Modélisation Sinusoïdale à Long Terme du Signal de Parole

Mohammad Firouzmand

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : **Signal, Images, Parole, Télécoms**

préparée au laboratoire : **Institut de la Communication Parlée, UMR CNRS 5009**

dans le cadre de **l'Ecole Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal**

présentée et soutenue publiquement par

Mohammad Firouzmand

le 6 avril 2007

Modélisation Sinusoïdale à Long Terme du Signal de Parole

Directeur de Thèse : Laurent GIRIN (MCF/HDR INPG/ICP)

JURY

M. Pascal PERRIER	Président
M. Frédéric BIMBOT	Rapporteur
M. Gaël RICHARD	Rapporteur
Mme Myriam DESAINTE-CATHERINE	Examinatrice
M. Olivier ROSEC	Examinateur
M. Laurent GIRIN	Directeur de thèse

Remerciements

Je voudrais tout d'abord remercier chaleureusement mon directeur de thèse Laurent. Sans doute, ce travail aurait été impossible sans toi. Je voudrais te remercier aussi pour le temps et la patience que tu m'as accordés tout au long de ces années, et pour les conseils que tu m'as divulgués tout au long de la rédaction. Encore un grand merci. Je penserai toujours à toi !

Je voudrais également remercier tous les membres de mon jury pour avoir pris le temps de lire et de critiquer ce manuscrit : Frédéric Bimbot et Gaël Richard pour avoir apporté votre caution et vos remarques en rapportant mon travail, Pascal Perrier pour avoir accepté de présider le jury, et Oliver Rosec et Myriam Desainte-Catherine pour avoir bien voulu examiner ce travail.

Je voudrais particulièrement remercier Madame Gaude pour ses précieux conseils, pour avoir relu attentivement les premiers brouillons de ce manuscrit, et corrigé les nombreuses fautes d'orthographe et de syntaxe.

Je tiens aussi à remercier spécialement Jean-Luc Schwartz, Directeur de l'ICP, qui m'a accepté dans son laboratoire. Je remercie également tous les membres de l'ICP qui m'ont chaleureusement accueilli, notamment Pierre Badin, Pascal Perrier, Gérard Bailly, Annemie Van Hirtum, Nadine Bioud, Dalila, Nino Medves, Christian Bulfone, Monique Revil, et j'en oublie certainement…

Je voudrais aussi profiter de cette occasion pour remercier mes collègues thésards et jeunes chercheurs du labo : Bertrand Rivet, Xavier Grandchamp, Antoine Serrurier, Anahita Basirat, Virginie Attina, Guillaume Gibert, Nicolas Ruty, Pauline Welby, et Albert Rillard.

Un grand merci à ma famille, et enfin je termine ces remerciements en remerciant le Dieu qui m'a donné cette occasion de passer une partie de ma vie avec toutes ces personnes.

Merci à toutes et merci à tous.

Mohammad
06 avril 2007

Table des matières

Glossaire

AAC : *Advanced Audio Coding*
CELP : *Code Excited Linear Predictive*
DCT : *Discrete Cosine Transform*
DCT-2D : *Discrete Cosine Transform in two dimensions*
F_e : Fréquence d'échantillonnage du signal
F_0 / ω_0 : Fréquence fondamentale du signal dans le cas de sons de parole voisés
FFT : *Fast Fourier Transform*
HB : *Harmonic Bandwidth* (largeur de bande harmonique)
ICP : Institut de la Communication Parlée
LPC : *Linear Prediction Coding*
LSF : *Line Spectral Frequencies*
MBE : *Multi-Band Excitation coder*
MCD : Modèle en Cosinus Discrets
MCDL : Modèle en Cosinus Discrets plus terme Linaire
MCDP : Modèle en Cosinus Discrets plus Polynômes
MCSD : Modèle en Cosinus et Sinus Discrets
MCSDL : Modèle en Cosinus et Sinus Discrets plus terme Linaire
MDCT : *Modified Discrete Cosine Transform*
MFS : Modulation de Fréquence Sinusoïdale
MM : Multi-Modèles
MMSE : *Minimum Mean Square Error*
MP : Modèle Polynomial
MPEG : *Moving Picture Expert Group*
NST : *Non Square Transform*
SMS / MSP : *Sinusoidal Model of Speech* / Modèle Sinusoïdal de Parole
RSB : Rapport Signal à Bruit
SPL : *Sound Pressure Level*
STC : *Sinusoidal Transform Coder* (codeur basé sur le MSP)
STFT / TFCT : *Short-Term Fourier Transform* / Transformée de Fourier à Court Terme
TFD : Transformée de Fourier Discrète
TFDI : Transformée de Fourier Discrète Inverse
VQ / QV : *Vector Quantization* / Quantification Vectorielle
WMMSE : *Weighted Minimum Mean Square Error*
1D : (en) une dimension
2D : (en) deux dimensions

Introduction

Cette thèse s'inscrit dans le cadre général de la modélisation paramétrique des signaux de parole en particulier, et des signaux audio en général. Plus particulièrement, nous nous plaçons dans le cadre des modèles paramétriques spectraux et en particulier dans celui du modèle sinusoïdal/harmonique. Ce modèle est usuellement défini « à court terme », c'est-à-dire sur des trames successives de signal d'une durée sur laquelle ce dernier peut être considéré comme stationnaire (de l'ordre de 10 à 30 ms). Cette thèse apporte une contribution nouvelle à ce domaine en ajoutant à ce niveau traditionnel de la modélisation spectrale un niveau de modélisation supplémentaire le long de l'axe temporel : dit très brièvement, on cherche à modéliser les trajectoires de paramètres sinusoïdaux sur des durées significativement plus longues que celles des trames à court terme. Comme précisé par la suite, cette approche se fonde sur la corrélation naturelle entre les valeurs de paramètres correspondant à des trames à court terme successives. L'objectif sous-jacent est bien sûr d'exploiter et de décrire efficacement cette corrélation.

Ce principe de modélisation temporelle n'est pas nouveau en soi : dans les systèmes d'analyse-synthèse usuels, les paramètres spectraux correspondant aux fenêtres d'analyse à court terme successives sont déjà interpolés temporellement sur des fenêtres de synthèse successives (voir le Chapitre 1). Mais ces fenêtres de synthèse sont aussi des fenêtres à court terme : les paramètres sont interpolés d'une mesure à la suivante, ce qui limite l'exploitation de leur corrélation temporelle. L'objectif de cette thèse est l'élaboration et l'implantation de modèles temporels paramétriques des descripteurs du signal (eux-mêmes paramétriques) à l'échelle de ce qu'on appelle dans ce document le « long terme », c'est-à-dire sur des durées de l'ordre de la centaine de millisecondes, voire de la seconde, plutôt que de la dizaine de millisecondes. Nous considérerons donc l'évolution du signal sur une fenêtre temporelle significativement plus large que les fenêtres à court terme utilisées généralement pour suivre les non-stationnarités du signal. De ce fait, on modélise l'évolution du signal globalement sur un ou plusieurs sons de parole (*i.e.* les réalisations acoustiques des phonèmes), plutôt que sur une partie d'un son. Les modèles à long terme que nous proposons sont ainsi plus généraux, et d'une certaine façon, plus simples et plus souples que les modèles temporels à court terme traditionnels. Cette « modélisation à long terme » doit permettre de répondre aux objectifs suivants :

- Elle doit d'abord fournir une représentation du signal en accord avec la physique des signaux, notamment avec leur caractère à la fois non stationnaire et suffisamment régulier (du fait de la corrélation temporelle), du moins sur les sections de signal considérées. Autrement dit, le modèle temporel considéré est un modèle de trajectoire en constante évolution mais avec des variations relativement modérées, comme l'est le signal naturellement. Par exemple, les trajectoires de fréquence modélisées par cette approche à long terme sont décrites par des

fonctions définies en chaque point, sans contrainte locale de stationnarité comme dans les modèles à court terme d'ordre zéro où la fréquence est localement constante, ce qui est, en toute rigueur, une hypothèse abusive. De plus, par rapport aux modèles à court terme qui peuvent relier des mesures faussées car bruitées, nous verrons que les modèles à long terme proposés sont des modèles intrinsèquement « lisses » dans le sens où les trajectoires qu'ils encodent sont des courbes lisses avec une évolution régulière au cours du temps. En effet, d'une part les fonctions mathématiques qui composent ces modèles sont elles-mêmes lisses : par exemple, nous utiliserons typiquement des fonctions trigonométriques ou polynomiales. Et d'autre part, ces modèles reposent sur un ajustement global suivant l'ensemble des mesures du début à la fin d'une trajectoire avec un relâchement des contraintes, c'est-à-dire sans forcément vérifier le passage exact par tous les points de mesure (comme c'est le cas pour un certain type de fonctions *splines* par exemple). Cette approche doit ainsi permettre de contribuer à résoudre implicitement le gommage des erreurs de mesure inévitablement fournies par la phase d'analyse des paramètres[1].

En plus de fournir en sortie une bonne synthèse du signal modélisé au sens de la fidélité par rapport au signal original grâce au respect de cette évolution naturelle, ces modèles doivent permettre d'aller vers une meilleure compréhension de l'utilisation de la représentation du signal sur laquelle ils sont appliqués (en l'occurrence la représentation sinusoïdale ou harmonique dans cette thèse). Ceci devrait être rendu possible par l'utilisation d'un cadre non-stationnaire « généralisé » par rapport aux approches non-stationnaires à court terme.

- La modélisation à long terme doit fournir en même temps une représentation du signal « parcimonieuse » dans le sens où ce modèle temporel à long terme est lui-même un modèle paramétrique contrôlé par un nombre réduit de paramètres (par rapport au nombre de mesures sur lesquelles il est appliqué). Ceci ouvre bien sûr des possibilités nouvelles dans le domaine du codage : un système exploitant la modélisation à long terme permettra un codage efficace du signal si le nombre de bits nécessaire à la quantification des paramètres du modèle à long terme (où à la quantification d'une représentation équivalente de celui-ci) est significativement inférieur au nombre de bits nécessaire à la quantification des paramètres spectraux initiaux. Le décodeur peut alors reconstruire les paramètres spectraux originaux des trames successives à partir des paramètres du modèle à long terme quantifiés[2].

Bien entendu, le codage à long terme de trames significativement plus longues que la dizaine de millisecondes impose un délai de codage/décodage qui peut limiter drastiquement l'usage d'un tel codeur dans des applications nécessitant une forte interactivité, notamment en communication *full-duplex*. À l'inverse, ce codeur pourrait être exploité dans n'importe quelle application *offline*, notamment celles liées au stockage et à la synthèse de la parole.

---

[1] Bien entendu, il ne s'agit pas non plus de lisser abusivement les trajectoires si le bruit qu'elles comportent est une composante naturelle du signal. Ce point est discuté au Chapitre 6.

[2] On peut mettre ici en parallèle la problématique du codage avec le caractère lisse des modèles mentionné plus haut. Cet aspect lisse permet de filtrer implicitement le bruit de mesure. Or, du point de vue du codage, ce bruit de mesure est une composante à la fois inutile et particulièrement coûteuse à coder, notamment du fait de sa structure temporelle complexe.

Par ailleurs, pour satisfaire la contrainte de parcimonie, nous utiliserons des modèles psychoacoustiques qui permettent d'ajuster la quantité d'information à coder à la sensibilité de l'oreille humaine. L'interaction de la modélisation du signal avec les modèles psychoacoustiques prend un éclairage nouveau dans ce cadre à long terme : comme on le verra dans ce document, nous partirons des critères psychoacoustiques utilisés dans les traitements à court terme pour en proposer une version adaptée au traitement temporel à long terme.

- Les modèles à long terme doivent être suffisamment flexibles pour permettre de réaliser facilement toute une série de transformations du signal, par exemple un étirement ou une compression temporelle, ou encore un changement de *pitch*, tout en garantissant une synthèse du signal de haute qualité. Il est important de noter que ce type de représentation répond par conséquent à un niveau de demande intermédiaire entre les objectifs de transformation du signal et ceux de codage. Ce niveau, intéressant dans la pratique, est peu abordé dans la littérature. En effet, on se rend compte que d'une manière générale en traitement de parole et de signaux audio, plus la représentation d'un signal et les techniques associées sont adaptées au codage au sens où elles concentrent bien l'information dans peu de coefficients, moins elles sont adaptées à la manipulation et la transformation des signaux[3]. Les modèles proposés dans cette thèse pourraient s'avérer être une représentation suffisamment flexible pour remplir les deux objectifs à la fois, ce qui n'est pas si courant.

- Enfin, et il s'agit d'un point un peu particulier, cette modélisation à long terme doit apporter un support efficace aux techniques de tatouage paramétrique développées récemment à l'Institut de la Communication Parlée (ICP) de Grenoble, le laboratoire « hôte » de cette présente thèse, en collaboration avec le Laboratoire Bordelais de Recherches en Informatique (LaBRI) [Girin & Marchand, 2004]. Le tatouage d'un signal (*watermarking* en anglais) consiste à insérer dans ce signal une information binaire « cachée », c'est-à-dire de façon imperceptible à l'utilisateur, tout en étant robuste aux attaques et aux transformations appliquées sur le signal hôte [Kim, 2003]. Les techniques de tatouage développées à l'ICP et au LaBRI sont fondées sur la modulation des trajectoires fréquentielles des composantes du modèle sinusoïdal. Elles nécessitent que ces trajectoires soient suffisamment lisses pour pouvoir correctement greffer et détecter le signal de tatouage (le *watermark*). C'est pourquoi le caractère lisse des modèles à long terme proposés dans cette thèse devrait pouvoir être exploité efficacement par ces techniques de tatouage audio. Ce point est aussi discuté au Chapitre 6.

Pour résumer l'approche développée dans cette thèse, nous dirons que la notion de modèle y prend un sens « à double niveau » un peu particulier. Il ne s'agit plus seulement des modèles spectraux qui représentent directement le signal sur une base à court terme (ici le modèle sinusoïdal) et que l'on qualifie ici de *modèles de premier niveau*. Il s'agit aussi de *modèles de second niveau* qui vont représenter les trajectoires

[3] On voit par exemple toute la difficulté qu'il y a à réaliser la transformation des signaux dans le domaine compressé, c'est-à-dire directement sur les coefficients résultant du codage des signaux [Levine, 1998]. Ce problème précis dépasse le cadre de travail de cette présente thèse.

des paramètres des modèles de premier niveau au cours du temps. On pourrait rétorquer que la synthèse à court terme qui relie usuellement deux mesures à court terme (voir la Section 1.3 de ce document) est aussi une modélisation de second niveau. Mais cette modélisation à court terme est plus proche dans l'esprit du problème de l'interpolation de paramètres, à la fois exacte et localisée, que de leur description « globale » par un modèle « ayant pris un certain recul », comme c'est le cas dans cette thèse[4]. En résumé, dans la présente étude, il s'agit donc de trouver une représentation du signal à la fois spectrale (modélisation de niveau 1) et temporelle (modélisation de niveau 2), et capable de capturer les non-stationnarités du signal de façon à la fois suffisamment fine (mais pas trop !), suffisamment globale (c'est-à-dire sur une longue portion de signal) et suffisamment efficace (au sens de la parcimonie). Cette représentation peut être exploitée dans une série d'applications technologiques telles que le codage, la synthèse de haute qualité incluant la transformation des signaux, et le tatouage.

Conformément au contexte et aux objectifs de cette thèse, ce document est organisé comme suit :

- Comme cette thèse focalise plus précisément sur le modèle sinusoïdal/harmonique de la parole et des signaux audio, en tant que modèle spectral de niveau 1 sur lequel s'applique la modélisation à long terme, nous commençons par décrire ce contexte général d'étude. Le Chapitre 1 est donc dédié à la description du modèle sinusoïdal de la parole et des signaux audio (et ses déclinaisons). Nous y décrivons aussi les principes généraux de la psychoacoustique et des modèles associés pour préparer leur exploitation dans notre étude à long terme.

- Le Chapitre 2 présente les principes généraux de la modélisation à long terme telle que nous l'avons étudiée dans cette thèse. Nous présentons dans ce chapitre un bref état de l'art des techniques de modélisation et de codage prenant particulièrement en compte la dimension temporelle dans les traitements. Nous y donnons la définition de notre approche de la modélisation à long terme, en précisant le type de sections de signal à long terme que nous avons considérées, et nous présentons les modèles à long terme proprement dit. Ensuite, le principe général de l'adaptation des modèles psychoacoustiques à notre problématique est présenté. Partant de là, nous présentons de façon générique un algorithme capable de réaliser l'ajustement optimal des modèles à long terme aux jeux de paramètres des modèles de premier niveau (dans notre étude, amplitudes et phases du modèle sinusoïdal/harmonique).

[4] Notons que dans le même esprit que notre étude, cette modélisation à deux niveaux peut se généraliser à une modélisation multi-niveaux avec des niveaux encore plus élevés. Par exemple, il est possible de modéliser l'évolution d'une trajectoire de fréquence comportant un vibrato par une somme de polynômes (pour la trajectoire de fond) et d'un cosinus (pour décrire le vibrato). Puis on peut encore modéliser l'évolution de la fréquence et de l'amplitude du vibrato/cosinus par de nouveaux modèles. Les perspectives offertes par cette généralisation sont explicitement décrites dans une étude préliminaire sur les signaux de musique par Marchand et Raspaud [Marchand & Raspaud, 2004]. Dans cet article, les termes exacts sont « modèles d'ordre 1 » et « modèles d'ordre 2 » (*order* 1 et *order* 2) à la place de « modèles de premier niveau » et « modèles de deuxième niveau ». La terminologie « niveau » (*level*) à été utilisée ensuite dans [Raspaud *et al*., 2005], une autre étude récente menée au LaBRI, dans laquelle l'ICP a été impliquée.

- Dans les deux chapitres suivants, nous présentons l'application des techniques présentées de façon générale au Chapitre 2, au problème particulier de la modélisation des trajectoires de paramètres sinusoïdaux/harmoniques. Ainsi, le Chapitre 3 est consacré à la modélisation des trajectoires d'amplitude et le Chapitre 4 est consacré à la modélisation des trajectoires de phase. Dans chaque cas, nous donnons le détail de la description du modèle psychoacoustique associé, et nous précisons la forme exacte de l'algorithme d'ajustement introduit au Chapitre 2. Une campagne d'expérimentations portant sur des signaux de parole réelle est présentée pour chaque type de paramètres. Des résultats sont fournis en terme de précision de l'ajustement des modèles, de qualité des signaux modélisés et synthétisés, et de débit de coefficients nécessaires.

- Dans le Chapitre 5, nous généralisons la notion de modélisation sinusoïdale à long terme à la notion de modélisation sinusoïdale en deux dimensions (2D) : pour cela un nouveau niveau de modélisation est proposé et testé, en accord avec la notion de modélisation multi-niveaux mentionnée précédemment. Dans ce cas, ce nouveau niveau concerne la dimension fréquentielle : nous modélisons ainsi d'abord l'enveloppe spectrale de chaque trame à court terme avant d'effectuer la modélisation à long terme des paramètres résultant de la modélisation de l'enveloppe. En d'autres termes, nous remplaçons les paramètres d'amplitude du modèle sinusoïdal/harmonique par des paramètres d'enveloppe spectrale. De nouveaux résultats obtenus avec cette approche sont présentés et discutés.

- Enfin, le Chapitre 6 est dédié à une discussion autour des perspectives ouvertes par ce travail. Des éléments résultant de notre étude ou reliés à celle-ci sont passés en revue. Ainsi, nous discuterons de divers points comprenant notamment l'application de la modélisation à long terme à la transformation des signaux, à leur codage à bas débit, et à leur tatouage (*watermarking*). Nous discuterons aussi largement de l'extension de l'approche à long terme au cadre d'autres modèles spectraux (notamment le modèle LPC), ainsi qu'à la partie bruitée du modèle harmonique + bruit.

Chapitre 1

1. Modélisation sinusoïdale des signaux de parole/audio : principes et contexte de l'étude

Les modèles à long terme qui sont le cœur de cette thèse, et qui seront présentés dans le chapitre suivant, ont été implantés dans le cadre de la modélisation sinusoïdale de la parole. En d'autres termes, dans cette thèse, ces modèles à long terme ont été proposés et utilisés pour modéliser la trajectoire temporelle de paramètres du type amplitudes et phases des composantes sinusoïdales (ou harmoniques) du signal. Une généralisation de l'emploi de ces modèles à long terme pour d'autres types de représentation du signal est bien sûr possible. Elle est d'ailleurs à l'étude à l'ICP, et une série de propositions concernant ce point est donnée au Chapitre 6. Mais dans cette thèse on focalise sur des implantations et des expérimentations réalisées plus spécifiquement dans le cadre de la représentation sinusoïdale. C'est pourquoi ce premier chapitre est consacré à la présentation relativement détaillée de ce cadre de travail. On y décrit ainsi les principes de base de la modélisation sinusoïdale des signaux de parole et des signaux audio, et des phases d'analyse et de synthèse associées.

Plus précisément, ce chapitre est organisé comme suit. Partant du vocodeur de phase, on présente les principes généraux et la (ou plutôt les) définition(s) du modèle sinusoïdal, ainsi que le processus d'analyse-synthèse à court terme exploitant cette représentation du signal. Les contextes de signaux (et donc de modèles) stationnaires et non-stationnaires sont discutés. Les principales méthodes d'analyse et de synthèse sont ensuite présentées. On donne ensuite une brève présentation des principales variantes de ce modèle : la version harmonique et la version dite « sinusoïdale plus bruit » intégrant une composante stochastique pour modéliser d'éventuelles composantes bruitées en plus des composantes sinusoïdales. A la fin du chapitre, les principaux phénomènes psychoacoustiques pouvant intervenir dans le cadre de la perception de signaux sinusoïdaux sont introduits. Ceci permet de préparer l'exploitation ultérieure des propriétés du système auditif humain dans la modélisation à long terme que nous proposons dans ce cadre sinusoïdal.

1.1. Principes généraux

1.1.1. De la transformée de Fourier au vocodeur de phase

La *transformée de Fourier* (voir Section 1.2.1.1) permet de décomposer un signal en une somme de sinusoïdes. Les *fréquences*, les *phases* et les *amplitudes* de ces sinusoïdes sont les résultats de cette transformation. Celle-ci permet donc de passer de la

représentation temporelle d'un son à sa représentation spectrale, en indiquant les variations d'amplitude des composantes spectrales en fonction des fréquences, pour définir le *spectre d'amplitude*. Le *spectre de phase* renseigne lui sur la synchronisation relative des différentes composantes spectrales. Comme la transformée de Fourier repose sur une hypothèse de stationnarité du signal, et que le signal évolue généralement au cours du temps, plusieurs petits segments temporels de ce signal (appelés *fenêtres* ou bien *trames*) sont successivement étudiés : c'est le principe de la *transformée de Fourier à court terme* (TFCT). Pour les signaux numériques, la *transformée de Fourier discrète* (TFD) est employée. Un algorithme rapide (*FFT* pour *Fast Fourier Transform*) permet son application optimisée en calculs. Nous reviendrons plus en détail par la suite sur ces principes généraux lorsqu'ils interviendront directement dans notre étude.

Le *vocodeur de phase* est un modèle développé à partir de la transformée de Fourier à court terme [Flanagan & Golden, 1966]. Il représente le son par une succession de spectres (d'amplitude et de phase) calculés en appliquant la TFD sur des fenêtres temporelles successives du son considéré, avec généralement un certain recouvrement d'une fenêtre à l'autre. Ce traitement n'est destiné qu'à un certain type de sons notamment ceux dont les composantes ne varient pas trop au cours du temps. Ce modèle a de nombreuses applications [Dolson, 1986], notamment du point de vue musical [Moorer, 1978], et en traitement de la parole [Portnoff, 1980]. Les plus connues sont les étirements (dilatations) et compressions dans le temps et les changements de hauteur [M. H. Serra, 1997]. Le principe est que le traitement à la base de ces transformations est réalisé dans le domaine spectral avant de revenir au domaine temporel par TFD inverse de chaque fenêtre traitée.

1.1.2. Du vocodeur de phase au modèle sinusoïdal : définition de base

Le *modèle sinusoïdal* de la parole (et des signaux audio en général) peut être vu comme une amélioration du vocodeur de phase. En effet la base de ce modèle est à nouveau la transformée de Fourier à court terme et la manipulation du signal dans le domaine fréquentiel. Mais à la différence du vocodeur de phase, toutes les valeurs du spectre (discret) résultant ne sont pas prises en compte. Seuls les maxima locaux ou *pics spectraux* (c'est-à-dire les valeurs des fréquences dont les amplitudes associées sont significativement supérieures aux valeurs voisines) sont pris en considération et supposés être représentatifs des composantes fréquentielles présentes dans le signal. Cette idée initiée par [Moorer, 1978] [Portnoff, 1980] et [Dolson, 1986] est notamment à la base des nombreux développements réalisés au milieu des années 80. On peut citer principalement ceux réalisés par McAulay et Quatieri d'une part [McAulay & Quatieri, 1986], avec une série d'études plutôt destinées à la modélisation de la parole, et ceux réalisés par Smith et Serra d'autre part [Smith & Serra, 1987], avec des études plutôt destinées à modéliser les sons musicaux (instruments isolés et scènes polyphoniques). Il résulte de cette sélection de maxima locaux que le modèle se compose alors d'une somme d'un nombre réduit I de sinusoïdes (I étant petit par rapport au nombre de canaux spectraux de la transformée de Fourier discrète initiale) :

$$\hat{s}(n) = \sum_{i=1}^{I} A_i \cos(\omega_i n + \varphi_i) \qquad (1.1)$$

Les paramètres du modèle sont respectivement les amplitudes A_i, les fréquences ω_i, et les phases φ_i des I sinusoïdes[5].

Cette forme basique du modèle sinusoïdal correspond à un signal strictement stationnaire, composé d'une somme de sinusoïdes dont les paramètres sont fixes sur la portion de temps considérée. Il permet donc sous cette forme de modéliser une portion de signal « régulière » au sens où elle est bien composée de telles sinusoïdes, ou de sinusoïdes s'approchant de ces hypothèses. Une telle représentation repose bien sûr sur la nature vibratoire du signal. Dans le cas de la production de parole par exemple, une vibration des cordes vocales est la source des sons voisés pour lesquels ce modèle est localement bien adapté (si le signal est proche d'un signal périodique sur la section considérée). On peut observer un exemple de section quasi-périodique de signal de parole sur la Figure 1.1.

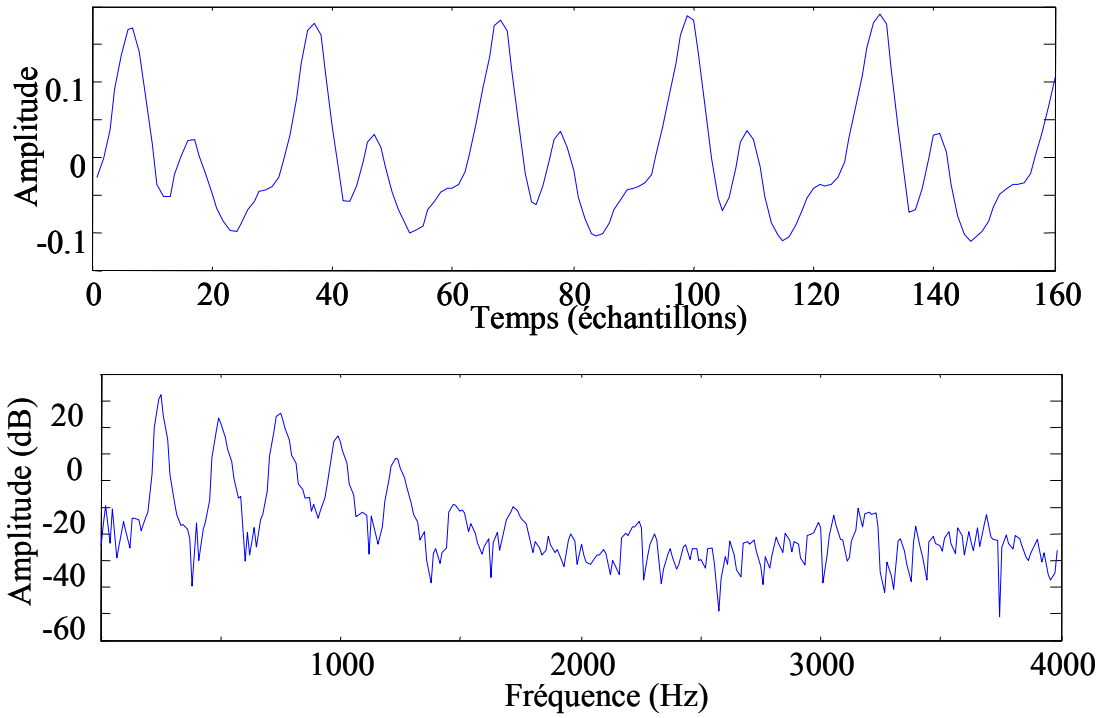

Figure 1.1 : Représentation d'une section quasi-périodique de signal de parole ; en haut : signal temporel ; en bas : spectre d'amplitude correspondant à cette portion de signal obtenu par *FFT* (le signal est échantillonné à la fréquence $F_e = 8$ kHz). La structure de raies spectrales est bien visible dans la première moitié du spectre, et les différentes composantes sont situées à des multiples de la fréquence fondamentale.

