du pitch T_0 (Diminuer la fréquence fondamentale $F_0=\frac{1}{T_0}$), on doit changer la valeur de la T_0 de cette trame, qui représente l'écartement entre chaque marque. On doit alors chercher les marques d'écriture qui nous permettent de réaliser cette transformation à l'intérieur de la trame actuelle qui est représenté sur la figure 2. La nouvelle période est notée T_{mod} avec $F_{mod} = \frac{1}{T_{mod}}$.

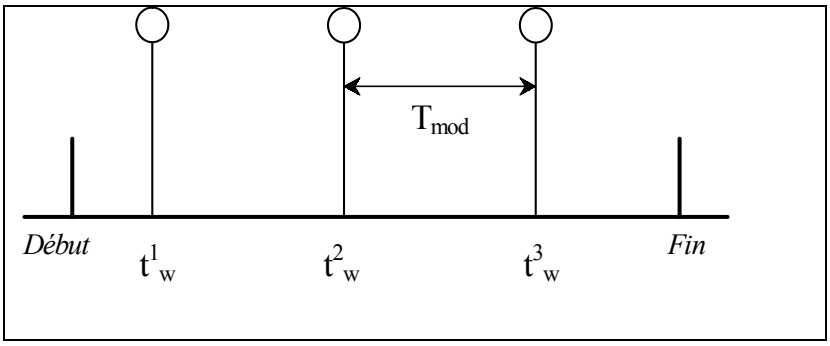


Figure 2 : Marques d'Ecriture.

On a maintenant 3 marques d'écriture sur cette trame, alors Ns=3. En appliquant l'organigramme de la figure 4.6 (voir chapitre 4) nous devons éliminer Na-Ns marques qui est cette fois égale à 1 (4-3=1). Nous devons ainsi éliminer une seule marque, celle qui est située à la fin de la trame d'analyse et correspondre les marques restantes une par une (Figure 3).

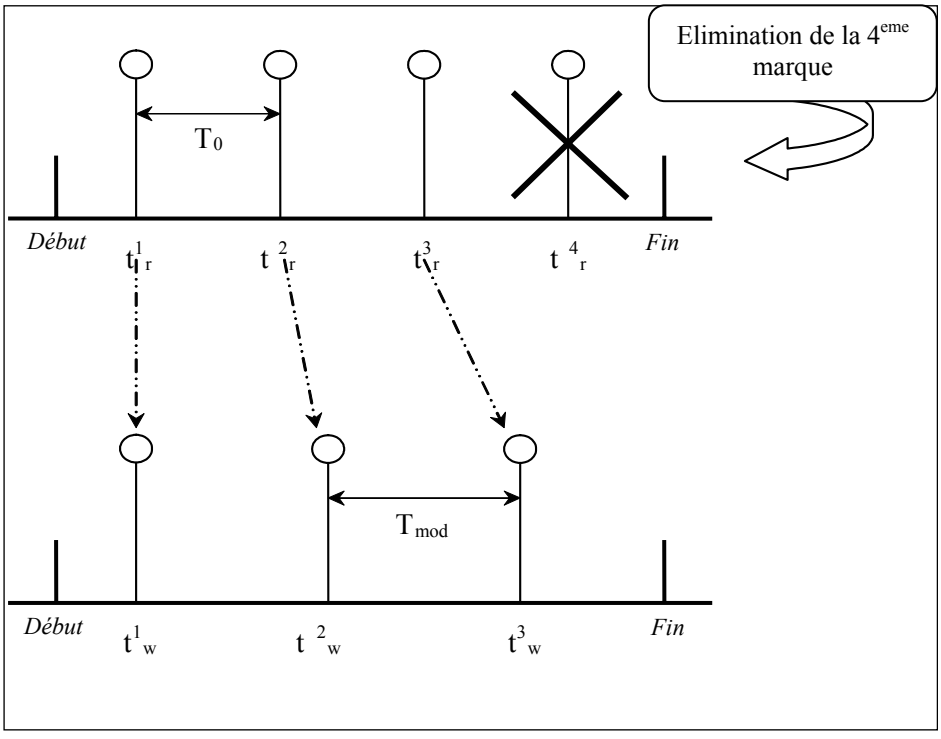


Figure 3 : Procédure de correspondance des marques Lecture/Ecriture

B.2) Règle de duplication

B.2.1) Cas d'une duplication sans élimination (Qos ≤ 2)

Soit aléatoirement une trame j du signal de départ dont le nombre de marques lecture/écriture est donné respectivement par Na=3, Ns=5 (Figure 4).

