N° d'ordre: 03/2007-M/EL

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

جامعة هواري بومدين
للعلوم و التكنولوجيا
U S T H B

FACULTE D'ELECTRONIQUE ET D'INFORMATIQUE

MEMOIRE

Présenté pour l'obtention du diplôme de MAGISTER

EN : ELECTRONIQUE

Spécialité : ELECTRONIQUE DES SYSTEMES

Par **Abdelhakim OUICHER**

Thème

ÉTUDE DU MODELE HARMONIQUE PLUS BRUIT POUR L'ANALYSE / SYNTHESE DES SIGNAUX DE PAROLE

Soutenu le : 18 / 12 / 2007, devant le jury composé de :

M^me BELHADJ-AISSA A.	Professeur	USTHB	Président
M^r A. HOUACINE	Professeur	USTHB	Directeur de thèse
M^r D. BERKANI	Professeur	ENP	Examinateur
M^r B. BOUDRAA	Maître de Conférences	USTHB	Examinateur

Table des matières

CHAPITE II : *Modélisation du signal de parole*

CHAPITE III : *Analyse par HNM du signal de parole*

CHAPITE IV : *Synthèse par le modèle harmonique plus bruit*

REMERCIEMENTS

Je dois d'abord remercier mon directeur de thèse Mr Amrane HOUACINE pour m'avoir soutenu et encouragé tout au long de mon cursus.

Je remercie vivement Mme BELHADJ-AISSA A. pour avoir présider le jury de ma soutenance tant espérée, je remercier également Mr D. BERKANI et Mr B. BOUDRAA qui ont honoré par leurs présence le jury,

Je veux aussi remercier mes collègues et amis pour leurs supports et conseils.

Merci à mes parents pour m'avoir soutenu et encouragé à étudier, ma grand- mère, mes frères et sœurs, cousins, beaux- frères ainsi que tous les membres de ma grande famille.

Enfin merci à mon épouse pour son amour, sa patience et sa foi.

LISTE DES ABRÉVIATIONS

3-D:	Trois Dimensions
ADPCM :	Adaptive Differential Pulse Code Modulation
AMDF :	Average Magnitude Differential Function
AR :	Autoregressive
ASCII:	American Standard Code for Information Interchange
CODECS :	Coder-decoders
FFT:	Fast Fourier Transform
FP :	Fonction de Périodicité
GSM :	Global System for Mobile
HNM :	Harmonic plus Noise Model
IP :	Internet Protocol
LPC :	Linear Predictive Coding
OLA:	Overlap-Add
PSOLA :	Pitch-Synchronous Overlap-Add
RIF :	Réponse Impulsionnelle Finie
TFCT :	Transformée de Fourier à Court Terme
TTS :	Text-To-Speech

Liste des figures

Introduction générale

Introduction générale :

La parole est l'un des moyens les plus importants de communication des humains, ce moyen a connu une présence considérable dans l'environnement de communication entre l'homme et la machine, après une série de recherches sur les techniques de traitement de la parole relatives à ce type de communication (dialogue homme-machine). L'établissement d'une telle communication devient de plus en plus nécessaire vu l'accroissement des applications d'ordinateurs dans la vie quotidienne, et comme le champ de communication (homme-machine) devient de plus en plus indispensable dans les activités où les informations proviennent d'une base de données sonore ou visuelle, la nécessité de meilleure qualité est d'autant plus croissante.

Les domaines des techniques du traitement de la parole concernés par ce type de problème sont la reconnaissance vocale et la synthèse vocale. Le premier consiste en l'extraction du message d'information dans le signal de parole afin de commander les actions de la machine suite aux commandes parlées. Par contre, le deuxième domaine est le procédé qui doit permettre de générer un signal vocal pour transmettre l'information de la machine à la personne.

Notre travail se situe dans le champ de la synthèse de la parole, dont le but est de permettre à la machine de transmettre des informations orales à son utilisateur.

Les recherches entreprises ces dernières années ont permis une amélioration sensible de la qualité de la parole synthétique. De ce fait, on peut considérer que la synthèse de la parole a atteint un niveau suffisant pour pouvoir être intégrée dans de nombreux services. Cependant, son déploiement à grande échelle exige une diversification des voix. Le but serait donc de pouvoir proposer un catalogue de voix suffisamment riche pour personnaliser les services vocaux.

On peut alors imaginer un système de synthèse "idéal" permettant:

- Une voix synthétisée intelligible naturelle;
- L'habilité de créer des messages de larges champs de vocabulaire;
- Un équipement de synthèse dont la complexité est mesurée en terme de capacité de stockage et de calcul.

Il existe deux types de systèmes de synthèse: Les systèmes par simple concaténation (systèmes concaténant plusieurs enregistrements sonores) et les systèmes de synthèse de parole "complets" (systèmes TTS : text-to-speech systems).

La première classe possède une base de données de messages préenregistrés (vocabulaire limité), et peut être utilisée dans les systèmes à caractère informationnel pour certaines applications comme l'heure locale, le temps (météo) , etc. Ces systèmes ne sont pas complexes et produisent une bonne qualité de parole synthétique du fait qu'ils utilisent des mots et des phrases préenregistrées, par contre, ils présentent l'inconvénient du vocabulaire limité, et ne peuvent prononcer que la combinaison des mots contenus dans leurs bases de données. Par ailleurs, quand le message change souvent au cours du temps (vocabulaire illimité), nous aurons besoin de système de la seconde classe (systèmes de synthèse de parole complets) dont le potentiel est plus intéressant.

Dans le même temps, la création de nouvelles voix pour un système de synthèse de la parole est une opération coûteuse. En effet, les systèmes TTS actuels permettant de traduire un texte en un signal de parole sont basés sur la concaténation de segments de parole sélectionnés dans une grande base de parole, dite dictionnaire acoustique. La création d'une base acoustique se fait par un locuteur parlant suivant un style d'élocution bien défini.

Le texte synthétisé peut provenir de plusieurs sources différentes comme par exemple le clavier, la reconnaissance d'écriture ou l'Internet. Le synthétiseur n'a alors aucune connaissance préalable du texte qu'il devra synthétiser, mais il est capable de produire un message vocal correspondant au texte reçu. Malgré le fait que le synthétiseur TTS génère une voix moins naturelle que la méthode de la concaténation mot à mot (système de la première classe), il reste toujours plus intéressant puisqu'il apporte une solution réelle pour les systèmes regroupant une grande variété d'informations. L'amélioration de la qualité de la synthèse obtenue par ces systèmes nécessite le perfectionnement des trois points suivants:

- l'analyse linguistique
- les règles de la prosodie
- les modèles de la synthèse

Ce dernier point, auquel on s'intéresse dans notre travail, est celui où interviennent les techniques de traitement du signal.

Les systèmes de synthèse basés sur la concaténation d'unités acoustiques exigent des passages non brutaux d'une unité acoustique à l'autre; cela implique la modification de certaines caractéristiques de ces unités, telles que la durée et la hauteur (modifications prosodiques). Une autre propriété désirée des systèmes de synthèse consiste en leur capacité à changer d'identité de la voix synthétique. Dans le cas des systèmes de synthèse

par concaténation d'unités acoustiques, on ne peut changer l'identité de la voix synthétique que par le ré-enregistrement de la base de données par un nouveau locuteur. Cette solution est en pratique coûteuse et irréaliste. Une autre solution est d'avoir recours à la conversion de la voix qui consiste à modifier le signal de parole d'un locuteur de référence appelé aussi locuteur source, d'une façon telle qu'il semble, à l'écoute, être prononcé par le locuteur désiré.
Les techniques permettant de modifier le signal de parole trouvent d'autres applications dans de nombreux contextes tels que ceux de doublage sonore de films, de traduction automatique destinée à permettre les conversations téléphoniques entre locuteurs ne parlant pas la même langue (application de téléphonie interprétée) [YS98, ASK90] ou dans le codage de la parole à très bas débit [SNB96].

La puissance et la flexibilité d'un système de synthèse vocale permettant de synthétiser une voix de haute qualité et en même temps de changer l'identité du locuteur exigent la disponibilité de systèmes d'analyse et de synthèse de signaux de parole efficaces. En effet, elle implique des modifications fines des caractéristiques du signal de parole. Pour cela, il est nécessaire d'extraire du signal des paramètres permettant d'avoir accès aux caractéristiques spectrales et prosodiques. Inversement, il faut être capable de restituer un signal de parole de bonne qualité à partir des paramètres modifiés.

Ce travail met l'accent sur les modifications prosodiques de haute qualité de la parole. Notre objectif est de fournir, en utilisant les techniques de traitement du signal, un synthétiseur qui permettra, d'une façon flexible de synthétiser et modifier la parole tout en gardant la qualité de la voix synthétique, en adoptant un modèle paramétrique qui serait capable:

- D'extraire du signal de parole, les paramètres appropriés pour notre objectif (analyse du signal) ;
- Utiliser ces paramètres pour faire une reproduction du signal de haute qualité (synthèse sans modification) ;
- Modifier les paramètres prosodiques (durée et hauteur) pour la modification du signal de parole.

La suite du document est organisée comme suit. Dans le premier chapitre, nous introduisons les propriétés fondamentales du signal de parole. Le deuxième chapitre

introduit le modèle de parole adopté, et donne une description la démarche suivie dans la thèse.

Dans le troisième chapitre, nous étudions la phase d'analyse du signal de parole par le "HNM" (Harmonic plus Noise Model) par l'estimation de chacun des paramètres régissant le modèle.

Dans le quatrième chapitre, nous présentons les deux types de modification qu'on est appelés à faire à savoir la transformation de la fréquence fondamentale et la modification de la durée du signal, ainsi que les deux modifications des deux paramètres simultanément. Ce chapitre est entre autres une mise en œuvre de la transformation du signal de parole par HNM.

Pour conclure ce document, nous présentons une synthèse des résultats du travail effectués, et proposons quelques perspectives de recherche.

Chapitre I

Notions générales et définitions

1. Généralités sur le signal vocal :

1.1. Production de la parole naturelle :

Le signal de parole est produit par l'excitation du conduit vocal (cavités orale et nasale) par des impulsions glottiques et les ondes de pression créées par modulation du flux d'air en provenance des poumons.

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

1.2. Classification des sons :

Suivant les organes de l'appareil phonatoire mis en jeu et leur excitation, on peut classer les sons produits en différentes classes. On distingue deux classes importantes de sons : les sons voisés et les sons non-voisés.

1.2.1. Les sons voisés :

Les sons voisés sont produits par l'excitation du conduit vocal par des impulsions périodiques dues aux vibrations des cordes vocales. Ce sont des sons ayant une forme quasi-périodique dont la représentation temporelle est illustrée sur la figure 1.1.a. Cette catégorie comprend, en plus des voyelles, d'autres groupes de phonèmes comme par exemple les consonnes nasales, les fricatives et les plosives.

Le spectre d'un signal voisé (figure 1.1.b) présente la particularité d'avoir une fréquence appelée « fondamentale » ou pitch. Les autres correspondent aux harmoniques du pitch, dont l'enveloppe de ces raies présente des maxima appelés formants.

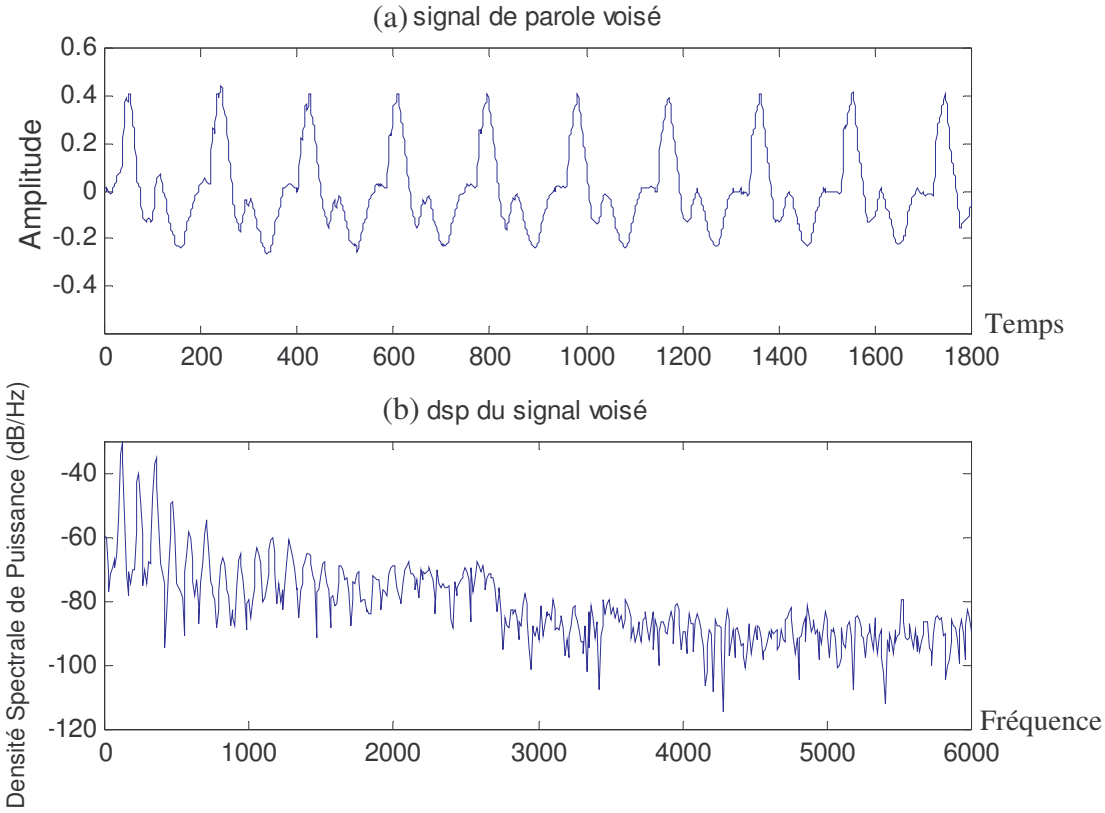

Figure 1.1: a)Signal de parole voisé, b) Spectre d'amplitude

Une des principales caractéristiques des sons voisés est de contenir la puissance des harmoniques dans une enveloppe de puissance caractéristique. Cette enveloppe de puissance est reliée au filtre créé par la forme du conduit vocal lors de la vibration des cordes vocales qui produisent les harmoniques.

1.2.2. Les sons non voisés :

Les sons non-voisés sont des signaux transitoires (comme les plosive) ou des signaux qui n'ont pas de composante périodique. Ils sont produits dans le premier cas par les lèvres et dans le deuxième cas par les turbulences de l'air passant par le conduit vocal. Dans ce dernier cas, la position des parties du conduit vocal modifie le spectre du signal non périodique.

Ces sons ne présentent pas de structures périodiques (figure 1.2.a) et peuvent être considérés comme un bruit blanc filtré par la transmittance de l'appareil phonatoire. On remarque aussi que leur spectre ne présente pas de structure de pitch (Figure 1.2.b)

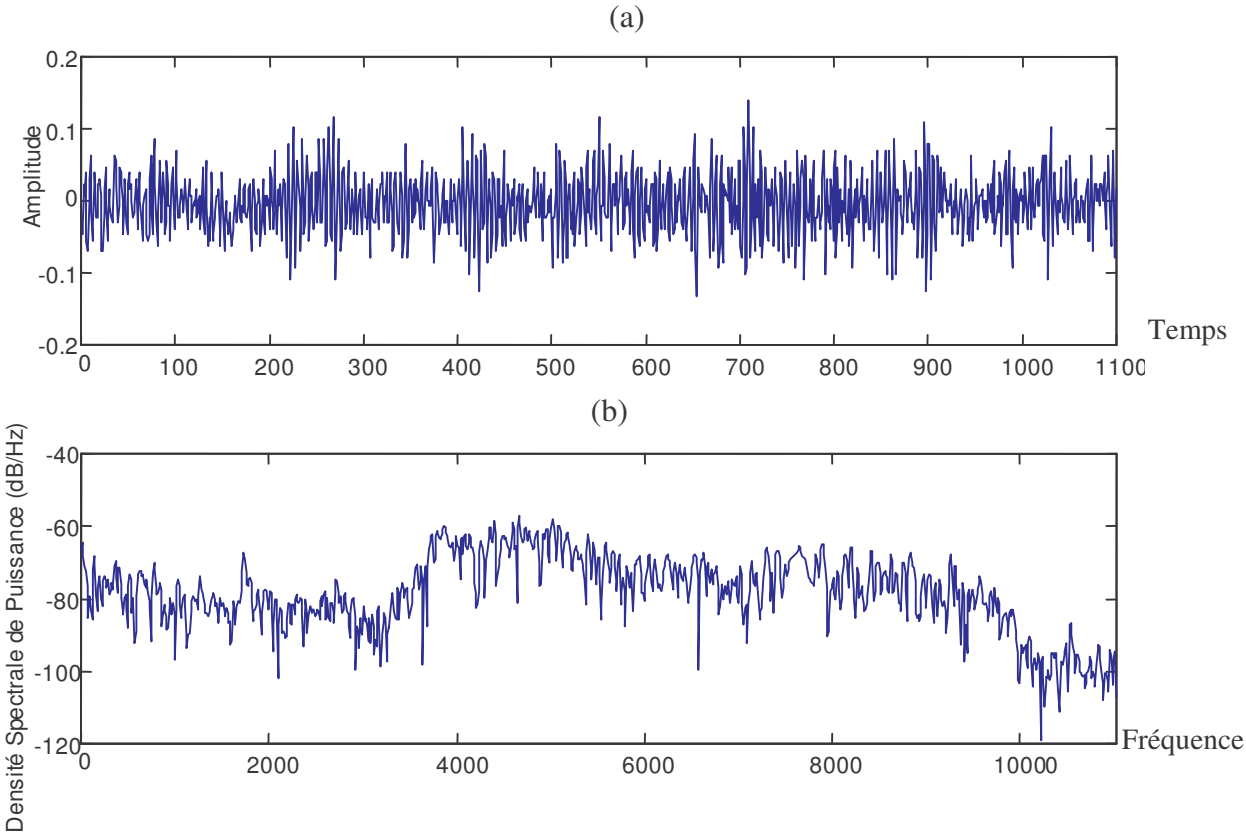

Figure 1.2: a)Signal de parole non voisé, b) Spectre d'amplitude

1.3. Propriétés spectrales du signal de parole

1.3.1. Analyse à court terme :

La représentation à court terme dans le domaine temporel du signal consiste à extraire une représentation trame par trame du signal sans effectuer de transformations sur ces trames. Dans certains cas, le signal est supposé stationnaire sur la trame donnée. La taille de la trame est un paramètre important. Elle est choisie de sorte à contenir au moins deux périodes du signal [DOV94]. Généralement, pour les signaux de parole continue, il est raisonnable de considérer que les voix d'hommes descendent jusqu'à 50 Hz (période égale à 20 ms), et les voix de femmes jusqu'à 100 Hz (période égale à 10 ms) ce qui donne des tailles de fenêtre d'environ 40 ms pour les hommes et de 20 ms pour les femmes.

La description du signal par la donnée de sa transformée de Fourier est une représentation fréquentielle du signal. Une représentation temps-fréquence d'un signal est une description qui est localisée à la fois en temps et en fréquence. Parmi les représentations les plus utilisées, la séquence des transformées de Fourier calculée sur le produit du signal par une fenêtre glissante

est une représentation temps-fréquence du signal. Cette représentation est appelée Transformée de Fourier à Court Terme (TFCT). Une propriété intéressante d'une telle représentation est qu'elle permet de reproduire l'information contenue dans le signal exactement. La TFCT ne vérifie cette propriété que sous certaines conditions telles que le choix du taux de recouvrement entre trames (espacement entre trames successives).

1.3.2. Spectrogramme :

Le spectrogramme est une représentation 3-D (amplitude/fréquence/temps) de l'évolution temporelle du spectre à court terme d'un signal, l'amplitude du spectre y apparaît en niveaux de gris dans une représentation temps fréquence. Il permet de voir la puissance d'un spectre de puissance varier dans le temps.

Le spectrogramme est dit à large bande ou à bande étroite selon la durée de la fenêtre de pondération (figure 1.3).

- Bande large : obtenu avec des fenêtres de pondération de faible durée (typiquement 10 ms); et mettent en évidence l'enveloppe spectrale du signal et donc la possibilité de visualiser l'évolution des formants; les périodes voisées y apparaissent sous forme de bandes verticales sombres.

- Bande étroite : ils sont moins utilisés et mettent en évidence la structure fine du spectre. Ainsi les harmoniques du signal dans les zones voisées apparaissent sous la forme de bandes horizontales.

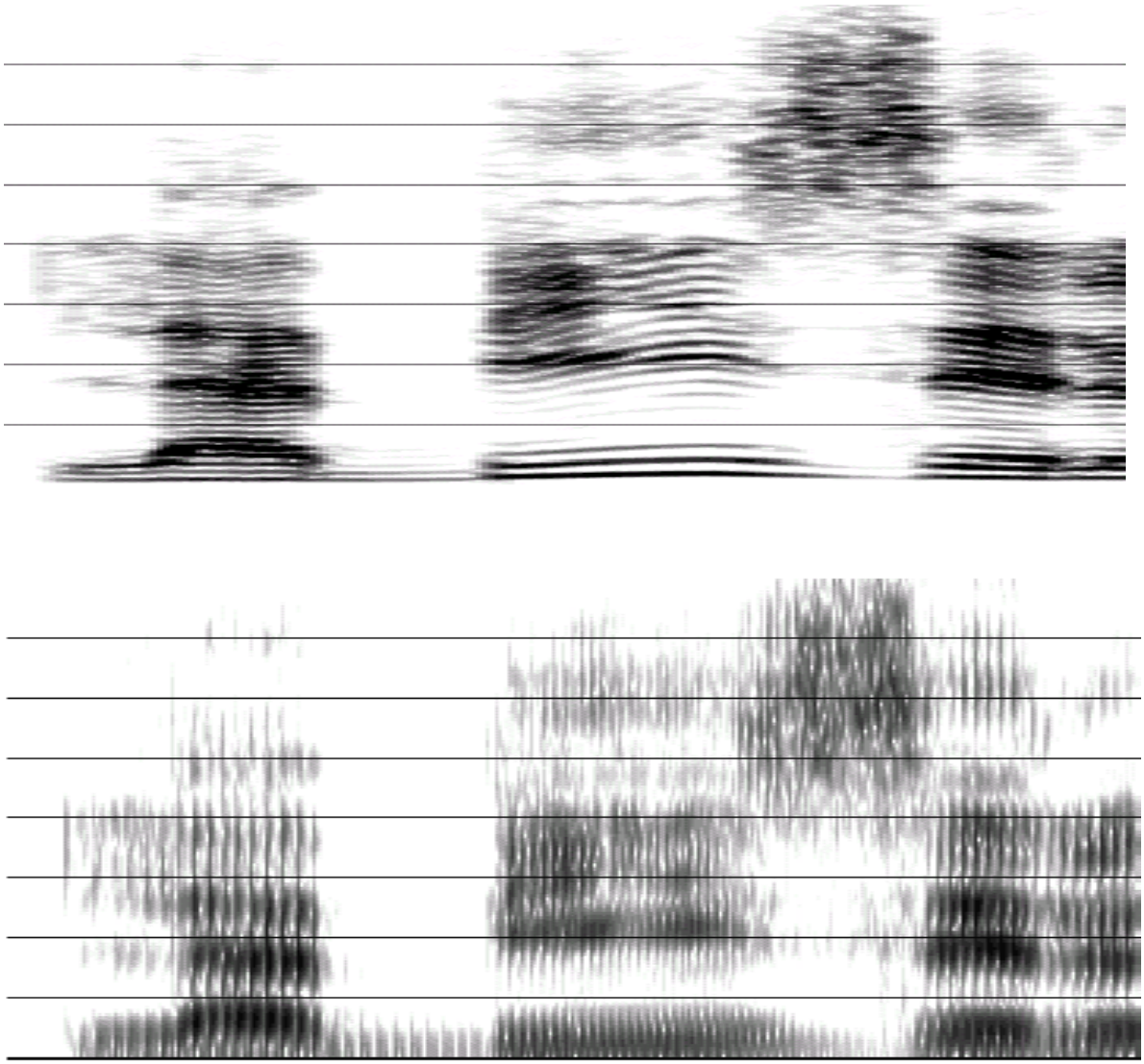

Figure 1.3: Spectrogrammes à bande étroite (en haut) et à large bande (en bas), (calculés avec des fenêtres de Hamming de 30 et 10 ms respectivement). [BBD00]

1.3.3. Enveloppe spectrale :

L'enveloppe spectrale est la réponse en amplitude du filtre modélisant le conduit vocal. Elle véhicule l'information formantique du signal.