[5] Dans la suite du document, nous conserverons ces notations pour ces paramètres et nous ne les redéfinirons pas, sauf si le contexte est particulier. Comme tous les traitements sont réalisés sur des signaux échantillonnés (destinés à être reconvertis en analogique après traitement avec la même fréquence d'échantillonnage), nous restons dans tout le document dans un cadre formel numérique. Nous faisons quelques liens de façon très isolée avec le formalisme des signaux analogiques si nécessaire. Par ailleurs, en toute rigueur, le terme de *fréquences* est ici abusif, les ω_i étant plus précisément les *pulsations* (numériques). Les fréquences (numériques) f_i correspondantes sont en réalité données par la relation $\omega_i = 2\pi f_i$. Dans ce document, on emploie ces deux variables selon le contexte, mais dans les deux cas on les désigne comme des fréquences, comme c'est souvent le cas dans la littérature.

Nous étudierons dans les sections suivantes des versions plus raffinées prenant en compte l'évolution temporelle des caractéristiques du signal. De même on verra aussi la possibilité de raffiner ou d'enrichir ce modèle pour décrire des signaux de type aléatoires (voir Section 1.4.2). On peut noter que bien que stationnaire, ce modèle sous la forme de l'équation (1.1) peut décrire des signaux qui ne sont pas forcément strictement périodiques : les différentes fréquences ne sont pas contraintes à être multiples d'une fréquence donnée, dite *fréquence fondamentale*. Si tel est le cas, on parle alors de *modèle sinusoïdal harmonique* ou de *modèle harmonique* tout simplement (voir Section 1.4.1) et la décomposition du signal selon le modèle sinusoïdal peut alors être identifiée à celle réalisée au sens des *séries de Fourier*.

1.1.3. Principe de base de l'analyse-synthèse sinusoïdale et application à la transformation et au codage des sons

Comme son nom l'indique un *système d'analyse-synthèse* fondé sur le modèle sinusoïdal consiste à extraire d'abord les paramètres du modèle, amplitudes, fréquences et phases des différentes composantes, à partir de la section de signal considérée. La synthèse peut ensuite être réalisée simplement en appliquant l'équation (1.1) avec les valeurs des paramètres mesurées[6]. Dans cette étude on appelle *fenêtre d'analyse* la portion de signal sur laquelle on extrait les paramètres du modèle et *fenêtre de synthèse* la portion de signal sur laquelle on resynthétise le signal à partir des paramètres mesurés (et éventuellement transformés). Ces notions seront précisées par la suite et on décrira en détail dans les Sections 1.2 et 1.3 les principales méthodes d'analyse et de synthèse respectivement. Entre les deux phases d'analyse et de synthèse, le modèle permet de réaliser toute une série de transformations permettant de modifier les caractéristiques du signal. Par exemple, on peut changer la valeur des amplitudes pour une certaine gamme de fréquence, ce qui correspond à un filtrage à phase nulle (certaines composantes sont amplifiées ou atténuées sans que leur synchronisation temporelle ne soit modifiée). De même, on peut appliquer l'équation (1.1) sur un nombre d'échantillons plus long ou plus court que le nombre initial, ce qui permet de réaliser un étirement ou une compression temporelle du signal.

C'est cette possibilité de transformer le signal de multiples façons qui a rendu le modèle sinusoïdal et ses dérivés très populaires à partir des années 80 dans les communautés de chercheurs en traitement des signaux de musique et de parole. En effet, d'un côté les chercheurs en informatique musicale ont développé et exploité de tels modèles spectraux pour transformer les signaux de musique selon des objectifs liés à la *composition* (par exemple changer la hauteur ou la durée d'une note [Serra & Smith, 1997]) et plus généralement à la *création musicale* (par exemple générer des effets musicaux, tels que le vibrato ou le trémolo [Serra, 1997] [Zölder, 2002]). D'un autre côté, les chercheurs en signal de parole ont développé ces techniques pour les utiliser dans des applications spécifiques à partir des possibilités offertes par les transformations de base [Quatieri & McAulay, 1992]. Parmi ces applications on trouve les systèmes de *synthèse de la parole*, notamment à partir du texte ou d'entrées phonétiques. En effet, le signal de parole synthétisé doit être le plus naturel possible. Or, il est généralement

[6] On verra par la suite des techniques plus sophistiquées dans le cas où le signal est non-stationnaire.

fabriqué par la concaténation d'unités pré-stockées, de la taille de la syllabe[7]. Il s'agit alors de pouvoir transformer le signal pour l'aligner sur des contraintes prosodiques (c'est-à-dire liées à la mélodie, au rythme, et à l'intensité) déterminées par l'application cible de la synthèse [Bailly & Benoit, 1992] [Boëffard & d'Alessandro, 2002] [Moulines & Charpentier, 1990]. Notons que toutes les transformations possibles peuvent se faire selon deux grandes lignes directrices : soit en modifiant directement les valeurs des paramètres avant d'appliquer les équations de synthèse, soit en modifiant les équations de synthèse elles-mêmes en gardant les valeurs initiales des paramètres. On peut bien sûr aussi combiner les deux procédés. Nous reviendrons sur ce principe par la suite.

Parallèlement, la modélisation sinusoïdale permet une représentation paramétrique efficace du signal et connaît ainsi plusieurs applications dans le domaine du *codage* des signaux audio et de la parole [McAulay & Quatieri, 1992, 1995], celui de la *conversion de voix* [Stylianou, 1996], et plus récemment pour une application de *tatouage* [Girin & Marchand, 2004]. A titre d'exemple, la problématique du codage est de représenter le signal avec le moins de ressource binaire possible et les modèles paramétriques sont traditionnellement de bons candidats pour arriver à cet objectif. Même si la tendance en codage générique de signaux audio est à l'utilisation de techniques de codage par transformée assez généralistes, telle que la MDCT (*Modified Discrete Cosine Transform ;* voir [Princen & Bradley, 1986] [Oppenheim & Schaffer, 1989]) largement utilisée dans les normes MPEG, l'approche paramétrique s'est révélée extrêmement efficace et très largement employée pour le codage des signaux de parole. Parmi les représentations paramétriques particulièrement efficaces, la modélisation par prédiction linéaire (dite communément modélisation LPC pour *Linear Predictive Coding*) arrive en tête des succès [Markel & Gray, 1976] [Gersho & Gray, 1992] [Makhoul, 1975] [LeBlanc *et al*., 1993]. Cependant la modélisation sinusoïdale et ses dérivées continuent d'être explorées comme une voix prometteuse, y compris pour le codage de scènes audio complexes et pour la musique [McAulay & Quatieri, 1995] [Serra, 1997] [Poli *et al*., 1991] [Roads *et al*., 1997] [Roads, 1996].

1.1.4. Prise en compte de la non stationnarité des signaux

1.1.4.1. Au niveau du signal

En général, les signaux sonores du monde réel ne sont pas strictement stationnaires, c'est-à-dire que leurs caractéristiques spectrales ou statistiques[8] évoluent en fonction du temps, plus ou moins régulièrement. En effet, si on prend l'exemple des signaux de parole, on peut dire de façon assez sommaire que la parole est une suite continue de sons de différente nature articulatoire. Par conséquent, il existe dans tout signal de parole des sections peu stationnaires : par exemple, les transitions entre sons différents[9]. Il existe même des sections fortement non-stationnaires, telles que les consonnes

[7] Généralement, il s'agit plus précisément de diphones, c'est-à-dire de transitions du milieu stable d'un son vers le milieu stable d'un autre son.

[8] Le modèle sinusoïdal est un modèle intrinsèquement spectral. La caractérisation statistique concerne plutôt les composantes aléatoires du signal, telles qu'on les aborde à la Section 1.4.2.

[9] D'un point de vue un peu plus quantitatif, le débit moyen de la parole est typiquement de 10 à 15 phonèmes par seconde selon le locuteur, le style, la langue, etc. Cela signifie que la durée typique des phonèmes est d'environ de 50 ms à 100 ms.

plosives par exemple : on peut voir un exemple de son [k] sur la Figure 1.2 entre les échantillons 1 et 500 environ. Par ailleurs, même pour les sections de parole relativement stables, telles que les voyelles par exemple, on devrait plutôt dire que le signal est « localement quasi-stationnaire », c'est-à-dire que ses caractéristiques spectrales et/ou statistiques locales (sur des sections de quelques millisecondes à quelques dizaines de millisecondes) évoluent lentement dans le temps, en tout cas par rapport à la vitesse d'évolution des échantillons de la forme d'onde. Ceci est illustré par exemple par la voyelle de la Figure 1.2, en l'occurrence un [a] : le noyau de la voyelle semble relativement stable mais on remarque nettement la modulation d'amplitude, rapidement croissante au début de la voyelle et plus lentement décroissante à la fin. Ce type de paramètres, ainsi que d'autres tels que la fréquence fondamentale par exemple, évoluent continûment au cours du temps selon la « mélodie de la parole »[10]. Pour la musique, on peut retrouver le même type d'évolution et le même type d'opposition entre sons « stables », tels que des notes soutenues par exemple, et des sons plus localisés en temps, tels que les sons de percussions.

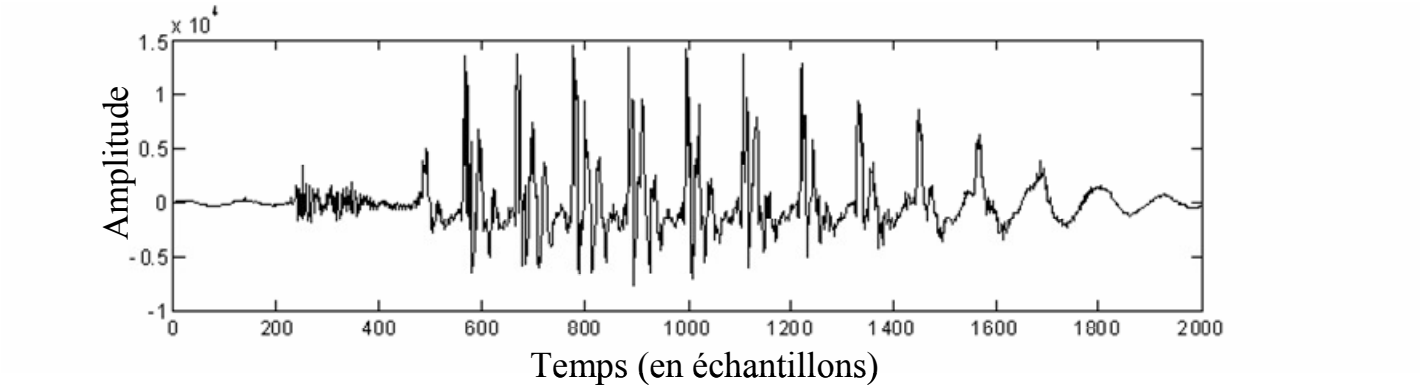

Figure 1.2 : Syllabe [ka] prononcée par un locuteur adulte mâle (sur l'axe des X, l'indice des échantillons ; sur l'axe des Y, leur amplitude ; la fréquence d'échantillonnage est de 16 kHz).

1.1.4.2. Au niveau du modèle : principe de l'analyse-synthèse à court terme

Compte tenu de l'évolution du signal au cours du temps, considérer les paramètres des composantes du modèle sinusoïdal, amplitudes, fréquences et phases, comme constants pendant toute la durée d'un signal audio (de parole ou de musique par exemple), comme le fait l'Equation (1.1) est généralement peu pertinent pour décrire efficacement le signal. Par conséquent, on cherche dans les systèmes d'analyse-synthèse des sons à suivre cette évolution, c'est-à-dire à la caractériser, à « l'inscrire » dans les paramètres extraits et à la restituer en synthèse.

[10] L'exemple de la variation de la fréquence fondamentale illustre le fait que l'*harmonicité* d'un signal ne garantit pas sa *stationnarité* : la fréquence fondamentale peut varier au cours du temps, de même que l'amplitude des différentes composantes, alors même que les fréquences des différentes composantes restent les multiples de cette fréquence fondamentale à chaque instant. Si les variations de la fréquence fondamentale sont lentes par rapport à l'évolution de la forme d'onde du signal, on parle de *quasi-périodicité*. La *quasi-stationnarité* est une notion plus générale de l'évolution lente des caractéristiques du signal, issue de l'étude statistique des signaux, et englobant les sons non-voisés : elle peut décrire une lente modulation d'amplitude d'un bruit par exemple.

Comme on l'a déjà mentionné rapidement à propos de la TFD, une démarche usuellement adoptée pour cette tâche consiste à segmenter le signal à modéliser en petits morceaux successifs, généralement de taille fixe et régulièrement espacés, sur lesquels on considère que l'approximation par un modèle stationnaire reste valide[11]. On met ainsi à jour les paramètres du modèle sur ces *fenêtres à court terme* qui sont généralement d'une durée de l'ordre de la dizaine de millisecondes (de 10 à 30 ms en pratique)[12]. On a donc une succession de *fenêtres d'analyse à court terme* pour l'extraction des paramètres et une succession de *fenêtres de synthèse à court terme* pour la resynthèse du signal à partir de ces paramètres. Dans la suite de ce chapitre, nous effectuerons une brève revue des méthodes d'analyse à court terme à la Section 1.2, puis nous nous intéresserons à la synthèse à court terme dans la Section 1.3.

Notons que cet aspect trame-à-trame du processus d'analyse-synthèse ne remet pas en cause le principe de transformation et de codage du signal entre les processus d'analyse et de synthèse, comme annoncé à la Section 1.1.3. Les transformations et le codage éventuels sont aussi mis à jour au fur et à mesure de l'évolution du signal. Par exemple on peut choisir d'étirer temporellement une certaine partie du signal comportant un certain nombre de trames et de laisser intacte une autre partie.

1.1.4.3. Evolution vers un modèle non-stationnaire

Parallèlement au principe d'analyse-synthèse à court terme, une autre conséquence de l'évolution du signal au cours du temps est qu'on peut explicitement reformuler le modèle sinusoïdal dans une version plus évoluée où les paramètres du modèle deviennent des fonctions (discrètes) du temps. Ainsi, ces paramètres sont définis pour chaque indice n : $A_i(n)$ pour les amplitudes, $\omega_i(n)$ pour les fréquences, et $\varphi_i(n)$ pour les phases (pour la sinusoïde d'indice i). Le signal est alors représenté comme suit :

$$\hat{s}(n) = \sum_{i=1}^{I} A_i(n)\cos[\varphi_i(n)] \tag{1.2}$$

avec

$$\varphi_i(n) = \sum_{m=1}^{n} \omega_i(m) + \varphi_i(0) \tag{1.3}$$

[11] En réalité, cette démarche est surtout valable sur les portions de signal à évolution relativement lente. On peut considérer qu'il existe des portions de signal où le modèle sinusoïdal (ou tout autre modèle stationnaire) n'est pas le plus approprié, par exemple les sons transitoires. On pratique alors une segmentation préalable du signal plus précise associée à une détection de ces transitoires, et on utilise un modèle plus approprié sur ces sections transitoires. Ce point est plus largement discuté à la Section 1.4.3.

[12] En ce qui concerne la parole, on a vu que la durée typique d'un phonème est de l'ordre de 50 à 100 ms (voir la note de bas de page 9). Cependant, les trames d'analyse-synthèse à court terme doivent être plus courtes que cette durée pour deux raisons : d'abord, de nombreuses réalisations de phonèmes ou de sections de phonèmes sont beaucoup plus courtes que la durée mentionnée ci-dessus ; ensuite, pour qu'un système d'analyse-synthèse à court terme produise un signal de haute qualité et intelligible, il faut non seulement extraire et transmettre les caractéristiques acoustiques de la position centrale (la partie stable) de ces phonèmes, mais il faut aussi représenter correctement les parties transitoires. Ces contraintes requièrent une taille d'environ 10 à 30 ms pour les trames successives. D'un autre côté, des valeurs plus petites impliqueraient que les trames successives soient très corrélées et montreraient une trop grande redondance de l'information, ce qui est pénalisant en terme de coût de traitement (pour le codage de parole par exemple).

Comme des sinusoïdes peuvent apparaître et disparaître au cours de l'évolution du signal, le nombre I de sinusoïdes est en toute généralité une fonction du temps $I(n)$. Nous reviendrons sur ce point par la suite[13].

Les amplitudes, les fréquences et les phases sont des fonctions à bande étroite comparée à la bande du signal, c'est-à-dire que leur évolution est relativement lente par rapport à l'évolution des échantillons du signal. C'est pourquoi nous cherchons précisément dans ce travail de thèse à modéliser leur évolution sur une base à long terme. La modélisation à long terme des amplitudes et des phases sera détaillée respectivement dans le Chapitre 3 et le Chapitre 4 de ce document.

Comme nous allons nous intéresser à l'évolution des paramètres de phase à long terme, il est très important de noter que la notion de phase prend un sens différent de la définition prise dans le cadre stationnaire. Il s'agit ici de la *phase instantanée* ou *phase absolue*, définie en toute généralité comme l'argument du cosinus d'une composante sinusoïdale en chaque indice du temps n. Comme le montre l'équation (1.3), cette phase instantanée est reliée à la fréquence correspondante par une relation de différenciation / sommation (qui est une relation de dérivation / intégration dans le cas continu). La fréquence est elle-même une *fréquence instantanée* définie en chaque indice et cette fréquence est la dérivée de la phase instantanée en cet indice[14]. Comme les fréquences instantanées sont des fonctions positives qui représentent physiquement la notion de cycles dans le signal, les phases instantanées sont des fonctions strictement croissantes du temps : leur évolution est globalement linéaire avec des variations autour de cette forme linéaire traduisant les variations de la fréquence instantanée de la composante considérée (voir par exemple la Figure 1.5).

La phase instantanée ne doit donc pas être confondue avec la phase dite *phase relative*, c'est-à-dire le paramètre de phase tel qu'on l'a défini dans l'équation (1.1). Cette phase relative est en effet un paramètre de déphasage constant défini uniquement dans le cadre stationnaire où la fréquence est elle-même constante. On a bien toujours dans ce cas stationnaire une relation de dérivation entre phase et fréquence : l'argument de la fonction cosinus est bien donné par $\varphi_i(n) = \omega_i n + \varphi_i(0)$. Mais le terme « phase » utilisé jusqu'ici désigne le coefficient constant du terme de droite de cette équation (c'est pourquoi on parle aussi de *phase à l'origine* pour désigner la phase dans ce cadre stationnaire : ce paramètre décrit les décalages relatifs des composantes du signal entre elles à l'instant d'origine de la fenêtre[15] ; cette notion est illustrée sur la Figure 1.3). Une des raisons pour laquelle on peut avoir tendance à confondre facilement la phase absolue et la phase relative est que les méthodes d'analyse couramment employées

[13] Voir les Sections 1.1.4.4 et 1.3.2.1 pour plus de précisions sur le problème du suivi des sinusoïdes au cours du temps.

[14] En réalité, les sommations/différences discrètes sont des approximations des intégrations/dérivations continues. La version continue du modèle sinusoïdal (1.2) sur laquelle sont originellement définis les termes de fréquence instantanée et phase instantanée est :

$$\hat{s}(t) = \sum_{i=1}^{I(t)} A_i(t)\cos[\varphi_i(t)] \qquad \frac{d\varphi_i(t)}{dt} = 2\pi f_i(t)$$

[15] Notons qu'on peut choisir de définir les décalages relatifs des composantes à un instant particulier du signal choisi arbitrairement. Pour les sons voisés de parole, cet instant peut être le moment n_0 du pitch, c'est-à-dire l'indice correspondant au pic d'excitation maximum où le phasage des composantes du signal est maximal. On a alors pour chaque composante, $\varphi(n) = 2\pi f(n-n_0)+\varphi_{n0}$ et φ_{n0} est la *phase de dispersion*. Pour mesurer φ_{n0}, il faut bien sûr estimer préalablement l'instant du pitch.

faisant l'hypothèse de stationnarité locale (voir Section 1.2.1) renvoient une valeur de phase qui correspond à un échantillon précis de la fenêtre d'analyse. Dans ce cas, les notions de phase absolue (valeur argument du cosinus) et phase relative (décalage relatif des composantes sinusoïdales) sont très proches, voire confondues si l'échantillon en question est l'origine $n = 0$. Notons pour être complet que même si la phase absolue est une fonction strictement croissante du temps qui peut prendre très vite de grandes valeurs par rapport à π, les procédures d'analyse à court terme fournissent généralement des valeurs de phase définies à 2π près, et comprises entre 0 et 2π, ou entre $-\pi$ et π selon la méthode utilisée. Ceci s'explique par le fait que ces procédures ne connaissent pas l'évolution passée du signal en dehors de la fenêtre d'analyse courante. Si à partir de ces mesures, on veut reconstruire la vraie trajectoire de la phase absolue au cours du temps sans « cassure » (c'est-à-dire des sauts de valeurs multiples de 2π), il faut estimer les « vraies » valeurs de phase à partir des valeurs modulo 2π, en leur ajoutant un nombre de fois approprié cette valeur 2π. Cette procédure s'appelle un dépliement de phase (*phase unwrapping* en anglais)[16].

Cette différenciation entre *phase absolue* (ou *phase instantanée*) et *phase relative* (ou *phase à l'origine*) est importante pour deux raisons. D'une part elle permet de généraliser la notion de phase très souvent considérée dans le cadre stationnaire à ce nouveau cadre intrinsèquement non-stationnaire. D'autre part, il sera nécessaire dans notre étude à long terme de considérer les paramètres du modèle sinusoïdal et en particulier les paramètres de phase comme des fonctions du temps en constante évolution. A ce titre, la procédure de dépliement de phase sera par exemple absolument nécessaire pour retrouver la « vraie » trajectoire à partir des mesures modulo 2π. La modélisation à long terme de la phase prise au sens de fonction au cours du temps, et donc de phase absolue ou instantanée, fera l'objet du Chapitre 4 de ce document.

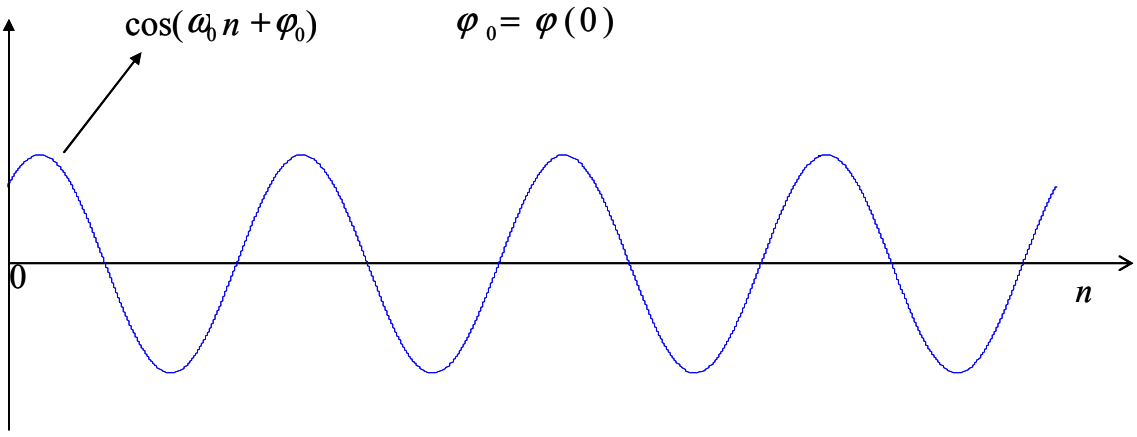

Figure 1.3 : Illustration de la phase à l'origine : $\varphi_0 = \varphi(0)$

[16] Attention, on peut utiliser la même expression de *dépliement de phase* pour désigner le même procédé le long de l'axe des fréquences pour assurer cette fois la continuité de la phase comme fonction de la fréquence (sur un intervalle de temps donné), mais il ne s'agit pas du même objectif. Dans notre étude de modélisation à long terme, sauf indication contraire, le dépliement désignera toujours le réajustement des valeurs de la phase le long de l'axe temporel.

1.1.4.4. Une dichotomie correspondante des principales méthodes d'analyse-synthèse sinusoïdale à court terme

Compte tenu à la fois du principe de suivi à court terme de l'évolution du signal et de la réécriture du modèle sinusoïdal de la sous-section précédente, les méthodes d'analyse-synthèse se divisent d'un certain point de vue en deux grandes familles : les méthodes appelées ici *analyse-synthèse additive avec recouvrement* (*overlap-add analysis-synthesis* en anglais) d'une part, et les méthodes appelées ici *analyse-synthèse par interpolation* (*interpolative analysis-synthesis*) d'autre part. On donne juste ici une description des grands principes qui régissent ces méthodes. Chacune des briques de ces méthodes sera présentée plus en détail à la Section 1.3.

Généralement, les méthodes du type *overlap-add* se contentent de resynthétiser des portions de signal successives indépendamment les unes des autres, en appliquant localement l'équation du modèle stationnaire (1.1) (ou un modèle non-stationnaire plus sophistiqué comme celui de l'équation (1.2) mais toujours considéré « localement »), puis à assurer la continuité du signal par un lissage adapté entre trames successives. Ce principe est plus largement détaillé dans la Section 1.3.1.

L'alternative donnée par les techniques par interpolation est, comme le nom l'indique, d'interpoler les valeurs des paramètres extraits entre les trames d'analyse consécutives à chaque indice discret du temps n. Cette interpolation est réalisée de façon exacte, c'est-à-dire en assurant la continuité des valeurs aux jonctions entre fenêtres de synthèse considérées ici sans recouvrement, c'est-à-dire simplement concaténées les unes à la suite des autres. Le signal est alors synthétisé à partir des valeurs interpolées des paramètres, et la continuité du signal est assurée par la continuité des paramètres. Les méthodes appliquant ce principe seront présentées de façon plus détaillée à la Section 1.3.2. Avec ce type de méthode, les paramètres du modèle sinusoïdal deviennent donc des fonctions du temps. Ce type de méthode colle donc parfaitement à la formulation du modèle dans sa version évoluée de la Section 1.1.4.3, et peut donc être intrinsèquement considérée comme une technique du type non-stationnaire (exception faite d'une interpolation par des modèles d'ordre zéro, c'est-à-dire à fréquence constante, comme on le verra par la suite)[17].

Notons qu'une difficulté majeure de cette technique consiste à associer les valeurs des paramètres d'une fenêtre d'analyse aux valeurs correspondant à la fenêtre suivante avant de réaliser l'interpolation proprement dite. En effet, le nombre de paramètres peut varier d'une fenêtre à l'autre et leur trajectoire dans le plan temps-fréquence peut être plus ou moins régulière. Il s'agit donc d'appairer les paramètres entre trames successives de façon cohérente avant de réaliser leur interpolation. Une trajectoire temporelle cohérente d'une sinusoïde donnée du modèle sinusoïdal étant appelé *partiel* dans la littérature du traitement audio, cette tâche a pour nom *suivi de partiels* (*partial-tracking* en anglais) et est plus largement décrite à la Section 1.3.2.1.

[17] De ce point de vue, ce type de méthode peut exploiter efficacement les informations extraites par les méthodes d'analyse non-stationnaire comme celles présentées à la Section 1.2.2. Mais elle peut également s'accommoder des mesures fournies par les méthodes stationnaires. C'est d'ailleurs dans le cadre d'une analyse stationnaire qu'elle a été développée à l'origine [McAulay & Quatieri, 1986] [Serra & Smith, 1990].