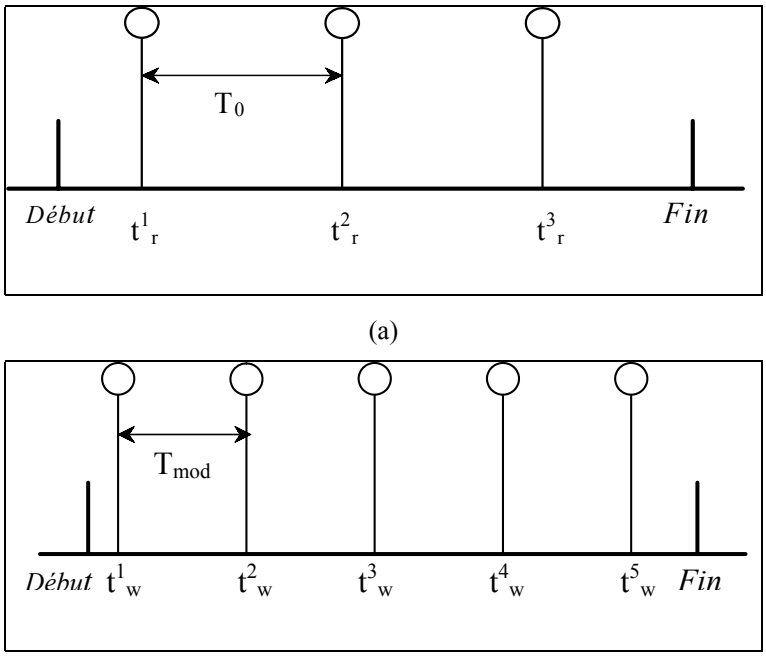


Figure 4 : (a) Marques de Lecture.
(b) Marques d'Ecriture.

Comme il a été décrit au chapitre 4, cette règle est utilisée dans le cas d'une diminution de la période du pitch qui correspond à une augmentation de la F_0. Suivant l'organigramme 3.7 nous avons :

Diff=5-3=2 ; Qos= $\frac{5}{3} \equiv 2$. Alors on duplique 2 marques ; soit les deux premiers et on fait une correspondance des marques lecture/écriture en les plaçant cette fois suivant la nouvelle cadence voulue (Figure 5).