L'enveloppe spectrale d'un signal dans laquelle on observe les formants forme le timbre d'un son (figure 1.4). La particularité d'un timbre est due à des propriétés physiologiques, principalement aux formes des cavités de la bouche.

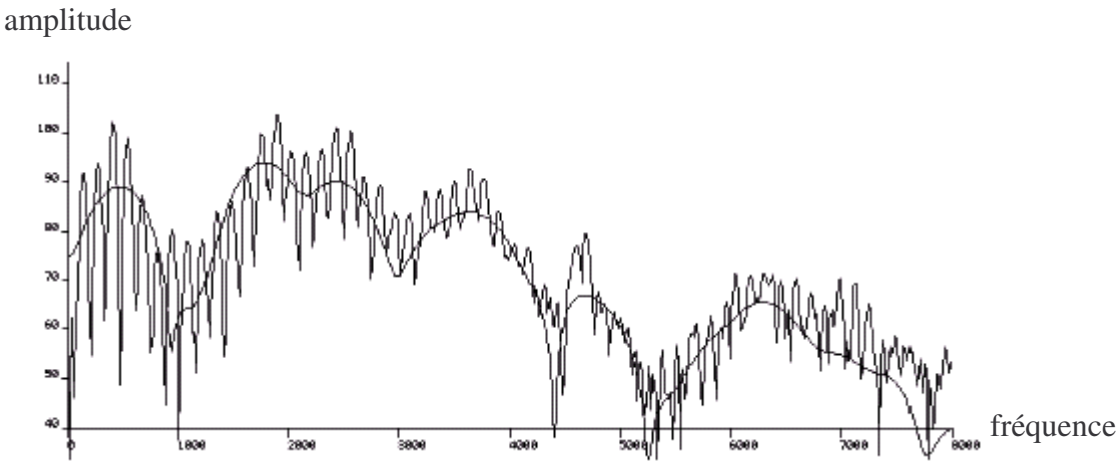

Figure 1.4: Représentation de l'enveloppe spectrale

1.4. Les paramètres prosodiques :

La prosodie recouvre les phénomènes de variations dans l'actualisation des phonèmes. Du point de vue acoustique, la prosodie se définit au moyen des paramètres de la fréquence du fondamental (estimation du son laryngien à un instant donné sur le signal), de la durée (intervalle de temps entre deux points sur le signal) et de l'intensité (énergie contenue dans le signal).

1.4.1. La fréquence fondamentale (hauteur) : l'estimation de la fréquence laryngienne à partir du signal acoustique à un instant donné (la fréquence de vibration des cordes vocales).

1.4.2. La durée : Cette notion comprend le débit de parole, la durée et la répartition des pauses, les allongements syllabiques. Pour beaucoup d'auteurs, le paramètre de la durée est difficile à calculer car il ne dépend d'aucun corrélat biologique, contrairement à la fréquence fondamentale et à l'intensité (qui dépendent respectivement de la tension des cordes vocales et de la pression sous-glottique). Pour calculer la durée d'un phonème, il faudrait se fixer deux événements qui délimitent ses repères initial et final. Sur un signal de parole, cette tâche incombe au processus de segmentation qui, pour beaucoup de systèmes actuels, est basé sur le phonème.

L'intensité : c'est l'énergie contenue dans le signal de parole durant un intervalle de temps donné (énergie à court terme). Elle résulte de la pression sous-glottique et

mesurée à la sortie de l'ensemble lèvres-narines et vibration des parois sur des portions de signal allant de 5 à 10 ms.

2. Techniques de traitement de la parole :

Le domaine du traitement de la parole peut se diviser en trois catégories: la compression, la reconnaissance de caractéristiques et l'analyse-synthèse.

La compression de la voix est un sujet de recherche pour lequel il y a une grande littérature. D'abord étudié pour la transmission de la voix sur les premiers systèmes de télécommunication numériques, le sujet s'est enrichi par les besoins issus de la généralisation du téléphone cellulaire et la perspective du transport de voix sur des réseaux utilisant le protocole Internet (IP). L'objectif de la compression est d'encoder la voix avec le moins d'informations possibles, tout en la gardant intelligible. Une bonne compression permet d'utiliser plus efficacement les liens de télécommunication.

La reconnaissance de caractéristiques a aussi été motivée en partie par la téléphonie. La reconnaissance vocale est utilisée dans divers systèmes de réponse automatique (comme l'annuaire téléphonique ou la confirmation de vols d'avion). De même que dans les logiciels de dictée automatique. Elle a pour objectif d'identifier les phrases et les mots prononcés par un locuteur.

L'analyse-synthèse est l'analyse du signal vocal dans le but de faire des modifications, de la segmentation ou de la concaténation pour la synthèse de voix. Cette catégorie inclue la synthèse de la parole en général (à partir d'un texte ou non). Les voix de synthèse sont créées en représentant une voix réelle par un modèle qui se compare parfois à ceux utilisés pour la compression. Cependant, le modèle pour la synthèse ne vise pas nécessairement à réduire l'information nécessaire à restituer la voix, il veut surtout pouvoir la restituer dans plusieurs contextes en y appliquant des modifications.

3. La synthèse vocale :

3.1. Définition :

La synthèse vocale est une technologie qui permet d'automatiser la production d'une parole artificielle par une machine. C'est l'implémentation d'un processus qui permet de transformer un message symbolique ou un ensemble de paramètres de commandes, en message acoustique.

3.2. Les systèmes de synthèse de la parole à partir du texte (TTS) :

Tout système TTS nécessite une entrée spécifiant la parole à synthétiser, il s'agit d'un texte écrit en format ASCII. Le processus qui effectue la conversion de cette entrée en parole est appelé synthèse à partir du texte. Une caractéristique importante de ces systèmes, est la possibilité de générer une parole arbitraire à partir de phrases qui ne soient pas connues à

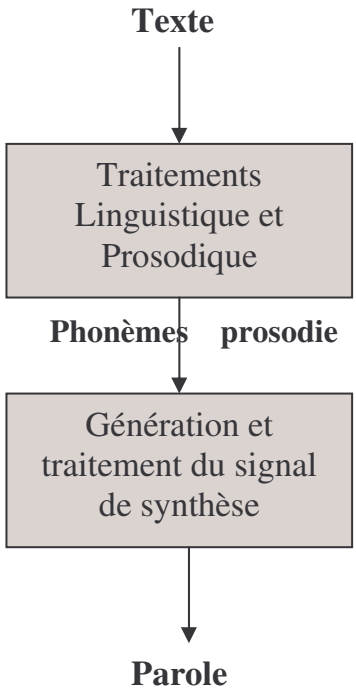

Figure 1.5: Schéma général d'un système de synthèse à partir du texte.

l'avance comme dans les dispositifs de lecture de messages, ou même dans l'enseignement de la langue. Un système de synthèse vocale à partir du texte peut se décomposer en deux parties :

3.2.1. Traitement du langage naturel :

La première étape concerne la transformation du texte en une chaîne de phonèmes avec des marqueurs prosodiques et fait appel à des concepts, des théories et donc des modélisations (syntaxiques, phonologiques…) dépendant de la langue dans laquelle le texte et le signal acoustique sont entrés et émis.

3.2.2. Génération du signal de synthèse :

Cette étape est chargée de transformer cette chaîne de phonèmes et avec des marqueurs prosodiques, en données acoustiques à transmettre au synthétiseur. Cette partie utilise des connaissances indépendantes du langage, elle dépend plutôt de la technique de synthèse choisie. La figure 1.5 illustre l'organisation globale d'un tel système.

Classiquement, les deux modules s'enchaînent linéairement, la sortie du premier module correspondant à l'entrée du second. Un texte est donné en entrée du premier module afin d'être analysé, sur la base de connaissances syntaxiques et prosodiques attachées à la langue dans laquelle il est écrit. A l'issu de ce traitement symbolique, une chaîne contenant des phonèmes et des marqueurs prosodiques est produite en fonction des résultats des différentes analyses effectuées. Cette chaîne est elle-même analysée et mise en relation avec une modélisation acoustique de la parole.

3.3. Méthodes de génération du signal de synthèse :

Il existe deux méthodes principales qui ont servi au développement des systèmes de synthèse de la parole:

3.3.1. *La synthèse par règles:*

La synthèse par règles prend pour hypothèse de décrire de manière explicite l'évolution en fonction du temps d'un ensemble de paramètres représentant le signal de parole. Il s'agit d'un ensemble de règles construit à partir d'une connaissance explicite des phénomènes de production de la parole, du moins à un niveau acoustique.

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. 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 [POL 90].

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 règles contrôle alors quelques dizaines de paramètres situés au niveau d'un modèle acoustique [STE 90].

Si des règles indispensables sont assez complexes à élaborer, l'ensemble des unités nécessaires est quant à lui relativement restreint. Le jeu de règles peut être plus ou moins important selon le niveau de modélisation souhaité.

3.3.2. *La synthèse par concaténation d'unités acoustiques:*

La synthèse par concaténation d'unités acoustiques rejette l'hypothèse de connaissance explicite des phénomènes de production et propose un principe de boîtes noires. Ce qu'il n'est pas possible de modéliser explicitement à partir du signal de parole, on l'encapsule dans une unité qui assure le lien entre le niveau linguistique et acoustique. Cette unité ne tire son existence que de caractéristiques externes (sa description phonétique par exemple).

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. Si l'on prend l'exemple du français, une quarantaine de phones ou modèles de phones seraient suffisants; chacun d'eux représentant une durée moyenne d'une centaine de millisecondes.

Elle consiste à synthétiser le signal par concaténation d'unités acoustiques, c'est à dire de segments de parole pré-enregistrés. Cette technique, qui repose sur l'utilisation de segments de signaux extraits de la parole naturelle permet de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain.

Ces synthétiseurs ont généralement une connaissance très limitée du signal qu'ils mettent en forme. Ces connaissances sont stockées dans les unités de parole mises en œuvre par le synthétiseur qui procède en mettant bout à bout les segments acoustiques

déjà co-articulés extraits d'une base de données d'un signal de parole. La principale difficulté dans cette technique reste de calculer un signal final dans lequel les transitions entre unités ne sont pas perceptibles.

Les algorithmes de synthèse de la parole construisant un modèle de la voix a partir de vrais segments d'une voix utilisent un modèle intermédiaire qui permet de re-synthétiser la voix avec le rythme et la hauteur (fréquence) désirés. Une fois tous les segments modélisés, il est possible de les synthétiser et de les concaténer pour créer l'expression désirée. Le domaine des télécommunications appelle ces modèles codecs (coder-decoders) lorsqu'ils sont utilisés pour la compression de voix. Ils ne sont pas alors construits pour une seule voix comme dans le cas de la synthèse, et sont souvent organisées en dictionnaires (codebooks), ce qui permet d'envoyer des codes minimaux durant la transmission pour une reconstitution intelligible.

3.4. Les 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.

Il est tout d'abord important de rappeler qu'un système de synthèse de la parole diffuse une information via le canal acoustique. Une caractéristique importante de ce canal est qu'il permet difficilement le multiplexage de voies de communication. Dans une conversation, le canal acoustique est le plus souvent utilisé de manière exclusive par un émetteur. Un système de synthèse de la parole sera un bon choix technologique si l'application diffusant des messages est pratiquement sûre de disposer pour elle seule de l'espace acoustique qui environne le système de diffusion.

Un prototype idéal d'un service à base de synthèse de la parole est un service devant diffuser une quantité d'information importante et difficilement prévisible dans une ergonomie où le canal acoustique reste le plus pratique.

Dans le secteur des télécommunications, différents opérateurs proposent deux services phares à base de synthèse de la parole: la lecture des courriers électroniques par téléphone (notamment les téléphones mobiles, pour lesquels la surface de lecture est limitée), la lecture vocale: annuaire inverse où l'utilisateur obtient des informations sur une personne à partir de son numéro de téléphone, consultation des télécopies par voie orale, lecture vocale de journaux, consultation de la messagerie écrite, etc.

Dans le cadre des projets visant le dialogue homme-machine, des prototypes de système de dialogue existent en laboratoire où la synthèse vocale est couplée à un système de reconnaissance et exploite ainsi différentes ressources (informations syntaxiques, sémantiques...) utiles à la génération acoustique du signal et permet par exemple de proposer des outils d'aide à la navigation pour les véhicules ou encore des systèmes plus spécifiques comme la lecture de compte-rendu de dysfonctionnement dans le cadre du contrôle de systèmes. Enfin, il faut citer un des cadres applicatifs les plus anciens qui est l'aide à la lecture aux personnes handicapées: le système de TTS joue le rôle de lecteur de textes sur écran ou à distance via un serveur vocal, avec la possibilité pour l'utilisateur de configurer le système, en choisissant la voix de synthèse (masculine ou féminine), le débit de la parole et même la langue souhaitée si le document existe en plusieurs langues.

Chapitre II

Modélisation du signal de parole

Introduction:

En général, le modèle mathématique utilisé dépend de l'application visée. Dans le domaine de la synthèse de la parole, le modèle de parole doit satisfaire au besoin de modifications prosodiques et spectrales complexes et être en mesure de produire une grande variété de voix intelligibles, aussi bien que naturelles et précises en ce qui concerne l'identité du locuteur.

Que l'on adopte une méthode de synthèse fondée sur un pouvoir explicatif (approche par règles) ou seulement descriptif (approche par 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 vocal et à définir des paramètres mieux appropriés aux traitements de la parole.

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 sa forme temporelle (c'est-à-dire sans modèle), ou bien à l'aide de codeurs, comme les méthodes de prédiction linéaire.

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

- Les modèles s'appuyant sur une hypothèse de production; ces modèles décrivent la fabrication d'un signal de parole à partir de paramètres physiologiques. Il s'agit de reproduire le comportement de l'appareil phonatoire humain;
- Les modèles placés sous une hypothèse de perception; ces modèles décrivent comment fabriquer un signal de parole qui pourra être perçu comme un signal de parole humaine peut l'être.

L'hypothèse de perception est celle qui est la plus répandue dans la réalisation concrète des systèmes de synthèse de la parole car elle est propice à une formalisation simple. Cependant, l'hypothèse de production reste certainement celle qui apporte un plus grand fondement à l'explication des mécanismes de production de la parole.

De nombreux modèles décrivant l'évolution temporelle du signal de parole ont été proposés. 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 de la perception.

1. Modélisation du signal de parole

Parmi les modèles s'appuyant sur une hypothèse de production nous pouvons citer l'hypothèse de décomposition source-filtre qui considère que le signal de parole résulte de la combinaison d'une énergie sonore (interaction des poumons et du larynx) couplée à une fonction de transfert déterminée par la forme des cavités supraglottiques Fant [FAN60]. Le modèle articulatoire pour sa part [MAE 79] [RIC95], permet de produire un signal de parole à partir d'une description physiologique (détermination de la configuration du conduit vocal) Dans l'approche par formants [BBD00], le procédé de synthèse se trouve lié à une hypothèse d'interprétation à la fois phonétique et acoustique. L'analyse d'un signal acoustique naturel consiste, à partir de son spectre, à caractériser ses résonances par leurs positions fréquentielles et leurs bandes passantes. La synthèse, inversement, part d'une séquence finie de représentations de formants, position, amplitude et bande passante, pour générer un signal de parole à partir d'un spectre artificiel. Ces techniques restent relativement complexes et ne donnent pas de bonne qualité du signal.

1.1. Modélisation autorégressive (AR) :

Fant introduisit en 1960 une modélisation linéaire de l'évolution temporelle du signal de parole [FAN60]. Il ajoute à l'hypothèse source-filtre une hypothèse d'indépendance de l'onde glottique et du conduit vocal. Le conduit vocal est modélisé par un filtre tout-pôle ou filtre autorégressif. L'onde de glotte, quant à elle, est modélisée de manière grossière 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.

Le codage par prédiction linéaire (Linear Predictive Coding ou simplement LPC) est une représentation qui permet de reconstruire le signal vocal en supposant un modèle source/filtre du conduit vocal. La source dans ce cas est le mouvement des cordes vocales, et le filtre est un filtre à réponse impulsionnelle finie (FIR) représentant le conduit vocal [RAS 78]. Les p paramètres a_i du filtre sont recalculés à chaque fois de sorte à respecter un critère de quasi stationnarité. La voix est synthétisée en filtrant une excitation, d'une période et puissance variables avec les filtres correspondant aux phonèmes cibles.

Un signal original est échantillonné, chaque échantillon est ensuite prédit à l'aide d'une combinaison linéaire des échantillons précédents. Un coefficient est calculé pour chaque échantillon afin de minimiser les erreurs entre l'original et la copie.

Cette technique est fondée sur l'hypothèse qu'un échantillon *s(nT)* est prédit approximativement à partir d'une combinaison linéaire des échantillons qui le précédent selon l'équation suivante :

$$s(n) = \sigma\, u(n) - \sum_{i=1}^{p} a_i s(n-i) \qquad (2.1)$$

où $u(n)$ est le signal à l'entrée du filtre et σ est le gain.

La modélisation autorégressive consiste à estimer le modèle décrivant le conduit, en connaissant le signal d'excitation qui peut être un bruit blanc de moyenne nulle et de variance unité pour un son non voisé ou un signal impulsionnel d'amplitude unité dans le cas d'un son voisé [BBD00]. C'est un modèle mathématique dont la transmittance est un filtre polynomial tous pôles de la forme : *1/ A(z).*

Il est courant d'utiliser 10 à 13 coefficients pour le filtre. Cela permet d'avoir une approximation raisonnable des formants et donne une voix suffisamment intelligible pour être utilisée dans des applications commerciales.

Ce type de modélisation est largement répandu en traitement de la parole et notamment en codage. L'approximation grossière de l'onde de glotte fait que le signal de synthèse est de qualité médiocre. Cependant, le nombre de paramètres du modèle est relativement modeste et la modification des paramètres prosodiques en contexte de synthèse de la parole est relativement aisée. En effet, pour modifier la durée d'un phone, il suffit d'allonger ou de raccourcir le signal d'excitation et la modification du fondamental est opérée de manière triviale par une simple modification de la période élémentaire du train des impulsions.

En plus d'être utilisé couramment dans les systèmes de synthèse de voix, l'encodage LPC est utilisé dans plusieurs systèmes de compression de voix comme le standard GSM pour la téléphonie cellulaire.

Cette méthode a l'avantage de réduire la quantité d'informations nécessaire pour produire la parole.

Le schéma suivant illustre les différentes étapes de cette technique.

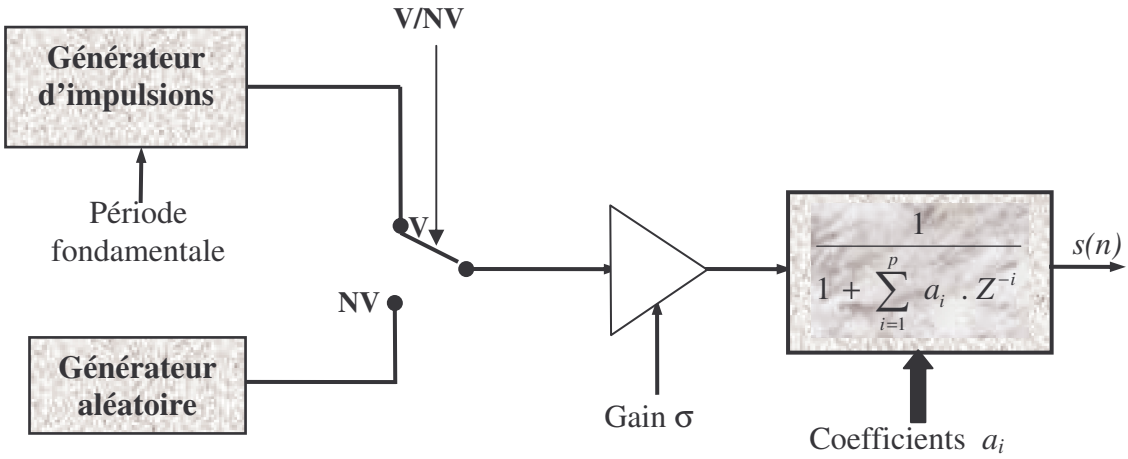

Figure 2.1: Modèle autorégressif de production de la parole

Cette modélisation a été largement utilisée jusqu'à la fin des années 1980 et qui a cédé la place à des techniques plus complexes et offrant une meilleure qualité du signal comme par exemple le modèle harmonique et bruit.

1.2. Modification des paramètres prosodiques par PSOLA :

PSOLA (Pitch Synchronous Overlap and Add) est 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.

Son principe consiste à découper le segment de parole en plusieurs trames courtes en utilisant une fenêtre de Hamming synchronisée sur le pitch dans la partie voisée, et à des intervalles fixes dans la partie non voisée du signal. Après le calcul des instants de synthèse, les trames sont recombinées pour produire le signal synthétisé. La valeur du pitch augmente ou diminue en faisant varier l'espacement entre les trames, tandis que la répétition ou la suppression des trames du signal provoque la modification des durées.