Dans la suite, nous faisons une revue plus détaillée des briques composant les méthodes d'analyse-synthèse introduites dans cette section. Dans la pratique, les tâches d'analyse et de synthèse peuvent être considérées séparément[18]. D'ailleurs, les deux grandes familles d'analyse-synthèse par *overlap-add* et par *interpolation* se différencient principalement par la méthode de *synthèse*, alors qu'elles peuvent utiliser des méthodes d'analyse communes. Pour cette raison, nous faisons d'abord la revue des différentes méthodes d'analyse, puis nous détaillons les techniques de synthèse. Pour ce qui est de l'analyse, nous différencions clairement le cas stationnaire du cas non-stationnaire.

1.2. Analyse des paramètres

1.2.1. Analyse avec hypothèse de stationnarité locale

Conformément à notre dichotomie de la section précédente, la première grande classe de méthodes d'analyse est celle qui repose d'abord sur l'hypothèse de stationnarité locale du signal. On rappelle que dans un modèle stationnaire, on considère que les paramètres d'amplitude, de fréquence et de phase des sinusoïdes du modèle (1.1) sont constants sur la fenêtre de signal considérée. Pour extraire ces paramètres, la majeure partie des implantations logicielles se base sur la TFD. On présente donc d'abord les méthodes reposant sur cette transformée. Un résultat commun de ces méthodes est qu'en cas de variation effective des paramètres sur la fenêtre d'analyse, les estimations retournées correspondent aux valeurs moyennes des paramètres sur l'ensemble de la fenêtre d'analyse (du moins pour les amplitudes et les fréquences). Ces valeurs peuvent alors être considérées comme celles prises au centre de la fenêtre d'analyse.

1.2.1.1. Analyse par transformée de Fourier discrète et peak-picking

Basiquement, la transformée de Fourier discrète consiste en une projection du signal sur une base orthonormée d'exponentielles complexes de taille N. Les modules de ces exponentielles forment le spectre d'amplitude et les arguments forment le spectre de phase. Comme on travaille avec une portion de signal à court terme de taille N, cette sélection de signal est opérée en multipliant le signal noté $s(n)$ par une fenêtre notée $w(n)$ de taille N, pour obtenir un signal fenêtré $s_w(n)$ (l'origine est ici arbitraire). Appliquée sur le signal fenêtré, la TFD et son inverse (TFDI) sont des opérations mathématiques définies comme suit (l'indice m représente les canaux fréquentiels de la TFD ; pour des signaux réels, on calcule la TFD pour des valeurs de m allant de 0 à $N/2$ (on suppose N pair), et m est relié à la fréquence f analogique correspondante par $f = mF_e/N$ où F_e est la fréquence d'échantillonnage du signal) :

$$S_w(m) = \frac{1}{N}\sum_{n=0}^{N-1} s_w(n) e^{-j2\pi\frac{nm}{N}} \tag{1.4}$$

$$s_w(n) = \sum_{m=0}^{N-1} S_w(m) e^{j2\pi\frac{nm}{N}} \tag{1.5}$$

[18] A condition toutefois de rester relativement cohérent : par exemple, il n'est pas très pertinent de réaliser une analyse de paramètres de type non-stationnaire, par exemple une dérivée de fréquence, si on n'utilise que des paramètres de type stationnaires à la synthèse.

Rappelons brièvement que l'on obtient ainsi la convolution du spectre « théorique » du signal dans sa version « idéale » (un spectre composé de raies correspondant aux composantes sinusoïdales) avec le spectre de la fenêtre d'analyse. Ceci se traduit notamment sur le spectre d'amplitude par la transformation des raies spectrales en des lobes échantillonnés à F_e/N et dont la forme dépend de la fenêtre choisie. Il existe une large gamme de fenêtres possibles. Son détail et l'étude de l'influence de ces fenêtres sur l'analyse est un domaine classique de l'analyse spectrale qui peut être trouvé par exemple dans [Kay, 1988] [Oppenheim & Schaffer, 1989] [Harris, 1978] et qui dépasse largement le cadre de ce manuscrit.

Une fois le spectre calculé, la sélection des sinusoïdes pertinentes pour le modèle sinusoïdal se fait généralement par un algorithme de détection des maxima locaux du spectre d'amplitude (les pics spectraux) appelé *peak-picking* [McAulay & Quatieri, 1986] [Serra, 1987]. Cet algorithme se conjugue ensuite avec la mesure des fréquences, amplitudes et phases correspondant aux positions de ces pics. Généralement, une première détection grossière des pics est réalisée en détectant les passages par zéros (du positif vers le négatif pour un maximum) de la dérivée du spectre d'amplitude en fonction de la fréquence. Ces valeurs ne sont cependant pas très précises, puisqu'elles correspondent à des points discrets de la TFD, la résolution fréquentielle étant égale à F_e/N. En d'autres termes, le « vrai » maximum qui correspond à la version continue sous-jacente du spectre discret peut être situé aux environs du maximum discret, au pire à $F_e/2N$. Cette première approximation doit donc généralement être améliorée.

Ceci peut être fait en augmentant N, la taille de la TFD, mais on souhaite dans le même temps conserver un nombre d'échantillons faible pour respecter l'hypothèse de stationnarité locale du signal (voir Section 1.2.1.3). De manière à calculer plus de points fréquentiels sans augmenter le nombre d'échantillons observés, on peut utiliser la méthode dite du *zero-padding* qui consiste à compléter les N échantillons de signal par des valeurs nulles, et effectuer la TFD sur ce nouvel ensemble de points. Il s'agit d'une méthode très classique d'analyse spectrale que nous ne détaillerons pas ici (voir par exemple [Oppenheim & Schaffer, 1989]). Une autre solution consiste à interpoler le spectre d'amplitude à proximité du maximum local initialement estimé au canal m, en utilisant une parabole passant par les points $|S_w(m-1)|$, $|S_w(m)|$ et $|S_w(m+1)|$ (sur une échelle log) [Press *et al.*, 1992]. Une telle approximation est particulièrement adaptée à l'utilisation d'une fenêtre gaussienne car le lobe principal de cette fenêtre est exactement une parabole sur l'échelle dB. Elle fournit toutefois de bons résultats pour les autres types de fenêtres. Cette procédure est illustrée sur la Figure 1.4. On peut montrer qu'avec une telle interpolation, la fréquence estimée affinée et l'amplitude correspondante sont données respectivement par :

$$\hat{f} = m\frac{F_e}{N} + \frac{1}{2}\frac{|S_w(m-1)| - |S_w(m+1)|}{|S_w(m-1)| - 2|S_w(m)| + |S_w(m+1)|} \tag{1.6}$$

$$\hat{A} = \frac{2}{W(0)}\left(|S_w(m-1)| - \frac{\left[1.5|S_w(m-1)| - 2|S_w(m)| + 0.5|S_w(m+1)|\right]^2}{2\left[|S_w(m-1)| - 2|S_w(m)| + |S_w(m+1)|\right]}\right) \tag{1.7}$$

avec

$$W(0) = \sum_{n=0}^{N-1} w(n) \tag{1.8}$$

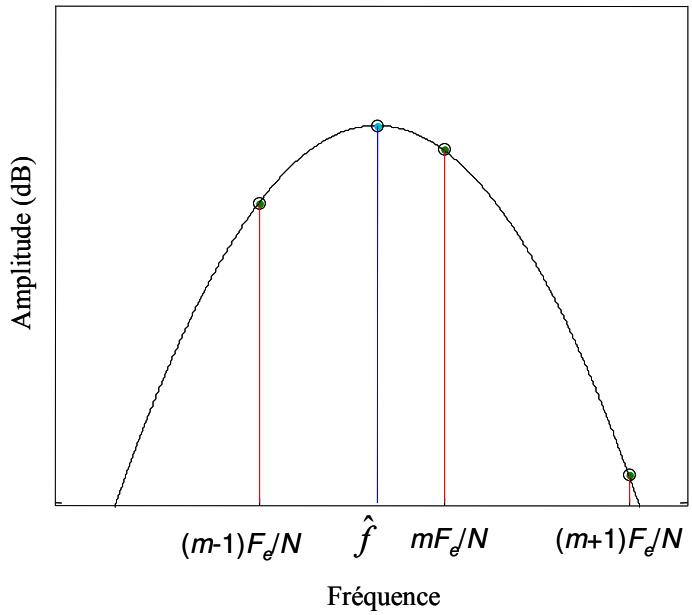

Figure 1.4 : Correction de la détection d'un pic spectral par interpolation parabolique du point d'amplitude maximum et de ses deux voisins.

Notons que la méthode de l'interpolation parabolique et la méthode du *zero-padding* peuvent être combinées pour un meilleur compromis complexité/précision [Serra, 1997].

Le même principe d'interpolation peut être utilisé pour estimer la phase des sinusoïdes par la valeur du spectre de phase à la valeur de la fréquence estimée corrigée. Tout comme pour les amplitudes, et peut-être même encore plus, il faut faire attention à l'influence de la fenêtre d'analyse qui peut être majeure dans ce cas [Harris, 1978]. Une solution pour s'affranchir de ce problème consiste à utiliser une version spéciale de la fenêtre d'analyse associée à un décalage circulaire de $N/2$ échantillons des valeurs fenêtrées du signal (N est supposé pair). La fenêtre doit être symétrique et de longueur impaire $N-1$ avant l'ajout d'un échantillon à zéro en terme initial (on arrive ainsi à une taille N). On peut montrer que cet ensemble de conditions garantit d'avoir une fenêtre à phase nulle dans le lobe principal du spectre, et donc ne perturbant pas l'analyse de la phase du signal au niveau du maximum d'amplitude détecté. Par ailleurs, cette modification de la fenêtre par rapport à la version classique initiale n'a quasiment aucun effet sur la forme du spectre d'amplitude. On peut noter qu'un avantage de cette méthode est en outre de fournir une estimation du paramètre de phase au centre de la fenêtre d'analyse (précisément à l'échantillon d'indice $N/2$ si l'origine est indicée à 0). Dans le cas où les paramètres d'amplitude et de fréquence varient légèrement sur la fenêtre d'analyse, cette propriété est cohérente avec le fait que les valeurs estimées de ces paramètres sont des valeurs moyennes sur l'ensemble de la fenêtre d'analyse (et donc des valeurs proches de celles prises au niveau du centre de la fenêtre en cas d'évolution de type linéaire ou plus généralement symétrique). Enfin, notons que le fait que la fenêtre soit à phase nulle sur l'ensemble de son lobe central permet de s'affranchir de la phase d'interpolation entre les trois valeurs correspondant au pics maximum et ses voisins, car la phase d'un cosinus fenêtré de cette façon sera constante dans cette région. Une faible erreur d'estimation de la fréquence, inférieure à la moitié de la largeur du lobe, ne faussera donc pas l'estimation de la phase.

Alternativement, une fois que les fréquences des sinusoïdes sont estimées (par exemple avec les procédés d'interpolation/suréchantillonnage vus ci-dessus), on peut calculer directement le spectre du signal fenêtré pour chaque fréquence estimée à l'aide d'une *transformée de Fourier discrète en un point* :

$$S_w(\hat{f}) = \frac{1}{N}\sum_{n=0}^{N-1} s_w(n) e^{-j2\pi\frac{n\hat{f}}{F_e}} \tag{1.9}$$

L'amplitude estimée de la sinusoïde correspondante est alors :

$$\hat{A} = 2\left|\frac{S_w(\hat{f})}{W(0)}\right| \tag{1.10}$$

Sa phase est donnée par l'argument du spectre recalculé à la fréquence considérée (en supposant la fenêtre d'analyse à phase nulle) :

$$\hat{\varphi} = \angle S_w(\hat{f}) \tag{1.11}$$

Notons enfin que d'autres méthodes plus récentes ont été proposées pour affiner la mesure de la fréquence des sinusoïdes fournie par la première détection grossière sur le spectre TFD. L'une d'entre elles est la méthode appelée *réassignement spectral* [Auger & Flandrin, 1995] [Röbel, 2002]. En plus du spectre du signal pondéré par la fenêtre d'analyse, cette méthode utilise le spectre du signal pondéré par la dérivée $w'(n)$ de cette fenêtre, noté $S_{w'}(m)$. La fréquence du maximum local est alors estimée par (Im(x) dénote la partie imaginaire de x) :

$$\hat{f} = m\frac{F_e}{N} - \mathrm{Im}\left(\frac{S_{w'}(m)}{S_w(m)}\right)\frac{F_e}{2\pi} \tag{1.12}$$

Il est aussi possible d'améliorer la précision d'estimation de la fréquence du maximum local en considérant la dérivée du signal au lieu de la dérivée de la fenêtre d'analyse [Desainte-Catherine & Marchand, 2000]. Cette méthode est appelée *méthode de la dérivée*. Dans le cas discret, qui nécessite une approximation discrète de la dérivée du signal, l'estimateur de fréquence est :

$$\hat{f} = \frac{F_e}{\pi}\arcsin\left(\frac{1}{2F_e}\frac{S^1(m)}{S(m)}\right) \tag{1.13}$$

où $S^1(m)$ est la composante d'indice m du spectre de la dérivée du signal.

Ces estimateurs, ainsi que ceux décrits dans d'autres études plus récentes, telle la méthode trigonométrique proposée dans [Lagrange *et al.*, 2005], sont comparés dans [Marchand & Lagrange, 2006]. Cette étude montre que, pour une large part, nombre de ces estimateurs sont équivalents d'un point de vue théorique, et qu'ils sont de plus très proches d'un point de vue expérimental.

1.2.1.2. Analyse par synthèse

Un autre type de méthode d'analyse, dite analyse-par-synthèse, consiste à régler les paramètres du modèle sinusoïdal de façon à ce qu'on minimise une certaine erreur entre le signal et le modèle. Les paramètres optimaux ainsi réglés correspondent aux paramètres estimés. On décrit ici plus particulièrement cette méthode, bien qu'elle ne soit pas aussi populaire que celles basées sur la TFD, car on utilisera dans nos études à long terme aux chapitres suivants une variante de cette méthode adaptée à notre problème.

Le critère d'erreur entre le signal et le modèle est généralement un critère des moindres carrés moyens pondérés, qui s'exprime de la façon suivante :

$$\varepsilon = \sum_{n=0}^{N_a-1} w^2(n)\left[s(n) - \hat{s}(n)\right]^2 \tag{1.14}$$

où $s(n)$ est le signal original, $\hat{s}(n)$ est le modèle sinusoïdal défini par (1.1), et $w(n)$ est la fenêtre d'analyse de longueur N_a[19]. Les valeurs de cette fenêtre viennent donc pondérer la contribution des échantillons d'indice correspondants dans le calcul de l'erreur. Comme l'analyse consiste à chercher les paramètres qui minimisent cette erreur, la procédure d'estimation est dite « minimisation d'erreur moyenne au sens des moindres carrés pondérés » (*WMMSE* pour *weighted minimum mean square error* ; voir par exemple [George & Smith, 1997] [Stylianou, 1996]).

Ce problème de minimisation traité conjointement en fonction de l'ensemble des trois paramètres, amplitude, fréquence et phase, est un problème non-linéaire difficile à résoudre tel quel. La difficulté est en effet liée au fait que le modèle est une fonction non-linéaire des paramètres de fréquences et phases. Par contre, le problème se simplifie grandement si on connaît les différentes fréquences des composantes. En effet, dans ce cas, on peut linéariser l'écriture du modèle en fonction des amplitudes et des phases :

$$\hat{s}(n) = \sum_{i=1}^{I} a_i \cos(n\omega_i) - b_i \sin(n\omega_i) \quad \text{avec} \quad \begin{cases} a_i = A_i \cos(\varphi_i) \\ b_i = A_i \sin(\varphi_i) \end{cases} \tag{1.15}$$

Ce sont alors les nouveaux paramètres a_i et b_i qui sont estimés par une procédure des moindres carrés[20] (décrite plus loin dans cette sous-section), et les amplitudes et phases sont données par :

$$\begin{cases} A_i = \sqrt{a_i^2 + b_i^2} \\ \varphi_i = -\arctan(b_i / a_i) \end{cases} \tag{1.16}$$

Cette méthode peut donc être combinée avec une des méthodes d'estimation des fréquences vues précédemment. Cette simplification fait aussi que cette méthode est

[19] Pour des raisons de simplicité, on décrit la méthode pour une trame de signal donnée, en plaçant arbitrairement l'origine de cette trame à zéro. Dans la pratique, on a une succession de trames de signal à analyser, décalées d'un certain nombre d'échantillons, avec un possible recouvrement.

[20] On ne considère ici que les indices i à partir de 1. La composante continue a_0 peut être estimée directement par la moyenne (pondérée ou non par $w(n)$) du signal sur la fenêtre d'analyse, ce qui correspond bien d'ailleurs à une minimisation au sens des moindres carrés sur ce coefficient particulier.

particulièrement adaptée au cas des signaux harmoniques, lorsqu'on réalise au préalable l'estimation de la fréquence fondamentale : dans ce cas les fréquences des composantes sont les multiples de la fréquence fondamentale estimée (voir les Sections 1.4.1 et 1.4.2.3). C'est ce cas que nous allons considérer dans notre étude de modélisation à long terme des signaux de parole, et c'est pourquoi, comme on l'a déjà mentionné, nous utiliserons cette méthode d'analyse dans les Chapitres 3 et 4, avec quelques adaptations liées aux particularités de notre étude. Dans notre étude, on aura donc $\omega_i = i\omega_0$, mais on garde la notation ω_i ci-dessous pour conserver la généralité de la présentation : la méthode reste valable sans l'hypothèse d'harmonicité, pourvu qu'on réalise au préalable l'estimation des différentes fréquences ω_i.

Pour présenter maintenant de façon concise la procédure d'estimation des coefficients a_i et b_i par minimisation des moindres carrés, nous adoptons les notations matricielles suivantes :

- $\boldsymbol{S} = [s(0)\ s(1)\ \ldots\ s(N_a-1)]$ est le vecteur (ligne) des échantillons du signal à analyser correspondant à la fenêtre d'analyse.
- $\boldsymbol{w} = [w(0)\ w(1)\ \ldots\ w(N_a-1)]$ est le vecteur (ligne) des échantillons de la fenêtre d'analyse, et $\boldsymbol{W}$ est la matrice carrée $N_a \times N_a$ diagonale dont les éléments de la diagonale sont les éléments de $\boldsymbol{w}$.
- $\boldsymbol{C}_i = [a_i\ b_i]$ est le vecteur (ligne) des coefficients du modèle mis sous la forme (1.15) correspondant à la composante harmonique d'indice i.
- $\boldsymbol{C} = [\boldsymbol{C}_1\ \ldots\ \boldsymbol{C}_i\ \ldots\ \boldsymbol{C}_I]$ est le vecteur (ligne) des coefficients du modèle complet, c'est-à-dire pour toutes les composantes.
- $\boldsymbol{M}_i$ dénote la matrice de taille $2 \times N_a$ qui contient les termes de la i-ème composante du modèle mis sous la forme (1.15), ces termes étant évalués aux indices d'échantillons le long de la fenêtre d'analyse, c'est-à-dire de 0 à N_a-1 :

$$\boldsymbol{M}_i = \begin{bmatrix} 1 & \cos(\omega_i) & \cos(2\omega_i) & \cdots & \cos([N_a-1]\omega_i) \\ 0 & \sin(\omega_i) & \sin(2\omega_i) & \cdots & \sin([N_a-1]\omega_i) \end{bmatrix} \tag{1.17}$$

- Enfin $\boldsymbol{M}$ dénote la matrice de taille $2I \times N_a$ qui contient les termes de toutes les composantes du modèle mis sous la forme (1.15), ces termes étant évalués aux indices d'échantillons le long de la fenêtre d'analyse, c'est-à-dire de 0 à N_a-1. En d'autres termes, $\boldsymbol{M}$ est la concaténation en lignes des matrices $\boldsymbol{M}_i$ pour $i = 1$ à I :

$$\boldsymbol{M} = \begin{bmatrix} 1 & \cos(\omega_1) & \cos(2\omega_1) & \cdots & \cos([N_a-1]\omega_1) \\ 0 & \sin(\omega_1) & \sin(2\omega_1) & \cdots & \sin([N_a-1]\omega_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos(\omega_i) & \cos(2\omega_i) & \cdots & \cos([N_a-1]\omega_i) \\ 0 & \sin(\omega_i) & \sin(2\omega_i) & \cdots & \sin([N_a-1]\omega_i) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos(\omega_I) & \cos(2\omega_I) & \cdots & \cos([N_a-1]\omega_I) \\ 0 & \sin(\omega_I) & \cos(2\omega_I) & \cdots & \sin([N_a-1]\omega_I) \end{bmatrix} \tag{1.18}$$

Avec ces notations, le vecteur (ligne) des N_a échantillons du modèle sur la fenêtre d'analyse est donné par :

$$\hat{\boldsymbol{S}} = \boldsymbol{CM} \tag{1.19}$$

et l'erreur quadratique moyenne pondérée peut être reformulée par :

$$\varepsilon = \left\| \boldsymbol{W}\left(\hat{\boldsymbol{S}} - \boldsymbol{S}\right) \right\|^2 \tag{1.20}$$

On cherche donc le vecteur de coefficients qui minimise ε, soit :

$$\boldsymbol{C}_{opt} = \underset{C \in R^{2I}}{\arg\min} \left[(\boldsymbol{CM} - \boldsymbol{S})\boldsymbol{W}(\boldsymbol{CM} - \boldsymbol{S})^T \right] \tag{1.21}$$

On suppose que le nombre de coefficients du modèle est inférieur au nombre de points de mesures, c'est-à-dire $2I<N_a$. Dans ce cas, le vecteur optimal de coefficients au sens des moindres carrés pondérés est donné par :

$$\boldsymbol{C}_{opt} = \boldsymbol{SWM}^T (\boldsymbol{MWM}^T)^{-1} \tag{1.22}$$

Il s'agit d'un résultat classique en algèbre linéaire qu'on retrouve tout aussi classiquement dans les domaines du traitement du signal, de la théorie de l'estimation et des statistiques par exemple. Nous ne détaillerons donc pas ici les calculs qui peuvent être trouvés dans de nombreuses références [Levinson, 1947] [Wiener, 1949] [Papoulis, 1977] [Kay, 1993] [Golub & van Loan, 1983]. Notons juste que ce résultat est obtenu en dérivant ε par rapport à chaque coefficient du modèle, puis en annulant le résultat de cette dérivation.

En pratique, si le nombre de composantes est très grand, ce calcul peut devenir assez lourd. C'est pourquoi, à une époque où la puissance de calcul des machines n'était pas ce qu'elle est aujourd'hui, certains auteurs ont proposé d'effectuer un calcul des paramètres de façon itérative composante par composante, plutôt que le calcul global de (1.22). Ainsi, dans la technique proposée par George et Smith [George & Smith, 1997], on commence par estimer les coefficients pour la première composante par :

$$\boldsymbol{C}_1 = \boldsymbol{SWM}_1^T (\boldsymbol{M}_1\boldsymbol{WM}_1^T)^{-1} \tag{1.23}$$

Ce qui correspond à minimiser :

$$\varepsilon_1 = \left\| \boldsymbol{W}\left(\hat{\boldsymbol{S}}_1 - \boldsymbol{S}\right) \right\|^2 \tag{1.24}$$

où $\hat{\boldsymbol{S}}_i$ désigne le modèle réduit à la composante i (c'est-à-dire $\hat{\boldsymbol{S}}_i = \boldsymbol{C}_i\boldsymbol{M}_i$). Puis on retire du signal cette composante estimée avec :

$$\boldsymbol{R}_1 = \boldsymbol{S} - \hat{\boldsymbol{S}}_1 = \boldsymbol{S} - \boldsymbol{C}_1\boldsymbol{M}_1 \tag{1.25}$$

Ensuite on réalise l'estimation de la composante suivante sur ce signal résiduel, selon le même principe de minimisation de l'erreur quadratique entre cette nouvelle composante et le signal résiduel :

$$\varepsilon_2 = \left\| \boldsymbol{W}\left(\hat{\boldsymbol{S}}_2 - \boldsymbol{R}_1\right)\right\|^2 \tag{1.26}$$

et ainsi de suite pour l'ensemble des composantes. Ainsi pour tout $i \in [2, I]$, on itère selon :

$$\boldsymbol{C}_i = \boldsymbol{R}_{i-1}\boldsymbol{W}\boldsymbol{M}_i^T\left(\boldsymbol{M}_i\boldsymbol{W}\boldsymbol{M}_i^T\right)^{-1} \tag{1.27}$$

$$\boldsymbol{R}_i = \boldsymbol{R}_{i-1} - \hat{\boldsymbol{S}}_i = \boldsymbol{R}_{i-1} - \boldsymbol{C}_i\boldsymbol{M}_i \tag{1.28}$$

A la fin des itérations, on a donc bien les paramètres du modèle estimés pour toutes les composantes, et le signal modélisé est donné par :

$$\hat{\boldsymbol{S}} = \sum_{i=1}^{I}\hat{\boldsymbol{S}}_i = \sum_{i=1}^{I}\boldsymbol{C}_i\,\boldsymbol{M}_i \tag{1.29}$$

Dans la pratique, pour un nombre modéré de composantes, tel que c'est le cas pour les signaux pseudo-harmoniques tels que la parole et les signaux de musique monophonique (un seul instrument), les deux versions, globale et itérative, de l'estimateur *WMMSE* donnent des résultats très similaires. La deuxième version est un peu sous-optimale par rapport à la première au sens où elle fournit une erreur quadratique globale (c'est-à-dire entre le signal et toutes les composantes considérées ensemble) un peu supérieure à la première version. Cependant, elle apporte une information fiable sur la valeur du spectre aux fréquences harmoniques pour un coût de calcul inférieur, et dans nos études, nous avons utilisé cette méthode (encore une fois, avec des réglages spécifiques que nous détaillerons dans la Section 3.3.1.2). La Figure 1.5 illustre la comparaison entre les résultats de cette méthode et ceux obtenus par une analyse du type transformée de Fourier discrète + *peak-picking* (voir Section 1.2.1.1) pour le paramètre de phase. Au-delà du simple tracé, les résultats sont très proches, et cette procédure fournit donc une estimation des paramètres très précise avec un coût de calcul très bas.

Notons que d'après les relations de Parseval, l'erreur quadratique moyenne entre le signal et le modèle sinusoïdal définie ici dans le domaine temporel est égale à l'erreur quadratique moyenne dans le domaine spectral, c'est-à-dire entre les spectres de ces signaux. Par conséquent, la minimisation de cette erreur peut être réalisée dans le domaine spectral : elle correspond alors aux méthodes d'ajustement spectral qui consistent à trouver les paramètres qui « fittent » au mieux le spectre du modèle sur celui du signal. Toutefois, des critères plus ou moins sensiblement différents peuvent être utilisés dans ce domaine spectral : par exemple on peut réaliser un ajustement au sens des moindres carrés des modules du spectre seulement, sans tenir compte des phases. Quoi qu'il en soit, il existe des liens théoriques et expérimentaux forts entre l'approche temporelle et l'approche spectrale de ce type d'analyse (y compris avec l'approche du type TFD + *peak-picking*). C'est pourquoi, on ne détaillera pas plus cette approche spectrale ici. Pour plus de détails, on peut se reporter par exemple à [Makhoul, 1975], [Griffin & Lim, 1988], et [McAulay & Quatieri, 1986].