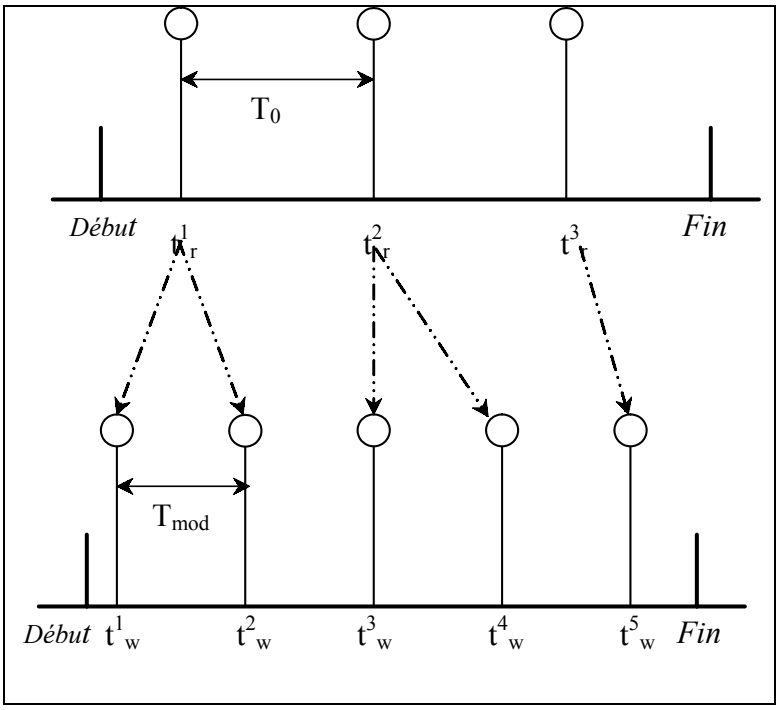


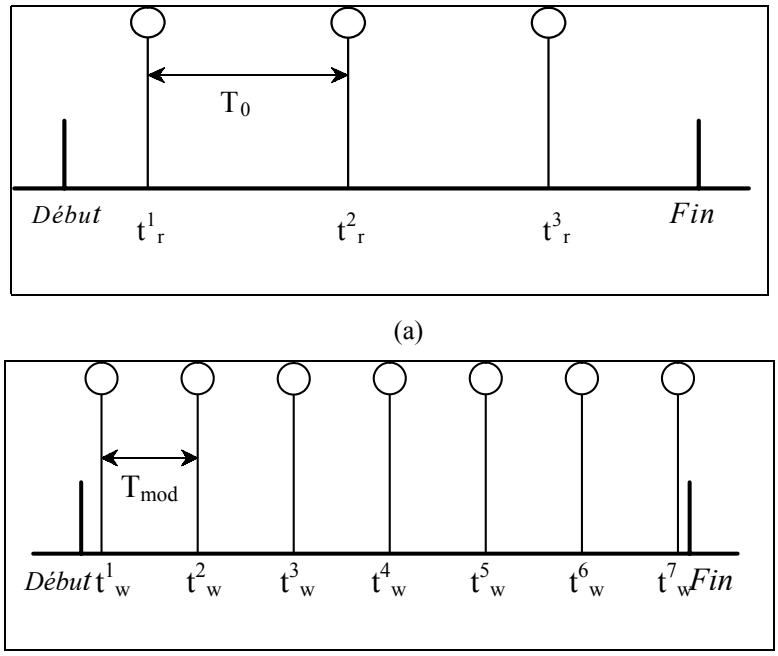
Figure 5 : Procédure de correspondance des marques Lecture/Ecriture

B.2.2) Cas d'une duplication avec élimination (Qos ≥ 2)

Soit aléatoirement une trame j du signal de départ dont le nombre de marques lecture/écriture est donné respectivement par Na=3, Ns=7 (Figure 6).

Comme pour le premier cas, cette règle est utilisée dans le cas d'une diminution de la période du pitch qui correspond à une augmentation de la F_0. Suivant l'organigramme 3.7 nous avons :

Diff=7-3=4 ; Qos= $\frac{7}{3} \equiv 3$. Alors, on reproduit 3 fois les marques de cette trame ; puis ils sont placés suivant la nouvelle cadence.



(a)

(b)

Figure 6 : (a) Marques de Lecture.
(b) Marques d'Ecriture.

Le nombre exact des marques de lecture en suivant cette reproduction est égale à 9 ($Na^{'}$=9) maques au total. Alors nous devons éliminer deux marques soit, $Na^{'}$-Ns pour pouvoir réaliser une correspondance exacte entre marques d'écriture / lecture (Figure.7).