Le signal de parole étant décomposé, par fenêtrage synchrone, en signaux élémentaires qui se recouvrent, les modifications prosodiques sont alors effectuées sur le flux de ces signaux élémentaires, par modification du taux de recouvrement et par la suppression ou la répétition de certains d'entre eux.

Comparée à la technique LPC, la technique PSOLA nécessite un volume de mémoire plus important, mais produit une qualité sonore beaucoup plus naturelle.

1.3. Modèles sinusoïdaux:

Les modèles sinusoïdaux [KLP95] [QUA 01] permettent de synthétiser ou de compresser des voix avec une performance généralement supérieure aux algorithmes basés sur la prédiction linéaire. Cependant, la quantité de calcul nécessaire pour analyser et re-synthétiser la voix a retardé leur application dans les systèmes de téléphonie; l'encodage des modèles sinusoïdaux entraîne un délai d'entre 10 et 25 ms dans les communications. Ce délai gène la conversation, particulièrement dans le cas où la ligne téléphonique est sujette à un écho important (un écho dont le délai est de plus de 20 ms déplaît aux usagers des systèmes téléphoniques). Par ailleurs, l'encodage de la voix n'est pas la seule étape responsable des délais de transmission, c'est pourquoi les codeurs de style LPC ou ADPCM sont généralement préférés.

Le contexte qui nous intéresse, la synthèse de voix réaliste, est moins sensible au temps de calcul et au délai. Les modèles sinusoïdaux sont donc intéressants à étudier dans le cadre de notre recherche.

Les modèles sinusoïdaux se basent sur le principe de la transformée de Fourier qui dit que tout signal peut se représenter comme étant une somme de sinus. La modélisation de la voix peut donc être faite en trouvant un ensemble de fonctions sinusoïdales représentant l'essentiel de l'énergie du signal vocal. Il suffit ensuite de reproduire ces sinusoïdes principales pour re-synthétiser la voix. La principale caractéristique du modèle sinusoïdal que nous présentons est de contraindre les sinus à avoir une relation harmonique, c'est à dire que leur fréquence est multiple d'une fréquence fondamentale. Même si cela ne semble pas toujours être le cas des sons voisés [KLP 95].

Le modèle prend donc la forme mathématique suivante :

$$s(t) = \sum_k A_k \cos(\varphi_k(t)) \tag{2.2}$$

où $A_k(t)$ et $\varphi_k(t)$ sont respectivement l'amplitude et la phase de l'harmonique k au temps t et $s(t)$ est la valeur du signal harmonique au temps t.

D'autres modèles n'assument pas que les sinusoïdes ont une relation harmonique. Certains algorithmes choisissent les k sinusoïdes les plus puissantes de chaque fenêtre d'analyse de la FFT [QUA 01]. D'autres suivent l'apparition et la disparition de sinusoïdes en définissant certains critères d'existence (puissance soutenue, continuité de la fréquence de la sinusoïde [SER 97]. Il est aussi possible d'exprimer les puissances des sinusoïdes de fréquence f_k avec un filtre d'ordre 10 à 22 [KLP95]. Le modèle sinusoïdal harmonique est également la base du modèle harmonique plus bruit présenté dans l'étude suivante; son étude est donc une bonne introduction à ce dernier modèle.

1.3.1. Modélisation des sons voisés :

La modélisation des sons voisés se fait en trouvant la fréquence fondamentale et en calculant la puissance et la phase de chaque harmonique. L'analyse est faite généralement sur des fenêtres de signal contenant de deux à quatre périodes fondamentales pondérées par des fonctions qui atténuent les effets de bords.

1.3.2. Modélisation des sons non-voisés :

Les sons non-voisés (transitoires ou bruités) ne sont pas sinusoïdaux à la manière des sons voisés, mais peuvent tout de même être modélisés comme une somme de sinus ayant une relation harmonique. Comme dans le cas de la modélisation des sons voisés, nous chercherons une série de fonctions sinusoïdales expliquant la majeure partie de l'énergie du signal vocal.

Nous voulons également que la fréquence fondamentale utilisée soit inférieure à 150 Hz, pour éviter que les sons bruités aient la sonorité résonnante caractéristique des signaux périodiques. Utiliser une basse fréquence fondamentale dans le cas de sons bruités permet d'approcher l'approximation sinusoïdale de la transformée de Fourier. La transformée de Fourier a une résolution de f_e/N pour une fenêtre d'analyse de N points et un signal échantillonné à f_e Hz (donc environ 43 Hz pour une fenêtre de 1024 points à 44,1 kHz). Lorsque la fréquence fondamentale de l'approximation harmonique du bruit est proche de cette valeur (p.e. 43 Hz), le modèle s'approche de la représentation complète du signal de la transformée de Fourier.

1.3.3. Synthèse avec un modèle sinusoïdal :

Pour re-synthétiser le signal vocal, il suffit de faire la somme des k sinusoïdes de fréquence instantanée $kf_0(t)$ et de puissance instantanée $A_k(t)$. La phase $\varphi_k(t)$ du sinusoïde k est calculée en

additionnant une différence de phase $2\pi f_0(t)/f_e$ à la phase précédente $\varphi_k(t-1)$. Les fréquences et puissances instantanées peuvent être interpolées linéairement entre les centres des fenêtres d'analyse. Dans [MOC 90], on suggère d'utiliser la méthode de recouvrement et d'addition (overlap-and-add), qui consiste à synthétiser une fenêtre à la fois pour éviter une accumulation de l'erreur sur la phase due à l'analyse et l'interpolation. Les fenêtres se recouvrent aux extrémités et sont pondérées de manière à reproduire le signal original sans modification de puissance.

1.3.4. Evaluation du modèle sinusoïdal :

Le modèle sinusoïdal que nous venons de décrire comporte quelques lacunes dues à sa simplicité. Premièrement, le modèle néglige toujours une partie du signal (sauf dans le cas purement périodique et harmonique), ayant comme conséquence une perte de qualité inévitable du signal vocal. De plus, alors qu'il permet des modifications du signal comme le changement de hauteur ou la dilatation temporelle, il ne permet pas de jouer sur le ratio d'harmonicité de la voix, par exemple pour rendre une voix plus suave ou pour transformer une voix en murmure. Alors que ces limites sont sans conséquence grave pour les systèmes téléphoniques de basse qualité (p.e. les téléphones cellulaires ou portables), elles deviennent plus grave dans le cas de la synthèse de voix réaliste, où le locuteur doit non seulement être intelligible mais reconnaissable et naturel.

1.4. Modélisation harmonique plus bruit:

Le principe original du modèle harmonique plus bruit consiste à décrire formellement le signal de parole comme la superposition d'une composante harmonique et d'une composante de bruit

La modification des paramètres prosodiques de la partie harmonique est relativement aisée dans la mesure où la variable temporelle et la période de distribution des harmoniques sont explicites dans le modèle.

2. Description du Modèle Harmonique plus Bruit :

La figure 2.2 donne un exemple d'un signal de parole d'une voix masculine prononçant le mot "sa", échantillonné à 16khz.

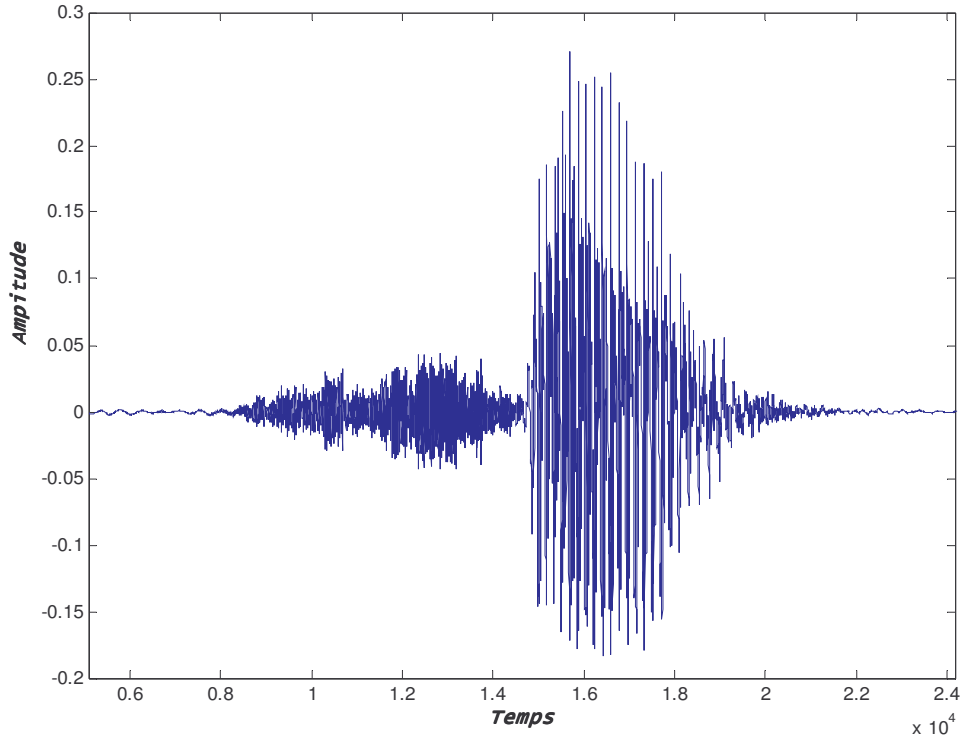

Figure 2.2: Représentation temporelle d'un signal de parole

En prenant une trame pour analyse dans la partie voisée du signal (figure 2.3.a), le caractère périodique de ce signal est facilement observé dans le domaine fréquentiel (figure 2.3.b). Le spectre de la trame est formé de pics (raies) régulièrement espacés, ces harmonique sont d'amplitudes différentes et séparées l'une de l'autre par un intervalle égale à la fréquence fondamentale. Cette périodicité ne couvre pas toute la gamme de fréquence (de 0Hz à la moitié de la fréquence d'échantillonnage : *fe/2*), elle disparaît au-delà d'une certaine fréquence appelée "fréquence maximale de voisement :*fm*". De ce fait, il nous est possible de décomposer ce spectre en deux bandes séparées par la fréquence maximale de voisement.

Les signaux correspondant à chacun des spectres des deux bandes (0- f_m et f_m- $f_e/2$) sont tracés dans les figures 2.3.c et 2.3.d respectivement.

Il est intéressant de constater la forme du signal correspondant à la bande haute (figure 2.3.d), le signal représenté par cette figure présente un bruit, ayant des pics synchronisés à la periode fondamentale. Le bruit des trames du signal voisé résulte du bruit de friction et de la variation

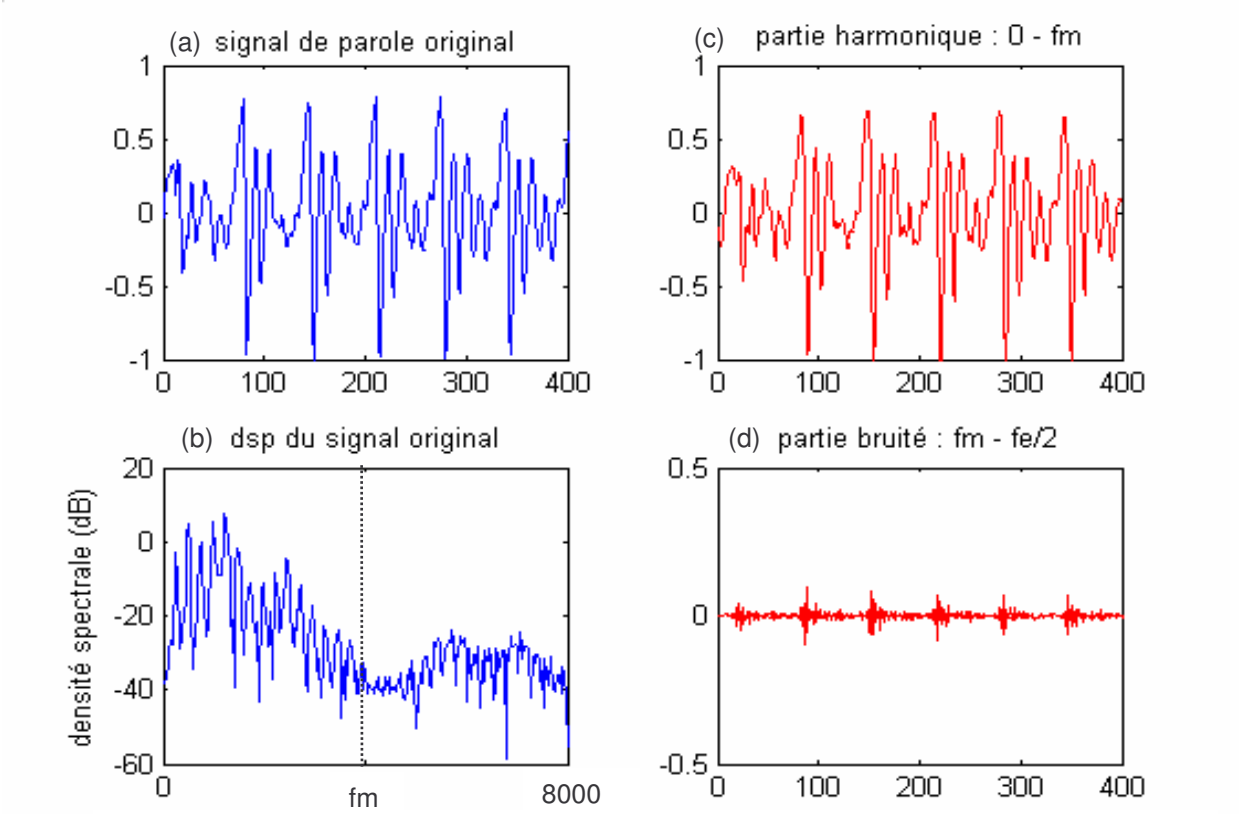

Figure 2.3: Décomposition Harmonique plus Bruit d'un signal de parole

de l'excitation glottale d'une période à l'autre. La fréquence (f_m) limitant les deux bandes est donc un paramètre variant dans le temps.

Pour certains sons fricatifs, la partie bruitée couvre une bande beaucoup plus importante que dans celle de notre exemple.

Afin d'obtenir une "bonne qualité" de synthèse de la parole, on doit considérer que le signal de parole pour les sons voisés est composé de deux composantes distinctes: une composante reflétant le caractère périodique du signal, et une autre composante englobant toutes les formes du bruit contenu dans ce signal. Les modèles de parole adoptant cette technique de décomposition sont appelés modèles hybrides ou modèles de sinusoïdes+bruit. Si les fréquences de ces sinusoïdes sont des multiples de la fréquence fondamentale le modèle est appelé "modèle harmonique plus bruit" [STY96].

Le Modèle Harmonique plus Bruit (HNM : Harmonic plus Noise Model), est un système d'analyse/modification/synthèse de la parole qui suppose que le signal de parole *s(n)*

peut être décomposé en une partie harmonique *h(n)* et en une partie bruité *b(n)*. La partie harmonique modélise la composante quasi-périodique du signal de parole; la partie bruitée modélise la composante non-périodique du signal, c'est-à-dire le bruit de friction et les variations de l'excitation glottale d'une période à l'autre. Le spectre est divisé seulement en deux bandes et uniquement pour les trames voisées. Les deux bandes se délimitent par la "fréquence maximale de voisement" qui est en général un paramètre variant dans le temps. Cette décomposition permet d'appliquer les modifications sur le signal de parole sans lui affecter sa qualité audible.

3. Representation du signal:

L'analyse du signal de la parole se présente dans un contexte particulier : le signal de parole n'est pas stationnaire, on va devoir travailler sur des trames de temps de durée limitée avec recouvrement (figure 2.4). Cette représentation à court terme du signal de parole est obtenue par le produit du signal par une fenêtre de pondération "glissante". En pratique les fenêtres, appelées fenêtres d'analyse, sont des fonctions positives à support borné, en général symétriques et choisies pour leurs propriétés spectrales. Le calcul des paramètres consiste alors à extraire du signal discret des segments d'échantillons de taille fixe, et régulièrement espacés, appelés trames.

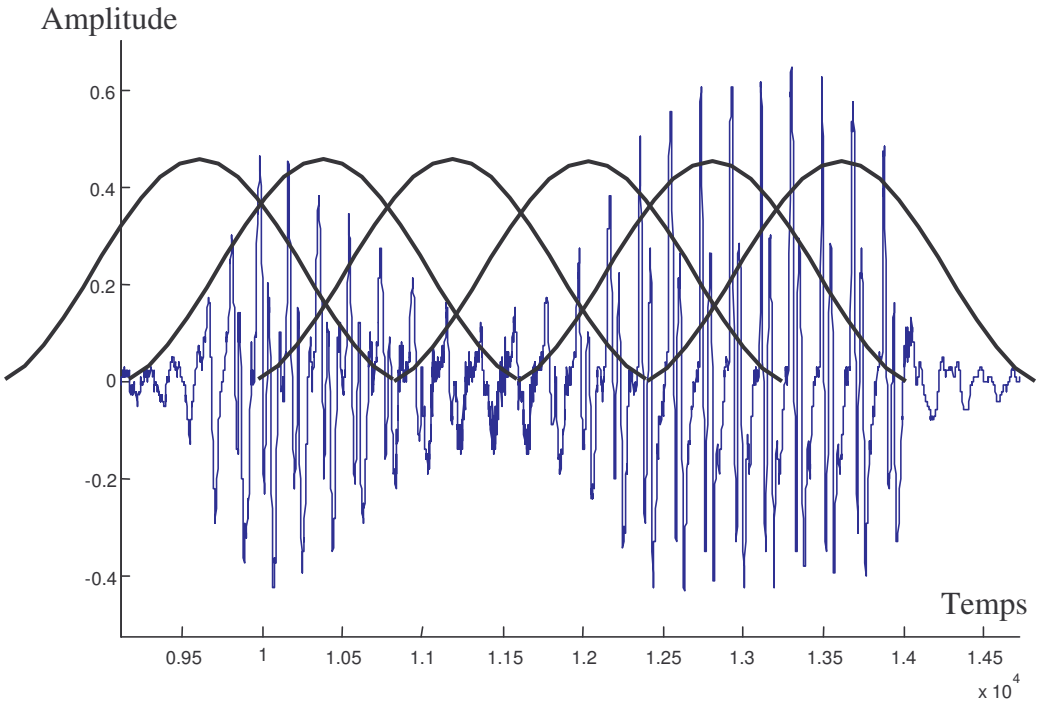

Figure 2.4 : Découpage du signal de parole en trames avec recouvrement et pondération

Une telle représentation du signal se justifie par le fait que la forme du conduit vocal n'évolue pas rapidement. En général, on considère que l'appareil vocal est quasi-stationnaire sur un intervalle de temps de l'ordre d'une vingtaine de millisecondes. On parle donc ici de statistique du signal à court terme.

La procédure pratique choisie consiste à calculer le modèle sur des trames successives de durée suffisante pour empiéter plusieurs période du fondamental pour les sons voisés, soit de durée de 25 ms décalées de 10 ms.

3.1. Fenêtres de pondération:

Le but du fenêtrage est de découper le signal de parole en petites tranches où il peut être considéré localement comme quasi- stationnaire.

En outre, et pour profiler de l'évolution lente du signal vocal, le fenêtrage permet le traitement en temps réel et facilite aussi l'analyse des signaux sur la machine.

Du fait que le signal de parole est non-stationnaire, il faut mettre en évidence les dépendances qui existent entre les échantillons, pour ce faire, deux tranches adjacentes doivent avoir une région commune (de 5 à 12 millisecondes environ [BBD00]).

Il existe plusieurs types de fenêtres d'analyse : Rectangulaire, Blackman, Kaiser, Hanning,… chacune ses caractéristiques temporelles et spectrales. Pour le cas de l'analyse par HNM du signal de parole, il est préférable de travailler avec la fenêtre de Hamming.

- **Fenêtre de hamming :**

C'est une fenêtre "en cloche " (Figure 2.5). Elle est définie par:

$$w_{hamm}(k) = \begin{cases} 0.54 + 0.46\cos\left(\dfrac{2\pi k}{M-1}\right) & si\ k \in \left[-\dfrac{M}{2}, \dfrac{M}{2}-1\right] \\ 0 & ailleurs \end{cases} \qquad (2.3)$$

avec M : longueur de la fenêtre.

Sa FFT est définie par:

$$w_{hamm}(f) = 0.54\frac{\sin(\pi.f.n)}{\sin(\pi.f)} + \frac{0.46}{2}\left(\frac{\sin N.\pi(f-1/N)}{\sin.\pi(f-1/N)} + \frac{\sin N.\pi(f+1/N)}{\sin.\pi(f+1/N)}\right) \qquad (2.4)$$

Cette fenêtre présente un lobe principal assez étroit, et des lobes secondaires plus atténués (perte de plus de 40 dB vis à vis du lobe principal).

Les figures suivantes représentent une trame de signal voisé (figure 2.5) et le résultat donné après pondération par la fenêtre de Hamming (figure 2.6).