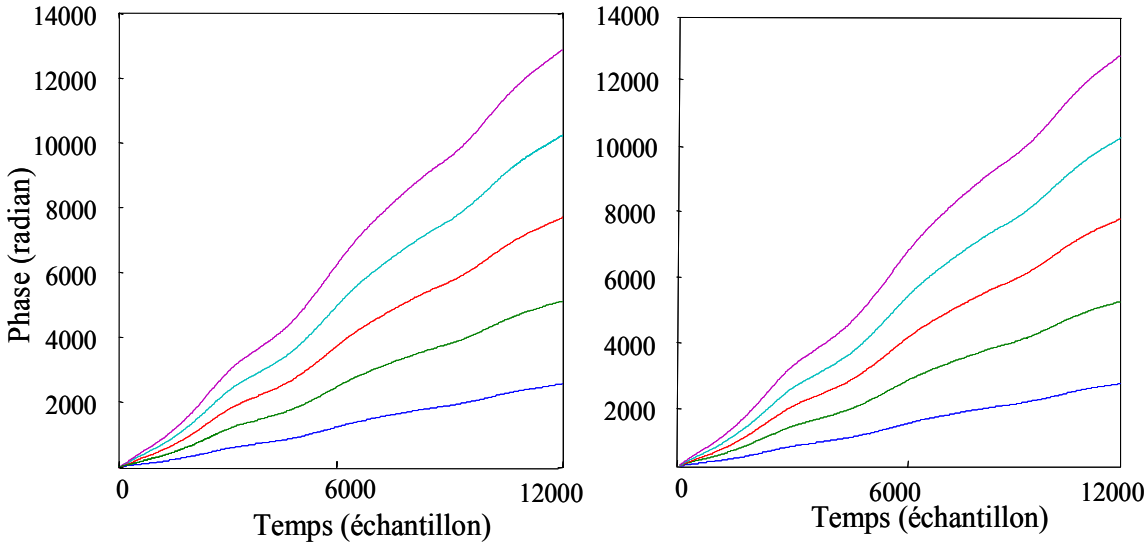

Figure 1.5 : Comparaison des résultats de mesure de la phase extraite par la méthode d'analyse-par-synthèse (MMSE) et TFD + *peak-picking*. L'analyse est réalisée ici en déplaçant la fenêtre d'analyse échantillon par échantillon sur une section de parole relativement longue (environ 1,5s). Le dépliement de phase est réalisé au fur et à mesure. On a représenté les résultats pour les cinq premières harmoniques.

1.2.1.3. Limites fondamentales de ces méthodes

Comme on l'a déjà mentionné, dans le modèle sinusoïdal avec hypothèse de stationnarité locale, les paramètres sont supposés localement constants. L'utilisation de ce modèle nécessite donc une bonne résolution temporelle, c'est-à-dire que le nombre d'échantillons N nécessaires à l'estimation spectrale à base de TFD doit être le plus faible possible pour que les paramètres puissent être raisonnablement considérés comme constants sur la fenêtre considérée. Parallèlement, l'analyse spectrale de sons comprenant potentiellement un grand nombre de composantes implique une bonne résolution fréquentielle, c'est-à-dire un grand nombre de points pour décrire le spectre discret. Le nombre d'échantillons temporel N étant égal au nombre d'échantillons spectraux (multiplié par deux si on ne considère que les fréquences positives, et sans *zero-padding*), N doit donc être grand. On voit donc ainsi naître une limite classique de ces méthodes opposant les contraintes en résolution temporelle et résolution spectrale, et nécessitant toujours un compromis entre ces deux contraintes[21]. Comme on l'a déjà mentionné précédemment, en pratique, une fenêtre de quelques dizaines de millisecondes est le plus souvent utilisée pour les signaux audio et la parole.

Pour pallier ces limites, une alternative possible consiste à considérer un modèle non-stationnaire où les paramètres des sinusoïdes varient dans l'intervalle considéré. Dans la section suivante, on donne un bref aperçu des méthodes d'analyse dans le cadre non-stationnaire.

[21] Même si on ne passe pas directement par la TFD pour les estimations des amplitudes et phases dans la méthode présentée à la Section 1.2.1.2, on a le même souci de résolution temporelle/spectrale pour construire les estimateurs de fréquence utilisés pour l'estimation des amplitudes et phases.

1.2.2. Méthodes d'analyse prenant en compte la non stationnarité à l'échelle locale

On a vu à la Section 1.1.4.1 qu'en pratique, les signaux sonores évoluent plus ou moins régulièrement au cours du temps. L'hypothèse de stationnarité locale est donc en toute rigueur abusive et en tout cas mal adaptée au cas où les variations temporelles des caractéristiques des signaux sont significatives (par rapport aux variations de la fonction d'onde). Dans ce cas, on peut chercher à raffiner les méthodes d'analyse pour prendre en compte cette non-stationnarité. Dans cette section on présente rapidement quelques méthodes d'analyse faisant l'hypothèse relativement réaliste (sur des fenêtres suffisamment courtes) que les paramètres du signal évoluent selon un modèle simple sur la fenêtre d'analyse : les fréquences évoluent linéairement et les amplitudes évoluent selon une loi exponentielle (c'est-à-dire linéairement sur une échelle logarithmique). En gardant les notations des sections précédentes et en notant respectivement δ_ω et δ_A les facteurs de modulation de fréquence et d'amplitude[22], supposés constants sur la fenêtre d'analyse, on a avec ce modèle :

$$\begin{cases} \hat{s}(n) = \sum_{i=1}^{I} A_i(n)\cos[\varphi_i(n)] \\ \varphi_i(n) = \varphi_i(0) + n\omega_i + \dfrac{\delta_{\omega_i} n^2}{2} \\ A_i(n) = A_i(0)10^{\frac{\delta_{A_i}}{20}n} \end{cases} \tag{1.30}$$

Il s'agit donc d'un cas particulier des équations non-stationnaires générales de la Section 1.1.4.3. Nous ne faisons ci-dessous qu'une présentation brève des méthodes d'analyse associées, car dans notre série d'études aux chapitres suivants nous n'utiliserons pas de telles méthodes d'analyse purement non-stationnaire, mais plutôt une adaptation de la méthode stationnaire de la Section 1.2.1.2.

1.2.2.1. Méthode d'analyse de Marques et Almeida

D'une façon générale, il n'est pas possible de calculer analytiquement la transformée de Fourier à court terme d'une sinusoïde dont la fréquence varie linéairement, pour n'importe quel type de fenêtre de pondération (en tout cas, pas pour les fenêtres usuelles en analyse spectrale, telles que les fenêtres de Hamming, Hanning, triangulaire, etc. …). Il n'est donc pas possible de généraliser directement les résultats de l'analyse stationnaire au cas non-stationnaire. Cette observation a une exception : le calcul donne un résultat analytique lorsque la fenêtre de pondération est une fenêtre gaussienne[23]. Dans [Marques & Almeida, 1986], il est proposé d'utiliser cette propriété, et les auteurs montrent qu'il existe une relation entre le facteur de modulation δ_ω et le lobe principal du spectre du signal. En effet, comme on l'a déjà mentionné à la Section 1.2.1.1, le lobe

[22] Ces facteurs sont définis en réalité pour chaque composante *i*. Lorsque ce n'est pas nécessaire, on omet par la suite l'indice *i* pour simplifier la présentation.

[23] Ce résultat repose sur le fait que la transformée de Fourier d'une fenêtre gaussienne en temps est une fenêtre gaussienne en fréquence.

principal du spectre d'amplitude d'une sinusoïde pondérée par une fenêtre gaussienne de variance σ^2 est une parabole sur une échelle log :

$$S_g(\omega) = \alpha\omega^2 + \beta\omega + \gamma \tag{1.31}$$

L'estimation de la fréquence et du facteur de modulation est donc réalisée à partir de celle des paramètres de la parabole. On a alors :

$$|\delta_\omega| = \pm\frac{1}{2\sigma^2}\sqrt{\frac{-\sigma^2}{2\alpha} - 1} \tag{1.32}$$

$$\hat{\omega} = \frac{-\beta}{\alpha} \tag{1.33}$$

Le signe de δ_ω est donné par la concavité ou la convexité du spectre de phase autour de la fréquence estimée. Après avoir estimé ces paramètres non stationnaires de fréquence et de facteur de modulation, les auteurs de [Marques & Almeida, 1986] fournissent les équations permettant de calculer les paramètres d'amplitude et de phase correspondants. Malheureusement, la fenêtre gaussienne n'est pas une fenêtre performante du point de vue de ses propriétés spectrales (moins bonne résolution que les fenêtres de Hamming ou Hanning par exemple), ce qui limite l'intérêt de cette méthode pour l'analyse de sons comportant un grand nombre de composantes.

1.2.2.2. Méthode d'analyse de Masri

L'utilisation de la transformée de Fourier à court terme sur des signaux non-stationnaires, associée à un *peak-picking* (comme vu à la Section 1.2.1.1) permet d'obtenir des valeurs de paramètres de type stationnaire, amplitudes et fréquences qui sont les moyennes de ces paramètres sur la fenêtre d'analyse. La TFCT semble ainsi moyenner les non-stationnarités du signal. Pourtant, cette transformée est inversible et la transformée inverse reconstruit parfaitement le signal non-stationnaire. Les non-stationnarités sont donc « cachées quelque part » dans le spectre résultant de la transformée : elles résident dans les distorsions fines du spectre par rapport au même spectre obtenu sur un signal stationnaire ayant des valeurs de paramètres égales aux moyennes des valeurs non-stationnaires.

Basé sur cette observation, Masri propose dans [Masri, 1996] de caractériser empiriquement les relations entre les modulations de fréquence et d'amplitude du modèle (1.30) et les déformations du spectre correspondantes pour pouvoir ensuite estimer les facteurs de modulation à partir de mesures spectrales. En particulier, il est montré que le spectre de phase est particulièrement sensible aux modulations. On a vu à la Section 1.2.1.1 que lorsqu'on utilise une fenêtre de type périodique avec décalage pour fenêtrer le signal (un fenétrage dit « à phase nulle »), le spectre de phase est constant dans la région correspondant au lobe principal. Une modulation d'amplitude linéaire en échelle log introduit une distorsion antisymétrique du spectre de phase, centrée sur l'indice correspondant au maximum local du spectre d'amplitude (indice le plus proche de la fréquence de la composante sinusoïdale analysée). De plus, l'orientation (pente positive ou négative) de la distorsion est reliée au signe de la modulation. L'effet de la modulation sur le spectre d'amplitude est généralement plus limité. De même, une modulation de fréquence linéaire introduit une distorsion symétrique du spectre de phase, toujours centrée sur l'indice du pic local, et

l'orientation (incurvation positive ou négative) de la distorsion dépend également du signe de la modulation. Masri propose alors une série d'abaques permettant de relier les valeurs des facteurs de modulation (d'amplitude et de fréquence) aux approximations des dérivées premières et secondes du spectre de phase en l'indice *m* de maximum local. Ces dérivées sont calculées à partir de la valeur du spectre de phase en cet indice *m* et en les deux indices voisins $m-1$ et $m+1$. Ces valeurs peuvent ensuite être utilisées pour réestimer la valeur de la phase correspondant à la composante modulée (au centre de la fenêtre d'analyse) par simple moyennage[24].

1.2.2.3. Autres méthodes

Les modulations de fréquence et d'amplitude modifient par ailleurs le spectre d'amplitude, même si c'est généralement de façon moins spectaculaire que pour le spectre de phase. Dans [Master & Liu, 2003], les auteurs proposent ainsi de modéliser le comportement de la largeur du lobe principal d'une composante sinusoïdale dans le domaine spectral en fonction du facteur de modulation de fréquence δ_ω. Ce modèle est basé sur des intégrales de Fresnel et délivre un estimateur du facteur de modulation qui n'est valable que pour des grandes valeurs de ce paramètre. Alternativement, pour des petites valeurs de δ_ω, Master et Liu proposent d'utiliser un estimateur basé à la fois sur une approximation du signal par une série de Taylor et sur une mesure du degré de courbure du spectre de phase au niveau du lobe principal (en utilisant une fenêtre symétrique à phase nulle, cf Section 1.2.1.1), rappelant ainsi le principe de l'estimateur de Masri décrit à la section précédente. Notons que ces auteurs étudient aussi la transition entre les deux domaines de valeurs de δ_ω.

Dans [Röbel, 2002], il est proposé d'utiliser la technique de réassignement spectral pour estimer la variation linéaire de fréquence. Par rapport à la méthode basée sur la fenêtre gaussienne de [Marques & Almeida, 1986], dans cette méthode de réassignement on peut employer différentes fenêtres de pondération. Comme nous l'avons déjà mentionné, le calcul de la fréquence réassignée du maximum local exige le spectre du signal fenêtré, ainsi que le spectre du signal fenêtré par la dérivée de la fenêtre. Dans cette extension de la méthode au cas non-stationnaire, l'estimation de δ_ω nécessite en plus le spectre du signal fenêtré par la dérivée seconde de la fenêtre [Röbel, 2002].

On peut encore citer l'étude de [Lagrange *et al.,* 2002] dans laquelle les auteurs reprennent et étendent le principe des estimateurs de Masri basés sur la déformation du spectre de phase par les facteurs de modulation. Dans cette étude, les auteurs proposent de plus plusieurs techniques, le réassignement temporel et la modulation de spectre, pour corriger l'estimation des paramètres sinusoïdaux corrompue par les modulations de fréquence et d'amplitude.

Enfin, d'autres méthodes utilisant des transformées temps/fréquence ou des méthodes de minimisation au sens des moindres carrés existent dans le cas non-stationnaire [Nieuwenhuijse *et al*., 1998] [Boyer & Abed-Meraim, 2002]. Ces méthodes MMSE peuvent être vues comme des généralisations de la méthode présentée à la Section 1.2.1.2, en remplaçant le modèle stationnaire par le modèle non-stationnaire

[24] Notons que comme dans le cas stationnaire, on peut aussi utiliser une valeur de fréquence estimée corrigée par une méthode de raffinement à la place de l'indice « grossier » *m*.

dont on cherche à extraire les paramètres. On peut d'ailleurs dans ce cadre de méthodes complexifier le modèle par rapport à l'équation (1.30). Par exemple, on peut chercher à modéliser les transitoires par des sinusoïdes amorties et une analyse basée sur un critère MMSE permet d'extraire les paramètres de l'enveloppe temporelle de ces sinusoïdes en plus des paramètres usuels de fréquence et d'amplitude. Ces méthodes sont souvent complexes et coûteuse en calculs et peuvent être basées sur le principe d'analyse-par-synthèse : on teste une succession de valeurs paramètres candidates et on retient celles qui ajustent le mieux le modèle aux valeurs du signal au sens du critère choisi (par exemple MMSE).

1.3. Synthèse du signal

Dans cette section, on présente les principales approches développées pour la partie synthèse du processus d'analyse-synthèse basé sur la modélisation sinusoïdale. Comme on l'a déjà vu à la Section 1.1.4.4, les différentes méthodes de synthèse peuvent être regroupées en deux grandes classes, la synthèse par *overlap-add* et la synthèse par interpolation. On présente à présent plus en détail ces méthodes dans cet ordre.

1.3.1. Synthèse par *Overlap-Add*

La technique de synthèse additive avec chevauchement ou recouvrement des trames de signal, appelée *Overlap-Add*, a été décrite en détail dans plusieurs articles, notamment dans les travaux de George et Smith [George & Smith, 1997]. Comme déjà mentionné à la Section 1.1.4.4, elle consiste à générer le signal à partir des paramètres stationnaires du modèle sinusoïdal sur des fenêtres de synthèse successives qui se recouvrent en partie. Ensuite, il faut assurer la continuité du signal par une pondération adéquate des parties recouvrantes. En effet, les paramètres extraits sur chaque fenêtre d'analyse servent à la resynthèse du signal sur chaque fenêtre de synthèse correspondante. Comme ces paramètres sont ceux d'un modèle qui approxime plus ou moins bien le signal, les segments de signaux resynthétisés ne sont en général pas identiques aux segments originaux (sans compter le fait que ces paramètres peuvent être modifiés volontairement entre l'analyse et la synthèse si on veut transformer le signal). Par conséquent, la continuité exacte du signal entre deux fenêtres de synthèse n'est pas forcément assurée. Si on concatène directement les fenêtres de synthèse les unes à la suite des autres, les discontinuités engendrent généralement des artefacts très gênants sonnant comme des « clics ». Le rôle de la pondération est donc de lisser ces transitions inter-trames. On utilise pour cela des fenêtres de pondération avec croissance et décroissance régulières sur les bords de la fenêtre.

Plus formellement et avec les notations habituelles, dans ce type de méthode chaque trame de synthèse (indicée par k) est donnée par[25] :

$$\hat{s}^k(n) = \sum_{i=1}^{I_k} A_{i,k} \cos(\omega_{i,k} n + \varphi_{i,k}) \tag{1.34}$$

[25] Dans [George & Smith, 1997], les auteurs raffinent ce modèle en traitant séparément l'enveloppe temporelle globale du signal comme un paramètre multiplicatif supplémentaire, interpolé en chaque indice temporel n.

Le signal complet est donné par la sommation pondérée de ces trames :

$$\hat{s}(n) = \sum_{k=-\infty}^{\infty} w_s(n - kN_s)\hat{s}^k(n - kN_s) \tag{1.35}$$

Pour assurer la régularité de la transition inter-trames et en particulier la conservation correcte de l'enveloppe du signal, la fenêtre doit satisfaire la condition selon laquelle la somme de tous les coefficients correspondants aux différentes fenêtres se chevauchant doit être égale à 1 :

$$\sum_{k=-\infty}^{\infty} w_s(n - kN_s) = 1 \tag{1.36}$$

Plusieurs types de fenêtre de pondération sont bien sûr possibles. En pratique l'utilisation de fenêtres telles que la fenêtre de Hanning est en adéquation avec le processus d'analyse et permet une reproduction fidèle du signal (voir la Figure 1.6).

Plusieurs remarques peuvent venir compléter ce tableau. D'abord, en ce qui concerne la méthode employée pour la synthèse proprement dite du signal sur chaque fenêtre avant pondération, on peut noter qu'on peut employer directement une somme d'oscillateurs dans le domaine temporel, ou bien on peut réaliser une synthèse efficace par transformée de Fourier rapide (*FFT*) inverse [Freed *et al.*, 1993]. Une discussion détaillée autour des algorithmes optimaux pour la synthèse du signal dépasse le cadre de travail de ce manuscrit.

Ensuite, l'analyse-synthèse par *overlap-add* a été présentée jusqu'ici dans le contexte stationnaire : c'est-à-dire qu'on rappelle que le signal est supposé stationnaire sur la fenêtre d'analyse (et implicitement sur celle de synthèse). C'est la juxtaposition des fenêtres qui permet de retrouver la non-stationnarité du signal. On peut noter que cette méthode peut être appliquée sur une version non-stationnaire du modèle sinusoïdal : pour chaque fenêtre, on peut réaliser la synthèse des composantes avec des équations non-stationnaires semblables par exemple à l'équation (1.2) si les paramètres non-stationnaires correspondant ont été analysés, puis faire intervenir la pondération avec recouvrement entre les fenêtres successives de la même façon que dans le cas du modèle stationnaire ci-dessus. On cherche donc dans ce cas à caractériser et restituer l'évolution du signal à l'intérieur même des fenêtres d'analyse et de synthèse avant de réaliser la phase d'*overlap-add*, même si la taille de ces fenêtres est relativement petite (on rappelle qu'il s'agit de quelques dizaines de millisecondes).

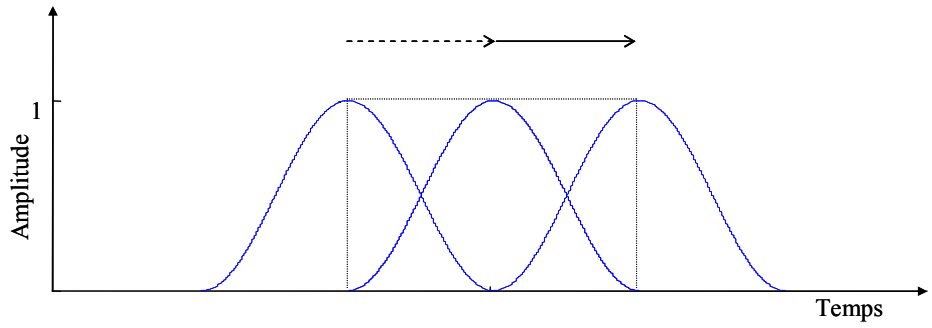

Figure 1.6 : Illustration du recouvrement des trames de synthèse avec une fenêtre de Hanning (figure réalisée d'après [Marchand, 2000]).

1.3.2. Synthèse par interpolation

Dans cette section on décrit de façon plus précise les techniques mises en œuvre dans la synthèse par interpolation. On décrit d'abord le suivi de partiels[26], puis les différents modèles utilisés dans la littérature pour réaliser l'interpolation des paramètres proprement dite.

1.3.2.1. Principe du suivi de partiels

Comme déjà brièvement mentionné à la Section 1.1.4.4, une fois que les sinusoïdes ont été détectées et leurs paramètres estimés pour chaque trame d'analyse, il est nécessaire avant de réaliser l'interpolation de leurs valeurs d'une trame de synthèse à l'autre, d'appairer d'abord les jeux de paramètres correspondants de façon cohérente. En effet, en toute généralité, le nombre de ces jeux de paramètres peut varier d'une trame à l'autre, du fait de l'évolution du signal. Certaines composantes sinusoïdales peuvent apparaître dans la nouvelle trame alors que d'autres de la trame précédente ont disparu, et inversement. De plus, les valeurs des fréquences (et amplitudes) varient au cours du temps. Certaines trajectoires de paramètres au cours du temps peuvent être relativement « lisibles » car régulières alors que d'autres sont plus « tumultueuses ». Pour permettre l'interpolation inter-trames, il est donc nécessaire d'effectuer ce qu'on appelle un suivi des trajectoires des paramètres le long des différentes trames : le suivi de partiels (*partial-tracking*). Son rôle est plus formellement de déterminer à quelle sinusoïde de la trame k+1 correspond la sinusoïde d'indice i dans la trame courante k.

La plupart des méthodes de suivi de partiels se basent sur la proximité des fréquences des composantes de trames adjacentes. Les principaux travaux faisant référence dans ce domaine sont ceux de McAulay et Quatieri d'une part [McAulay & Quatieri, 1986] et ceux de Serra et Smith d'autre part [Smith & Serra, 1987] [Serra & Smith, 1990]. On donne ici uniquement les grandes lignes des techniques possibles, et on se reportera à ces références pour plus de détails. En général, on associe à la sinusoïde i dans la trame k de fréquence $\omega_{i,k}$ la sinusoïde j dans la trame k+1 de fréquence $\omega_{j,k+1}$, la plus proche de $\omega_{i,k}$. C'est le principe de base de l'appariement. Cette recherche peut être effectuée en plusieurs passes. Elle peut aussi être effectuée en mode *forward*, c'est-à-dire de la trame k vers la trame k+1, mais aussi en mode *backward*, c'est-à-dire de la trame k+1 vers la trame k, ou en combinant les deux. Si une composante de la trame k ne trouve pas de successeur dans la trame k+1, le partiel auquel il appartient est déclaré « mort » et se termine donc à la trame k. Pour assurer la continuité du signal, on « éteint » progressivement cette composante en faisant tendre progressivement son amplitude vers zéro en conservant sa fréquence constante. Parallèlement, si un pic de la trame k+1 n'est pas lié à un pic de la trame k, un partiel « naît ». Pour assurer la continuité du signal, l'amplitude de ce partiel est cette fois progressivement augmentée à partir de zéro. Un résumé graphique du suivi de partiels ainsi réalisé est donné à la Figure 1.7.

[26] Notons que bien qu'on présente cette étape dans la partie synthèse, on pourrait plutôt penser que le suivi de partiel est du ressort de l'analyse du signal. On a fait ce choix uniquement parce que cette tâche est nécessaire pour les méthodes de synthèse par interpolation, mais pas pour la synthèse par recouvrement pour laquelle la « jonction » des composantes est réalisée « globalement » par la pondération. Il n'est donc pas surprenant de retrouver ce suivi de partiels dans la partie dédiée à la synthèse par interpolation.

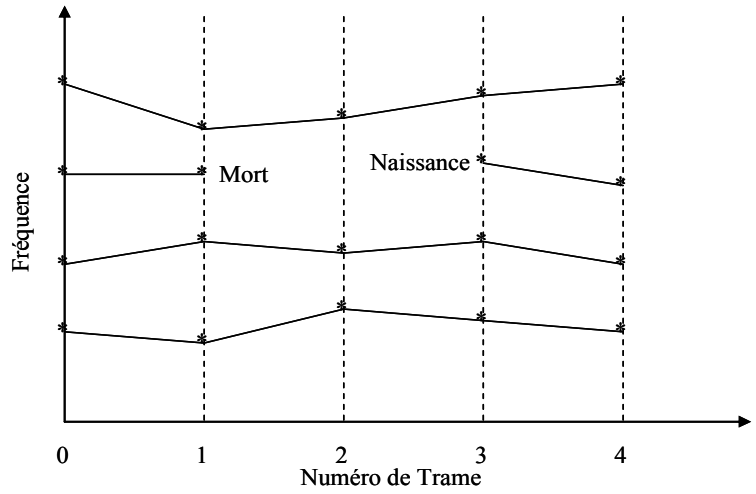

Figure 1.7 : Résultat d'un algorithme typique de suivi de partiels (d'après [McAulay & Quatieri, 1986]).

D'autres approches ont été proposées depuis les travaux de [McAulay & Quatieri, 1986] et [Smith & Serra, 1987] pour améliorer cette technique. Par exemple, une méthode récente proposée dans [Lagrange, 2004] consiste à utiliser la prédiction linéaire pour prédire le suivi de partiels. Lagrange introduit un critère additionnel qui permet de s'assurer que le prolongement possible pour un partiel donné n'engendre pas de hautes fréquences dans les vecteurs de fréquence et d'amplitude du partiel. Grâce à cette modélisation autorégressive plus fine, de nombreux pics de bruit peuvent être éliminés.

Notons pour finir que le suivi de partiels se simplifie énormément dans le cas où on fait l'hypothèse d'harmonicité puisque, dans ce cas, il se résume généralement à assurer la continuité entre les harmoniques de même rang (voir Section 1.4.1).

1.3.2.2. Interpolation des paramètres

Une fois que le suivi de partiels a été réalisé, il faut réaliser l'interpolation proprement dite, c'est-à-dire trouver une fonction de passage d'une valeur de paramètre à une autre assurant certaines propriétés de continuité. On décrit rapidement l'interpolation des paramètres d'amplitude avant de s'intéresser au cas plus délicat de l'interpolation des phases et des fréquences.

1.3.2.2.1. Interpolation des amplitudes

L'idée de base retenue par McAulay et Quatieri [McAulay & Quatieri, 1986] ou Smith et Serra [Smith & Serra, 1987] pour l'amplitude est une simple interpolation linéaire entre les trames k et k+1 :

$$\hat{A}_i(n) = A_{i,k} + \frac{A_{i,k+1} - A_{i,k}}{N} n \tag{1.37}$$

où N est le nombre d'échantillons entre les centres des trames k et k+1. Malgré sa simplicité, ce choix donne de bons résultats et l'emploi de ce modèle s'est généralisé en synthèse audio. On peut remarquer que cette interpolation linéaire peut être réalisée sur

les amplitudes prises sur une échelle elle-même linéaire, ou avec les amplitudes prises sur une échelle logarithmique. Elle correspond dans ce deuxième cas à une interpolation exponentielle sur les amplitudes linéaires, ce qui est relativement bien adapté à la dynamique des signaux audio (voir le modèle de la Section 1.2.2).

1.3.2.2.2. Interpolation des phases par le modèle de McAulay & Quatieri

L'interpolation des phases est un problème plus délicat que celui de l'interpolation des amplitudes. En effet, les paramètres de phases sont directement reliés aux paramètres de fréquence puisque les fréquences, en tant que fonction du temps, sont les dérivées des phases, elles-mêmes prises en tant que fonction du temps (voir Section 1.1.4.3). On a alors le choix entre plusieurs alternatives pour l'interpolation de la phase : soit on tient compte à la fois des mesures des phases et des fréquences, soit on ne tient compte que des unes ou des autres. Ces différents cas sont développés par la suite. Dans cette sous-section, on se place dans le premier cas, et on décrit le modèle le plus fameux réalisant ces conditions, qui est certainement celui proposé par McAulay et Quatieri dans [McAulay & Quatieri, 1986].