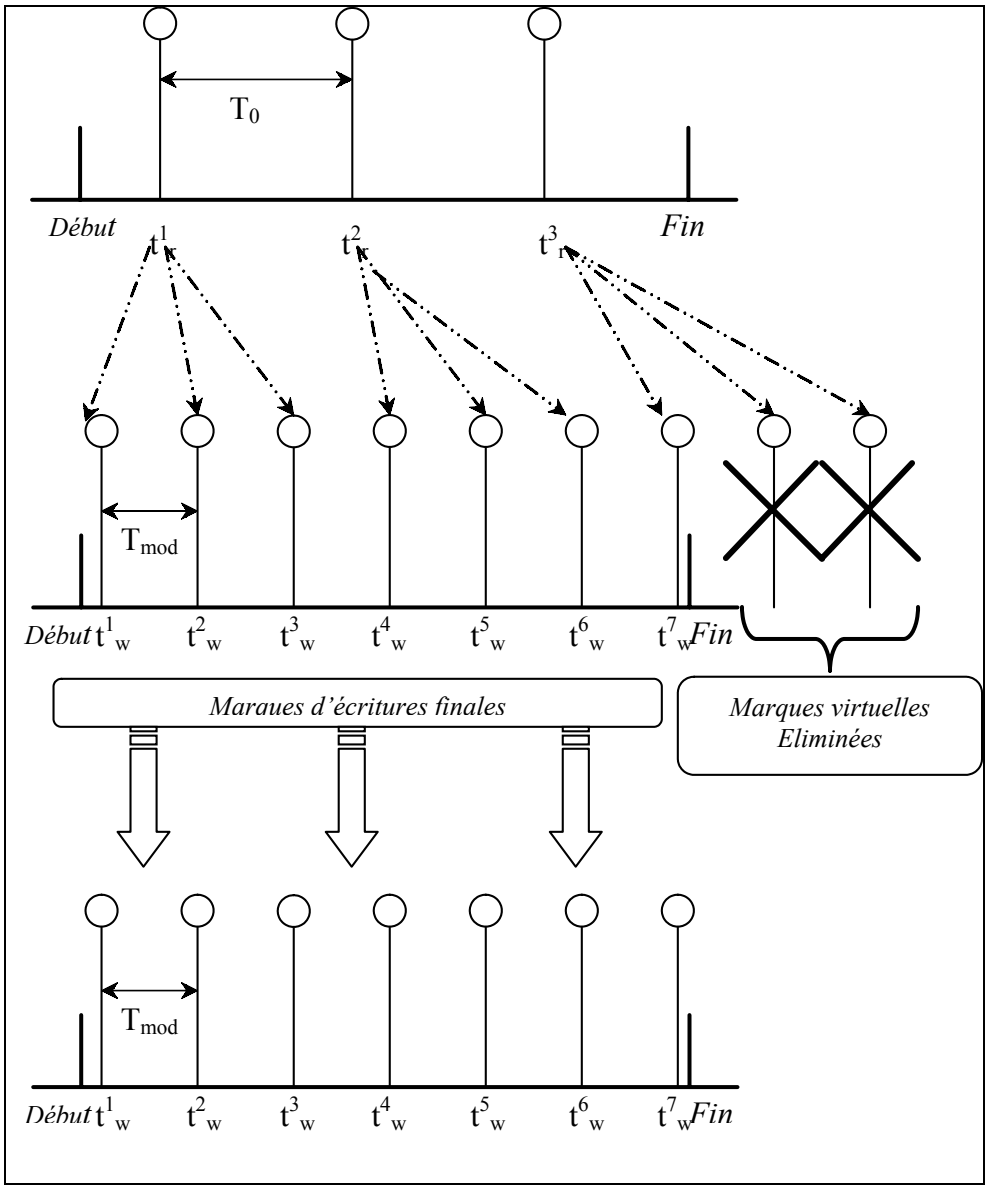


Figure 7 : Procédure de correspondance des marques Lecture/Ecriture

REFERENCES

1. R. Boite, H. Bourlard, T. Dutoit, “Traitement de la parole”. Collection électronique, Presses Polytechniques et Université Romandes, (1999).

2. G. Blanchet, M. Charbit, “Traitement numérique du signal”. Technique de l’Ingénieur, E 3 087, (2002), 8 – 11.

3. T. Dutoit, “Introduction au traitement numérique du signal”. Notes de cours, Première Edition, Faculté Polytechnique de Mons, TCTS Lab., (2000). Site Web: http://www.info.fundp.ac.be/~gde/dec.html.

4. T. Dutoit, “Je parle donc je suis”. Un bilan des développements récents en Traitement Automatique de la Parole ». Faculté Polytechniques de Mons, TCTS Lab., (2000). Site web: http://www.info.fundp.ac.be/~gde/dec.html.

5. A. Galvan-Rodriguez, “Amélioration d’un modèle articulatoire, normalisation du locuteur et récupération du lieu de constriction des plosives”. Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, Paris, France, (1997).