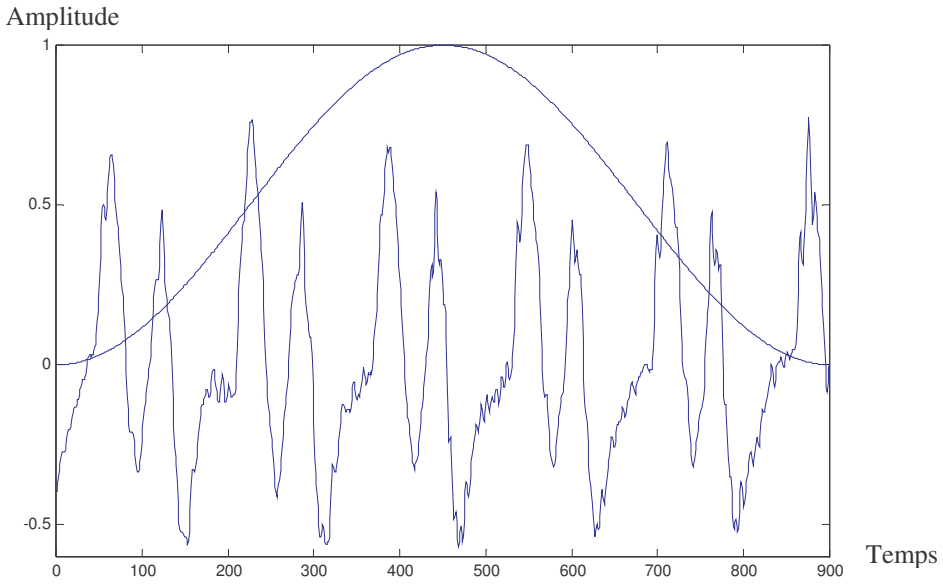

Figure 2.5 : Une trame d'analyse pondérée par la fenêtre de Hamming

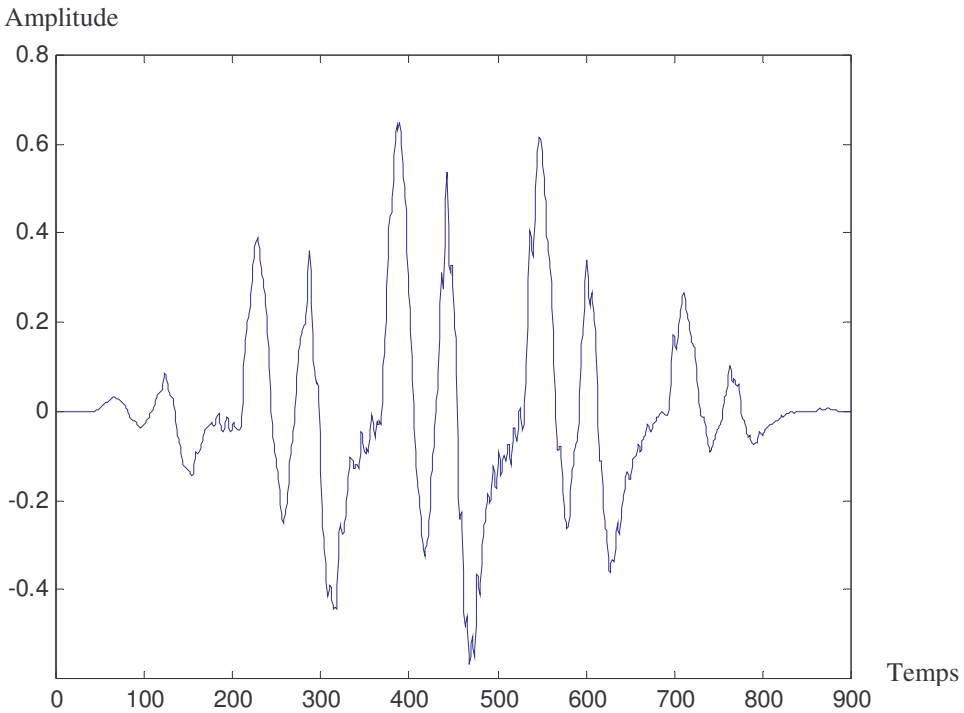

Figure 2.6 : La trame d'analyse après pondération

4. Analyse du signal

Le Modèle Harmonique plus Bruit consiste à décomposer le signal de parole *s(n)* en partie harmonique *h(n)* et partie bruitée *b(n)* :

$$s(n) = h(n) + b(n) \tag{2.5}$$

Pour les sons voisés, le spectre est décomposé en deux bandes distinctes, séparées par la fréquence maximale de voisement (f_m), qui est un paramètre variant dans le temps. La bande basse ($f < f_m$) est représentée par la partie harmonique, et principalement constituée de raies situées à des fréquences multiples de la fréquence fondamentale :

$$h(n) = \sum_{k=1}^{L} A_k \cos(2\pi f_0 kn + \varphi_k) \tag{2.6}$$

où :

A_k et φ_k sont respectivement l'amplitude et la phase initiale de la k^{ieme} harmonique ;

f_0 : Fréquence fondamentale du signal;

L : Nombre d'harmoniques.

Ces paramètres sont réactualisés aux instants d'analyse spécifiques, notés t_a^i *et* t_a^{i+1}.

La bande haute ($f > f_m$) représente la partie bruitée d'un son voisé, ou encore un son non voisé, et peut être modélisée comme un processus Auto-Régréssif.

La partie bruitée *b(n)* est obtenue en filtrant un bruit blanc Gaussien (*w*), par un filtre tout-pôles $G(z)$;

$$b(n) = g(n) * w(n) \tag{2.7}$$

w: Bruit blanc gaussien de variance σ^2 ;

g(n) filtre tout pôles d'ordre *p* de paramètres a_i.

Le schéma de la figure 2.7 résume les étapes de notre système de décomposition et modélisation harmonique plus bruit.

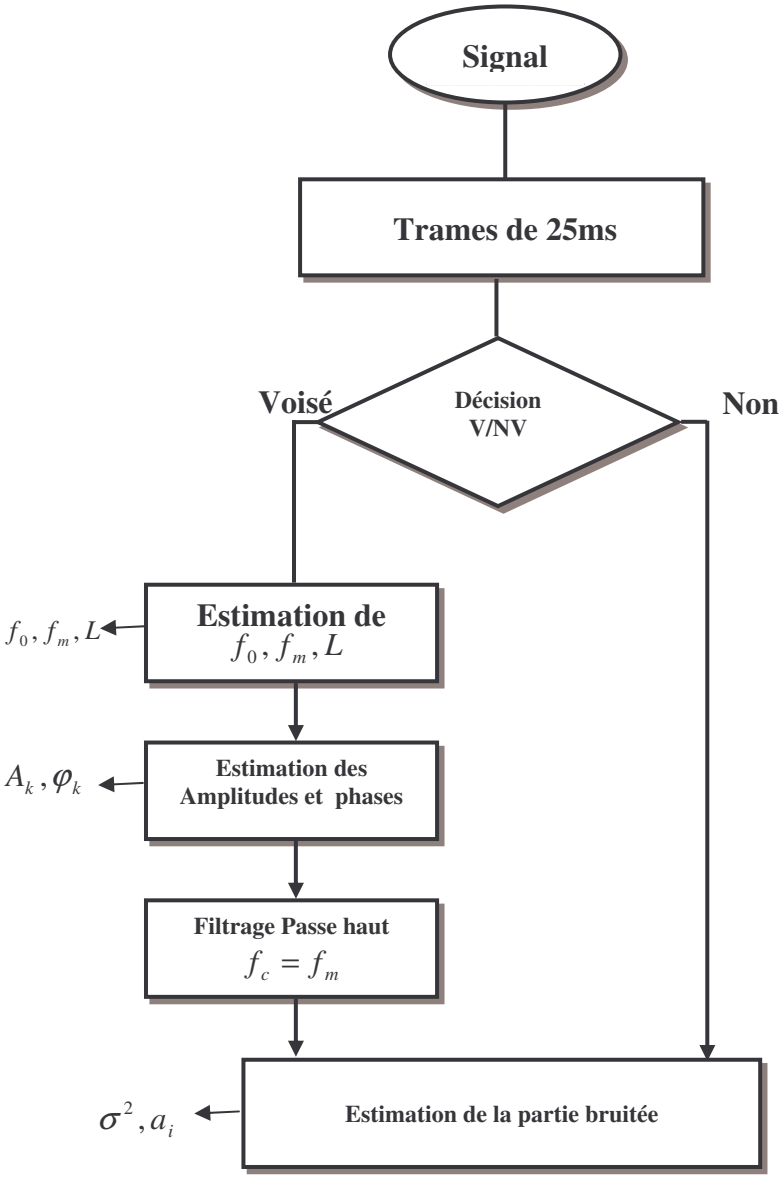

Figure 2.7 : Schéma de l'analyse par le "Modèle Harmonique plus Bruit" des signaux de parole

5. Modifications et synthèse:

Le signal de parole synthétique est la superposition de la partie harmonique avec la partie bruitée (équation 2.5):

La synthèse HNM (figures 2.8 et 2.9)se fait d'une manière additive pour la partie harmonique *h(n)*. On assure une évolution continue des paramètres par interpolation entre les instants d'analyse. La partie bruit est synthétisée par une technique de recouvrement-addition

(OLA). Les modifications sont appliquées directement sur les paramètres estimés lors de la phase d'analyse pour générer le signal synthétique désiré.

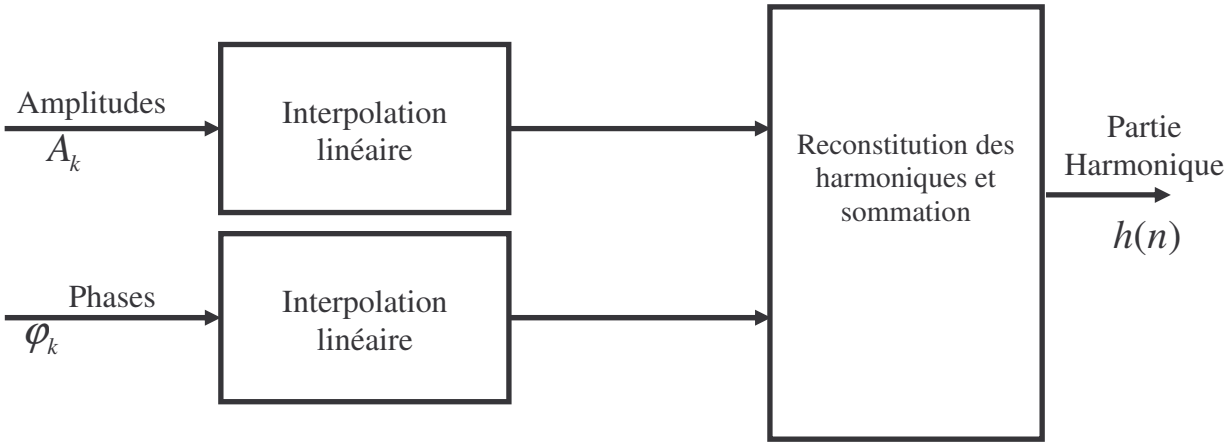

Figure 2.8: Synthèse de la Partie Harmonique

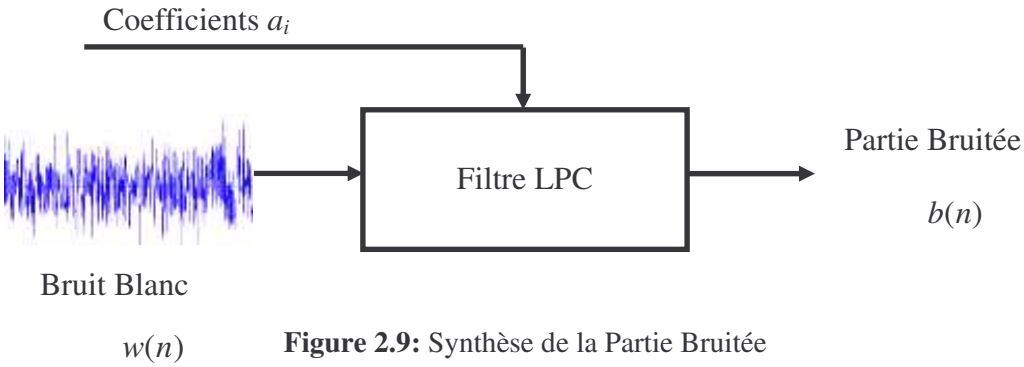

Figure 2.9: Synthèse de la Partie Bruitée

Chapitre III

Analyse par HNM du signal de parole

Introduction :

L'analyse HNM porte principalement sur la détermination d'un jeu de paramètres modélisant le signal de parole. Modéliser les caractéristiques d'un locuteur ou du signal de parole par quelques paramètres est un objectif autant pour les systèmes de reconnaissance et de vérification d'identité, que pour les systèmes de synthèse vocale.

La décomposition adoptée propose une modélisation du signal par certains paramètres dont l'estimation fait l'objet de la partie en cours.

Le signal discret observé *s* est donné par (2.5) reprise ci-dessous :

$$s(n) = \sum_{k=1}^{L} A_k \cos(2\pi f_0 kn + \varphi_k) + g(n) * w(n) \qquad (3.1)$$

Ce modèle complet du signal est décrit par un jeu de paramètres $\left(f_0, L, A_k, \varphi_k, f_m, g, \sigma^2\right)$.

où :

A_k et φ_k sont respectivement l'amplitude et la phase initiale de la k^{ieme} harmonique ;

f_0 : Fréquence fondamentale du signal;

L : Nombre d'harmoniques ;

f_m : Fréquence maximale de voisement ;

g : Filtre tout-pôles (auto-régréssif) ;

σ^2 : Variance du bruit blanc Gaussien *w*.

Ces paramètres doivent être réactualisés pour chaque trame d'analyse.

On défini des instants d'analyse uniformément répartis sur l'échelle des temps tels que

$\Delta t = t_a^{i+1} - t_a^i = 25ms$ (trame pour laquelle le signal de parole est quasi-périodique)

Les paramètres du modèle sont estimés autour de chaque instant d'analyse t_a^i.

1. Estimation de la fréquence fondamentale et détection de voisement:

La fréquence fondamentale (f_0) ou "pitch" est un paramètre très important du modèle. Sa détermination constitue la première étape à réaliser dans l'analyse du signal de parole.

L'estimation de la fréquence fondamentale est considérablement utilisée dans le codage, la reconnaissance et la synthèse de la parole. Malheureusement ce n'est pas une grandeur directement mesurable, ce qui rend sa détermination une tâche difficile. Cette difficulté est due principalement à la complexité du système phonatoire humain. La fréquence fondamentale d'un

signal vocal change constamment : elle change faiblement due à des microvariations de la tension des cordes vocales et de manière plus importante à cause de l'intonation et de l'accentuation. D'autres problèmes sont aussi à l'origine de cette difficulté:

- Les changements de f_0 peuvent être brusques (saut par octave) ;
- Les sous harmoniques de f_0 apparaissent souvent;
- Le voisement est très irrégulier au début et à la fin des parties voisées;
- Parfois, une zone voisée ne contient que peu de cycles glottaux.
- La présence du bruit rend parfois la tâche plus difficile.

Plusieurs recherches en traitement de la parole se sont focalisées sur le domaine de l'estimation de la fréquence fondamentale [HES 83], [DOV 94]. A cet effet, plusieurs méthodes et algorithmes ont été développés. Leur but est de fournir des estimateurs permettant d'extraire de façon automatique l'information sur la fréquence fondamentale des signaux voisés. Cette information peut être déterminée soit à partir de la périodicité du signal dans le domaine temporel ou à partir de l'espacement régulier entre les harmoniques dans le domaine fréquentiel.

1.1. Techniques fréquentielles:

Les algorithmes fréquentiels utilisent la structure harmonique des signaux périodiques dans l'estimation de la fréquence fondamentale. Cela se fait de plusieurs façons :

1.1.1. Histogramme d'harmonique :

Pour déterminer la fréquence du signal, dans le domaine spectral, on procède par plusieurs étapes:

- La première étape consiste à calculer la transformée de Fourier du signal sur une fenêtre de longueur donnée.
- Cette étape est suivie par une détection de pics destinée à localiser les harmoniques présentes dans le signal.
- Enfin, à partir des pics obtenus, on forme un histogramme des fondamentaux possibles: on divise chaque fréquence par 2, 3, 4, etc... et on ajoute leurs contributions, avec ou sans pondération.

Ainsi, le fondamental est le seul à recueillir les contributions de toutes les harmoniques présentes, l'harmonique "2" ne recueillant que celles des harmoniques paires etc...

Les sous-multiples du fondamental recueillent également quelques contributions, mais dans une moindre mesure. Au total, l'histogramme possède un pic prononcé à la fréquence fondamentale.

1.1.2. Sommes et produits spectraux :

Une généralisation de l'histogramme permet de se débarrasser de l'étape de la détection de pics. On calcule alors directement deux fonctions appelées la somme et le produit spectral.

Somme spectrale: La somme spectrale est définie par:

$$S(\omega) = \sum_{1}^{K} \left| X(e^{jk\omega}) \right|^2 \qquad pour \quad \omega < \frac{\pi}{K} \tag{3.2}$$

Produit spectral: Le produit spectral est défini par:

$$P(\omega) = \prod_{1}^{K} \left| X(e^{jk\omega}) \right|^2 \qquad pour \quad \omega < \frac{\pi}{K} \tag{3.3}$$

On voit que le principe est le même que celui de l'histogramme. Le spectre est compressé par un facteur k, puis les spectres obtenus sont ajoutés ou multipliés. Le fondamental se trouve ainsi considérablement renforcé, et apparaît alors sous la forme d'un pic prononcé facile à détecter.

Le produit spectral harmonique (et dans une moindre mesure la somme spectrale harmonique) est parmi les méthodes les plus fiables. Malheureusement, elles font intervenir un nombre élevé de calculs, car le spectre doit être estimé avec précision. C'est la raison pour laquelle les méthodes temporelles leur ont été préférées jusqu'à présent.

Le problème est que les méthodes d'analyse numérique supposent une certaine stabilité du spectre fréquentiel du signal à l'intérieur d'une fenêtre d'analyse, condition qui n'est pas toujours satisfaite.

La plupart des techniques supposent que les propriétés du signal restent constantes sur une durée d'au moins deux (pour les locuteurs avec une voix basse) à quatre (pour les locuteurs avec une voix haute) périodes fondamentales, ce qui veut dire concrètement une stabilité sur une période de 15 à 25 ms.(c.à.d travailler avec des fenêtres d'une longueur de 1024 points pour des signaux dont la fréquence d'échantillonnage est de 44,1 kHz, donc de 23,2 ms). Les variations brusques du pitch peuvent donc rendre les sommets du spectre de puissance "flous". Ce problème d'incertitude est en partie corrigé par une pondération de la fenêtre d'analyse qui atténue les extrémités (par exemple une pondération avec une fenêtre de Hanning ou de

Blackman), ce qui permet de concentrer l'analyse du signal sur ce qui se passe au milieu de la fenêtre.

1.2. Techniques temporelles:

Les techniques temporelles se basent sur l'idée de retrouver les périodes de répétition du signal de parole qui nous renseigne sur sa périodicité, elle permettent une estimation de la période P avec un délai minimal, et des calculs très simples.

1.2.1. Estimateurs à seuil :

- **Passage par zéro:**

C'est le plus élémentaire des détecteurs. Le signal est filtré par un filtre passe-bas, puis on détecte ses passages par zéro dans un sens bien précis (par exemple du négatif vers le positif). Lorsque le signal ne possède qu'une sinusoïde, cela fonctionne bien, mais en présence d'harmoniques multiples, on peut observer plus de deux passages par zéro pendant une période. Une condition nécessaire et suffisante pour n'en observer que deux par période, indépendamment des phases respectives des composantes, est la suivante:

$$R = 20\log\left(\frac{A(1)}{\sum_{1}^{K} k.A(k)}\right) > 0 \qquad (3.4)$$

où $A(k)$ représente l'amplitude de l'harmonique k. Cette condition n'est que rarement satisfaite dans le signal original, d'où la nécessité de filtrer par un filtre de pente très raide. Un des problèmes de cette technique est la nécessité d'adapter la fréquence de coupure du filtre à la fréquence fondamentale du signal (sous peine de voir disparaître le fondamental, ou d'atténuer insuffisamment les harmoniques).

- **L'estimateur avec seuil non nul:**

Pour améliorer les performances de l'estimateur par zéro, on peut détecter non pas les passages par zéro, mais par un seuil plus élevé. On voit que si ce seuil est correctement choisi, on évite certains cas de confusion. Cette technique permet de filtrer par un filtre moins raide que précédemment. En revanche, elle nécessite l'adaptation du seuil à l'amplitude à court terme du signal, et est sensible à la polarité du signal.

- **L'estimateur à deux seuils et mémoire:**

On se fixe ici deux seuils, généralement de signe opposé, et l'on détecte les passages par ces deux seuils en éliminant les passages consécutifs par un même seuil figure 3.1). Cette technique rend l'estimation plus robuste, et donc permet un filtrage plus doux. En revanche, il y a désormais deux seuils à adapter.

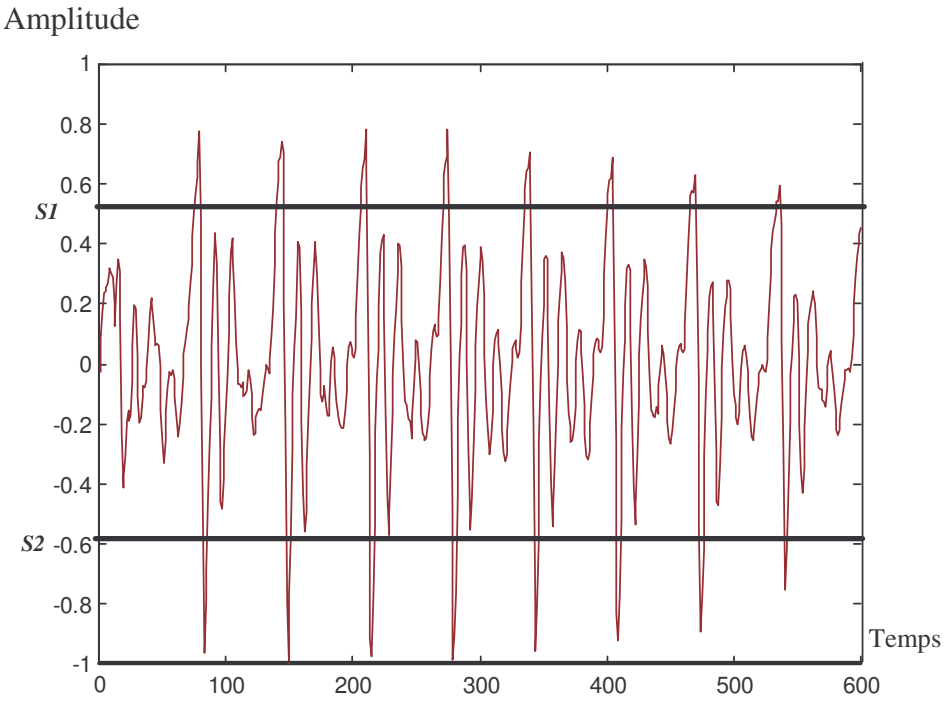

Figure 3.1: Exemple de détection du fondamental par estimateur à deux seuils et mémoire

Les principaux avantages de ces méthodes sont leur simplicité et leur rapidité. Elles fonctionnent bien dans un certain nombre de cas (parole, avec fondamental présent), mais peuvent donner des erreurs grossières (lorsque deux passages sont détectés au lieu d'un seul) ainsi que des erreurs plus fines: si l'amplitude du signal varie brusquement, les estimateurs à seuils non nuls présentent un biais significatif.

1.2.2. Algorithmes de type corrélation:

Les algorithmes de type corrélation travaillent dans le domaine temporel à court terme : le signal est extrait trame par trame et aucune transformation n'est appliquée sur ces trames. La taille des trames est un paramètre important : lorsqu'elle a une valeur fixe, elle contient

généralement deux à trois périodes de 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 et une version décalée de ce signal, décalage correspondant à la période cherchée. Cette corrélation peut aussi être remplacée par une fonction de dissemblance.