Le principe de base de la méthode de McAulay et Quatieri repose sur le fait qu'on possède quatre contraintes pour réaliser l'interpolation d'une trajectoire de phase d'une trame k à la suivante (pour chaque partiel i) : ces quatre contraintes sont les deux mesures de fréquence $\omega_{i,k}$ et $\omega_{i,k+1}$, et les deux mesures de phase[27] $\varphi_{i,k}$ et $\varphi_{i,k+1}$, correspondant aux trames k et k+1. On choisit par conséquent un modèle de phase ayant quatre degrés de liberté à ajuster aux mesures. Ce modèle est un polynôme d'ordre trois (avec donc quatre coefficients) :

$$\hat{\varphi}_i(n) = \varphi_{i,k} + \omega_{i,k} n + \alpha n^2 + \beta n^3 \tag{1.38}$$

Les deux premiers coefficients du polynôme sont donnés directement par les deux premières contraintes[28] :

$$\begin{cases} \hat{\varphi}_i(0) = \varphi_{i,k} \\ \left.\dfrac{d\hat{\varphi}_i(n)}{dn}\right|_{n=0} = \omega_{i,k} \end{cases} \tag{1.39}$$

Les deux autres contraintes de continuité des paramètres du modèle sont :

$$\begin{cases} \hat{\varphi}_i(N) = \varphi_{i,k+1} + 2\pi M \\ \left.\dfrac{d\hat{\varphi}_i(n)}{dn}\right|_{n=N} = \omega_{i,k+1} \end{cases} \tag{1.40}$$

[27] Les paramètres de phase sont ici les valeurs des phases *absolues* (voir Section 1.1.4.3) aux points de jonction des trames. Ces mesures sont fournies modulo 2π par la méthode d'analyse.

[28] On conserve abusivement les notations discrètes pour les expressions des dérivées pour simplifier la présentation et garder son homogénéité. Dans [McAulay & Quatieri, 1986], les développements mathématiques sont présentés dans un cadre formaliste continu.

M est un facteur de « dépliement de phase » prenant en compte le fait que les mesures de phase sont fournies à 2π près par le processus d'analyse alors que la phase absolue est en réalité incrémentée par sommation de la fréquence correspondante entre les deux instants de mesure. M est réglé dans [McAulay & Quatieri, 1986] selon un critère de minimisation de la dérivée seconde de la courbe de phase entre les deux instants de mesure considérés. Ce critère donne (e[.] désigne la partie entière) :

$$M = e\left[\frac{1}{2\pi}\left((\varphi_{i,k} - \varphi_{i,k+1}) + (\omega_{i,k} + \omega_{i,k+1})\frac{N}{2}\right)\right] \tag{1.41}$$

On peut alors montrer que les contraintes conduisent à déterminer les paramètres du polynôme selon :

$$\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = \begin{bmatrix} 3/N^2 & -1/N \\ -2/N^3 & 1/N^2 \end{bmatrix}\begin{bmatrix}\varphi_{i,k+1} - \varphi_{i,k} - \omega_{i,k}N + 2\pi M \\ \omega_{i,k+1} - \omega_{i,k}\end{bmatrix} \tag{1.42}$$

1.3.2.2.3. Autres modèles prenant en compte phases et fréquences

D'autres modèles d'interpolation ont été proposés qui prennent en compte à la fois les mesures de phase et de fréquence. Dans [Ding & Qian, 1997] un modèle polynomial quadratique est proposé pour interpoler la trajectoire de phase entre les points de mesure. Le critère d'ajustement de ce modèle est un critère des moindres carrés portant à la fois sur les mesures de phase et les mesures de fréquence, avec un facteur de pondération entre ces deux contraintes. C'est pourquoi ce modèle s'apparente à une interpolation de type *spline* [Unser *et al.,* 1993] [Cohen *et al.,* 2001]. Les auteurs de [Ding & Qian, 1997] affirment que ce modèle permet de résoudre certains artefacts du modèle de [McAulay & Quatieri, 1986], notamment en assurant un meilleur lissage des trajectoires de fréquence.

Dans la même optique, dans [Girin *et al.*, 2003], le polynôme décrivant la phase entre deux instants de mesure est élevé à l'ordre cinq, pour prendre en compte deux contraintes supplémentaires intégrant les mesures des *dérivées de fréquence* $\delta\omega_{i,k}$ aux jonctions de trame :

$$\hat{\varphi}_i(n) = \varphi_{i,k} + \omega_{i,k}n + \frac{\delta\omega_{i,k}}{2}n^2 + \alpha n^3 + \beta n^4 + \gamma n^5 \tag{1.43}$$

On se rapproche ainsi de l'esprit du modèle de la Section 1.2.2 afin de mieux tenir compte des non-stationnarités du signal extraites à la phase d'analyse. Les paramètres du polynôme sont calculés en fonction des valeurs mesurées selon une démarche similaire à celle de la méthode de McAulay et Quatieri. Le facteur de dépliement devient alors :

$$M = e\left[\frac{1}{2\pi}\left((\varphi_{i,k} - \varphi_{i,k+1}) + (\omega_{i,k} + \omega_{i,k+1})\frac{N}{2} + (\delta\omega_{i,k} - \delta\omega_{i,k+1})\frac{N^2}{40}\right)\right] \tag{1.44}$$

Les paramètres α, β, γ de (1.42) sont donnés par [Girin *et al*., 2003] :

$$\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix}=\begin{bmatrix}10N^3 & -4/N^2 & 1/(2N)\\ -15/N^4 & 7/N^3 & -1/N^2\\ 6/N^5 & -3/N^4 & 1/(2N^3)\end{bmatrix}\begin{bmatrix}\varphi_{i,k+1}-\varphi_{i,k}-N\omega_{i,k}-N^2\dfrac{\delta\omega_{i,k}}{2}+2\pi M\\ \omega_{i,k+1}-\omega_{i,k}-N\delta\omega_{i,k}\\ \delta\omega_{i,k+1}-\delta\omega_{i,k}\end{bmatrix} \quad (1.45)$$

1.3.2.2.4. Interpolation linéaire des phases ou des fréquences

Pour compléter cet éventail de méthodes d'interpolation, on peut citer l'interpolation linéaire de la phase. Réaliser une telle interpolation revient à considérer la fréquence comme constante sur la fenêtre de synthèse et on revient ainsi à un modèle intrinsèquement stationnaire sur la fenêtre de synthèse. Cependant, comme on veut conserver la continuité des phases mesurées pour conserver la continuité exacte du signal entre les points de jonction de trames consécutives, on appliquera l'équation suivante à partir des mesures de phase uniquement, sans considérer les mesures de fréquence (où M est le facteur de dépliement de phase défini précédemment) :

$$\hat{\varphi}_i(n)=\varphi_{i,k}+\frac{\varphi_{i,k+1}-\varphi_{i,k}+2\pi M}{N}n \quad (1.46)$$

Ce modèle ne tient donc pas compte des mesures de fréquence (en dehors de leur utilisation dans le facteur de dépliement bien sûr) et rien ne permet d'assurer que la pente de ce modèle (c'est-à-dire la fréquence correspondante) corresponde exactement aux valeurs de fréquence mesurées[29], même si dans la pratique, les valeurs sont généralement assez proches.

Alternativement, on peut reconstruire une trajectoire de phase continue uniquement à partir des fréquences mesurées, en appliquant un modèle linéaire sur les fréquences mesurées :

$$\hat{\omega}_i(n)=\omega_{i,k}+\frac{\omega_{i,k+1}-\omega_{i,k}}{N}n \qquad \text{et} \qquad \hat{\varphi}_i(n)=\hat{\varphi}_i(0)+\sum_{m=1}^{n}\hat{\omega}_i(m) \quad (1.47)$$

Dans ce cas, bien que la continuité de phase soit assurée, les valeurs des phases aux jonctions des trames ne correspondent pas forcément aux valeurs mesurées. On peut donc avoir un déphasage significatif (relativement stable ou bien évoluant selon la précision des erreurs de mesure de fréquence) entre le signal original et le signal resynthétisé. Comme ce déphasage n'a généralement que peu d'incidence sur la qualité du signal synthétisé, ce modèle a été utilisé dans bon nombre de travaux, dont les travaux pionniers de Serra et Smith qui seront plus largement abordés à la Section 1.4.2.2 [Serra, 1987] [Serra & Smith, 1990]. Une discussion plus approfondie sur l'influence du respect de la phase sur la qualité du signal de parole et de musique est menée à la Section 4.1.2.

[29] Ce point s'oppose au modèle linéaire de la Section 1.3.1, qui considère à la fois les phases et les fréquences mesurées, mais qui inversement n'assure pas la continuité exacte des phases à la jonction des trames, d'où la nécessité de l'*Overlap-Add*.

1.4. Variantes et raffinements du modèle sinusoïdal

Dans cette section, nous présentons une série de variantes et de raffinements qui ont été proposés dans la littérature afin de simplifier ou d'améliorer, selon les cas, le modèle sinusoïdal. Le but est de permettre à ce modèle de mieux « coller » aux différents types de signaux audio et d'améliorer ainsi sensiblement la qualité des signaux modélisés.

1.4.1. Modèle sinusoïdal harmonique

Jusqu'ici, on a présenté le modèle sinusoïdal sous sa forme générale dans le cas d'un signal non nécessairement harmonique. Dans le cas d'un signal périodique ou pseudo-périodique (voir la note de bas de page 10), l'analyse-synthèse sinusoïdale peut être grandement simplifiée [McAulay & Quatieri, 1986] [Stylianou, 1996] [George & Smith, 1997] [Oudot, 1998]. En effet, dans ce cas, on ne considère que les sinusoïdes positionnées aux multiples de la fréquence fondamentale ω_0 : la sinusoïde dont la fréquence est la plus proche de $i\omega_0$ est considérée comme l'harmonique de rang *i*. On « élimine » de fait toutes les sinusoïdes autres que les harmoniques[30], et donc on peut généralement significativement réduire le nombre de paramètres du modèle sinusoïdal par rapport à une version sans contrainte d'harmonicité. Dans ce cas, le signal est donc modélisé comme suit (les équations sont données respectivement en version stationnaire et non-stationnaire) :

$$\hat{s}(n) = \sum_{i=1}^{I} A_i \cos[i\omega_0 n + \varphi_i(0)] \tag{1.48}$$

ou bien

$$\hat{s}(n) = \sum_{i=1}^{I} A_i(n) \cos[\varphi_i(n)] \quad \text{avec} \quad \varphi_i(n) = \varphi_i(0) + \sum_{m=1}^{n} i\omega_0(m) \tag{1.49}$$

La validité de ce modèle peut être facilement confirmée sur de nombreux types de sons naturels, notamment ceux où une seule source est active. Par exemple, la parole dans les régions voisées est quasi-périodique et se prête bien à cette modélisation (voir la Figure 1.1 : la nature quasi-périodique du segment de parole voisée représenté est claire dans la forme d'onde du signal et dans la structure du spectre correspondant). Il en est de même pour de nombreux instruments de musique de part leur nature fondamentalement vibratoire.

Dans le cas où on utilise le modèle sinusoïdal harmonique à la place du modèle sinusoïdal général, on peut donc simplifier ou du moins adapter la phase d'analyse des paramètres. Ainsi, si on utilise une méthode d'analyse basée sur la TFD (voir Section 1.2.1.1), on peut restreindre la zone de recherche des pics aux zones environnant les multiples de la fréquence fondamentale. Bien entendu, cela nécessite

[30] Les autres pics du spectre sont alors considérés comme des artefacts de la méthode d'analyse ou bien comme des pics de bruit. Dans ce dernier cas, on peut associer au modèle sinusoïdal ou harmonique un modèle de bruit (voir la Section 1.4.2 ; le problème du « tri » entre pics de sinusoïdes et pics de bruit est brièvement présenté dans cette section).

d'abord une estimation de cette fréquence fondamentale[31], c'est pourquoi ce modèle harmonique est généralement utilisé en tandem avec un estimateur de fréquence fondamentale. On peut alors aussi utiliser une version simplifiée de la méthode d'analyse par la synthèse de la Section 1.2.1.2, où les fréquences sont imposées aux multiples de la fréquence fondamentale. Dans notre série d'études qui seront présentées dans les chapitres suivants, nous utilisons un tel procédé d'analyse.

Un point important émergeant lors de l'utilisation d'un modèle harmonique est la simplification du problème du suivi de partiels (voir Section 1.3.2.1) entre trames consécutives faisant partie d'une portion de signal considérée comme continûment pseudo-périodique. En effet, le suivi de partiels se réduit alors simplement à relier les harmoniques de même rang. Comme la fréquence fondamentale évolue au cours du temps, le nombre d'harmoniques n'est pas nécessairement constant, tout comme le nombre de composantes dans le cas général du modèle sinusoïdal. Mais ici le problème est localisé aux harmoniques en limite de la fréquence de Nyquist, et pour pallier ce problème, on a recours à nouveau à la possibilité de faire naître et mourir les harmoniques en question (en faisant tendre ou démarrer leur amplitude à zéro avec une fréquence constante pendant cette phase transitoire). A la jonction entre une portion pseudo-périodique de signal et une portion plus complexe (non harmonique), on retourne vers les techniques générales de gestion des transitions (par suivi de partiels et interpolation ou par *overlap-add* par exemple).

1.4.2. Modèle sinusoïdal + bruit

1.4.2.1. Principe et intérêt

Définir un son par une somme limitée de sinusoïdes dont les amplitudes et les fréquences varient lentement dans le temps implique une certaine restriction sur la catégorie de sons à considérer. En effet, la présence de bruits (comme les sons fricatifs de la parole et les composantes « soufflées » des instruments de musique à vent par exemple) ou de transitoires (comme les consonnes plosives par exemple) perturbe

[31] Le problème de l'estimation du pitch est un problème fondamental en traitement de parole du fait de l'importance de cette information dans nombre de systèmes de traitement de parole, que ce soit en analyse, codage, ou synthèse pour ne citer qu'eux. C'est pourquoi ce problème a été l'objet de très nombreuses études en traitement de la parole, depuis environ 40 ans. On ne présentera pas dans ce document l'ensemble des méthodes proposées dans la littérature, car certaines d'entres elles peuvent être vues comme des cas particuliers des méthodes d'analyse des paramètres de fréquences du modèle sinusoïdal vues à la Section 1.2. Ainsi, par exemple, on vient de mentionner que les méthodes de type *peak-picking* peuvent bénéficier de l'estimation de la fréquence fondamentale dans le cas de signaux harmoniques, mais inversement le *peak-picking* appliqué au premier lobe significatif d'un spectre peut servir de base à l'estimation de la fréquence fondamentale. On peut même chercher à mettre en correspondance un modèle de spectre harmonique complet avec le spectre du signal pour déterminer le fondamental. Une autre famille de méthodes très répandues d'algorithme d'estimation du pitch fonctionnant dans le domaine temporel est basée sur la recherche du maximum de la fonction d'autocorrélation du signal (après un filtrage passe-bas de celui-ci) [Gerhard, 2003] [Boersma, 1993]. Pour une revue assez globale et cependant relativement récente de ce problème d'estimation de fréquence fondamentale, aussi bien pour la parole que pour les signaux de musique, on peut citer entre autres [Paliwal, 1983] [Ahmadi & Spanias, 1999] [Manfredi *et al.*, 2000] [Liu & Lin, 2001] [Shimamura & Kobayashi, 2001] [de Cheveigné & Kawahara, 2002] [Gerhard, 2003]. On peut remarquer que les travaux récents dans ce domaine en parole se concentrent plus particulièrement sur le problème plus difficile de l'estimation de la fréquence fondamentale en milieu bruité.

généralement les différentes étapes de la modélisation. Par exemple, la procédure d'analyse peut considérer des zones de bruit relativement diffuses dans le spectre comme une série de pics locaux alors que ceux-ci ne correspondent pas à des partiels significatifs et ainsi entraîner de mauvaises connections dans le suivi de partiels. D'une manière plus générale, un bruit (additif) va perturber l'estimation des paramètres des sinusoïdes. L'étude de l'influence de ce bruit sur l'analyse est un pan important du domaine de l'analyse spectrale (qui dépasse largement le cadre de ce document) [Serra & Smith, 1990] [Rodet, 1997] [Stylianou, 1996]. Par ailleurs, à la synthèse, il est effectivement possible d'utiliser des sinusoïdes pour représenter des signaux de bruit, mais ceci est très coûteux en nombre de sinusoïdes car un spectre de bruit est généralement riche et diffus[32]. Pour plus d'efficacité et pour une meilleure complémentarité entre la modélisation des composantes sinusoïdales et des composantes bruitées d'un même son, il est donc proposé de modéliser le signal sous la forme d'un modèle additif sinusoïdal plus bruit, en rajoutant une composante additive stochastique à la somme de sinusoïdes. On a alors :

$$s(n) = \sum_{i=1}^{I} A_i(n)\cos[\varphi_i(n)] + b(n) \qquad (1.50)$$

1.4.2.2. Une étude de référence : Serra et Smith, 1990

Ce modèle additif appelé *modèle sinusoïdes + bruit* a été introduit originellement par Serra et Smith à la fin des années 1980 [Serra, 1987] [Serra & Smith, 1990]. Dans cette série d'études, Serra et Smith proposent de détecter et d'identifier d'abord les pics principaux du spectre supposés correspondre aux composantes sinusoïdales. Pour cela, ils utilisent d'abord un algorithme basé sur la transformée de Fourier à court terme, la détection de pics, et l'interpolation parabolique, tel que le procédé décrit à la Section 1.2.1.1. Puis cet algorithme est raffiné par une procédure de sélection des pics basée sur la prédominance de chaque maximum par rapport à son voisinage. Les pics significativement émergeants sont donc attribués aux composantes sinusoïdales alors que les maxima locaux plus diffus dans le spectre ou plus réduits en amplitude sont attribués aux zones de bruit[33]. On soustrait alors du spectre les pics principaux directement dans le domaine du spectre d'amplitude, et le spectre résiduel, défini sur toute la bande du signal, est supposé être celui du bruit :

$$|B(m)| = |S(m)| - |H(m)| \qquad (1.51)$$

où $|B(m)|$ représente le module de la transformée de Fourier discrète du signal de bruit résiduel $b(n)$, $|S(m)|$ celui du signal original, et où $|H(m)|$ est un spectre d'amplitude

[32] Cette possibilité a été proposée dans plusieurs études. Ainsi, dans [McAulay & Quatieri, 1995], les auteurs affirment que la représentation sinusoïdale reste valide pour les sons de parole non voisés, en employant le modèle sinusoïdal avec une phase aléatoire et quelques précautions. Ce principe est aussi à la base d'une étude par Macon [Macon, 1996] : le caractère aléatoire du paramètre de phase sur des segments à court terme assure le caractère aléatoire du signal synthétisé. Notons qu'il existe d'autres approches plus spécifiques pour modéliser les composantes de bruit de la parole (voir par exemple [Richard & d'Alessandro, 1996] [Hanna, 2003] [Hanna & Desainte-Catherine, 2005]).

[33] Un autre type de sélection de pics significatifs peut se baser sur la conformité de la forme du pic avec celle du lobe principal de la fenêtre d'analyse (rappelons que cette fenêtre est convoluée avec une raie spectrale pour une composante sinusoïdale).

« parcimonieux » composé des pics principaux détectés. Les auteurs proposent alors de modéliser le spectre résiduel par une enveloppe très simple à base de segments linéaires. Ce modèle d'enveloppe permet de « régulariser les trous » laissés par la soustraction spectrale. Il a l'avantage de permettre la modélisation du bruit de façon répartie sur toute la bande du signal[34].

A la synthèse, la partie bruitée du signal est resynthétisée en générant une série de sinusoïdes à phase aléatoire ayant ce modèle en segments linéaires pour enveloppe. En ce qui concerne la partie harmonique, les amplitudes sont interpolées linéairement entre les trames (voir la Section 1.3.2.2.1). Les auteurs choisissent d'interpoler les fréquences avec un modèle linéaire, et ne cherchent donc pas à respecter exactement la phase du signal (voir la Section 1.3.2.2.4). Ceci est cohérent avec le fait qu'aucune information de phase n'est utilisée dans l'analyse.

Cette technique s'est montrée très efficace pour modéliser une large gamme de sons de musique et de parole et pour permettre une large gamme de transformations préservant généralement bien la qualité du signal (étirement/compression temporelle, changement de pitch).

1.4.2.3. Modèle Harmonique + Bruit

Tout comme le modèle sinusoïdal peut être adapté en modèle harmonique, le modèle sinusoïdal + bruit a son pendant harmonique + bruit. Dans ce cas, on suppose que le signal $s(n)$ peut être décomposé en une partie harmonique $h(n)$ et une partie bruitée $b(n)$:

$$s(n) = h(n) + b(n) \tag{1.52}$$

avec

$$h(n) = \sum_{i=1}^{I} A_i(n)\cos[\varphi_i(n)] \qquad \text{et} \qquad \varphi_i(n) = \varphi_i(0) + \sum_{m=0}^{n} i\omega_0(m) \tag{1.53}$$

La partie harmonique modélise la composante quasi-périodique du signal et la partie bruitée modélise la composante aléatoire. Tout comme le parallèle déjà mentionné entre le modèle sinusoïdal et le modèle harmonique, le modèle sinusoïdal + bruit est plutôt un modèle généraliste, capable de modéliser une large gamme de sons polyphoniques, alors que la version harmonique + bruit est plutôt adaptée aux sons monophoniques, en particulier à certains instruments et en ce qui nous concerne plus précisément à la parole. Par exemple, pour les sons voisés, la composante aléatoire peut modéliser efficacement le bruit de friction et les variations irrégulières de l'excitation glottique d'une période à l'autre.

Notons qu'historiquement, une version spectrale de ce modèle a été proposée dès 1988 par Griffin et Lim dans [Griffin & Lim, 1988] sous l'appellation *Multi Band Excitation* (*MBE*). Dans cette étude, destinée à être appliquée au codage de la parole, les auteurs proposent de diviser le signal en sous-bandes spectrales. Dans chaque sous-bande, on classifie le signal en signal harmonique ou signal bruité pour ensuite le modéliser dans chaque sous-bande par un modèle adapté : soit un modèle spectral harmonique, soit un

[34] En cela il s'oppose par exemple au modèle « à deux bandes » de la Section 1.4.2.3.

modèle spectral stochastique. A la synthèse, on régénère un spectre mixte avec des paramètres quantifiés correspondant aux sous-bandes harmoniques et aux sous-bandes bruitées. Le signal temporel est resynthétisé à partir de ce spectre mixte.

Plus récemment, une version « simplifiée » de ce modèle en sous-bandes a été proposée pour les signaux de parole [Stylianou, 1996]. Dans cette série d'études, le spectre d'une trame de signal de parole voisée est divisé en deux bandes limitées par une fréquence variant dans le temps, $F_c(n)$ dite *fréquence maximale de voisement* ou fréquence de coupure (voir Figure 1.8). Au-dessous de la fréquence F_c, le signal est considéré comme étant purement harmonique et est représenté dans cette sous-bande par $h(n)$ dans l'équation (1.51). Au-delà de F_c, le signal est supposé être purement bruité et correspondre au filtrage d'un bruit blanc par le conduit vocal. Les trames non-voisées apparaissent comme un simple cas particulier de ce modèle où la fréquence de voisement est à zéro et où tout le spectre est considéré comme bruité[35]. Les paramètres de la partie harmonique $h(n)$ sont estimés par un processus d'analyse semblable à l'ajustement au sens des moindres carrés développé dans la Section 1.2.1.2 et déjà rappelé à la Section 1.4.1[36]. On synthétise alors le signal harmonique $h(n)$ et la partie bruitée est identifiée au signal résiduel $b(n)=s(n)-h(n)$[37]. Ce signal résiduel/bruité peut alors être modélisé par un processus stochastique comme un modèle auto-régressif (AR ou modèle LPC pour *Linear Predictive Coding*) variant dans le temps. A la synthèse, le signal de bruit $b(n)$ peut alors être regénéré en filtrant un bruit blanc $u(n)$ de bande inférieure limitée par F_c par le filtre tout-pôle à réponse impulsionnelle $g(n)$ obtenu lors de la modélisation AR du signal résiduel original.

Tout comme pour le modèle sinusoïdal + bruit pour une large gamme de signaux polyphoniques, un des points forts du modèle harmonique + bruit est que le signal de parole synthétisé avec ce modèle et le signal original sont presque indiscernables perceptuellement. En outre, il permet d'effectuer des traitements du signal de parole de haute qualité, en particulier des modifications du pitch et de la durée des signaux de parole montrant ainsi son utilité dans le cadre de la synthèse de la parole [Stylianou, 1996, 2001] [En-Najjary, 2005] [Macon, 1996]. Par ailleurs, cette représentation du signal de parole peut être considérée comme un pré-encodage (le signal est modélisé par un certain nombre de paramètres), et s'adapte facilement à des procédures de codage à débit réduit [McAulay & Quatieri, 1995].

---

[35] Notons que ce principe d'analyser et modéliser séparément la partie harmonique et la partie bruitée du signal correspond d'une façon générale à la nécessité de simplifier les méthodes d'analyse par rapport à une méthode globale cherchant à optimiser conjointement l'ensemble des paramètres des deux parties harmoniques et bruitées. L'analyse et la séparation de ces deux types de composants sans faire l'hypothèse d'un découpage en deux bandes distinctes (ou en plus de sous-bandes comme dans [Griffin & Lim, 1988]) est un problème beaucoup plus complexe. Il nécessite la mise en œuvre de techniques spécifiques dite de « séparation Harmonique/Bruit » qui dépasse largement le cadre de ce document [Yegnanarayana *et al.,* 1998] [Jackson & Shadle, 2000, 2001] [Girin, 2006].

[36] On rappelle que ceci nécessite l'estimation préalable de la fréquence fondamentale. De plus, il faut ici estimer la fréquence de voisement. Nous ne traiterons pas de ce dernier point dans ce document : on peut se référer par exemple à [Stylianou, 1996].

[37] D'une façon générale, une telle estimation de la partie bruitée du signal par soustraction au signal original de la partie sinusoïdale/harmonique dans le domaine temporel nécessite une synthèse de cette dernière avec une technique respectant la phase des composantes. Ceci s'oppose sur ce point précis à la technique de Serra et Smith (voir la Section 1.4.2.2).

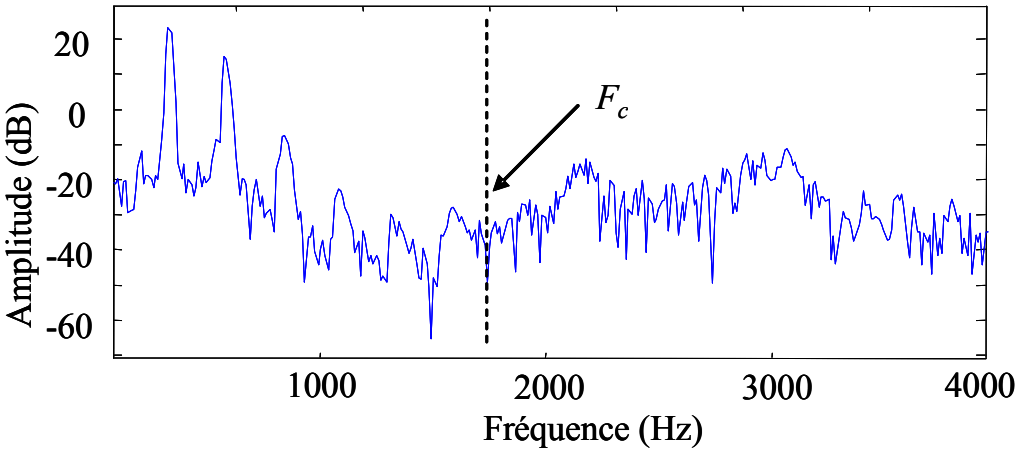

Figure 1.8 : Exemple d'un spectre d'un son de parole mixte voisé-bruité (pour une voix de femme) ; F_c est la fréquence de voisement choisie pour délimiter la partie harmonique et la partie bruité dans la modélisation en deux sous-bandes.

1.4.3. Un raffinement supplémentaire : Sinusoïdes/Harmoniques + Transitoires + Bruit

Le modèle sinusoïdal (ou harmonique) + transitoires + bruit (STB ou HTB) a été introduit assez récemment pour ajouter aux modèles sinusoïdes + bruit ou harmonique + bruit la possibilité de modéliser les signaux très localisés en temps et caractérisés par des variations d'énergie très brusques. Plus généralement on peut considérer ici les portions de signal avec des profils d'évolution temps-fréquence irréguliers sur des durées assez courtes (de l'ordre de la dizaine de millisecondes, voire moins). Il s'agit typiquement des sons percussifs en musique, et des sons transitoires en parole (comme les consonnes plosives par exemple). On peut observer à titre d'exemple sur la Figure 1.2 une réalisation du *burst* (« partie explosive ») du son [k] précédant un [a] approximativement entre les échantillons 200 et 500.