6. 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).

7. G. Droua-Hamdani, “Prédiction de la durée segmentale des phonèmes de l’Arabe Standard”. Mémoire de Magister, CRSTDLA, Alger, Algérie, (Février 2004).

8. M. Aissiou, “Application des Algorithmes génétiques en vue de la Reconnaissance Automatique des voyelles de l’Arabe Standard”. Mémoire de Magister, CRSTDLA, Alger, Algérie, (Février 2004).

9. P. Yves Le Meur, “Synthèse de la parole par unités de taille variable”, Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications, Paris, France, (1996).

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 texte”. Technique de l’Ingénieur, H 1 960, (2002), 5 – 10.

12. A. Benabbou, N. Chenfour, “Système de synthèse de la Parole par Règles pour la Langue Arabe”. 6^{éme} Conférence Magrébine des Sciences Informatiques, MCSEAT, Maroc, (2000), 397 - 404.

13. J. Mariani, “Analyse Synthèse et Codage de la parole”. Edition Hermes Lavoisier, (2002).

14. M. Guerti, “Contribution à la synthèse de la parole en Arabe Standard par diphones et technique de prédiction linéaire”. Thèse de Magister, ILP Alger, Algérie, (1984).

15. J. Farina, “La prosodie pour l’identification des langues”. Cours Doctorale en Informatique, Université Sabatier & Inpt par Pr R.Caubet, France, (1998).

16. G. Bailly, “Evaluation des systèmes d’analyse-modification-synthése de la parole” Actes des XXIII^{éme} JEP, Journées d’Etudes sur la Parole, (2000), 109 - 112.

17. P. Bastien, “Pitch shifting and Voice Transformation Techniques”. Technical report From Vocal Technologies TC ‘’Course Helicon’’. Acoustical Society of America, (2003).

18. P. Collein, “Techniques d’enrichissement de spectre des signaux audiométriques” Thèse de Doctorat en sciences. Ecole Nationale Supérieure des Télécommunications, Paris, France, (2002).

19. J. Travassos Romano, “Localisation de fréquences bruitées par filtrage adaptatif et implémentation d’algorithmes des moindres carrés rapides”. Thèse de Doctorat, Université de Paris sud, France, (1987).

20. A. Ouahabi, “Techniques avancées de traitement de signal et application”. Collection Sciences de l’Ingénieur. Alger, (1993).

21. E. Moulines, F. Charpentier, “Pitch Synchronous Waveform Processing Techniques for Text To Speech Synthesis using Diphones”. Speech Communication, V. 9, n° 5, (1990), 453 – 467.

22. J. Fernández, “Algorithmes d’acquisition, compression et restitution de la parole à vitesse variable, étude et mise en place”. Thèse de Doctorat. École Nationale Supérieure de l'Électronique et de ses Applications, Cergy-Pontoise (Paris), France, (Avril 1995).

23. A. Sidi Moussa, D. Youcef, “Segmentation automatique de la parole et application à l’annulation d’écho acoustique”. PFE, Université de Blida, Algérie, (Septembre 2000).

24. N. Schnell, G. Peeters, “Synthesizing a choice in real time using pitch synchronous overlap add (PSOLA) ”. ICMC, International Computer Music Conference, Berlin, (2000).

25. V. Collotte, Y. Laprie, “Amélioration de la précision de la resynthèse avec TD-PSOLA”. Actes des XXIV^{émes} JEP, Journées d’Etude sur la Parole, Nancy, France, (2002), 377 – 380.

26. G. Peeters, "Analyse et synthèse des sons musicaux par la méthode PSOLA", Actes des XXI^{émes} JIM, Journée d'Informatique Musicale, Agelonde, France, (1998), 302 - 306.

27. A. Tassa, J Lénaud, "A New approach to the evolution of vocal effort by the PSOLA method". The Electronic European Student Journal of Language and Speech "WEB-SLS", (2000).

28. C. d'Alessandro, "Notes de Cours Traitement Automatique de la Parole, Ecole d'été Ondelettes". LIMSI-CNRS, Toulouse, France, (1993).