- **Average Magnitude Differential Function**

Cette fonction est le complément de la fonction d'autocorrelation, elle mesure la similarité entre deux périodes du signal. On peut utiliser la formule suivante qui définit l'AMDF:

$$AMDF(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} \left| s(n) - s(n+k) \right| \qquad (3.5)$$

Si le signal est parfaitement périodique de période P_0, AMDF(iP_0) est bien sûr nul.
AMDF(k) peut être vu comme la mesure de l'amplitude absolue en sortie d'un filtre à réponse impulsionnelle finie (RIF) H_k dont la réponse impulsionnelle est:

$$h_n = \begin{cases} 1 & n = 0 \\ -1 & n = -k \\ 0 & ailleurs \end{cases} \qquad (3.6)$$

Ce filtre est communément appelé un filtre en peigne ou "Comb filter" en raison de sa réponse en fréquence en forme de peigne: $H(z) = 1 - z^{k}$, ce qui montre que le filtre s'annule pour toutes les fréquences multiples de $1/k$.
En filtrant le signal par les filtres correspondant à différentes valeurs de k, et en évaluant l'amplitude en sortie, on peut déterminer la valeur de k qui fournit la sortie la plus faible, et la retenir comme valeur de la période.
Cette méthode a surtout été utilisée pour sa simplicité numérique (pas de multiplication), mais elle se révèle très sensible au bruit.

- **Estimation de la période à l'aide de la fonction d'autocorrélation :**

L'utilisation de l'autocorrélation pour la détection de la fréquence fondamentale est très classique en traitement de la parole. Pour un signal $s(n)$, la fonction d'autocorrélation est définie par:

$$R\ (k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} s(n)s(n+k) \qquad (3.7)$$

où le paramètre N contrôle l'horizon sur lequel est calculée la fonction. On définit également la fonction d'autocorrélation normalisée *r(k)* par

$$r(k) = \frac{\sum_{n=0}^{N-k-1} s(n)s(n+k)}{\left(\sum_{n=0}^{N-k-1} s(n)^2\right)^{1/2}\left(\sum_{n=0}^{N-k-1} s(n+k)^2\right)^{1/2}} \qquad (3.8)$$

Les deux fonctions d'autocorrélation ci-dessus vérifient les propriétés suivantes:

- Si *s(n)* est périodique, *R(k)* et *r(k)* le sont aussi, avec la même période.
- $r(k) \le r(0) = 1 \quad \forall\, k$

  l'égalité n'ayant lieu que si $s(n) = \alpha.s(n+k) \quad \forall\, 0 \le n < N-k$

Ces deux propriétés impliquent que pour un signal périodique, la fonction d'autocorrélation possède des pics aux multiples de P. Les maxima de la fonction de périodicité correspondent bien à des multiples de la période fondamentale. Le premier pic (le décalage donnant la meilleure corrélation) indique la fréquence fondamentale.

1.3. Choix de l'approche par autocorrelation

L'estimation de la fréquence fondamentale du signal peut être faite par le calcul de l'autocorrélation (normalisée), suivi d'une recherche de maximum. Cette technique présente l'avantage de simplicité dans les calculs et permet également d'assembler les deux étapes: détection de voisement et estimation de la fréquence fondamentale.
Une estimation de la fréquence fondamentale est obtenue en utilisant un détecteur de pics de la fonction d'auto-covariance biaisée du signal.
Etant donné que nous avons considéré que le signal est harmonique pour les portions voisées, l'estimation de la fréquence fondamentale f_0 constitue la première étape d'analyse.

Rappelons le modèle du signal observé *s(n)* qui suppose qu'il est la somme d'un signal centré exactement périodique de période $P = 1/f_0$ entière, et d'un bruit blanc.

$$s(n) = \sum_{k=1}^{L} A_k \cos(2\pi f_0 kn + \varphi_k) + b(n)$$

où :

- $f_0 = 1/P$: fréquence fondamentale réduite du signal;

- *L*: Nombre d'harmoniques du signal;
- A_k : Amplitudes réelles strictement positives.

Tous ces paramètres sont fixés et inconnus. Par ailleurs, nous supposons que :

- Les phases φ_k sont des variables aléatoires indépendantes, de loi uniforme sur [0,2π];
- *b(n)* est un bruit blanc centré de variance σ^2 ;

Alors on vérifie que :

- $E[s(n)]=0$; car *b* est centré et $E[\cos\varphi]=E[\sin\varphi]=0$
- $E[s(n)s(n+m)]=\sum_{k=1}^{L}\sqrt{2}A_k^2\cos(2\pi f_0 km)+\sigma^2\delta(m)$ Ne dépend pas de *n*.

On déduit que $s(n)$ est un processus centré, stationnaire au sens large et que sa fonction d'auto-covariance :

$$R_s(m)=E[s(n)s(n+m)]=\sum_{k=1}^{L}\sqrt{2}A_k^2\cos(2\pi f_0 km)+\sigma^2\delta(m) \qquad (3.9)$$

On remarque que la fonction $\sum_{k=1}^{L}\sqrt{2}A_k^2\cos(2\pi f_0 km)$ est périodique et de période $P=1/f_0$, et que son maximum global est atteint pour tout multiple de P.

On peut donc estimer la période P en recherchant la plus petite valeur de m, pour laquelle $R_s(m)$ atteint son maximum sur un intervalle $[m_{\min}, m_{\max}]$ avec $m_{\min}>0$.

Malheureusement, on ne connaît pas dans la pratique la fonction d'auto-covariance théorique, et il est nécessaire de l'estimer à partir d'une réalisation du signal *s(n)* sur une fenêtre $n\in[0,...,N-1]$

Pour tout entier m, tel que : $-(N-1)\le m\le(N+1)$, on défini :

$$R_s(m)=\begin{cases}\dfrac{1}{N}\displaystyle\sum_{n=0}^{N-m-1}S(n)S(n+m) & si\ m\ge 0\\ R_s(-m) & sinon\end{cases} \qquad (3.10)$$

Comme l'espérance mathématique de cet estimateur est la fonction d'auto-covariance théorique multipliée par une fenêtre *f(m)*:

$$f(m)=\begin{cases}\dfrac{N-|m|}{N} & si\quad |m|<N\\ 0 & sinon\end{cases} \qquad (3.11)$$

La plus petite valeur de $m \in [m_{\min}, m_{\max}]$ pour laquelle $R_s(m)$ atteint son maximum correspond à l'unique maximum global de $R_s(m)$ sur cet intervalle. Il faut cependant remarquer que la multiplication de la fenêtre de pondération entraîne un biais sur l'estimation de P.

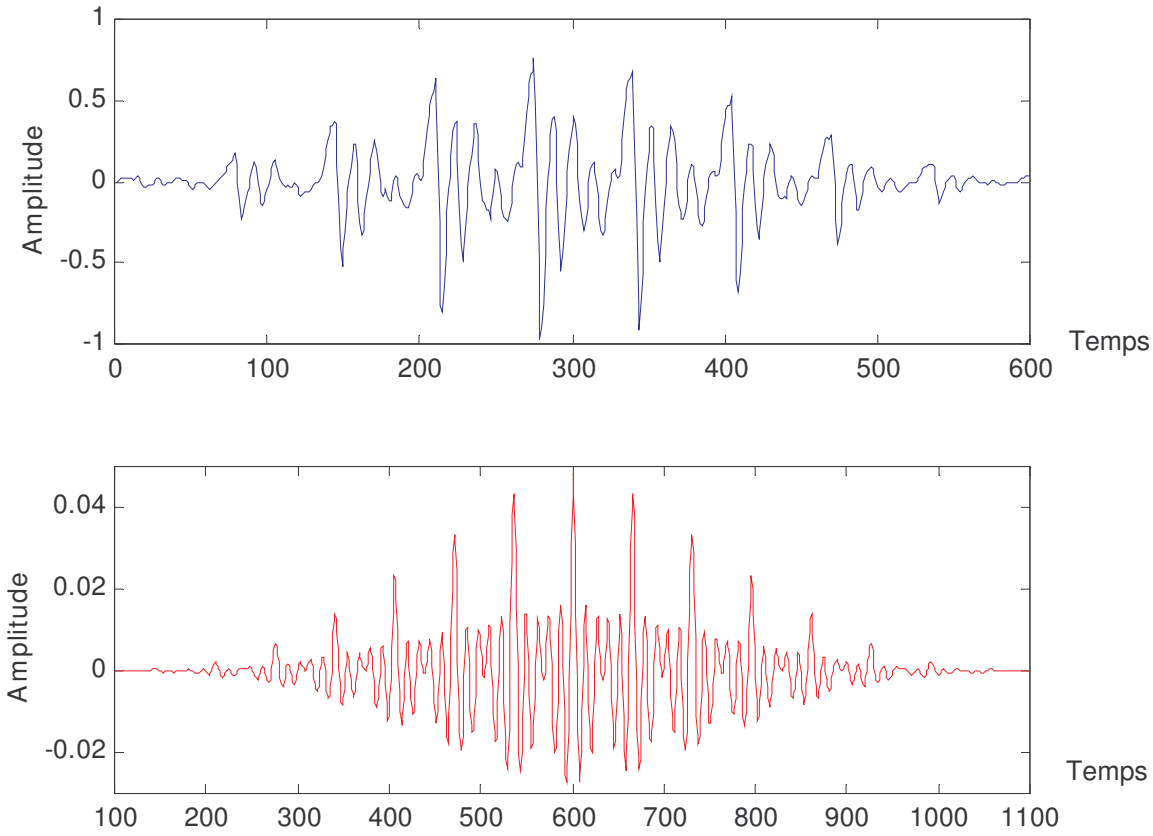

Figure 3.2: Signal de parole voisé (en haut) et sa fonction d'autocorrélation (en bas)

Les figures 3.2 et 3.3 illustrent l'emploi de la fonction d'autocovariance biaisée pour retrouver la fréquence fondamentale du signal de parole.

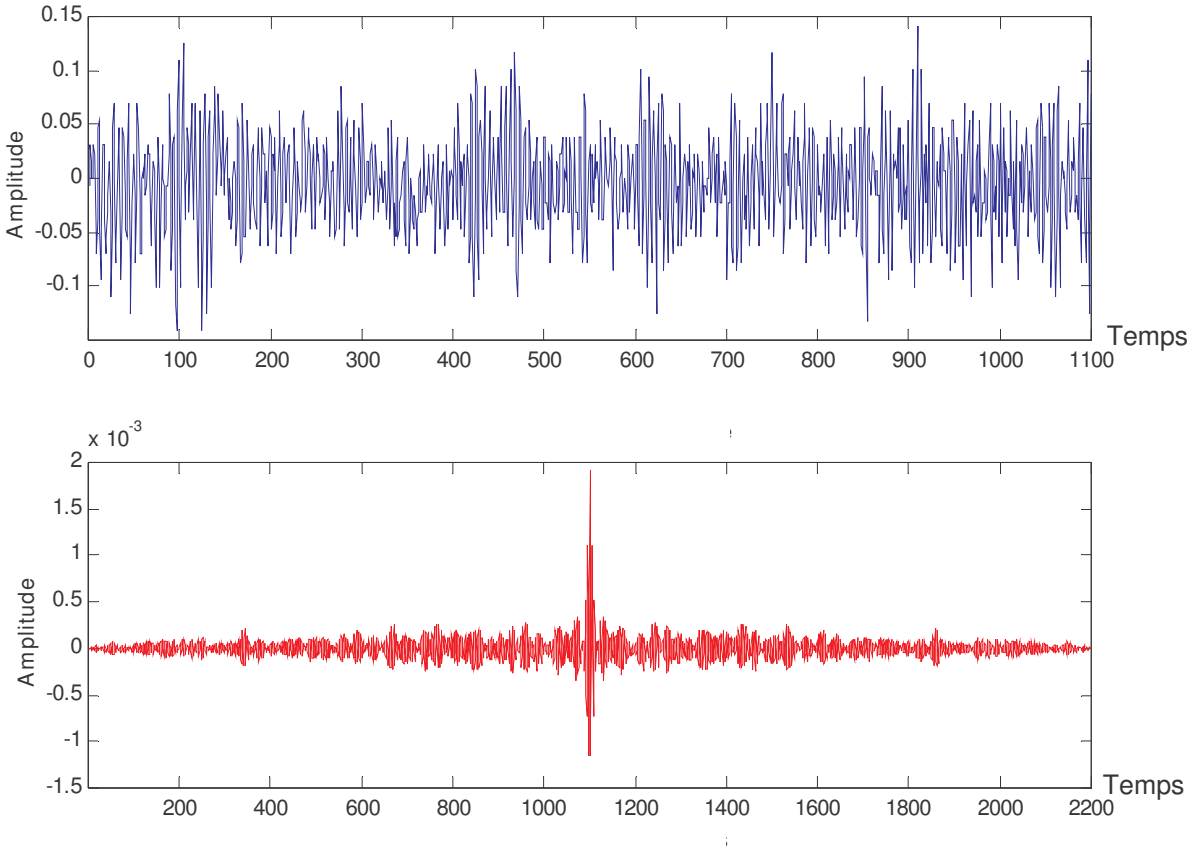

Figure 3.3: Signal de parole non voisé (en haut) et sa fonction d'autocorrélation (en bas)

- **Algorithme de l'estimation de la fréquence fondamentale (par autocorrelation) :**

1. Extraction d'un segment centré en t_a^i de durée Δt=25ms
2. Choisir une valeur s telle que: $s \in]0 \quad 1[$; (seuil de décision)
3. Calcul de la fonction d'auto-covariance biaisée *R(k)* du segment ;
4. Rechercher le maximum de *R(k)* sur l'intervalle $[2ms,20ms]$ (en considérant que la fréquence fondamentale du signal de parole voisé ne peut pas monter au dessus de 500Hz, ni descendre au dessous de 50Hz);
    - Si le maximum est inférieur à $s \times R(0)$, le signal est non voisé et sa partie harmonique est nulle.
    - Si le maximum est supérieur à $s \times R(0)$, le signal est voisé et sa période *P* est la valeur *k* qui est l'indice du maximum.

1.4. Détection de voisement:

La fonction d'autocorrélation normalisée donne une estimation du "degré de périodicité" du signal: une valeur $\max_k R(k)$ proche de 1 indique que le signal est très périodique (dans un sens "large": $s(n) = \alpha.s(n+k)$ avec éventuellement α différent de 1). La détection de voisement dépend essentiellement du chois du seuil s ($s \in]0 \quad 1[$). Cette décision se fait de manière synchrone à la détection de la fréquence fondamentale.

Les figures 3.4 et 3.5 illustrent l'emploi de la fonction d'autocorrélation pour la décision voisé/non voisé, pour les classes de sons.

La décision de voisement et l'estimation de la fréquence fondamentale se font pour toutes les trames du signal à analyser et ainsi nous obtenons l'évolution du pitch pour le signal complet, montrée dans la figure 3.6.

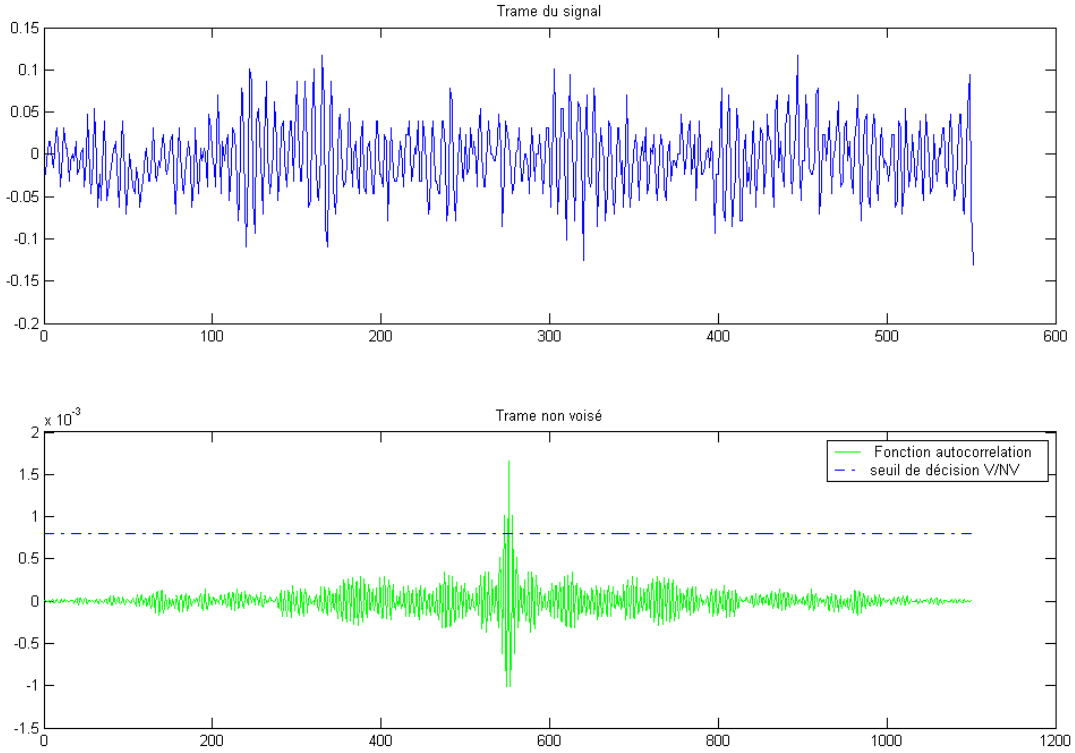

Figure 3.4: Décision de voisement pour une trame non voisé

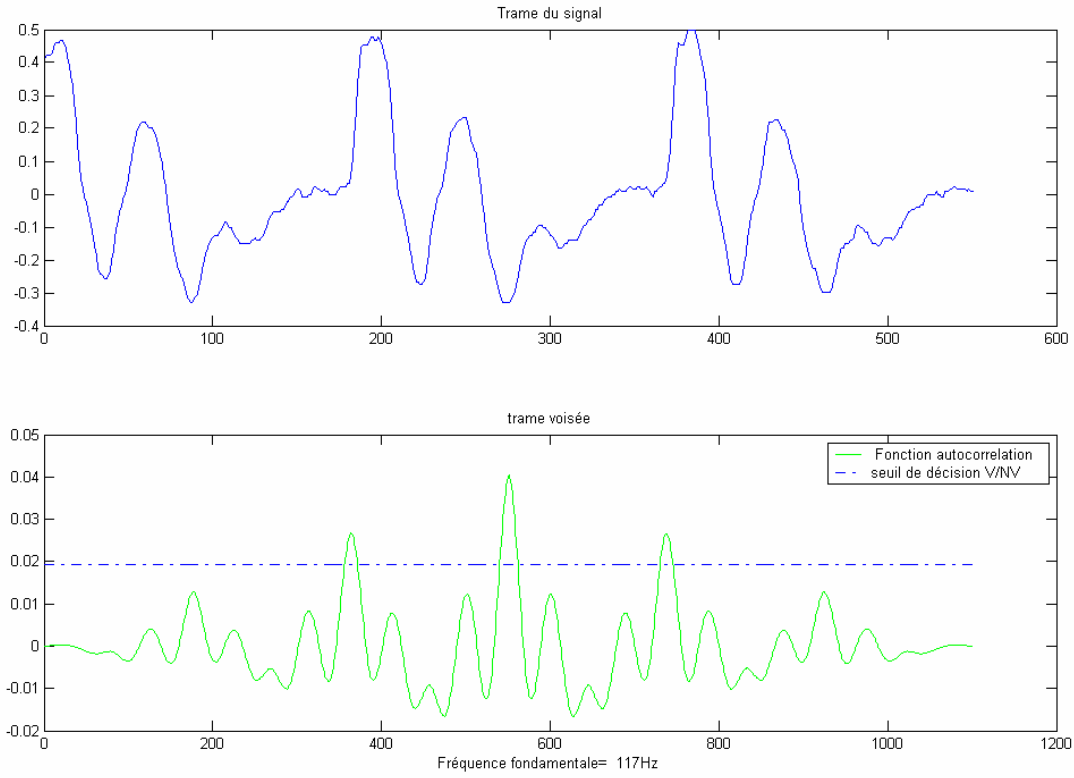

Figure 3.5: Décision de voisement pour une trame voisée

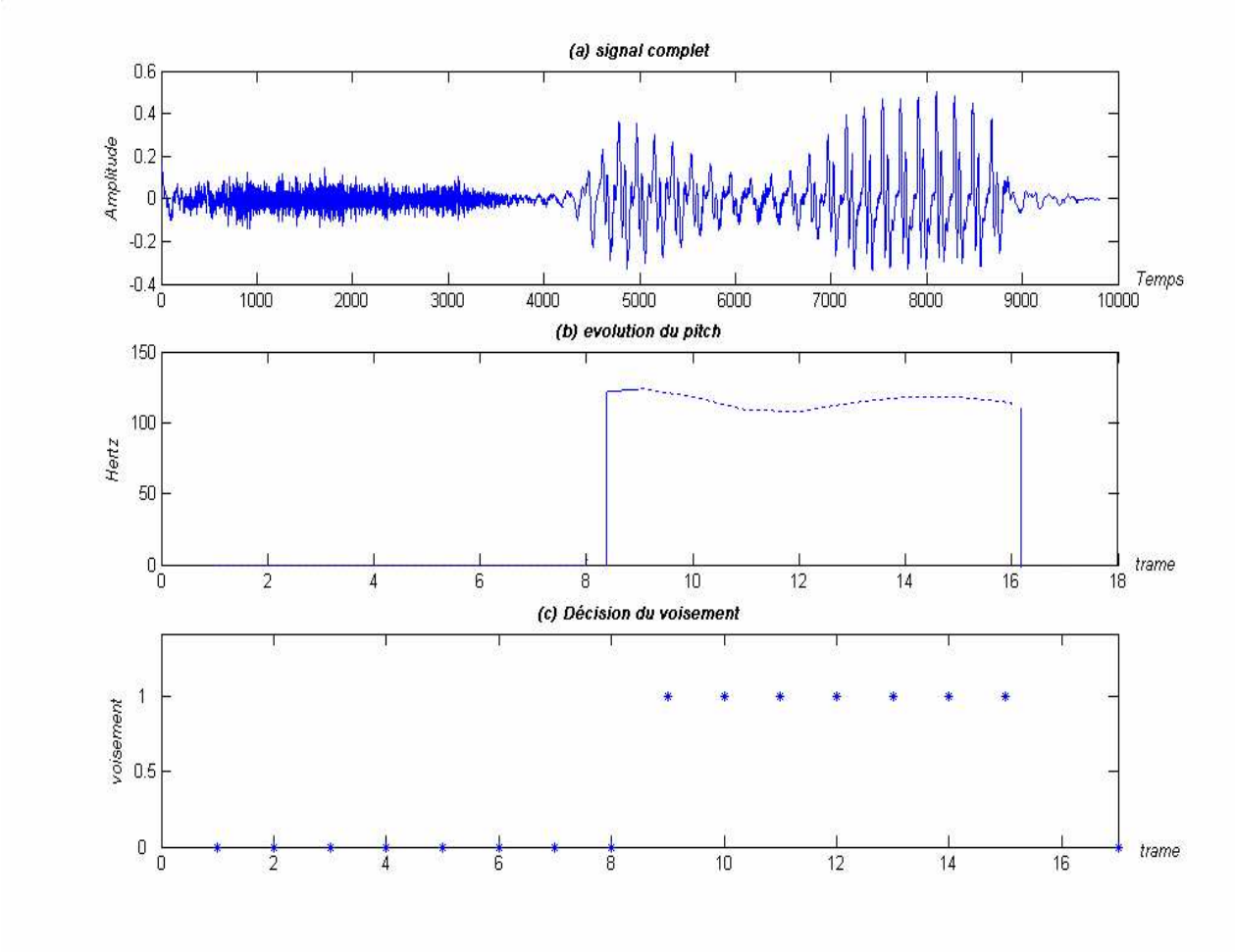

Figure 3.6: (a) Signal de parole, (b) Fréquence fondamentale
(c) Indicateur de voisement.