Dans ces modèles STB, on traite généralement de façon spécifique ces zones de signal transitoire, ce qui nécessite tout aussi généralement une phase de détection et de segmentation du signal : des modèles spécifiques sont alors appliqués à ces zones transitoires, alors que le modèle sinusoïdes + bruit ou harmonique + bruit est utilisé ailleurs. Afin de bien modéliser et coder le comportement temporel et fréquentiel des signaux transitoires, il existe plusieurs méthodes. On peut citer l'exemple de la méthode de Levine [Levine, 1998] qui propose de modéliser séparément les parties transitoires par une transformée de type MDCT (*Modified Discrete Cosine Transform* ; déjà mentionnée à la Section 1.1.3) et d'utiliser le modèle sinusoïdes + bruit pour les autres parties du signal sonore. De leur coté, pour modéliser les parties transitoires des signaux de parole, les auteurs de [Shlomot *et al.*, 2001] propose d'utiliser une technique de type CELP (analyse par synthèse dans le domaine temporel ; voir la note de bas de page n°68) en prenant un soin particulier à la synchronisation entre les zones transitoires et les zones harmoniques. On peut aussi citer l'approche de Collen [Collen, 2002] qui propose de diviser la trame d'analyse en plusieurs sous-trames de durée plus courte (5 ms par exemple) et d'estimer l'enveloppe spectrale pour chacune de ces sous-trames. Cette technique modélise plus efficacement l'enveloppe des signaux transitoires mais, comme celle de Levine, elle requiert en contrepartie un surcoût d'information

(transmission de plusieurs jeux de coefficients décrivant les différentes enveloppes spectrales) et donc un accroissement du débit. Une autre approche consiste à utiliser des modèles sinusoïdaux non stationnaires qui ont été proposés pour modéliser l'atténuation exponentielle de bon nombre de signaux transitoires [Nieuwenhuijse *et al*. 1998] [Prandom *et al.,* 1997]. On peut ainsi utiliser un modèle à base de sinusoïdes amorties retardées [Goodwin, 1997] [Boyer & Abed-Meraim, 2002]. La complexité accrue de ces modèles se traduit généralement par une complexité accrue des méthodes d'analyse correspondantes.

1.5. Prise en compte de la perception

Nous allons terminer ce chapitre par une section décrivant brièvement certaines propriétés perceptives de l'oreille. Ces propriétés relèvent du domaine de la psychoacoustique. L'objectif est de prendre en compte ces propriétés dans notre travail de modélisation à long terme pour rendre nos algorithmes plus efficaces (par exemple en utilisant moins de coefficients dans nos modèles si cela est rendu possible par des considérations psychoacoustiques, comme cela se fait en codage, voir ci-dessous). Bien entendu, la psychoacoustique est un domaine extrêmement vaste et complexe, et cette section reste par conséquent non exhaustive. Elle se limite aux principales propriétés telles qu'elles seront exploitées dans notre étude.

L'oreille humaine est un récepteur complexe, caractérisé essentiellement par un comportement très non-linéaire et une série de propriétés de nature sélective sur l'ensemble des composantes d'une scène sonore : en général, toutes les composantes d'un signal ne sont pas perçues, ou plus généralement ne sont pas perçues avec la même « précision ». Les travaux dans ce domaine remontent au 19ème siècle, avec par exemple les études de von Helmholtz [Helmholtz, 1877]. Les propriétés perceptives de l'oreille continuent de faire l'objet d'investigations de la part des psychoacousticiens qui proposent notamment un ensemble de modèles (mathématiques) de son comportement. Le pendant technologique de cet enjeu de connaissances sur les propriétés de l'oreille est leur exploitation dans les systèmes de traitement automatique du son (musique et parole principalement). En particulier, les systèmes de compression avec perte d'information sont concernés au premier plan : c'est précisément en grande partie la perte d'information autorisée dans ces systèmes qui permet une compression efficace. Or cette perte est supposée sans conséquences perceptives lorsque les exigences des modèles psychoacoustiques sont satisfaites.

Au-delà de l'objectif précis de compression (c'est-à-dire la réduction du débit de la ressource binaire représentant le signal), les considérations psychoacoustiques doivent permettre de guider les traitements en *modélisation* du signal, de façon à proposer des modèles appropriés au signal et avec un réglage adéquat de leurs paramètres. On espère ainsi parvenir à une représentation du signal simplifiée et « éclaircie » par rapport à un traitement sans contraintes perceptives[38]. Replacé dans le contexte du modèle sinusoïdal (ou sinusoïdal + bruit), cette problématique se concentre essentiellement autour de l'impact perceptif des différentes composantes spectrales : étant donnée une scène sonore, quelles sont les composantes sinusoïdales (ou de bruit) perçues par le système

[38] Ceci est cohérent avec le problème de la compression, dont la modélisation est souvent une première étape.

auditif ? Et avec quelle précision peut-on modéliser ces composantes en utilisant les outils présentés dans les sections précédentes (et les sections suivantes en ce qui concerne notre objectif de modélisation à long terme) ? Cette section du document propose d'aborder quelques éléments de réponse à ces questions. On décrit d'abord la sensibilité générale de l'oreille humaine aux différentes fréquences, puis la notion de bandes critiques. Enfin, les phénomènes psychoacoustiques de masquage pour l'amplitude et le problème de la perception de la phase/fréquence sont brièvement abordés. Ces points précis sont décrits de façon plus détaillée dans les Chapitres 3 et 4 respectivement, du fait qu'ils ont été spécifiquement exploités dans notre approche de la modélisation à long terme.

1.5.1. Sensibilité de l'oreille humaine en fonction de la fréquence

Dans cette section, nous reprenons les principaux résultats présentés par Zwiker et Feldtkeller dans [Zwicker & Feldtkeller, 1981], et Painter et Spanias dans [Painter & Spanias, 2000]. Le système auditif tient compte de nombreux facteurs comme les informations temporelles (enveloppe, durée) et fréquentielles (spectre, répartition de l'énergie). Par exemple, l'intensité perçue d'une sinusoïde pure dépend de sa fréquence car l'oreille est plus sensible dans certaines régions du spectre que dans d'autres. On considère que l'oreille humaine est capable de percevoir les sons compris entre 20 Hz et 20 kHz. L'oreille n'est toutefois sensible à un son pur, dans une ambiance parfaitement silencieuse, que si sa puissance est supérieure à un seuil appelé *seuil d'audition absolu* (Figure 1.9). Ce seuil d'audibilité dépend de la fréquence f du signal. Parmi les différents modèles proposés, on peut trouver par exemple l'équation (1.54) proposée par [Terhardt, 1979] et reprise dans [Painter & Spanias, 2000] qui donne une bonne approximation de ce seuil en dB SPL.

$$T_A(f) = 3.64\,(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \qquad (1.54)$$

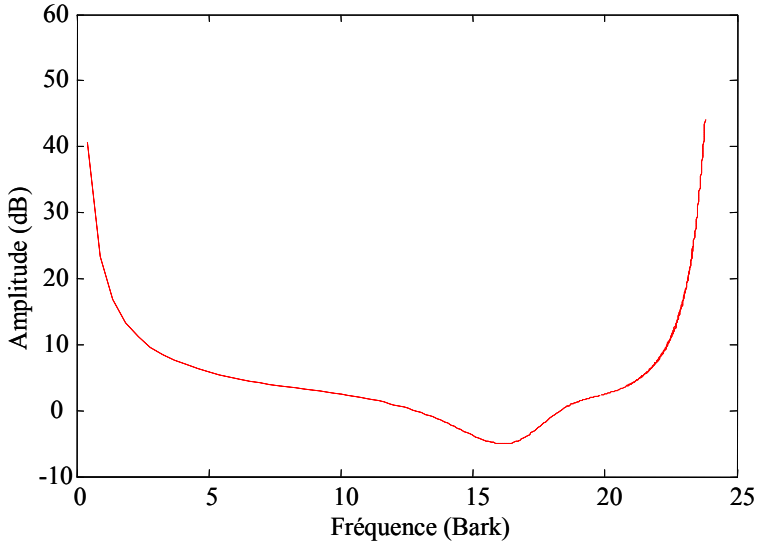

Figure 1.9 : Seuil d'audition absolu ; les fréquences sont données sur une échelle Bark qui est définie dans la section suivante [Terhardt, 1979] [Painter & Spanias, 2000].

1.5.2. Bandes critiques

L'oreille humaine a le pouvoir d'intégrer les contributions perceptives des différentes composantes spectrales dans certaines zones de fréquence découpées en bandes et appelées bandes critiques. En d'autres termes, ces bandes correspondent à un découpage du domaine fréquentiel, tel que dans une bande critique la puissance perçue par l'oreille est égale à la somme de toutes les puissances des composantes dans cette bande de fréquence. Ce découpage est « compatible » avec la notion de seuil d'audition : si la somme des contributions en puissance est supérieure au seuil d'audition alors le signal compris dans la bande considérée est audible, sinon il est masqué et donc inaudible.

Dans la littérature, les bandes critiques se divisent en environ 24 bandes (selon les modèles proposés) dont la largeur varie en fonction de la fréquence, sur une nouvelle échelle fréquentielle assez proche d'une échelle logarithmique et dite « échelle Bark » [Zwicker & Fastl, 1990]. L'oreille peut alors se modéliser sous la forme d'un banc de filtres s'échelonnant par ces bandes. Les valeurs correspondantes, telles qu'on les trouve dans [Zwicker & Fastl, 1990], sont données dans le Tableau 1.1. Notons qu'il n'existe pas de formule exacte pour passer de l'échelle linéaire des Hertz à celle des Barks, mais seulement une série de propositions de formules d'approximation permettant de déterminer les frontières de ces bandes. Par exemple, en notant $B(f)$ la bande Bark, et f une fréquence en Hz, on a d'après [Zwicker & Fastl, 1990] :

$$B(f) = 13 \tan^{-1}(0.00076 \ f) + 3.5 \tan^{-1}\left((f / 7500)^2\right) \quad (1.55)$$

Bande Bark	f_{inf}	f_{sup}	f_c	Δf	Bark	f_{inf}	f_{sup}	f_c	Δf
0 – 1	0	100	50	100	12 – 13	1720	2000	1850	280
1 – 2	100	200	150	100	13 – 14	2000	2320	2150	320
2 – 3	200	300	250	100	14 – 15	2320	2700	2500	380
3 – 4	300	400	350	100	15 – 16	2700	3150	2900	450
4 – 5	400	510	450	110	16 – 17	3150	3700	3400	550
5 – 6	510	630	570	120	17 – 18	3700	4400	4000	700
6 – 7	630	770	700	140	18 – 19	4400	5300	4800	900
7 – 8	770	920	840	150	19 – 20	5300	6400	5800	1100
8 – 9	920	1080	1000	160	20 – 21	6400	7700	7000	1600
9 – 10	1080	1270	1170	190	21 – 22	7700	9500	8500	1800
10 – 11	1270	1480	1370	210	22 – 23	9500	12000	10500	2500
11 – 12	1480	1720	1600	240	23 – 24	12000	15500	13500	3500

Tableau 1.1 : Caractéristiques des 24 bandes critiques de l'échelle Bark [Zwicker & Fastl, 1990], f_c, Δf, f_{inf} et f_{sup} sont respectivement le centre, la largeur et les limites inférieures et supérieures des bandes critiques.

1.5.3. Phénomènes psychoacoustiques de masquage

Les différentes composantes d'une scène sonore peuvent interagir entre elles au niveau perceptif. Ainsi, le phénomène de masquage se définit comme la capacité d'un son donné à rendre inaudible un autre son. Lorsque les deux sons sont simultanés, on a affaire au masquage fréquentiel. Lorsque les deux sons sont consécutifs, on a affaire au

masquage temporel. Dans le cadre du codage audio, le phénomène le plus exploité est le masquage fréquentiel : il est largement utilisé, notamment dans les codeurs de type transformé en fréquence. Comme d'une part le modèle sinusoïdal est un modèle de type décomposition spectrale, et comme d'autre part les effets du masquage temporel sont des effets relativement limités en temps, on s'intéresse principalement ici masquage fréquentiel. C'est pourquoi on décrit d'abord très rapidement le principe du masquage temporel avant de donner quelques premiers éléments sur le masquage fréquentiel, plus intéressant dans notre étude. Son exploitation dans le cadre de nos études sur la modélisation à long terme des amplitudes sera plus amplement détaillée au Chapitre 3.

1.5.3.1. Masquage temporel

Dans le contexte de l'analyse de signaux audio, les fortes variations du signal (par exemple, les sons de percussions d'un instrument musique) créent des zones de pré et de post masquage dans le temps où on ne perçoit pas les signaux faibles précédant ou suivant la zone de forte variation. Autrement dit, les seuils d'audibilité sont artificiellement augmentés avant (pré-masquage) et après (post-masquage) l'occurrence d'un signal masquant[39] (voir la Figure 1.10). On considère que le pré-masquage a un effet environ 5 ms avant le son masquant, tandis que le post-masquage persiste de 50 à 300 ms après le son masquant, selon la force et la durée du masqueur. Ce phénomène est difficile à modéliser et donc assez peu utilisé en codage audio. Il est toutefois exploité dans les codeurs par transformée (le codeur AAC de la norme MPEG4 par exemple) pour le traitement des signaux transitoires. Lors de l'apparition de tels signaux, la sélection de fenêtres d'analyse plus courtes est utilisée afin de réduire les phénomènes d'étalement d'un bruit non masqué[40].

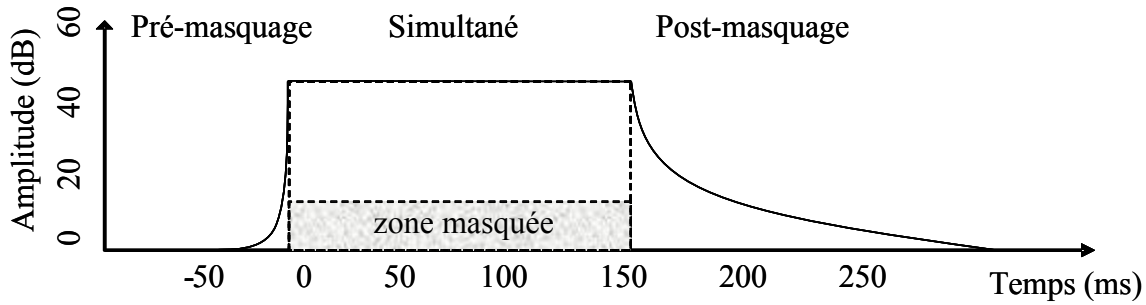

Figure 1.10 : Représentation schématique du phénomène de masquage temporel (d'après [Painter & Spanias, 2000] et [Zwicker & Fastl, 1990]).

1.5.3.2. Masquage fréquentiel

Le masquage fréquentiel simultané existe lorsque pendant une période donnée, deux sons sont émis avec des fréquences suffisamment proches l'une de l'autre et avec des amplitudes suffisamment inégales. Dans ce cas, seul le son dont l'amplitude est la plus

[39] L'augmentation du seuil d'audibilité *pendant* le son masquant correspond au masquage fréquentiel décrit dans la sous-section suivante.

[40] Après transformation d'une fenêtre de signal et transformation inverse, un bruit initialement localisé dans une partie restreinte de la fenêtre peut s'étendre sur la totalité de la fenêtre, avec des conséquences perceptives indésirables.

forte est entendu, le son le plus faible étant masqué, c'est-à-dire inaudible[41]. Dans ce cadre, le seuil perceptif global (à une fréquence donnée) est défini comme la valeur limite de la puissance d'un signal masqué (à cette même fréquence) rendu inaudible par le signal de puissance la plus forte. Le masque, c'est-à-dire la courbe qui délimite la zone des sons masqués pour un son masqueur donné, peut être approximé sur une échelle Bark/dB par un triangle dont le sommet correspond à la fréquence du son masquant [Zwicker & Fastl, 1990] [Schroeder *et al.*, 1979] [Jayant *et al.*, 1993] [Depalle *et al.*, 1993], comme illustré sur la Figure 1.11 et sur la Figure 1.12. Ce point sera plus amplement détaillé à la Section 3.1. La Figure 1.11 donne à titre d'exemple la courbe de masquage associée à deux sinusoïdes de fréquences respectivement 2000 Hz et 2200 Hz. Les sons situés dans la zone pointillée en dessous de la courbe de masquage déduite sont masqués.

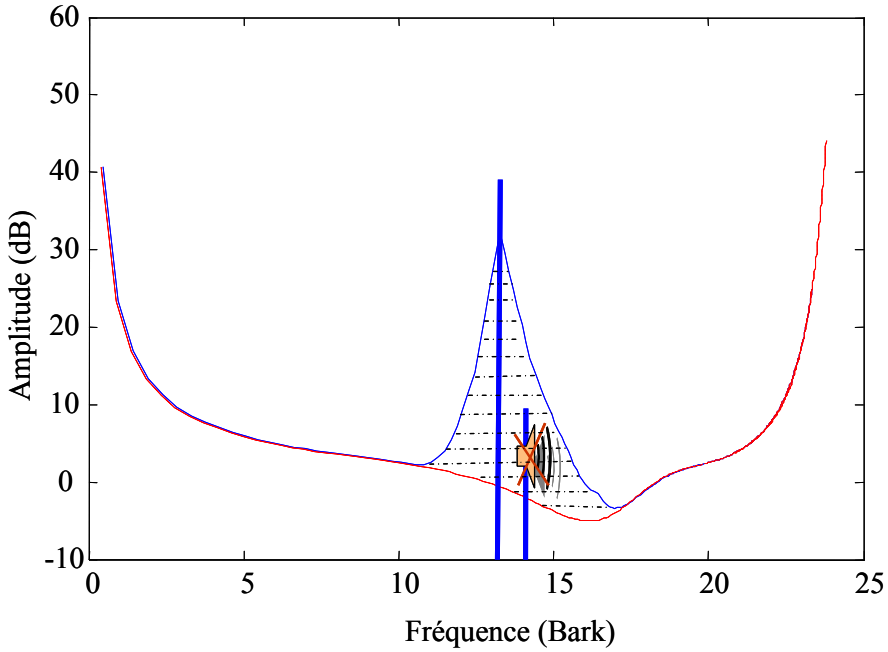

Figure 1.11 : Représentation du phénomène de masquage fréquentiel entre deux composantes sinusoïdales : la composante de plus grande puissance génère un seuil de masquage triangulaire autour de sa fréquence. Une composante plus faible présente sous ce seuil est inaudible.

Pour modéliser le phénomène de masquage global intervenant dans une scène sonore complexe composée d'un certain nombre de composantes fréquentielles, on fait usuellement l'hypothèse que les contributions individuelles des phénomènes de masquage élémentaires s'additionnent, en prenant en compte d'autres éléments de la

[41] De façon plus précise, ce cadre théorique et expérimental posé pour la modélisation du phénomène a été développé par les psychoacousticiens dans les cas où les sons masquant ou masqués élémentaires sont des sinusoïdes ou des bruits à bande étroite. On a donc quatre cas simples : une sinusoïde masquant une autre sinusoïde ou un bruit à bande étroite, ou bien un bruit à bande étroite masquant une sinusoïde ou un autre bruit à bande étroite. Ces cas élémentaires servent de base à l'exploitation globale du phénomène de masquage, c'est-à-dire en prenant en compte toutes les composantes d'une scène sonore considérée comme une somme de tels sons élémentaires (ce point est développé plus loin dans la section). On ne présente que le principe ici, indépendamment de la nature du son élémentaire pour plus de simplicité.

psychoacoustique (par exemple la notion de bandes critiques). On en déduit une procédure type de calcul d'une courbe de masquage globale valable sur une tranche de signal (supposé suffisamment stationnaire pour que ses caractéristiques fréquentielles soient stables). Cette courbe de masquage donne la courbe globale en-dessous de laquelle les sons ne sont pas perçus. Son calcul repose typiquement sur une séquence du type (certains traitements étant optionnels) :

- Estimation du spectre de puissance du signal (généralement par *FFT*) ;
- Classification de chaque composante fréquentielle en son pur (tonal) ou bruit (non tonal) selon des critères d'« isolement » et de puissance relative ;
- Intégration (sommation) des composantes les plus proches pour chaque type tonal et non tonal dans chaque bande critique ;
- Elimination des composantes inférieures au seuil d'audition absolu ;
- Pour les autres composantes, calcul des seuils de masquage individuels ;
- Calcul de la courbe de masquage totale par addition des contributions des composantes individuelles tonales et non tonales (plus le seuil d'audition absolu) pour chaque fréquence ; cette étape est illustrée sur la Figure 1.12.

Dans les systèmes de traitement automatique des sons, les courbes de masquage globales résultantes offrent le moyen non seulement d'éliminer les composantes inaudibles, mais aussi de conditionner l'erreur de modélisation et de codage des composantes audibles de façon à rendre inaudible cette erreur : le principe est de contraindre cette erreur à rester en-dessous du seuil donné par la courbe de masquage. Cette notion sera précisée à la Section 3.1 en ce qui concerne l'exploitation du phénomène de masquage fréquentiel proprement dit pour la modélisation à long terme des amplitudes. On verra alors une illustration du calcul et de l'utilisation pratique d'une telle courbe pour notre série d'études en modélisation à long terme.

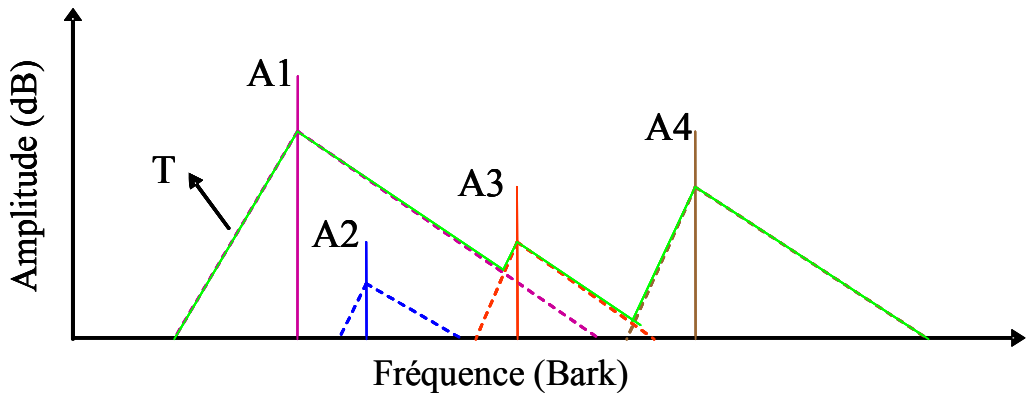

Figure 1.12 : Schéma représentant les masques élémentaires associés à quatre sinusoïdes représentées par les pics A1, A2, A3 et A4. Ces masques élémentaires contribuent au seuil de masquage global représenté en vert. Comme l'addition est effectuée sur une échelle linéaire et que la représentation est donnée sur une échelle log, la contribution dans la courbe globale des zones du triangle correspondant à A3 situées sous les triangles correspondants à A1 et A4 est très faible. De même pour le petit triangle bleu correspondant à A2, ce qui est cohérent avec le fait que cette zone est entièrement masquée par A1.

1.5.4. Perception de phase/fréquence

Le sujet de la perception de la phase est beaucoup plus délicat que dans le cas de l'amplitude. C'est pourquoi nous lui consacrons une section complète dédiée non seulement à sa description mais aussi aux réflexions que cela engendre dans le cadre de notre travail en modélisation à long terme. Pour plus de liant, nous avons choisi de présenter ces aspects ensemble, et ceci nous l'avons fait directement dans le Chapitre 4 : il s'agit de la Section 4.1 qui ouvre le chapitre. Nous invitons le lecteur à se reporter à cette section pour entrer dans ce domaine.

Chapitre 2

2. Une nouvelle approche : la modélisation à long terme

Dans ce chapitre, nous rentrons de plein pied dans le cœur du sujet de cette thèse avec l'étude de la modélisation à long terme des descripteurs du signal telle qu'elle a été présentée en introduction. Comme nous l'avons déjà mentionné, la modélisation à long terme peut s'appliquer sur diverses représentations du signal. Cependant, les expérimentations menées dans cette thèse portent spécifiquement sur le modèle sinusoïdal en version harmonique et les descripteurs du signal sont des paramètres d'amplitude et de phase des différentes harmoniques. Dans ce chapitre, nous gardons tant que possible une approche très générale, commune aux paramètres d'amplitude et de phase, sauf quand il s'avère nécessaire de dissocier ces deux cas. De plus, dans une certaine mesure, cette présentation est la plupart du temps suffisamment générale pour pouvoir englober potentiellement d'autres représentations du signal avec les adaptations nécessaires. Nous verrons dans le Chapitre 3 comment les principes développés ici s'adaptent plus spécifiquement aux cas des paramètres d'amplitudes du modèle sinusoïdal. De même, nous étudierons le cas des paramètres de phases dans le Chapitre 4. Notons aussi que l'analyse des paramètres telle que nous l'avons effectuée dans nos études (c'est-à-dire la procédure qui permet d'obtenir les séquences de valeurs à modéliser à long terme) n'est pas décrite dans ce chapitre : elle sera expliquée aux chapitres suivants dédiés aux expérimentations, sur la base de ce qui a déjà été décrit au Chapitre 1, Section 1.2.1.2. Ce choix est guidé par le fait que le principe de modélisation à long terme est quasiment indépendant de la procédure d'analyse et qu'il peut s'appliquer sur toute séquence temporelle de paramètres décrivant un signal quelle que soit la procédure d'estimation de ces paramètres.

Ce chapitre est organisé comme suit. On donne d'abord une définition générale de la modélisation à long terme telle qu'elle est abordée dans notre série d'études. Cette définition est ensuite replacée dans le contexte des approches temporelles déjà réalisées et décrites dans la littérature. Nous définissons alors la notion de trame à long terme et nous justifions pourquoi dans cette thèse on s'intéresse aux séquences de parole entièrement voisées. Ensuite, nous présentons les différents modèles à long terme que nous avons utilisés. Enfin, nous présentons l'algorithme que nous avons élaboré pour réaliser l'ajustement des modèles à long terme aux sections de données modélisées. Cet algorithme exploite l'adaptation de critères perceptifs usuellement définis dans un cadre à court terme. Comme déjà mentionné ci-dessus, ces principes et l'algorithme qui les met en œuvre sont présentés de façon générique et seront adaptés au cas des amplitudes et des phases du modèle sinusoïdal respectivement aux Chapitres 3 et 4.

2.1. *Définition générale de la modélisation à long terme*

Comme indiqué dans l'introduction, la modélisation dite dans ce document « à long terme » consiste à modéliser les trajectoires temporelles des paramètres d'un premier modèle spectral du signal, en l'occurrence dans cette thèse le modèle sinusoïdal décrit dans le chapitre précédent, sur des longues sections de parole. Le terme « longues » signifie que ces sections sont généralement significativement plus longues que la dizaine de millisecondes qui est l'ordre de grandeur pour les modèles à court terme (voir le Chapitre 1). En d'autres termes, il s'agit de sélectionner une large section de parole, d'appliquer sur cette section de parole une série de mesures à court terme successives pour extraire les valeurs des paramètres spectraux à modéliser à long terme (voir Figure 2.1), et d'appliquer sur chaque suite de paramètres cohérents (par exemple des amplitudes d'un partiel donné) un modèle mathématique paramétrique. Les modèles que nous avons implantés et testés dans cette thèse sont présentés à la Section 2.4. Le processus d'ajustement de ces modèles aux données est présenté à la Section 2.5.

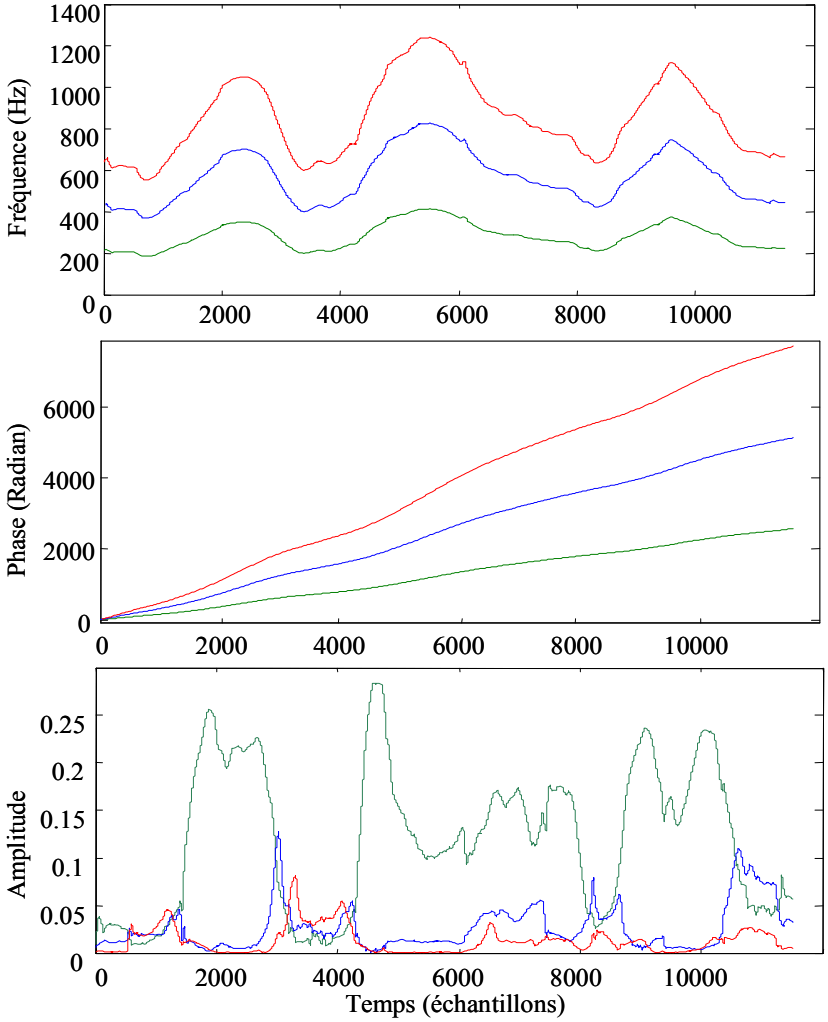

Figure 2.1 : Exemples de trajectoires temporelles de fréquence, de phase et d'amplitude pour les trois premières harmoniques d'une longue section de parole voisée de voix de femme (durée 1,4 s, $F_e = 8$ kHz). Les paramètres sont extraits sur une base à court terme : on a un jeu de paramètres par période de signal (voir Section 3.3.1.2).

Comme déjà mentionné en introduction, le principe général de la modélisation à long terme repose sur la corrélation existant entre paramètres spectraux successifs. Comme on peut le voir sur la Figure 2.1, ceux-ci décrivent en effet des trajectoires « cohérentes », au moins par morceaux. Cette corrélation a déjà été l'objet d'un certain nombre d'études en parole et en audio. Avant de poursuivre plus avant la description de notre travail, il est donc de rigueur d'effectuer une brève revue des techniques ayant un rapport avec notre approche, pour mieux la situer par rapport à l'existant.

2.2. Positionnement par rapport aux approches temporelles existantes

L'exploitation de la corrélation entre paramètres spectraux à court terme successifs a notamment été l'objet d'une attention particulière dans le domaine du codage, puisque le codage vise précisément à exploiter les différentes formes de redondance présentes dans les signaux. Prenons ainsi l'exemple de la modélisation par prédiction linéaire (LPC pour *Linear Predictive Coding*) que nous avons déjà abordé à la Section 1.1.3. L'usage de cette technique s'est généralisé en codage de la parole depuis plus de trente ans [Markel & Gray, 1976] [Gray & Gersho, 1992] [Makhoul, 1975]. Dans ce cadre, un ensemble de techniques a été développé pour exploiter la corrélation entre les vecteurs successifs des paramètres appelés *Line Spectral Frequencies* (LSF). Ces paramètres encodent l'enveloppe spectrale du signal issue de la technique de prédiction linéaire [Sugamura & Itakura, 1986] [Paliwal & Atal, 1993].

Ainsi, le codage différentiel et/ou adaptatif a par exemple été proposé par [Yong *et al.*, 1988] pour coder efficacement ces paramètres LSF : on encode la différence entre deux vecteurs de LSF consécutifs, plutôt qu'encoder séparément ces deux vecteurs. Du fait de la corrélation inter-trames de ces vecteurs, cette différence a une dynamique réduite par rapport à la dynamique des vecteurs de départ. De ce fait, elle est moins coûteuse à encoder [Gersho & Gray, 1992]. Dans ce cas précis, l'exploitation de la corrélation est limitée à deux trames à court terme consécutives.

Par ailleurs, la quantification matricielle est une généralisation de la quantification vectorielle [Gersho & Gray, 1992] qui permet d'exploiter la corrélation temporelle des vecteurs de LSF en encodant conjointement plusieurs vecteurs consécutifs : dans ce cas, les prototypes du dictionnaire qui remplacent les données sont des matrices [Tsao & Gray, 1985] [Xydeas & Papanastasiou, 1999]. Cette technique est plus performante que la quantification vectorielle du point de vue du rapport distortion/débit qui est un critère classique en codage. C'est une propriété à la fois théorique et vérifiée en pratique [Gersho & Gray, 1992]. Cependant, la quantification matricielle demande un délai accru (proportionnellement à la taille temporelle des matrices en jeu) et une complexité nettement plus importante. De fait, dans les études de [Tsao & Gray, 1985] [Xydeas & Papanastasiou, 1999], la taille temporelle des matrices prototypes est limitée à quatre vecteurs consécutifs. Avec des vecteurs de paramètres LSF extraits toutes les 20 ms, ceci représente une fenêtre temporelle de 80 ms. Nous reviendrons sur cette question du délai un peu plus loin dans cette section.

Dans un autre registre, mais toujours dans ce contexte de matrices de coefficients LSF, nous pouvons citer ici l'étude de [Farvardin & Laroia, 1989]. Dans cette étude, les

auteurs proposent d'appliquer une technique de codage largement utilisée en traitement d'images, la DCT-2D, pour encoder les paramètres LSF du modèle LPC. Comme son nom l'indique, la DCT-2D est une version en deux dimensions de la DCT dont on a déjà parlé à la Section 1.1.3. Sans entrer dans les détails techniques ici[42], on peut dire que le principe est d'abord de grouper L vecteurs de paramètres LSF successifs (de taille $p = 10$) en matrices, comme on le fait pour la quantification matricielle citée plus haut dans cette section. Puis, on applique une DCT-2D de taille $p \times L$ sur les matrices de LSF. On se retrouve dans un espace dual où les coefficients issus de la transformation sont largement décorrélés par rapport aux valeurs successives des vecteurs LSF[43]. Ces coefficients sont alors quantifiés (par une quantification scalaire uniforme dans [Farvardin & Laroia, 1989], avec un algorithme d'allocation de bits permettant de minimiser l'erreur quadratique moyenne sur l'ensemble des coefficients), puis on repasse dans le domaine des paramètres LSF par une transformation DCT-2D inverse. Du fait de la décorrélation des coefficients dans le domaine transformé, la quantification dans ce domaine est plus efficace qu'une même quantification effectuée dans le domaine original, c'est une propriété générale des systèmes de compression par transformée [Gersho & Gray, 1992] [Jayant & Noll, 1984]. Dans l'étude de [Farvardin & Laroia, 1989], les vecteurs LSF sont extraits toutes les 10 ms et les auteurs donnent des résultats en terme de distorsion spectrale et débits correspondants pour des tailles temporelles de matrice L allant de 1 à 10 (ils montrent par exemple que cette technique fournit une distorsion spectrale de 1 dB, généralement considérée comme la limite de transparence de codage [Paliwal & Atal, 1993], pour un débit de 2,1 bits par paramètre avec $L = 10$). Les auteurs limitent ainsi le délai d'encodage à 100 ms maximum.

Un point commun à l'ensemble de ces études en codage LPC décrites ci-dessus est qu'elles limitent généralement de façon assez drastique la fenêtre temporelle sur laquelle on exploite la corrélation. On vient ainsi de voir un ordre de grandeur maximum de 80-100 ms pour les méthodes de quantification matricielle et DCT-2D. Cette limite est généralement justifiée par les contraintes d'une communication interactive (sans doute aussi à l'époque de certaines de ces études se posaient des contraintes au niveau du temps et du coût de calcul). En d'autres termes, cette limitation est volontaire : elle vise à conserver un délai de codage relativement compatible avec les contraintes du pseudo-temps réel des systèmes de communication par la voix (notamment la téléphonie et plus récemment la voix sur IP). En tentant d'être un tout petit peu plus précis quantitativement, les études sur ce point révèlent que le délai devient significativement gênant pour une communication interactive à partir de 400 ms environ, de larges variations de cette estimation pouvant apparaître suivant le type de communication [Guéguin, 2006] [Cermak, 2002] [Kitawaki & Itoh, 1991] [Möller, 2002] [Hammer *et al.*, 2005]. Or, au délai de codage dû au pur fenêtrage du signal doit s'ajouter le délai algorithmique des procédures de compression-décompression ainsi que le délai de transmission. De plus, un délai « significativement génant » est certainement déjà bien trop pénalisant pour les clients des systèmes de télécommunications. Ceci

[42] Nous reparlerons de cette technique dans ce document au Chapitre 5, lorsqu'on généralisera notre modélisation à long terme des paramètres spectraux d'amplitudes à une approche à deux dimensions. Dans le cas où on utilise un modèle de type cosinus discrets pour cette modélisation, il existe de fortes connexions théoriques et pratiques avec cette DCT-2D.

[43] Cette corrélation des valeurs de paramètres LSF existe d'ailleurs aussi bien dans le sens temporel que dans le sens fréquentiel, c'est-à-dire au niveau des composantes LSF successives d'un même vecteur. C'est pourquoi la DCT-2D est efficace pour capter cette double corrélation.

explique que les délais de compression tolérés en terme de fenêtrage du signal dans les codeurs usuels sont en fait inférieurs à la centaine de ms. Par conséquent, ceci explique aussi le faible intérêt qui a été porté jusqu'ici à une approche à plus long terme (*i.e.* sur des fenêtres temporelles significativement plus larges) du codage du signal de parole, et intrinsèquement de sa modélisation.

Toujours dans le contexte du codage LPC de la parole, il existe à notre connaissance une « exception » à cette contrainte de traitement à court terme. Il s'agit de la série de travaux portant sur la technique dite de *Décomposition Temporelle*. Cette technique, initiée par Atal en 1983 [Atal, 1983] et reprise ensuite par plusieurs chercheurs [Ahlbom *et al*., 1987] [Van Dijk-Kappers & Marcus, 1989] [Van Dijk-Kappers, 1989], [Bimbot & Atal, 1991] [Ghaemmaghami & Deriche, 1991] [Ghaemmaghami *et al*., 1997] consiste à décomposer les trajectoires de vecteurs de paramètres spectraux en une séquence de fonctions-cibles temporelles associées de façon univoque à des vecteurs-cibles correspondants. Ces fonctions-cibles sont localisées temporellement le long de la trajectoire des paramètres et les fonctions successives se recouvrent en partie. Cette modélisation, si elle bien réalisée, est censée représentée la séquence de gestes articulatoires produite lors de la production du signal de parole analysée. En effet, les vecteurs-cibles représentent les cibles phonétiques idéales du signal de parole analysé, associées à des positions articulatoires idéales. Du fait de la coarticulation (recouvrement des gestes articulatoires), ces cibles ne sont pas forcément atteintes, elles peuvent être en quelque sorte des formes pondérées de cibles idéales successives, et cette pondération est modélisée dans cette technique par la forme et le recouvrement des fonctions-cibles successives. Cette technique a donc un double intérêt. Elle peut être appliquée à une problématique de codage[44] (elle a d'ailleurs été originellement proposée dans ce cadre). Et elle est aussi un outil pour analyser et comprendre l'organisation spectro-temporelle de la parole. Par rapport aux techniques de codage LPC usuelles, la décomposition temporelle se démarque donc par la recherche d'une représentation particulièrement parcimonieuse de l'évolution temporelle de l'information spectrale LPC. En termes concrets, on a à coder un vecteur-cible et une fonction-cible par événement phonétique[45]. Elle donne ainsi en pratique une modélisation spectrale à

[44] A ce titre, on peut noter que la décomposition temporelle peut être appliquée sur différents types de vecteurs spectraux, pourvu que leur échantillonnage capture correctement l'évolution spectrale du signal (typiquement des fenêtres d'analyse de taille et de décalage de l'ordre de 10 ms doivent être utilisées). Plusieurs études comparatives ont montré que, d'une façon globale, le jeu de paramètres spectraux qui donnaient les meilleurs résultats (à la fois selon des critères d'adéquation de la décomposition au contenu phonétique du signal et des critères de qualité du signal resynthétisé à partir des valeurs de paramètres fournis par la décomposition) étaient des paramètres dits Log-Area Ratios (LAR) [Van Dijk-Kappers, 1989] [Bimbot & Atal, 1991] [Ghaemmaghami *et al*., 1997]. Comme les paramètres LSF, ces paramètres représentent le spectre LPC du signal [Makhoul, 1975] [Markel & Gray, 1976] et ont été largement utilisés dans les techniques de codage utilisant le modèle LPC (même si aujourd'hui les paramètres LSF les ont généralement remplacés).

[45] L'intérêt en codage provient aussi du fait que, comme on l'a déjà dit, les fonctions cibles sont temporellement localisées, ce qui participe évidemment de la parcimonie. Ceci dit, la multiplicité des formes possibles pour ces fonctions fait que la problématique de leur quantification n'est pas triviale. De plus, la séquence de paramètres spectraux modélisée par cette technique et rééchantillonnée aux instants d'analyse de la séquence de paramètres originale sur laquelle on a appliqué la décomposition n'est généralement pas suffisamment fidèle à cette séquence originale pour fournir un codage transparent. De fait cette technique a été peu utilisée en codage pur. On peut aussi se demander dans quelle limite la contrainte de délai à l'échelle du phonème (et même bien au-delà en réalité puisque on doit considérer l'ensemble de la séquence pour réaliser la décomposition elle-même avant son encodage) n'a pas considérablement diminué l'intérêt porté à cette technique en codage pur.

l'échelle du phonème. Pour la situer par rapport à notre étude, elle fournit donc un excellent exemple de compromis entre ce que nous avons appelé dans ce document le court terme (10-30 ms, soit une taille généralement inférieure au phonème, voire à un événement acoustique/articulatoire sub-phonémique) et le long terme (plusieurs centaines de ms / plusieurs phonèmes). Elle apparaît aussi comme une mise en cascade de plusieurs modèles localisés à l'échelle du phonème, alors que dans notre approche nous verrons que nous allons considérer un seul modèle global pour toute une séquence de paramètres / de phonèmes.

Si on cherche maintenant à faire un bilan de ce bref état de l'art pour le positionner par rapport à l'approche que nous proposons dans cette thèse, on remarque d'abord que toutes les études citées jusqu'ici dans cette section concernent la modélisation et l'encodage de paramètres du modèle LPC. D'autre part, ces études sont relativement limitées au niveau de l'exploration temporelle des trajectoires de paramètres : un ordre de grandeur de la centaine de millisecondes maximum pour les premières, et eventuellement un peu plus pour la décomposition temporelle. Or, notre travail se place dans le cadre du modèle sinusoïdal, et avec une approche temporelle à plus long terme. On peut donc terminer ce bref état de l'art en le recadrant sur le modèle sinusoïdal. Si on revient sur ce modèle, on peut dire que le principe d'exploration des trajectoires des paramètres sur du long terme n'est pas nouveau en soi : en réalité, il est au cœur de la technique de *partial tracking* présentée à la Section 1.3.2.1, si on considère un traitement *offline*, c'est-à-dire que l'analyse-synthèse est réalisée globalement sur l'ensemble du signal sans contrainte de temps-réel. C'est notamment le cas pour la transformation d'un morceau de musique avant l'écoute par exemple [Lagrange, 2004] [Rodet & Depalle, 1993]. Dans ce cas, le *partial tracking* fournit le résultat de cette exploration à long terme en termes de suites de connexions entre mesures consécutives[46]. Ce qui est nouveau dans notre étude par rapport au contexte du modèle sinusoïdal, c'est de *remplacer la suite des valeurs de paramètres connectées par le partial tracking par un modèle paramétrique*. De plus, cette opération se fait en exploitant les propriétés perceptives de l'oreille par l'adaptation au traitement temporel des modèles psychoacoustiques usuellement exploités à court terme ou dans le cadre stationnaire. Cette approche psychoacoustique à long terme est, à notre connaissance, tout à fait originale. Enfin, pour revenir sur le contexte du codage LPC, et ainsi « boucler la boucle », nous verrons les liens entre notre étude à long terme dans le contexte du modèle sinusoïdal avec une possible étude similaire dans le contexte LPC au Chapitre 6 de ce document.

2.3. Choix des trames à long terme dans cette étude

Par rapport aux études citées dans la section précédente, notre approche se caractérise par la volonté d'étendre le domaine d'exploration de l'évolution des paramètres spectraux sur des fenêtres potentiellement beaucoup plus larges que quelques trames. Nous allons préciser maintenant sur quel type de fenêtre de signal nous allons appliquer la modélisation à long terme telle que définie dans la Section 2.1 et quelles sont les conséquences de ce choix.

[46] Dans [Lagrange, 2004], la phase de *partial tracking* est d'ailleurs précisément appelée « approche à long terme » mais cette dénomination prend donc un sens différent de celui de la présente étude.

2.3.1. Choix des sections voisées

Dans cette thèse, nous nous concentrons sur des signaux de parole, bien que les principes développés puissent s'adapter aux composantes des signaux audio en général et des signaux de musique en particulier. Les longues sections de parole que nous considérons pour la modélisation à long terme, appelées « trames à long terme », sont les sections de parole entièrement voisées. Ainsi, dans le traitement que nous proposons, la parole est d'abord segmentée en sections continûment voisées et continûment non voisées (par des techniques classiques qui n'ont pas été développées dans le cadre de cette thèse ; une illustration du résultat d'une telle segmentation sur un signal de parole continue est donnée à la Figure 2.2). Puis le modèle sinusoïdal est appliqué sur chacune des sections continûment voisées, sur une base traditionnelle à court terme (c'est-à-dire que les paramètres sont extraits sur des fenêtres d'analyse à court terme successives). Enfin, un modèle à long terme est employé pour représenter la trajectoire entière de chaque paramètre d'amplitude et de phase sur chaque section voisée.

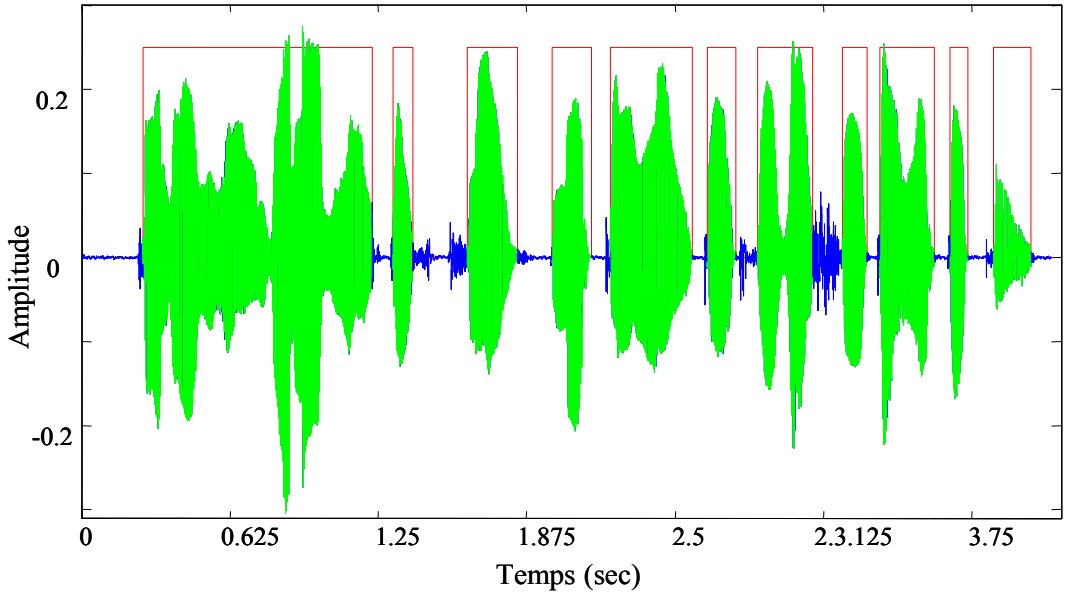

Figure 2.2 : Illustration de la segmentation d'une phrase en sections continûment voisées et continûment non voisées. La phrase est « Collision de l'avion espion : le ton monte entre Washington et Pékin » prononcée par une femme. Les sections voisées sont représentées en vert (avec les frontières marquées par les rectangles rouges) et les sections non voisées sont représentées en bleu.

2.3.2. Justification de ce choix

Ce choix est justifié par les considérations suivantes qui permettent par ailleurs de mieux comprendre cette nouvelle approche. Tout d'abord, les sections entièrement voisées de parole apparaissent généralement comme des signaux quasi-périodiques faiblement non-stationnaires. Ceci signifie que les caractéristiques spectrales de ces sections évoluent de façon particulièrement « lente et régulière », notamment par rapport aux sections non-voisées du type transitoires (voir la Section 1.1.4.1 et voir plus loin). Ceci est particulièrement vrai pour les sons vocaliques et les transitions lentes entre ceux-ci. L'exemple de la voyelle [a] de la Figure 1.2 illustre bien cette idée : le

noyau de la voyelle semble relativement stable mais on remarque nettement la modulation d'amplitude, rapidement croissante au début de la voyelle et plus lentement décroissante à la fin. Or les modèles à long terme utilisés dans cette étude sont à même, par leur caractère lisse déjà mentionné, de capter efficacement les trajectoires de paramètres suffisamment régulières. Le type d'évolution lente de la voyelle [a] mentionné ci-dessus est ainsi un exemple typique de caractéristique pouvant être capturée par une approche à long terme. Même pour les transitions voisées voyelle à consonne voisée ou consonne voisée à voyelle plus rapides, le voisement garantit dans une certaine mesure une certaine régularité dans l'évolution, en tout cas par rapport aux transitions encore plus rapides et plus irrégulières impliquant des sons non voisés. L'approche à long terme semble donc aussi justifiée dans ce cas, et devrait permettre de capturer des formes d'évolution relativement riches.

Inversement, on comprend qu'on ne s'intéresse pas dans cette étude aux transitions impliquant des phonèmes non voisés à évolution très localisée dans le temps et en contrepartie généralement très diffuse en fréquence (voir l'exemple du phonème [k] précédent le [a] dans la Figure 1.2). On l'a déjà mentionné à la Section 1.4.3, des modèles spécifiques ont été récemment proposés dans la littérature pour décrire de telles portions de signal : par exemple dans [Shlomot *et al.*, 2001], les auteurs proposent de permuter entre un modèle harmonique (à court terme) pour les sections de parole régulières et un modèle de type CELP adapté pour les sons transitoires. En musique, ces sons transitoires, souvent de nature percussive, sont aussi l'objet d'une attention spécifique (voir la Section 1.4.3).

En revanche, notre approche à long terme peut très bien s'appliquer aux sections de sons non voisés régulières, telles que les fricatives par exemple. Dans ce cas, l'évolution spectrale et/ou statistique[47] du signal est sans doute suffisamment régulière pour être capturée efficacement par un modèle d'évolution lisse à long terme, à condition toutefois que ce modèle s'applique sur un modèle spectral de premier niveau adapté. Comme ce n'est pas vraiment le cas du modèle sinusoïdal, en tout cas tel que nous l'utilisons dans cette étude[48], nous ne considèrerons pas la modélisation à long terme dans le contexte sinusoïdal pour de telles sections non voisées de parole. En revanche, la possibilité d'adapter la modélisation à long terme à d'autres modèles spectraux, et notamment à ceux permettant de décrire efficacement les sections non voisées (par exemple le modèle LPC) sera discutée au Chapitre 6.

2.3.3. Conséquences importantes

2.3.3.1. Exploitation de l'harmonicité

Une conséquence du choix des sections voisées pour notre étude de modélisation à long terme est de grandement simplifier le processus entier d'analyse-modélisation-synthèse. En effet, on considère que sur ces sections, le signal est pseudo-périodique (voir Section 1.4.1). Comme nous l'avons déjà mentionné à la Section 1.4.1, les partiels se réduisent alors aux harmoniques régulièrement espacés. Du coup, leur analyse est

[47] Voir les notes de bas de page 8 et 10.

[48] Il existe cependant des adaptations du modèle sinusoïdal pour les sons non voisés : voir la Section 1.4.2.1 et en particulier la note de bas de page 29.

facilitée, et les mesures de paramètres du même rang peuvent être directement liées. En un mot, le *partial tracking* est évité et nous ne traiterons donc pas de cette technique dans ce travail. Notre problème est plutôt de plaquer un modèle d'évolution à long terme sur chaque harmonique pris en tant que suite ordonnée (d'abord spectralement puis temporellement pour le traitement à long terme) des valeurs de paramètres de type amplitudes ou phases.

Notons qu'on ne considère pas directement les composantes de signal bruitées éventuellement présentes dans le signal et pouvant se superposer aux harmoniques (voir Section 1.4.2). On a déjà fait part dans la section précédente de la possibilité de modéliser à long terme ces composantes bruitées. L'interaction de la modélisation à long terme des composantes harmoniques avec la modélisation à long terme des composantes bruitées est discutée au Chapitre 6.

2.3.3.2. Des trajectoires multiformes

Les sections de parole entièrement voisées considérées dans cette étude sont donc des sections qui peuvent éventuellement se limiter à un seul phonème, mais dont la taille peut aussi souvent aller au-delà du phonème ou du niveau de la syllabe (dans certains cas, cela peut même aller jusqu'à une phrase entière assez longue). Comme on l'a déjà mentionné précédemment, on peut donc avoir à modéliser des sections longues de plusieurs centaines de millisecondes.

Plus généralement la longueur et la forme des sections voisées peuvent varier considérablement d'une section à l'autre. En effet, outre la séquence de phonèmes[49], qui est un facteur purement phonétique et phonologique, d'autres facteurs interviennent dans cette variabilité : des facteurs liés au locuteur, notamment son débit de parole, des facteurs linguistiques, notamment une prosodie plus ou moins marquée (on pense aux trajectoires de fréquences), et enfin on peut citer des facteurs acoustiques, notamment le rang de l'harmonique modélisé.

Du fait de cet aspect multiforme des jeux de paramètres à encoder, nous devons à la fois utiliser des modèles à long terme suffisamment souples pour pouvoir s'adapter aux différentes configurations possibles, et proposer conjointement une méthode d'ajustement de ces modèles à ces jeux de paramètres. Les deux sections suivantes traitent respectivement de ces deux thèmes.

2.4. Les différents types de modèles à long terme

2.4.1. Généralités sur les modèles à long terme utilisés

Dans cette section, on décrit les différents modèles que nous avons considérés dans cette thèse pour remplir la fonction de modélisation à long terme pour chaque séquence de paramètres représentant une section de signal entièrement voisée. Il s'agit de modèles de type linéaire au sens où ils sont tous définis par une combinaison linéaire de fonctions de l'indice temporel. Les paramètres du modèle à ajuster sont les coefficients

[49] Voir à ce propos les Sections 1.1.4.1 et 1.1.4.2, et notamment les notes de bas de page 9 et 12.

de cette combinaison linéaire. Cependant, les fonctions composant la combinaison linéaire ne sont pas forcément elles-mêmes linéaires, bien au contraire, puisqu'elles doivent décrire des trajectoires complexes et variées. Ce choix repose sur le fait que l'ajustement de ce type de modèle aux données peut être réalisé très facilement par une procédure des moindres carrés, comme on le verra à la Section 2.5.4. Notons que ce type de procédure a déjà été abordé dans la Section 1.2.1.2 dans le cadre de l'analyse des paramètres du modèle sinusoïdal.

Différents types de fonctions peuvent être proposés pour composer la combinaison linéaire du modèle. Dans cette étude, nous avons choisi des modèles très populaires à la fois simples et efficaces (populaires car simples et efficaces !), comme : le modèle en cosinus discrets (MCD), un modèle combinaison de fonctions cosinus et sinus (MCSD pour modèle en cosinus/sinus discrets), et le modèle polynomial (MP). Nous proposons aussi l'utilisation de diverses combinaisons des modèles mentionnés : on peut par exemple mixer des termes en cosinus avec des termes polynomiaux. Nous décrivons séparément chacun des modèles listés ci-dessus dans les sections suivantes.