2. Estimation de la composante harmonique:

2.1. Amplitudes et phases des harmoniques (A_k *et* φ_k) :

L'expression du signal observé (2.11) peut être écrite :

$$s(n) = \sum_{k=1}^{L} c_k \cos(2\pi f_0 kn) + \sum_{k=1}^{L} s_k \sin(2\pi f_0 kn) + c_0 + b(n) \qquad (3.12)$$

avec :

$$c_k = A_k Cos\varphi_k ;$$

$$s_k = -A_k Sin\varphi_k \qquad (3.13)$$

où :

$$A_k = \sqrt{c_k^2 + s_k^2}\ ;$$

$$\varphi_k = -arctg(s_k / c_k) \qquad (3.14)$$

avec

L est le nombre d'harmoniques, choisi en veillant à ce que la fréquence la plus aigue ($L\omega_0 / 2\pi$), reste inférieure à la moitié de la fréquence d'échantillonnage (pour éviter des problèmes de recouvrement);

c_0 est la contribution de la composante continue.

L'égalité précédente peut être condensée sous forme matricielle :

$$\underline{S}_{(Nx1)} = \underline{\underline{D}}_{(Nx2L+1)} . \underline{\theta}_{(2L+1x1)} + \underline{B}_{(Nx1)} \qquad (3.15)$$

où :

$\underline{\underline{D}}$: ne dépend que de ω_0 ;

$\underline{\theta}$: Vecteur de paramètres: c_0, c_k et s_k .

Ces paramètres sont estimés par la méthode des moindres carrés qui consiste à minimiser le critère ε donné par l'équation [STY96] suivante :

$$\varepsilon = \sum_{n=t_a^i-N}^{t_a^i+N} w^2(n)[s(n) - h(n)]^2 \qquad (3.16)$$

où:

s(n): Signal original;

h(n): Partie harmonique;

w(n) : fenêtre de pondération;

N: Nombre entier de périodes fondamentales locales $P(t_a^i)$.

La trame d'analyse est centrée autour de l'instant d'analyse t_a^i et elle est de longueur *M=2N+1*.

L'écriture matricielle de la partie harmonique est donnée par

$$\underline{H}_{(Nx1)} = \underline{\underline{D}}_{(Nx2L+1)} \underline{\theta}_{(2L+1x1)} \qquad (3.17)$$

La minimisation du critère ε revient à résoudre l'équation normale suivante [STY96]:

$$D^t W^t W D\theta = D^t W^t W S \qquad (3.18)$$

où $\underline{\theta}_{(2L+1x1)}$: Vecteur de paramètres à estimer :

$$\underline{\theta}^t = [c_1 \quad s_1 \quad c_2 \quad s_2 \quad \ldots\ldots \quad c_L \quad s_L \quad c_0\ ]$$

où D matrice de $M \times (2L+1)$ donnée par:

$$D = \begin{pmatrix} \cos\omega_0(-N) & Sin\omega_0(-N) & ... & .. & .. & .. & .. & \cos L\omega_0(-N) & Sin\omega_0(-N) & 1 \\ \cos\omega_0(-N+1) & Sin\omega_0(-N) & .. & .. & .. & .. & .. & \cos L\omega_0(-N+1) & Sin\omega_0(-N) & 1 \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. & .. & .. & .. \\ \cos\omega_0(N) & Sin\omega_0(-N) & .. & .. & .. & .. & .. & \cos L\omega_0(N) & Sin\omega_0(-N) & 1 \end{pmatrix}$$

ou encore

$$D = [dc_1\ ds_1\ dc_2\ ds_2\ \ldots\ldots\ldots dc_L\ ds_L\ 1]$$

Avec :

$$dc_i^t = [\cos i\omega_0(-N) \ldots\ldots\ldots \cos i\omega_0(N),]$$

$$ds_i^t = [\sin i\omega_0(-N) \ldots\ldots\ldots \sin i\omega_0(N),]$$

$$1^t = [1 \ldots\ldots\ldots\ldots\ 1]$$

$W(M\times M)$: matrice diagonale donnée par:

$$W = \begin{pmatrix} w(-N) & 0 & 0 & .. & .. & .. & .. & 0 \\ 0 & w(-N+1) & 0 & .. & .. & .. & .. & 0 \\ 0 & 0 & ... & .. & .. & .. & .. & 0 \\ .. & .. & .. & ... & .. & .. & .. & .. \\ .. & .. & .. & .. & ... & .. & .. & .. \\ .. & .. & .. & .. & .. & ... & .. & .. \\ .. & .. & .. & .. & .. & .. & ... & 0 \\ 0 & 0 & .. & .. & .. & .. & 0 & w(N) \end{pmatrix}.$$

La diagonale de cette matrice est un vecteur représentant la fenêtre de hamming.

$S(M \times 1)$: Vecteur contenant les données originales (Signal original).

$S^t = [S(-N)\ S(-N+1) \ldots\ldots\ldots\ldots S(-N)]$

A titre d'exemple, la figure suivante (3.7) montre les amplitudes des harmoniques correspondantes obtenues pour une trame d'un signal de parole voisé.

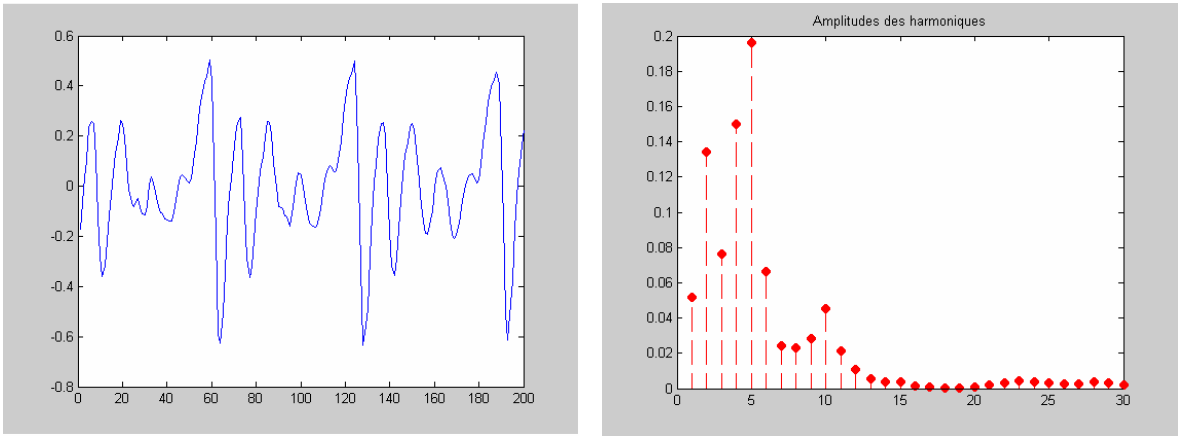

Figure 3.7: (a) Trame voisée, (b) Amplitudes des harmoniques correspondantes

2.2. Fréquence maximale de voisement :

L'estimation de la fréquence maximale de voisement (f_m) pour les trames voisées consiste à estimer le nombre de sinusoïdes ou du nombre de composantes fréquentielles significatives dans la trame du signal de parole. Malheureusement l'estimation de f_m est un problème délicat.

Si L est le nombre de sinusoïdes réellement présentes dans la partie harmonique, on pourra supposer que la fréquence maximale de voisement est égale à $(L + 1)f_0$ (fréquence de la première harmonique absente). Or jusqu'à présent le paramètre L a été fixé a priori, et donc volontairement surévalué. Les harmoniques les plus aigus estimés à l'étape précédente sont

donc vraisemblablement noyés dans la partie bruitée. Il faut donc déterminer de manière heuristique la véritable valeur de L, dont on pourra déduire (f_m). Pour cela, on utilisera un critère simple : L sera choisi comme étant la plus petite valeur pour laquelle $\sum_{k=1}^{L_{reel}} A_k^2 \geq s \times \sum_{k=1}^{L} A_{k-1}^2$ où le seuil s doit être très proche de 1 [STY96].

Une autre approche basée sur la recherche de pics dans le spectre de ces trames et s'appuie sur l'hypothèse que le niveau des harmoniques est sensiblement égal au niveau de bruit aux fréquences autres que les fréquences des sinusoïdes. Le premier pic le plus important est localisé dans l'intervalle [$f_0/2$, $3f_0/2$]. La recherche des pics est ainsi suivie jusqu'au dernier sommet, qui représente la fréquence maximale de voisement de la trame voisée (figure 3.8).

Ces deux méthodes donnent des résultats très similaires.

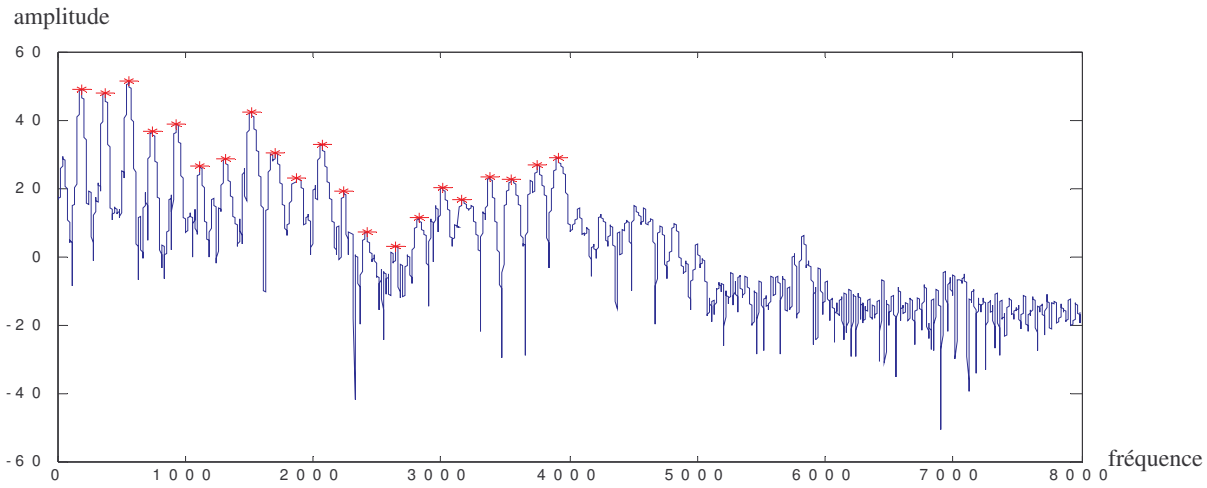

Figure 3.8: Détection de pics du spectre d'une trame voisée

Ces paramètres vont se calculer trame par trame pour tout le signal de parole et ainsi, nous obtiendrons l'évolution de la fréquence fondamentale et aussi, la fréquence maximale de voisement (figure 3.9).

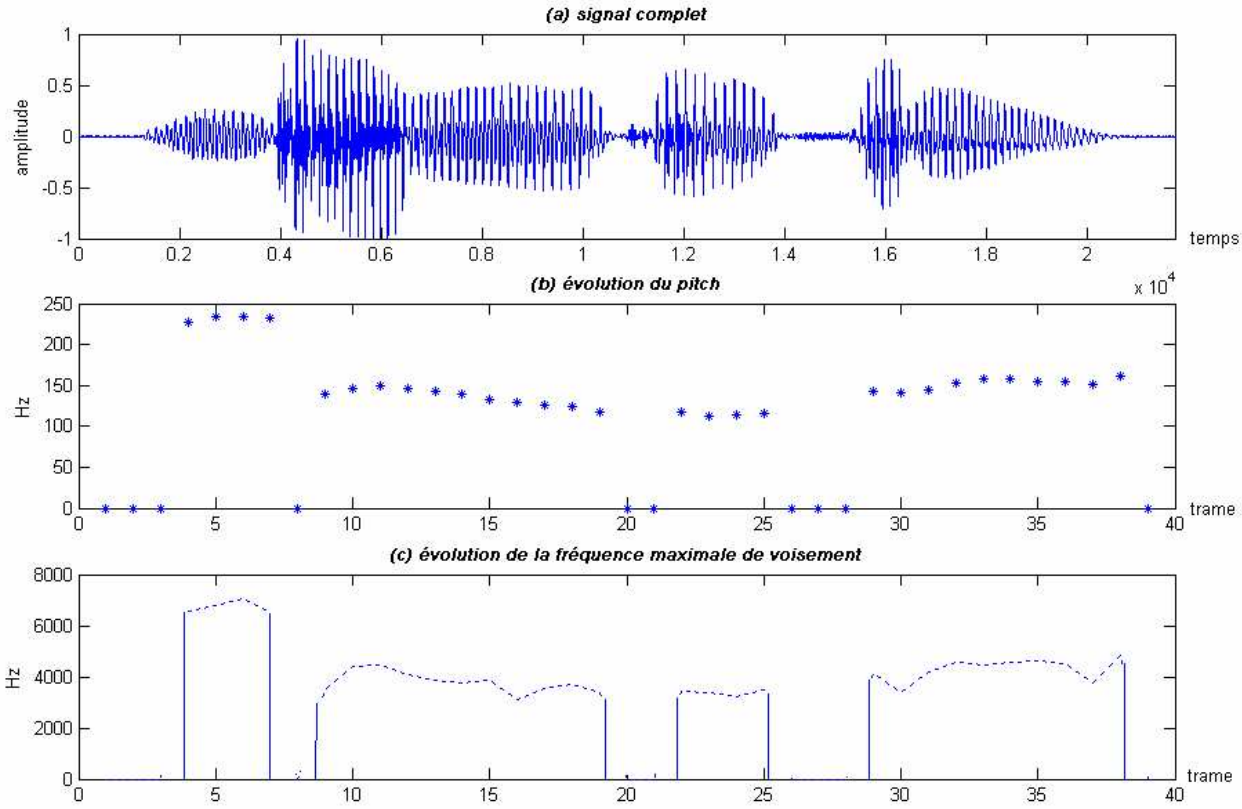

Figure 3.9: (a) Signal de parole, (b) Fréquence fondamentale
(c) Fréquence maximale de voisement.

3. ESTIMATION DE LA PARTIE BRUIT :($g(z)$ et σ^2):

La dernière étape d'analyse consiste à estimer la partie bruitée.

3.1. Extraction de la composante bruit :

3.1.1. Cas d'un signal voisé :

Pour les signaux voisés, la partie bruitée est obtenue par filtrage passe haut à la fréquence maximale de voisement.

La figure 3.11 représente la composante bruit, extraite de la trame voisée (figure 3.10), après un filtrage passe haut à la fréquence maximale de voisement (f_m).

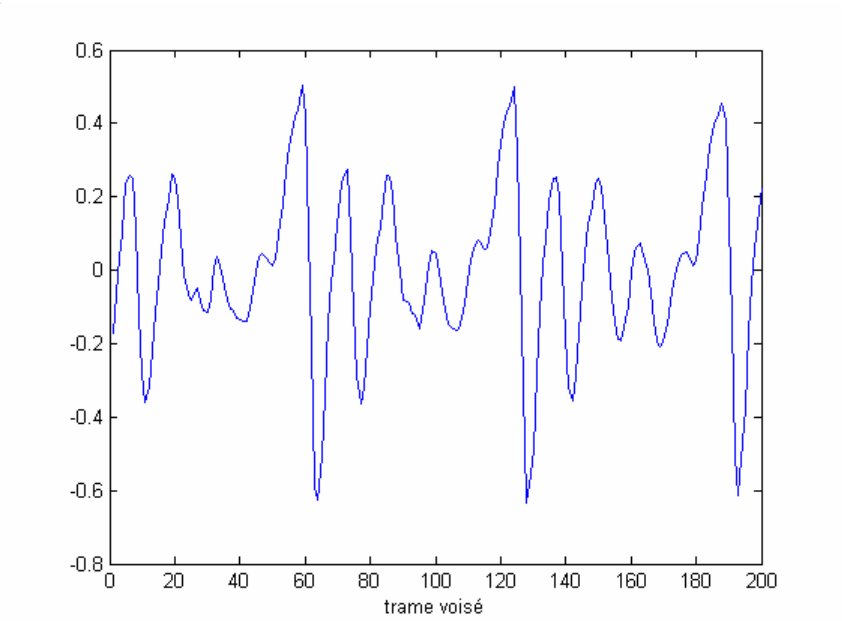

Figure 3.10 : Trame voisée

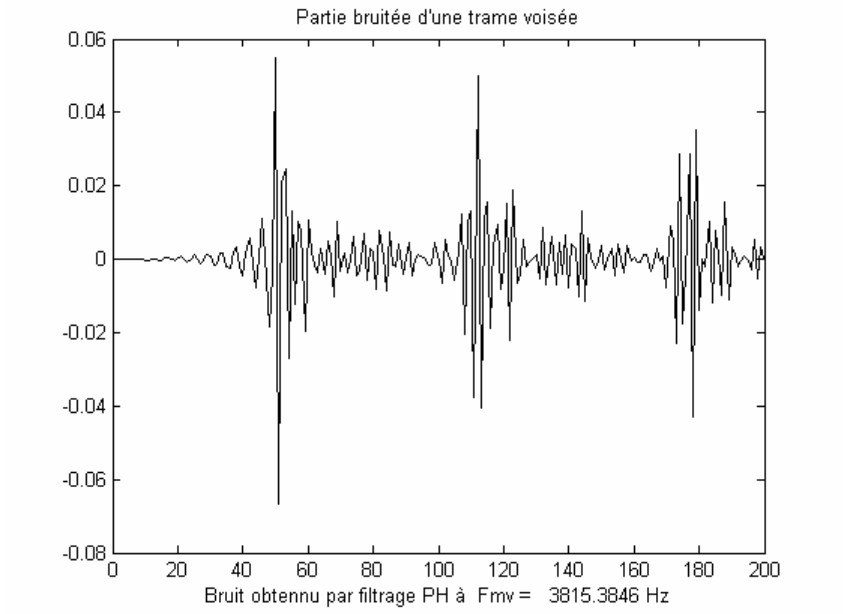

Figure 3.11: Bruit obtenu après filtrage passe haut à f_c=f_m

3.1.2. Cas d'un signal non voisé :

Dans le cas où le signal est non voisé, la partie harmonique est nulle, le signal est donc complètement modélisé comme du bruit ($s(n) = b(n)$).

La composante bruit n'est autre que le signal original (figure 3.12).

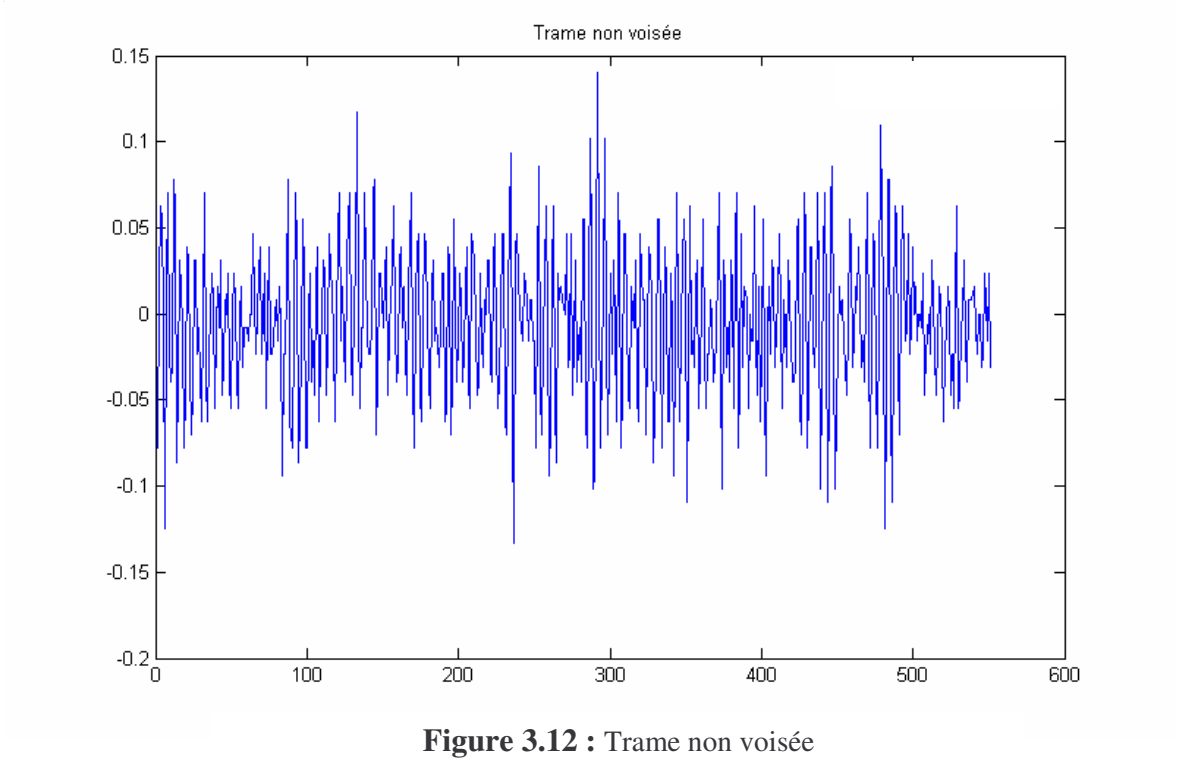

Figure 3.12 : Trame non voisée

3.2. Modélisation autorégressive

Pour chaque trame d'analyse, le bruit est modélisé comme un processus auto-regressif (AR) d'ordre p.

$$b(n) = g(n) * w(n) \qquad (3.20)$$

où

$G(z) = \dfrac{1}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}}$ filtre tout-pôles (auto-régréssif)

w: Bruit blanc Gaussien de variance σ^2 ;

La modélisation de la partie bruitée revient à déterminer les coefficients a_i du filtre AR, ainsi que la variance σ^2 du bruit blanc à l'entrée du filtre.

La figure 3.13 présente un schéma simplifié de cette modélisation.

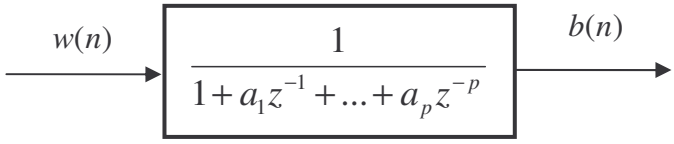

Figure 3.13: Modèle autorégressif pour la composante bruit

3.2.1. Estimation des coefficients (a_i) du filtre AR:

On peut définir une prédiction ou estimation de chaque échantillon *B(n)* à partir des *p* échantillons qui le précèdent :

$$\hat{B}(n) = \sum_{i=1}^{p} \hat{a}_i .B(n-i) \qquad (3.21)$$

où les $\hat{a}_i$ (i=1,..,p) sont les estimés des coefficients de prédiction a_i.

L'erreur commise par la prédiction vaut :

$$e(n) = B(n) - \hat{B}(n)$$

$$e(n) = B(n) - \sum_{i=1}^{p} \hat{a}_i .B(n-i) \qquad (3.22)$$

$$e(n) = \sum_{i=0}^{p} \hat{a}_i .B(n-i) \qquad avec \quad \hat{a}_0 = 1 \qquad (3.23)$$

On abandonnera désormais l'indice chapeau (^) pour désigner les coefficients estimés.

Estimer le modèle AR revient à estimer les coefficients a_i, c'est à dire trouver les coefficients optimaux.

On a $$e(n) = B(n) - \sum_{i=1}^{p} a_p(i) . B(n-i)$$

On définit l'énergie résiduelle de prédiction par la somme

$$E = \sum_{n=-\infty}^{+\infty} e^2(n) \qquad (3.24)$$

Le critère usuel pour l'optimisation est la minimisation de l'énergie résiduelle de prédiction, soit :

$$E = \sum_n e^2(n)$$

$$E = \sum_n \left[B(n) - \sum_{i=1}^{p} a_i \; . \; B(n-i) \right]^2 \qquad (3.25)$$

Trouver les a_i qui minimisent E revient à annuler les dérivées partielles de E par rapport à ces coefficients eux-mêmes:

$$\frac{\partial E}{\partial a_j} = -2 \sum_{n=-\infty}^{+\infty} \left[B(n) - \sum_{i=1}^{p} a_i \; B(n-i) \right] . B(n-i)$$

$$= -2 \sum_{n=-\infty}^{+\infty} B(n-j) \; . \left[B(n) - \sum_{i=1}^{p} a_p(i) \quad B(n-i) \right] \qquad pour \qquad j = 1 \ldots p.$$

$$= 2 \left\{ \sum_{i=1}^{p} \left[a_i \; . \sum_n B(n-i) \; . \; B(n-j) \right] - \sum_n B(n) \; . \; B(n-j) \right\}$$

$$= 2 \left\{ \sum_{i=1}^{p} a_i \; . \sum_m B(m) \; . \; B(m+i-j) - \sum_n B(n) \; . \; B(n-j) \right\}$$

$$\frac{\partial E}{\partial a_j} = 0 \Rightarrow \sum_{i=1}^{p} \left[a_i \; . \sum_m B(n) \; . \; B(m+i-j) \right] = \sum_n B(n) \; B(n-j) \qquad pour \; j = 1 \ldots p. \qquad (3.26)$$

Ce sont les équations de Yule-Walker.

On sait que pour un signal stationnaire, la fonction d'autocorrélation vérifie :

$$R(k) = R(-k) = \sum_l x(l) \; . \; x(l+k) \qquad (3.27)$$

Donc, les équations de Yule-Walker deviennent :

$$\sum_{i=1}^{p} a_p(i) \; . \; R(i-j) = R(j) \qquad pour \qquad j = 1 \ldots p. \qquad (3.28)$$

En utilisant la notation matricielle on aura :

$$\begin{bmatrix} R(0) & R(1) & \bullet & \bullet & \bullet & R(p-1) \\ R(1) & R(0) & & & & R(p-2) \\ \bullet & & & & & \bullet \\ \bullet & & \bullet & & & \bullet \\ \bullet & & & \bullet & & \bullet \\ R(p-2) & & & & \bullet & R(1) \\ R(p-1) & \bullet & \bullet & \bullet & \bullet & R(0) \end{bmatrix} \bullet \begin{bmatrix} a(1) \\ a(2) \\ \bullet \\ \bullet \\ \bullet \\ a(p-1) \\ a(p) \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \bullet \\ \bullet \\ \bullet \\ R(p-1) \\ R(p) \end{bmatrix}$$

Donc : $$R^{(p,p)} \bullet A = R^{(p)} \qquad (3.29)$$

La matrice est R(p,p) est dite Toeplitz puisqu'elle vérifie :

- La symétrie.
- Elle est définie positive.
- Les éléments situés sur une diagonale parallèle à la diagonale principale sont identiques.

Plusieurs algorithmes ont été proposés pour la résolution de ce système. Le plus utilisé est celui de Levinson-Durbin [BBD00].

La détermination des coefficients a_i du filtre AR revient à la résolution des équations de Yulle-Walker, pour un ordre de prédiction p.

La figure 3.14 montre les coefficients a_i du filtre AR obtenus pour un ordre de prédiction p=26.

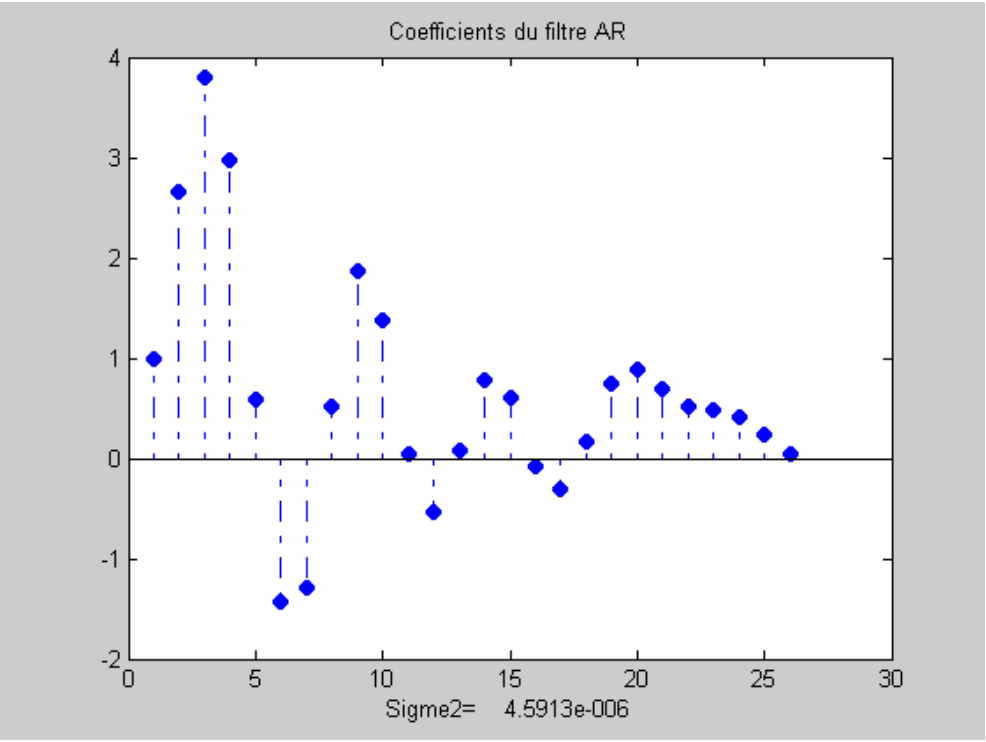

Figure 3.14: Coefficients a_i du filtre AR

3.2.2. Estimation de σ^2 :

La variance du bruit σ^2 est déduite par la relation:

$$\sigma^2 = R[0] + \sum_{k=1}^{p} a_k R[k] \qquad (3.30)$$

où R est la fonction d'auto-covariance biaisée de la partie bruité.

Le tableau suivant résume le nombre de paramètres du modèle HNM à estimer pour chaque trame du signal, selon le voisement. Il est à constater que pour les trames voisées, le nombre de paramètres à estimer change selon la fréquence fondamentale.

Paramètres	**Trame voisée**	**Trame non voisée**
f_0	*1*	*0*
f_m	*1*	*0*
A_k	*2L*	*0*
φ_k	*2L*	*0*
Modèle AR	*p*	*p*
Variance	*1*	*1*

Tableau 3.1 : Paramètres estimés pour chaque trame d'analyse

Chapitre IV
Synthèse par le Modèle Harmonique plus Bruit

Introduction

En phase de synthèse, les transformations sont appliquées au signal de parole afin de réaliser la conversion souhaitée.

La synthèse est une opération consistant à calculer les échantillons du signal de parole à partir des paramètres HNM modifiés ou non.

À partir du modèle harmonique plus bruit, certaines transformations peuvent être appliquées systématiquement sur le modèle afin d'obtenir une voix intelligible et caractéristique d'un locuteur.

Les étapes précédentes ont permis de déterminer, autour de chaque instant d'analyse, un jeu de paramètres $\left(f_0, L, A_k, \varphi_k, f_m, g, \sigma^2\right)$, il nous reste maintenant à construire le signal de synthèse *s(n)*, avec ou sans modifications.

1. Construction du signal de synthèse:

Une fois le modèle harmonique plus bruit construit, la synthèse se fait relativement simplement. Les harmoniques sont synthétisés point par point en calculant l'amplitude et la phase de chaque harmonique *k* au temps *t*. L'amplitude instantanée est calculée par une interpolation linéaire entre les points d'analyse. La phase est calculée en estimant la fréquence instantanée linéairement entre le milieu des fenêtres d'analyse originales.

Le bruit est synthétisé en filtrant un bruit blanc pour lui donner le spectre mesuré et régularisé de la partie bruitée originale. Cette composante du signal de parole est synthétisée fenêtre par fenêtre, en juxtaposant les segments en appliquant la même fenêtre de Hamming utilisée lors de l'analyse, pour la décomposition harmonique plus bruit. Cette technique dite "d'addition recouvrement", nous permet de retrouver la partie bruitée originale.

2. Synthèse sans modifications:

La synthèse est effectuée de manière synchrone à la fréquence fondamentale du signal. Quand aucune modification du signal de parole n'est à faire, les instants (marques) de synthèse (t^i_s) coïncident parfaitement avec les instants d'analyse (t^i_a) (Figure 4.1).

En utilisant les instants de synthèse (t^i_s), la partie harmonique est obtenue par la technique de la synthèse additive et la partie du bruit par la technique du recouvrement addition. Une fois les parties harmonique et bruitée calculées, le signal synthétique serait obtenu en additionnant simplement ces deux parties.

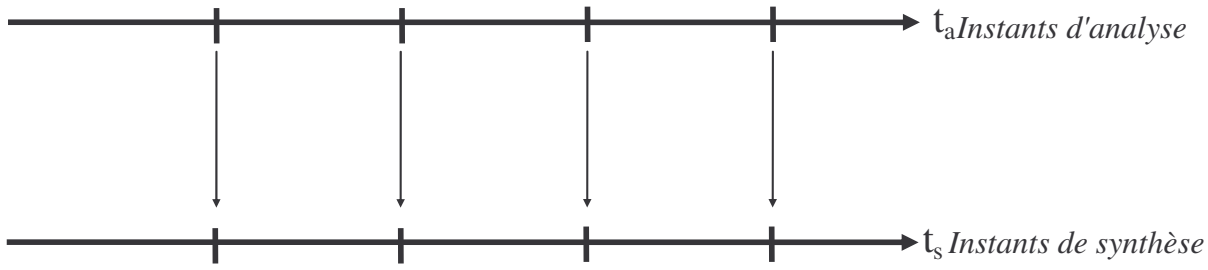

Figure 4.1: Coïncidence des instants de synthèse (t^i_s) et les instants s'analyse (t^i_a) lors de la synthèse sans modifications.

2.1 Synthèse de la partie harmonique:

La synthèse de la partie harmonique se fait échantillon par échantillon. Les amplitudes et les phases sont interpolées linéairement entre les instants de synthèse. Soient $[A^i_k, \varphi^i_k, f^i_0]$ et $[A^{i+1}_k, \varphi^{i+1}_k, f^{i+1}_0]$ les paramètres de la $k^{ème}$ harmonique aux instants de synthèse t^i_s et t^{i+1}_s. Les amplitudes instantanées $A_k(t)$ sont données par:

$$A_k(t) = A^i_k + \frac{A^{i+1}_k - A^i_k}{t^{i+1}_s - t^i_s} t \text{ , pour } \quad t^i_s \le t \le t^{i+1}_s \tag{4.1}$$

Pour ce qui concerne la phase, premièrement, la phase à l'instant t^{i+1}_s est prédite à partir de celle de l'instant t^i_s par

$$\hat{\varphi}^{i+1}_k = \varphi^i_k + k 2\pi \overline{f^i_0} (t^{i+1}_s - t^i_s), \tag{4.2}$$

Avec

$$\overline{f^i_0} = \frac{f^{i+1}_0 + f^i_0}{2}.$$

Puis cette phase est augmentée par un multiple M_k de 2π pour approcher la valeur prédite. M_k est donné par

$$M_k = \left\langle \frac{1}{2\pi} \left(\hat{\varphi}^{i+1}_k - \varphi^{i+1}_k \right) \right\rangle, \tag{4.3}$$

où le symbole $\langle \bullet \rangle$ représente l'entier le plus proche. La phase instantanée est alors donnée par une simple interpolation linéaire:

$$\varphi_k(t) = \varphi_k^i + \frac{\varphi_k^{i+1} + 2\pi M_k - \varphi_k^i}{t_s^{i+1} - t_s^i} t \quad \text{pour} \quad t_s^i \le t \le t_s^{i+1} \tag{4.4}$$

Dans cette analyse, nous avons supposé que la trame *i* et la trame *i+1* sont toutes deux voisées. Si la trame *i* est voisée et la trame *i+1* est non voisée, les amplitudes de la trame *i+1* sont mises à zéro, i.e., $A_k^{i+1} = 0, \forall k$, tout en gardant la même fréquence fondamentale, i.e., $f_0^{i+1} = f_0^i$. Les phases sont alors données par:

$$\varphi_k^{i+1} = \varphi_k^i - k2\pi f_0^i (t_s^{i+1} - t_s^i) . \tag{4.5}$$

Dans le cas où la trame i est non voisée et la trame *i+1* est voisée, $A_k^{i+1} = 0, \forall k$ et $f_0^{i+1} = f_0^i$, alors la phase est donnée par

$$\varphi_k^i = \varphi_k^{i+1} + k2\pi f_0^i (t_s^{i+1} - t_s^i) \tag{4.6}$$

Dans la procédure d'interpolation décrite ci-dessus, nous avons supposé que les deux trames ont le même nombre d'harmoniques. Or, comme la fréquence maximale de voisement et la fréquence fondamentale sont variables dans le temps, le nombre d'harmoniques n'est pas nécessairement le même. Pour contourner ce problème, on complète par des harmoniques d'amplitudes nulles pour avoir le même nombre d'harmoniques. Les phases sont définies comme précédemment par les équations (4.5) ou (4.6) selon les cas.

Ayant déterminé est amplitudes instantanées $A_k(t)$ et les phases instantanées $\varphi_k(t)$ la partie harmonique sera donnée par:

$$\hat{h}[t] = \sum_{k=1}^{L} A_k(t) \cos(\varphi_k(t)) . \tag{4.7}$$

La procédure de la synthèse de la partie harmonique est illustrée dans le schéma de la figure (4.2). Notons que la partie harmonique $\hat{h}[t]$ modifiée occupe la même bande de fréquences que la partie harmonique originale $h[t]$, ce qui fait que, dans le cas de modification de la fréquence fondamentale, elle ne contient pas le même nombre d'harmoniques. Le nouveau nombre d'harmoniques est obtenu par la division de la fréquence maximale de voisement f_m par la fréquence fondamentale f_0.

$$L = \left\langle \frac{f_m}{f_0} \right\rangle \tag{4.8}$$

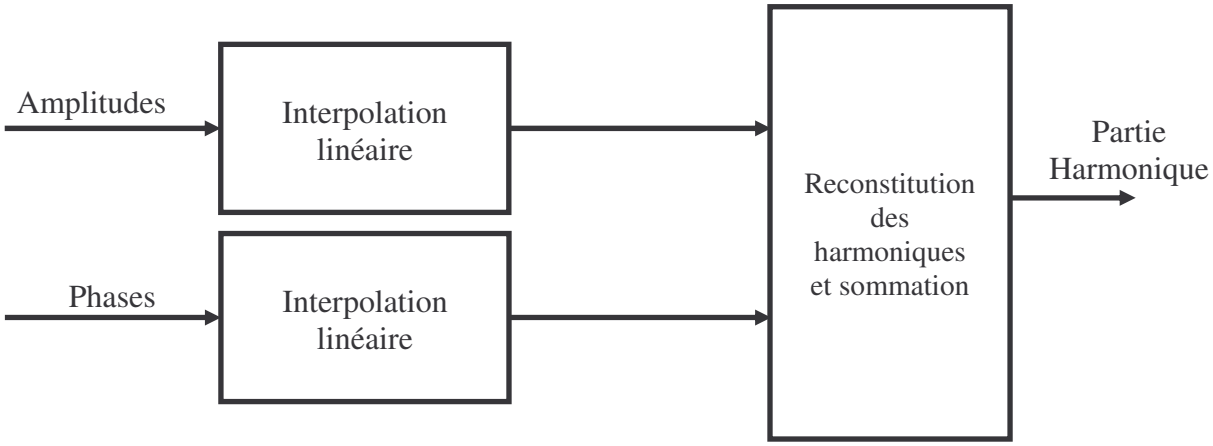

Figure 4.2: **Synthèse de la Partie Harmonique**

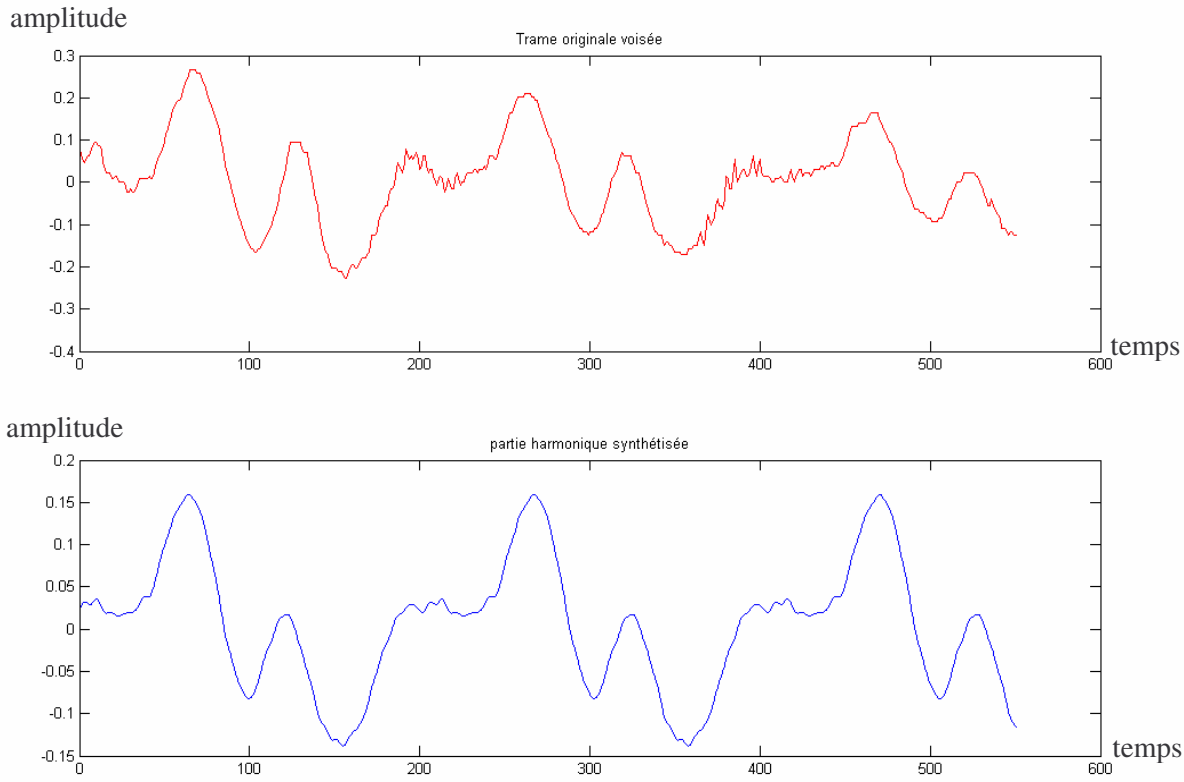

Figure 4.3: composante harmonique synthétisée (en bas) comparée au signal analysé (en haut).

La figure (4.3) permet la comparaison de la partie harmonique synthétisée avec le signal voisé original.

2.2 Synthèse de la partie bruit:

Le bruit est localisé sur la plage comprise entre la fréquence maximale de voisement f_m et la moitié de la fréquence d'échantillonnage $f_e/2$. Il s'agit de filtrer un bruit blanc gaussien par un filtre AR tel que présenté dans la figure (4.4). Le signal à court-terme résultant est alors agencé aux signaux précédents par la méthode d'addition-recouvrement.

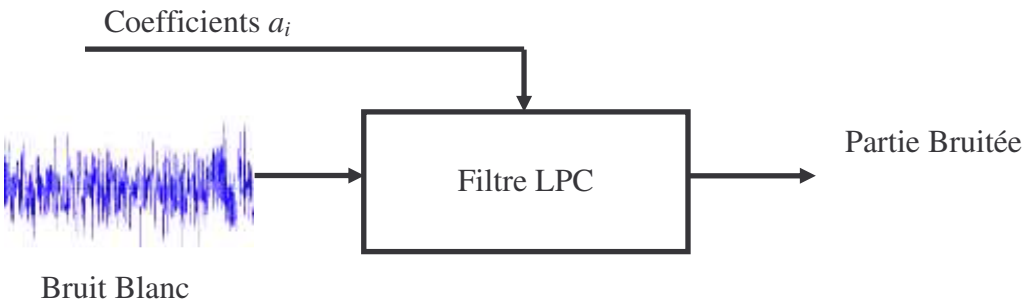

Figure 4.4: Synthèse de la Partie Bruitée

Dans la figure (4.5), on peut comparer le bruit contenu dans une trame voisé,ou encore la trame non-voisée avec le bruit synthétisé par la technique décrite précédemment.

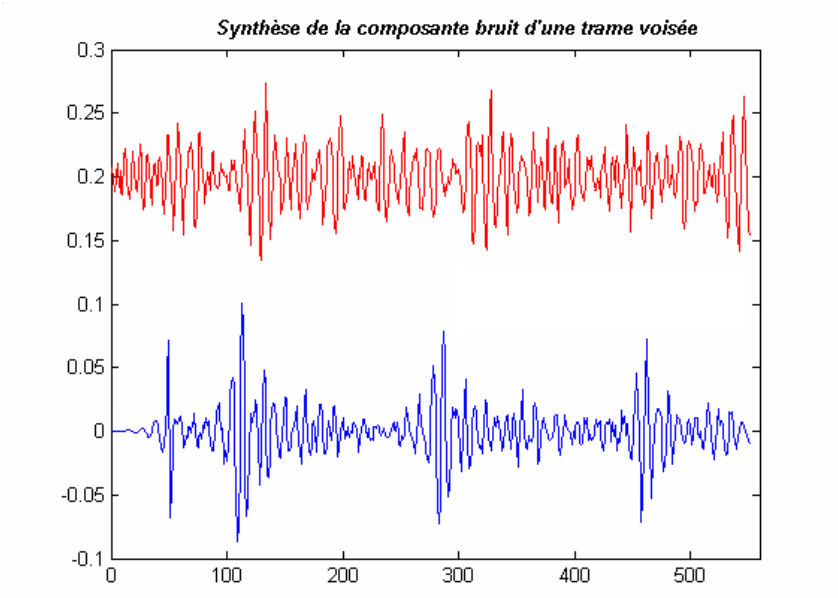

Figure 4.5: Bruit de synthèse (en haut) comparé à la composante bruit (en bas).

Le signal de synthèse global correspond à la somme des deux composantes obtenues, trame par trame (figure 4.6), et ainsi nous obtiendrons le signal complet identique au signal original (figure 4.7).

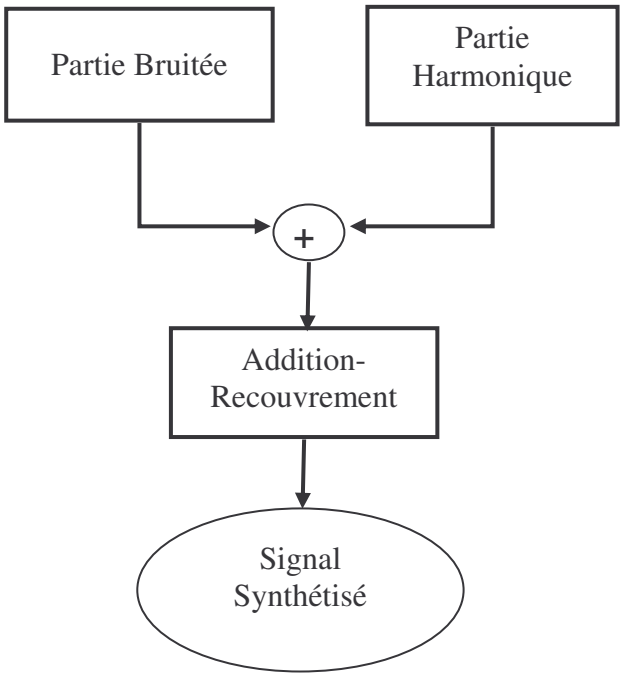

Figure 4.6: Construction du signal de synthèse

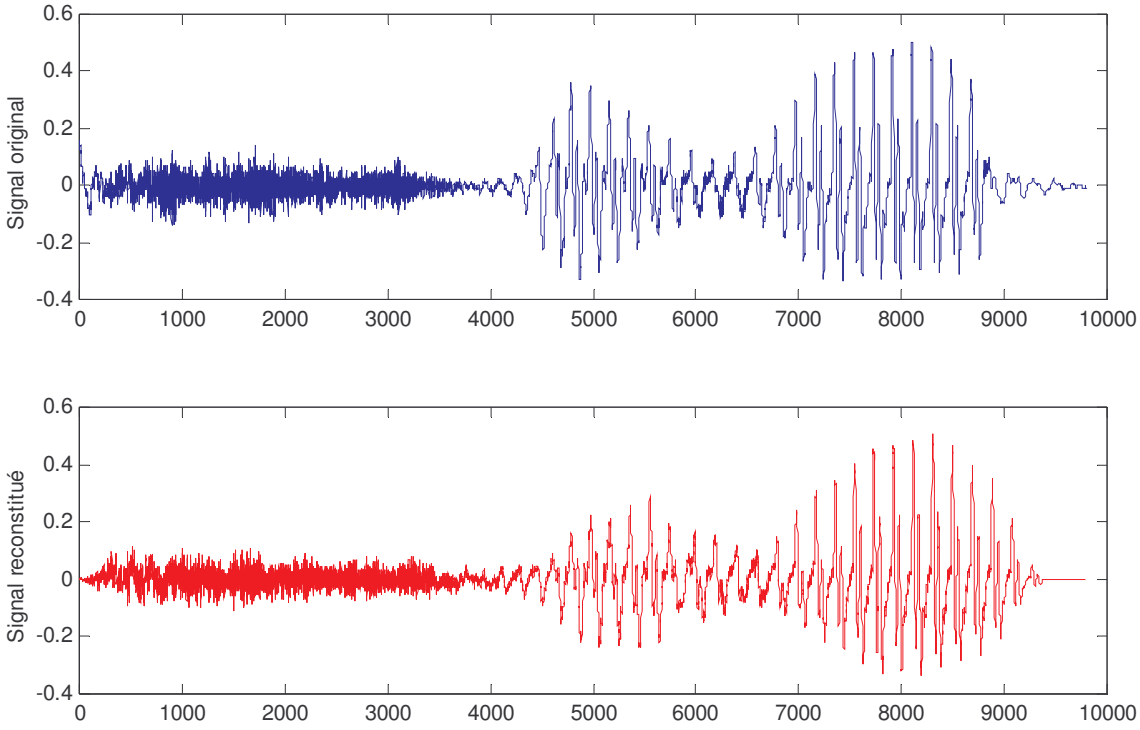

Figure 4.7 : Exemple comparatif d'un signal de parole original avec un signal de synthèse.

3. Synthèse avec modifications des paramètres prosodiques:

Les modifications prosodiques qu'on est amenées à faire sont les modifications de l'échelle temporelle et fréquentielle. Le schéma de la synthèse synchrone à la fréquence fondamentale du HNM prévoit l'usage d'une technique simple et flexible pour les modifications de l'échelle temporelle et du pitch. La première étape consiste à calculer les nouveaux instants de synthèse (t^i_s) (ou marques synthèse) à partir des instants d'analyse et des facteurs des modifications désirées. En utilisant les nouveaux instants de synthèse, le signal synthétique modifié est obtenu de la même manière que précédemment (synthèse sans modification).

3.1 Modification de l'échelle temporelle:

L'objectif de la modification de l'échelle temporelle est de changer le rythme d'articulation sans affecter le contenu spectral, autrement dit, l'évolution du fondamental et la structure formantique devraient être conservées.

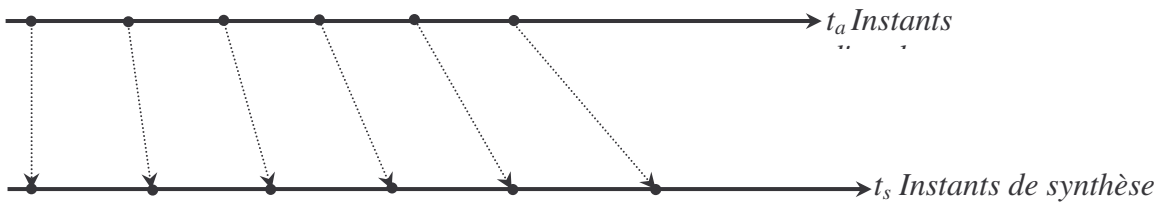

Figure 4.8 : Répartition des instants de synthèse (t^i_s) pour une modification de la durée par un facteur de 1.2

Le changement de rythme se fait en rapprochant ou en espaçant les centres des fenêtres d'analyse lors de la resynthèse (Figure 4.8). La partie harmonique est synthétisée normalement, en interpolant les fréquences et amplitudes entre les points d'analyse. La partie bruitée est synthétisée normalement, en ajustant la largeur des fenêtres de Hamming pour qu'elles correspondent à l'espacement des fenêtres. La figure (4.9) montre un exemple de ralentissement du signal de parole.

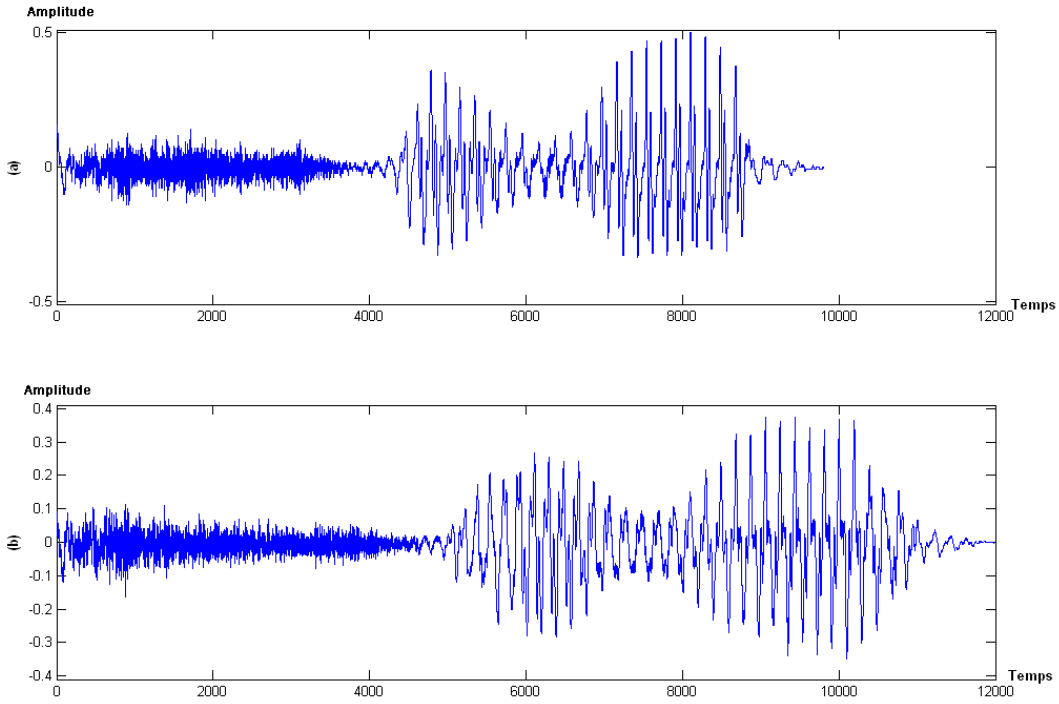

Figure 4.9: (a) Signal de parole original, (b) Signal de synthèse après modification de la durée (facteur 1.2).

3.2 Modification de la fréquence fondamentale :

Le changement de fréquence fondamentale par un coefficient α se fait en synthétisant des harmoniques de fréquence $\alpha k f_0$. L'amplitude de chaque harmonique doit suivre l'enveloppe de puissance harmonique lorsqu'elle est disponible, sinon l'amplitude est interpolée linéairement entre les mesures des amplitudes des harmoniques originaux. La partie bruitée est synthétisée sans modification (Figure 4.10).

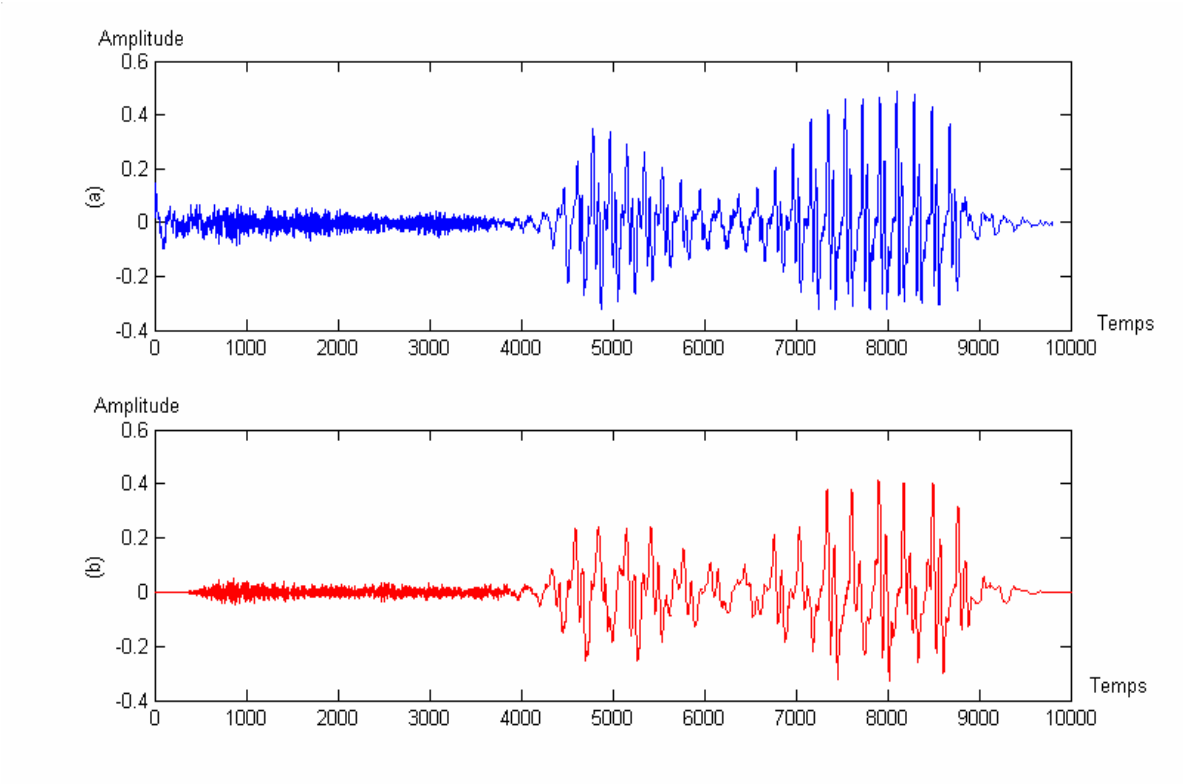

Figure 4.10: (a) Signal de parole original, (b) Signal de synthèse après modification du pitch.

3.3 Conservation de l'enveloppe spectrale :

La transposition de la fréquence fondamentale d'un facteur α se fait généralement sans altérer le timbre significativement lorsque α est entre 0,6 et 1,7. Lorsque ce coefficient est plus grand que 2, l'altération est souvent perceptible, car l'enveloppe de puissance harmonique diffère selon le pitch.

4 Autres types de modifications :

D'autres transformations de la voix qui n'entrent pas dans le cadre de notre recherche sur les voix réalistes sont possibles telles que la modification du locuteur.

Le changement de locuteur consiste à modifier un signal de parole prononcé par un locuteur d'une façon telle qu'il semble, à l'écoute, être prononcé par un locuteur différent. Le terme de modification (ou conversion) indique que l'objectif consiste à transformer les signaux de parole enregistrés par un locuteur, dit de référence, de telle façon qu'ils deviennent aussi proches que possible sur le plan acoustique de ceux enregistrés par un autre locuteur dit cible (ayant prononcé le même texte).

La modification du locuteur est fortement liée à la reconnaissance du locuteur (c'est à dire l'identification d'une personne à partir de sa voix. Il a été en effet possible de mettre en évidence les principaux éléments qui permettent de distinguer les voix de locuteurs différents [Dod85][Fur86] [RS91]. Parmi les caractéristiques des signaux de parole liées à l'identité du locuteur, on distingue, en général, les caractéristiques spectrales et les caractéristiques prosodiques.

La modification du locuteur est une application qui devient envisageable du fait de l'émergence de systèmes d'analyse/synthèse de signaux de parole de bonne qualité. En effet, la modification du locuteur implique des modifications fines des caractéristiques acoustiques du signal de parole. Pour cela, il est nécessaire de pouvoir extraire du signal des paramètres permettant d'avoir accès aux caractéristiques spectrales et prosodiques (étape d'analyse). Inversement,il faut être capable d'obtenir un signal de parole de bonne qualité à partir des paramètres modifiés (synthèse).

conclusion et perspectives

CONCLUSION ET PERSPECTIVES

Le travail réalisé au cours de cette thèse a porté sur le développement d'un système de modification des paramètres prosodiques par le modèle harmonique plus bruit.

Nous avons utilisé le modèle HNM (Harmonic plus Noise Model) pour représenter la voix. Ce modèle nous a semblé approprié pour reproduire une voix d'assez bonne qualité pour retrouver les caractéristiques acoustiques d'un locuteur. La qualité de la représentation de la voix est importante dans notre cas, parce que nous voulons que la synthèse soit réaliste et non seulement intelligible.

Le but de ce modèle est de séparer la voix en deux signaux : un signal périodique harmonique et un bruit. La partie harmonique permet de modéliser certaines caractéristiques acoustiques dues aux cordes vocales (comme la hauteur du son) et du conduit vocal. La partie bruitée permet de modéliser le reste du signal, principalement dû aux turbulences de l'écoulement de l'air des poumons dans le conduit vocal.

Le système est principalement composé de trois parties :

- Estimation de la fréquence fondamentale ;
- Mesure de la partie harmonique ;
- Mesure de la partie bruitée ;

Pour développer un tel système, une bonne connaissance des mécanismes de production de la parole et des paramètres acoustiques caractérisant le signal de parole, ainsi que des principes d'un système de synthèse vocal est nécessaire. Ce fut l'objet du premier chapitre.

Le deuxième chapitre a dressé un aperçu de quelques modélisations du signal de parole et a expliqué les principes des systèmes d'Analyse/Modification/Synthèse de la parole.

Dans le chapitre 3 nous avons étudié de plus près les étapes d'analyse des paramètres du modèle adopté qui consiste en l'extraction du jeux de paramètres régissant le modèle.

Dans le chapitre 4, nous nous sommes penchés sur la synthèse du signal de la parole par le modèle choisi.

La modélisation par les deux composantes harmonique et bruit conduit à une description par un nombre réduit de paramètres directement en relation avec les propriétés acoustiques du

signal de parole. D'autre part, une telle modélisation a permis un processus de synthèse simple et une bonne flexibilité lors de l'application de transformations de la voix telles que les modifications de prosodie. Cette représentation du signal de parole peut aussi être considérée comme un pré-encodage par le fait que le signal est modélisé par un nombre réduit de paramètres, et s'adapte à des procédures de codage à débit réduit.

D'autres transformations de la voix sont possibles telles que la modification du locuteur qui est une application qui devient envisageable et plus intéressante. En effet, la modification du locuteur implique des modifications fines des caractéristiques acoustiques du signal de parole. Pour cela, il est nécessaire de pouvoir extraire du signal en plus des paramètres prosodiques, des paramètres permettant d'avoir accès aux caractéristiques spectrales dans l'étape d'analyse. Inversement et lors de la synthèse, il faut être capable d'obtenir un signal de parole de bonne qualité à partir des paramètres modifiés d'une façon telle qu'il semble, à l'écoute, être prononcé par un locuteur différent.

Ce modèle constitue un outil efficace pour des applications multiples exigeant une bonne qualité de la voix.

Bibliographie

BIBLIOGRAPHIE

[ASK 90] M. Abe, K. Shikano et H. Kawabara - « Cross-language voice conversion », *Proc. of ICASSP*. p. 345-348, 1990.

[BBD 00] BOITE, R., BOURLARD, H., DUTOIT, T., HANCQ, J. et LEICH, H., "Traitement de la parole", Presses polytechniques et universitaires romandes, 2000.

[CHA89] CHARPENTIER, F., et MOULINES, E., "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones," », *Proc. of ICASSP*. p. 13-19, 1989.

[DOD85] G. Doddington. "Speaker recognition - Identifying people by their voices.", Proc. IEEE 73, no11, p.1651-1664, 1985.

[DOV 94] B. Doval." Estimation de la Fréquence Fondamentale des Signaux Sonores." PhD thesis, Université de Paris VI, Mars 1994.

[DUT 93] T. DUTOIT, H. LEICH, "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, Elsevier Publisher,

[SNB96] A. Schmidt-Neilsen et D. Brock - « Speaker recognizability testing for voice coders », *Proc. of ICASSP* 2. no. 1249-1152. 1996.

[FUR 86] S. Furui. "Research on indivituality features in speech waves and automatic speaker recognition techniques". Speech Communication 5, no.2, p.183-197, 1986.

[HAH 01] HUANG, X., ACERO, A. et HON, H.-W., "Spoken Language Processing : A Guide to Theory, Algorithm and System Development", Pearson Education ; 1st edition,2001.

[HES 83] W. Hess." Détermination du fondamental du signal de parole, algorithmes et méthodes" . In J. Mariani C. Berthommier, Y. Grenier, editor, Traitement du Signal de Parole, pp. 35-47, 1983.

[KLP 95] KLEIJN, W.B. et PALIWAL, K.K., "Speech Coding and Synthesis", Elsevier, Amsterdam. 1995.

[MAE 79] MAEDA.S “Articluar model of th tondue based on statistical analysis”, The journal of the Acoustical Society Of America,65(S22), 1979.

[MQ 95] R. McAulay et T. Quatieri “ Speech coding and synthesis”», Sinusoidal coding, elsevier science, p. 121-173, 1995.

[MOC 90] MOULINES, E. et CHARPENTIER, F., "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones," Speech Communication Vol. 9., numéro 5, pp 453-467, 1990.

[OUD 98] Oudot M., "Etude du modèle Sinusoïdes et Bruit pour le traitement des signaux de parole, Estimation robuste de l'enveloppe spectrale », Thèse de doctorat de l'ENST paris, France, Novembre 1998.

[POL 90] POLS L., “Does improved performance of a rule synthesizer also contribute to more phonetic knowledge?”, Proceedings of the ESCA tutorial day on speech synthesis, p. 50-54, 1990.

[QUA 01] QUATIERI T., "Discrete-Time Speech Signal Processing : Principles and Practice", Prentice Hall, 2001.

[RAS78] RABINER, L.R. et SCHAFER, W., "Digital Processing of Speech Signals", Pearson Education. 1978.

[RIC 95] RICHARD G., LIU M. SINDER D., DUNCAN H., LIN Q., FLANAGAN J., LEVINSON S., SLIMON S., “Numérical simulations of fluid flowing the vocal tract”, Proceedings of the Eurospeech Conference, p.1297-1300, 1995.

[RS 91] A. E. Rosenberg and F. K. Soong. Recent research in automatic speaker recognition.In S. Furui and M. Sondhi, editors, Advances in Speech Signal Processing, chapter 22, pages 701-738. Marcel Dekker, 1991.

[SER 97] X. Serra “Musical Sound Modeling with Sinusoids plus Noise”, C. Roads, S. Pope, A. Picialli, G. De Poli, editors. 1997.

[STE 90] STEVEN K. N., “Control parameters of synthesis by rule”, Proceedings of the ESCA tutorial day on speech synthesis, p. 27-37, 1990.

[STY 96] Y. Stylianou « Harmonic plus noise models for speech, combined with statistical methods for speech and speaker modifications », Thèse, Telcom. Paris, Janvier 1996.

[YS 98] P. Yang et Y. Stylianou “ Real time voice alteration based on linear prediction”, *Proc. of ICSLP*. p. 1667-1670. 1998.