Dans la suite de cette section, la variable V dénote de façon générale le paramètre à modéliser à long terme lorsque les définitions et propriétés des modèles sont indépendantes de la nature de ce paramètre. Lorsqu'il est nécessaire de préciser la nature du paramètre, nous emploierons comme dans le Chapitre 1, la notation A pour les paramètres d'amplitude et φ pour les paramètres de phase.

2.4.2. Modèle en cosinus discrets

Le premier modèle que nous présentons ici est le modèle en cosinus discrets (MCD). Nous insistons particulièrement sur ce modèle ici, car nous verrons par la suite, notamment dans les expérimentations des Chapitres 3 et 4, que ce modèle (et une de ses variantes proches) s'est révélé comme un très bon compromis entre efficacité et simplicité. Dans le cadre de cette thèse, ce modèle est défini comme la combinaison linéaire de fonctions cosinus suivante :

$$\hat{V}_i(n) = \sum_{p=0}^{P_i} c_{i,p} \cos\left(p\pi \frac{n}{N} \right) \tag{2.1}$$

Dans cette équation, i désigne l'indice de l'harmonique modélisé[50], P_i est un nombre entier positif définissant l'ordre du modèle et par conséquent P_i +1 est le nombre de coefficients du modèle. Ces coefficients notés $c_{i,p}$ sont tous réels puisque ce modèle sera utilisé pour modéliser une suite de valeurs réelles (des valeurs d'amplitude et de phase). N est la taille de la section de signal modélisée à long terme en nombre d'échantillons. Il est donc important de noter que ce modèle fournit bien une valeur de paramètre pour chaque indice temporel de la section considérée, indicée arbitrairement de 0 à $N-1$ lors du processus de modélisation. Le terme en $1/N$ dans l'argument des cosinus permet de normaliser la base de fonctions du modèle en fonction de la taille de la section.

[50] C'est le même indice que dans l'équation du modèle harmonique (1.49). On pourrait se passer de cet indice ici, pour la présentation des modèles proprement dite, mais on préfère le garder pour fixer l'idée qu'on modélisera plus tard les trajectoires des paramètres du modèle sinusoïdal harmonique par harmonique.

Le modèle en cosinus discrets tel que présenté ci-dessus est connu pour capturer de façon efficace les variations d'un signal. A ce titre, il est tout à fait similaire aux autres modèles en cosinus discrets développés dans d'autres domaines et sous d'autres appellations. Par exemple, il est proche de la transformée en cosinus discret (TCD) dont on a déjà cité une variante, la *Modified Discrete Cosine Transform* (MDCT), dans ce document (voir les Sections 1.1.3 et 1.4.3). Ce type de transformée est généralement appliqué directement sur les échantillons de signal audio pour encoder leur redondance (corrélation temporelle) sur des fenêtres à court terme (généralement de l'ordre de 256 à 2048 échantillons selon la fréquence d'échantillonnage et le type de signal). Notons que les auteurs de [Li *et al.*, 2001] montrent que la transformation en cosinus discrets est, à dimension fixée, la transformation linéaire fixe[51] la plus efficace en terme de minimisation de l'erreur de modélisation et de codage.

De même, le modèle MCD est analogue au modèle dit « cepstre discret » dans [Cappé *et al.*, 1995] [Galas & Rodet, 1990, 1991]. Dans ce cas, ce modèle est utilisé pour décrire l'enveloppe spectrale des signaux de parole et des signaux audio en général. La référence à l'appellation cepstre provient du fait que cette enveloppe est modélisée sur une échelle logarithmique, comme dans le cas du cepstre classique (le « cepstre *FFT* ») où les coefficients cepstraux sont fournis par une *FFT* inverse du logarithme du module de la *FFT* [Oppenheim & Schaffer, 1989] [Rabiner & Schaffer, 1978]. Bien qu'on s'éloigne (momentanément) de notre approche à long terme, on donne ici l'expression du MCD en tant que cepstre discret, car on se servira des développements et résultats donnés dans ce cadre [Cappé *et al.*, 1995] [Galas & Rodet, 1990, 1991] [Campedel-Oudot, 1998] [Campedel-Oudot *et al.*, 2001] pour notre approche de la modélisation à long terme en deux dimensions dans le Chapitre 5. Le cepstre discret est donné par :

$$A_c^k(f) = d_{0,k} + 2\sum_{l=1}^{L} d_{l,k}\cos(2\pi f l) \tag{2.2}$$

Ici, L est l'ordre du cepstre, f est la variable des fréquences réduites, allant de 0 à ½ et on rappelle que les amplitudes modélisées sont sur une échelle logarithmique. Dans ce cas on modélise l'enveloppe du spectre le long de l'axe des fréquences, alors que dans le cas de la modélisation à long terme nous modélisons des trajectoires temporelles des paramètres d'amplitude (et de phase) le long de l'axe du temps pour chaque partiel séparément. Dans notre modèle à long terme, on remplace donc dans les fonctions cosinus l'espace de variation des fréquences par celui des indices temporels, la forme de ces fonctions étant identique (voir Figure 2.3(a)). On remarque par exemple que la valeur limite de ½ pour la variable fréquence réduite dans le cepstre est remplacée dans notre version à long terme par la valeur limite N en termes d'échantillons temporels.

2.4.3. Combinaison de cosinus et sinus

Le deuxième modèle proposé pour modéliser à long terme les trajectoires des paramètres issus du modèle sinusoïdal est une combinaison linaire des fonctions cosinus du modèle MCD et de fonctions sinus correspondantes (c'est-à-dire de même argument ; voir Figure 2.3).

[51] C'est-à-dire dont les coefficients (ceux de la matrice de transformation) sont indépendants du signal.

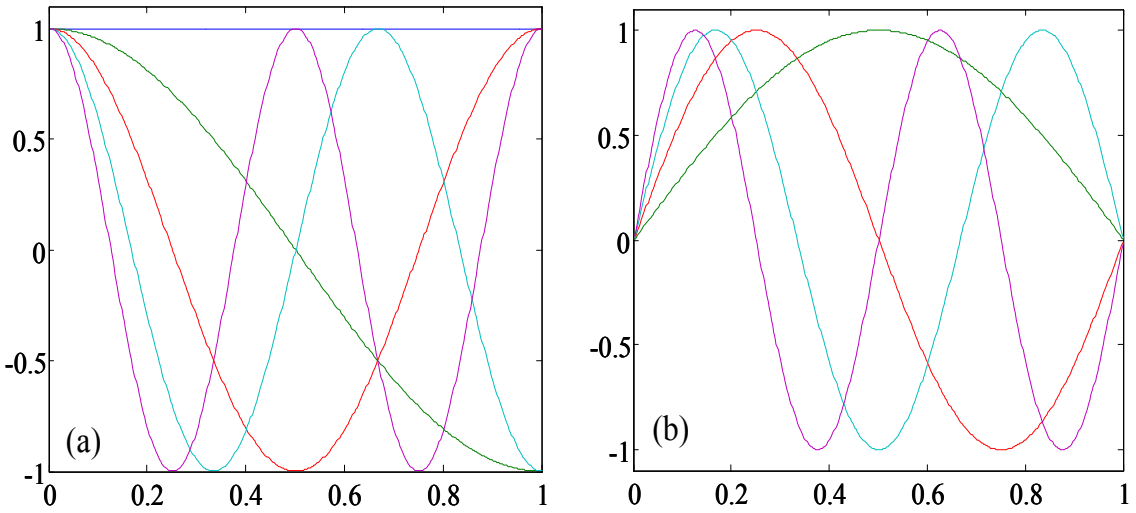

Figure 2.3 : (a) Les cinq premières fonctions cosinus composant le modèle MCD, (b) les fonctions sinus correspondantes pour le modèle combiné de fonctions cosinus et sinus (MCSD).

L'idée de base derrière cette combinaison de fonctions cosinus et sinus provient du fait que les fonctions cosinus du MCD sont contraintes au niveau de la phase : elles commencent toutes par la même phase à l'origine, en l'occurrence la valeur 0. Ceci signifie que le MCD n'a pas la liberté d'ajuster aux données les fonctions cosinus au niveau de la phase à l'origine (on ajuste seulement les amplitudes de ces cosinus). Un décalage de phase des fonctions cosinus peut être contrôlé en ajoutant les fonctions sinus de même argument au modèle MCD, en se basant sur une propriété trigonométrique très simple, déjà mentionnée et exploitée à la Section 1.2.1.2 de ce document :

$$\hat{V}_i(n) = c_{i,0} + \sum_{p=1}^{P_i/2}\left(c_{i,p}\cos\left(p\pi\frac{n}{N}\right) + c_{i,p+P_i/2}\sin\left(p\pi\frac{n}{N}\right)\right) \tag{2.3}$$

soit :

$$\hat{V}_i(n) = \sum_{p=0}^{P_i/2} b_{i,p}\cos\left(p\pi\frac{n}{N} + \theta_{i,p}\right) \tag{2.4}$$

avec

$$\begin{cases} \theta_{i,p} = -\arctan(c_{i,p+P_i/2} / c_{i,p}) \\ b_{i,p} = \sqrt{c_{i,p}^2 + c_{i,p+P_i/2}^2} \end{cases} \tag{2.5}$$

Dans ces équations, nous supposons que P_i est pair et nous prenons $P_i/2$ fonctions cosinus et $P_i/2$ sinus pour conserver le même nombre total de coefficients que pour le modèle MCD. Le modèle résultant est appelé ici modèle MCSD pour modèle en cosinus et sinus discrets.

2.4.4. Modèle polynomial

Un autre type de modèle relativement classique basé sur une combinaison linéaire de fonctions est le modèle polynomial. On peut représenter ce modèle par l'équation suivante, en rappellant que, ici aussi, P_i est l'ordre du modèle qui compte P_i +1 coefficients, tous réels :

$$\hat{V}_i(n) = c_{i,0} + c_{i,1}n + c_{i,2}n^2 \cdots + c_{i,P_i}n^{P_i} \tag{2.6}$$

Tout comme le modèle MCD, ce modèle est aussi efficace pour décrire des trajectoires lisses, du fait de la nature des fonctions employées. Cependant, il est caractérisé par un problème spécifique : on a dit à la Section 2.3.3.2 que les trajectoires à modéliser pouvaient être longues et de forme assez complexe, ce qui nécessite *a priori* un ordre relativement élevé pour les modèles à long terme. Or, un ordre élevé combiné à une longue taille de trame à long terme signifie pour le modèle polynomial une hétérogénéité dans l'ordre de grandeur de ses différents termes. Ceci peut entraîner des problèmes numériques lors de la détermination des coefficients. Ce point sera discuté plus particulièrement à la Section 3.3.5.4 du chapitre suivant.

2.4.5. Modèle combinaison de cosinus, sinus et polynômes

D'une façon générale, le choix des fonctions composant les modèles à long terme est arbitraire. Par conséquent, toute combinaison de fonctions bornées est possible et en particulier les combinaisons des fonctions déjà mentionnées. On peut par exemple proposer une combinaison de fonctions cosinus et polynomiales qu'on appellera MCDP pour modèle en cosinus discrets et polynômes :

$$\hat{V}_i(n) = c_{i,0} + \sum_{p=1}^{P_i/2}\left(c_{i,p}n^p + c_{i,p+P_i/2}\cos\left((p + P_i/2)\pi\frac{n}{N}\right)\right) \tag{2.7}$$

Dans cette équation, le nombre de fonctions polynomiales est identique au nombre de fonctions cosinus pour simplifier la présentation, mais cette contrainte peut bien sûr être levée et on peut choisir arbitrairement le nombre de fonctions de chaque type. De même, on pourrait rajouter à ce modèle des termes en sinus, comme précédemment. Ce type de combinaison peut par exemple permettre de résoudre les contraintes d'utilisation du modèle polynomial qui entraînent des problèmes numériques : on peut se restreindre aux termes polynomiaux d'ordre faible (1, 2 et 3 par exemple). C'est cette démarche qui a été ainsi adoptée dans [Raspaud *et al*., 2005]. A l'inverse, comme on le verra dans les expérimentations des chapitres suivants, le modèle MCD est très robuste pour modéliser les sections longues de parole et il n'a pas besoin *a priori* d'être limité en ordre.

Une fois que ce principe de combinaison a été posé, il est important de mentionner ici un cas particulier qui mérite une attention spéciale: il s'agit du terme linéaire dans le cas de la modélisation de la phase. Comme on l'a vu à la Section 1.1.4.3 et en particulier dans l'équation (1.3), les trajectoires de phase résultent de la sommation en temps des

trajectoires de fréquence. On rappelle alors que les trajectoires de phase sont des fonctions du temps possédant une forme générale linéaire croissante assez nette avec des variations plus ou moins fortes autour de cette forme de fond, ces variations étant dues aux variations de fréquence (voir la Section 1.1.4.3 et l'illustration à la Figure 2.1). C'est pourquoi, dans le cas de la modélisation des trajectoires de phase, on choisit dans nos modèles à long terme d'intégrer systématiquement dans le modèle un terme linéaire lorsque celui-ci n'est pas déjà présent dans la définition du modèle (comme c'est le cas par exemple pour le modèle polynomial quand celui-ci n'est pas réduit à un terme constant). Par exemple, le MCD ne sera pas utilisé tel quel pour modéliser les trajectoires de phase, mais sera transformé en modèle MCDL (modèle en cosinus discrets + linéaire) :

$$\hat{\varphi}_i(n) = \sum_{p=0}^{P_i-1} c_{i,p} \cos\left(p\pi \frac{n}{N} \right) + c_{i,P_i} n \qquad (2.8)$$

Ainsi, dans ce modèle, les fonctions cosinus sont utilisées pour modéliser les variations des trajectoires de phase autour de la forme linéaire de base censée être capturée par le terme linéaire du modèle. Notons que pour conserver la relation entre l'ordre P_i et le nombre de coefficients P_i+1, le terme linéaire remplace en fait ici le terme de cosinus le plus élevé du MCD. C'est pourquoi on limite la sommation de 0 à P_i−1 dans l'équation (2.8).

2.5. Ajustement des modèles aux données

2.5.1. Généralités

Dans cette section, nous présentons de façon générale la méthode que nous employons pour ajuster les modèles à long terme (présentés à la Section 2.4) aux trajectoires de paramètres des sections de parole considérées. On a vu à la Section 2.3.3.2 que les longueurs et les formes des trajectoires de paramètres à modéliser peuvent être très variables d'une section à l'autre. Or la complexité des formes possibles (en tant que fonction de l'indice temporel) prises par les modèles présentés dans la Section 2.4 peut aussi être très variable selon les valeurs des coefficients (à déterminer pour chaque section modélisée). Elle dépend aussi largement de l'ordre du modèle qui fixe le nombre de ses degrés de liberté. C'est pourquoi il s'agit pour chaque section de parole à modéliser, chaque type de paramètres sinusoïdaux, amplitude ou phase, et chaque harmonique, de trouver une méthode capable à la fois de fixer l'ordre optimal au sens d'un certain critère et d'estimer les coefficients correspondants. Comme nous avons un objectif de parcimonie dans la représentation des trajectoires, l'ordre optimal sera le plus petit ordre permettant de vérifier le critère en question. En pratique, les modèles à long terme proposés pourront être efficacement exploités dans le codage de parole à très bas débit si dans la pratique on constate généralement que l'ordre estimé P_i est significativement inférieur au nombre de mesures K des paramètres à modéliser sur la section de parole considérée. Ce point sera discuté plus en détail dans les Sections 3.3.4 et 4.3.3.4.

Le critère d'ajustement des modèles à long terme est un critère de type perceptif : il s'agit d'une adaptation le long de l'axe temporel de critères perceptifs définis

usuellement selon l'axe des fréquences (pour les amplitudes) ou bien dans le cadre stationnaire (pour les phases). Nous donnons le principe général de cette adaptation dans la Section 2.5.3. Les critères précis correspondant à la modélisation à long terme des amplitudes et des phases étant donnés respectivement aux Sections 3.1 et 4.3.1 des chapitres suivants. L'ajustement lui-même est réalisé par une régression au sens des moindres carrés pondérés, les poids de la régression étant déterminés par la contrainte des critères perceptifs. Cette procédure est associée à un algorithme itératif pour la recherche de l'ordre optimal. Pour plus de simplicité, nous présenterons cette technique en deux temps : le principe général des moindres carrés pondérés sera présenté à la Section 2.5.4. L'algorithme itératif faisant appel à cette procédure des moindres carrés sera présenté à la Section 2.5.5, toujours sous forme très générale. Nous verrons son adaptation aux cas respectifs des amplitudes, des phases, ainsi qu'à celui de la modélisation 2D, respectivement aux Chapitres 3, 4 et 5. Au final, la force de la méthode proposée est qu'elle assure conjointement une estimation optimale de l'ordre du modèle et l'ajustement perceptuel du modèle avec les données pour chaque section de parole modélisée.

Avant d'entrer dans le vif du sujet, nous donnons à la section suivante l'ensemble des notations utilisées dans la suite de ce document pour formaliser notre algorithme de modélisation à long terme.

2.5.2. Notations

Par souci de concision de la présentation, nous utilisons ici des notations vectorielles. Dans toute la suite du document, pour les vecteurs et les matrices, les lignes représentent la dimension temporelle pour les données ou la dimension des coefficients pour les modèles, alors que les colonnes représentent la dimension paramétrique des données (c'est-à-dire concrètement les différentes harmoniques). Ainsi, nous notons :

- K la taille temporelle des données, c'est-à-dire le nombre de paramètres sinusoïdaux à modéliser à long terme sur la portion de parole considérée. On note k l'indice temporel correspondant ;

- $\boldsymbol{V}_i = [V_{i,1}\ V_{i,2}\ \ldots\ V_{i,K}]$, le vecteur ligne de données de la i-ème harmonique que l'on se propose de modéliser à long terme (qui sera par la suite une valeur d'amplitude ou de phase) ;

- $\boldsymbol{N} = [n_1\ n_2\ \ldots\ n_K]$, le vecteur ligne contenant les indices d'échantillons des centres des trames d'analyse à court terme successives ; en d'autres termes, il s'agit des positions des valeurs mesurées $V_{i,k}$;

- $\boldsymbol{C}_i = \lfloor c_{i,0}\ c_{i,1}\ \ldots\ c_{i,P_i} \rfloor$, le vecteur ligne des coefficients du modèle à long terme considéré ;

- $\boldsymbol{M}_i$ la matrice $(P_i+1)\times K$ qui contient les termes du modèle à long terme utilisé[52], ces termes étant évalués aux composantes du vecteur $\boldsymbol{N}$, c'est à dire aux instants de mesure.

[52] Par simplicité, et comme on l'a déjà mentionné, l'ordre P_i du modèle est ici identifié au nombre de coefficients moins un (c'est-à-dire qu'on a dans tous les cas P_i+1 coefficients pour le modèle, ces coefficients étant indicés de 0 à P_i).

Ainsi, pour le modèle polynomial (MP), cette matrice a pour terme général $m_{p,k} = (n_k)^p$, c'est-à-dire qu'on a (voir l'équation (2.6)) :

$$\boldsymbol{M}_i = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ n_1 & n_2 & n_3 & \dots & n_K \\ n_1^2 & n_2^2 & n_3^2 & & n_K^2 \\ \dots & \dots & \dots & & \\ n_1^{P_i} & n_2^{P_i} & n_3^{P_i} & \dots & n_K^{P_i} \end{bmatrix} \tag{2.9}$$

Dans le cas du modèle MCD, le terme général de la matrice est $m_{p,k} = \cos(p\pi n_k/N)$, et $\boldsymbol{M}_i$ est alors donnée par (voir l'équation (2.1)) :

$$\boldsymbol{M}_i = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \cos\left(\pi\frac{n_1}{N}\right) & \cos\left(\pi\frac{n_2}{N}\right) & \dots & \cos\left(\pi\frac{n_K}{N}\right) \\ \cos\left(2\pi\frac{n_1}{N}\right) & \cos\left(2\pi\frac{n_2}{N}\right) & \dots & \cos\left(2\pi\frac{n_K}{N}\right) \\ \dots & \dots & \dots & \dots \\ \cos\left(P_i\,\pi\frac{n_1}{N}\right) & \cos\left(P_i\,\pi\frac{n_2}{N}\right) & \dots & \cos\left(P_i\,\pi\frac{n_K}{N}\right) \end{bmatrix} \tag{2.10}$$

Pour ne pas alourdir la présentation, on ne donne pas les matrices pour tous les modèles introduits à la Section 2.4. Disons simplement que pour le modèle MCSD, la deuxième partie des termes en cosinus est remplacée par des termes en sinus de même argument que les $P_i/2$ premiers cosinus. De la même façon, pour le modèle MCDP, on peut remplacer une partie des termes en cosinus d'ordres les plus élevés par des termes polynomiaux. En particulier, on peut présenter ici la matrice du modèle MCDL (modèle en cosinus discrets + linéaire) qui sera privilégiée dans le Chapitre 4 pour la modélisation des trajectoires de phase. Dans ce cas, le vecteur $\boldsymbol{N}$ est concaténé à la matrice du modèle MCD pour avoir une ligne supplémentaire de termes linéaires, et $\boldsymbol{M}_i$ devient (voir l'équation (2.8) dans la Section 2.4.5) :

$$\boldsymbol{M}_i = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \cos\left(\pi\frac{n_1}{N}\right) & \cos\left(\pi\frac{n_2}{N}\right) & \dots & \cos\left(\pi\frac{n_K}{N}\right) \\ \cos\left(2\pi\frac{n_1}{N}\right) & \cos\left(2\pi\frac{n_2}{N}\right) & \dots & \cos\left(2\pi\frac{n_K}{N}\right) \\ \vdots & \vdots & & \vdots \\ \cos\left((P_i-1)\pi\frac{n_1}{N}\right) & \cos\left((P_i-1)\pi\frac{n_2}{N}\right) & \cdots & \cos\left((P_i-1)\pi\frac{n_K}{N}\right) \\ n_1 & n_2 & \dots & n_K \end{bmatrix} \tag{2.11}$$

Avec ces notations, les valeurs des paramètres modélisés à long terme, échantillonnées aux instants de $\boldsymbol{N}$ sont données par :

$$\hat{\boldsymbol{V}}_i = \boldsymbol{C}_i\,\boldsymbol{M}_i \tag{2.12}$$

L'erreur de modélisation, en tant que série temporelle de valeurs, est définie par la différence entre le vecteur des échantillons du modèle et celui des paramètres originaux, c'est-à-dire :

$$\boldsymbol{E}_i = \hat{\boldsymbol{V}}_i - \boldsymbol{V}_i = \boldsymbol{C}_i \boldsymbol{M}_i - \boldsymbol{V}_i = \left[E_{i,1} \quad E_{i,2} \quad \ldots \quad E_{i,K}\right] \tag{2.13}$$

avec pour chaque composante i et pour chaque indice k (on rappelle que n_k est le centre de la trame d'analyse à court terme d'indice k) :

$$E_{i,k} = \hat{V}_i\ (n_k) - V_{i,k} = \hat{V}_{i,k} - V_{i,k} \tag{2.14}$$

2.5.3. Principe général de l'adaptation des critères perceptifs au traitement à long terme

Le principe général de l'utilisation de critères perceptifs dans le cadre de la modélisation à long terme consiste à adapter des critères existants déjà dans des cadres plus classiques, au nouveau cadre temporel non stationnaire de cette modélisation. On verra aux chapitres suivants que ces cadres classiques sont celui du masquage fréquentiel pour la modélisation des trajectoires d'amplitudes, et celui de la modulation de fréquence pour la modélisation des trajectoires de phase, tous deux étant définis généralement pour des signaux stationnaires. On suppose donc que dans chaque cas, ces critères « classiques » fournissent pour chaque trame k d'analyse à court terme de signal (considérée comme une portion stationnaire), pour chaque harmonique i et pour chaque type de paramètre (amplitude ou phase), une valeur de seuil perceptif $T_{i,k}$. Ce seuil est défini comme la limite en dessous de laquelle une certaine fonction de l'erreur de modélisation (par exemple sa puissance pour les amplitudes) est supposée inaudible.

L'adaptation de ce critère perceptif au cadre de la modélisation à long terme consiste à reconsidérer le problème globalement le long de l'axe temporel sur l'ensemble de la section à modéliser à long terme plutôt qu'en chaque trame à court terme. Pour cela, on considère globalement les valeurs de seuil selon l'axe temporel en définissant pour chaque harmonique le vecteur (ligne) de seuil perceptif :

$$\boldsymbol{T}_i = [T_{i,1}\ \ T_{i,2}\ \ \ldots\ \ T_{i,K}] \tag{2.15}$$

On a vu précédemment que l'erreur de modélisation est vectorisée selon l'axe temporel avec le formalisme de (2.13). Le critère perceptif est alors défini par une certaine fonction de ce vecteur qui prend ses valeurs en chacune de ses composantes :

$$f(\boldsymbol{E}_\mathrm{i}) = [f(E_{i,1})\ f(E_{i,2})\ \ \ldots\ f(E_{i,K})] \tag{2.16}$$

On verra la nature de ce vecteur dans le cas des amplitudes et des phases respectivement aux Sections 3.2 et 4.3.3.1. Disons juste ici qu'on se réserve la possibilité de modifier les valeurs de seuil par rapport à celles utilisées généralement dans le cadre stationnaire ou à court terme pour tenir compte de l'aspect dynamique de l'information traitée. On va voir maintenant comment insérer ces critères perceptifs « vectorisés » selon la dimension temporelle dans l'ajustement des modèles à long terme.

2.5.4. Ajustement au sens des moindres carrés pondérés

Pour n'importe lequel des modèles à long terme présentés à la Section 2.4, et pour un ordre de modèle fixé, l'ajustement proprement dit du modèle aux données est réalisé par une procédure très classique de minimisation de l'erreur de modélisation au sens des moindres carrés pondérés (ou *WMMSE* pour *Weighted Minimum Mean Square Error*). Nous avons déjà présenté cette procédure dans le cadre spécifique de l'analyse des paramètres du modèle sinusoïdal à la Section 1.2.1.2. Nous la présentons à nouveau brièvement dans ce nouveau cadre de l'estimation des paramètres du modèle à long terme. On suppose donc dans cette sous-section que l'ordre du modèle à long terme est fixé et l'estimation de cet ordre fera l'objet de la sous-section suivante.

Comme on l'a déjà mentionné, ce type d'ajustement est rendu possible par la nature même des modèles choisis qui sont des combinaisons linéaires de fonctions ne dépendant que de l'indice temporel. Les coefficients du modèle à déterminer sont les coefficients de la combinaison. Dans l'ajustement des moindres carrés, la notion de pondération provient de l'utilisation sous forme de « poids » des critères perceptifs, dans leur version vectorisée selon l'axe temporel. Ainsi, pour une séquence de paramètres $\boldsymbol{V}_i$, la forme quadratique pondérée par le critère perceptif associée à l'erreur de modélisation (2.13) est définie ici de la façon suivante :

$$\varepsilon_i = \left(\boldsymbol{C}_i \boldsymbol{M}_i - \boldsymbol{V}_i\right)\boldsymbol{W}_i \left(\boldsymbol{C}_i \boldsymbol{M}_i - \boldsymbol{V}_i\right)^T \tag{2.17}$$

Dans cette expression, $\boldsymbol{W}_i$ est une matrice diagonale de taille K contenant les poids qui dépendent du critère perceptif et donc de la fonction (2.16). Ces poids évoluent lors de l'exécution de l'algorithme itératif d'ajustement qui sera présenté dans la Section 2.5.5. Le contenu de cette matrice de poids sera alors présenté dans cette section. Dans le développement ci-dessous, on considère que cette matrice est fixée au moment où on applique la procédure des moindres carrés. On peut juste souligner ici que ces poids sont définis le long de l'axe de temps et ils servent donc à donner une importance relative aux différentes régions de la séquence de données dans le processus de modélisation (on précisera ce point dans les Sections 2.5.5, 3.2 et 4.3.2). Le processus de minimisation au sens des moindres carrés pondérés consiste ici à chercher le vecteur de coefficients du modèle qui minimise ε_i, soit :

$$\boldsymbol{C}_i = \arg\min_{C \in R^{P_i+1}} \left[\left(\boldsymbol{C}\boldsymbol{M}_i - \boldsymbol{V}_i\right)\boldsymbol{W}_i \left(\boldsymbol{C}\boldsymbol{M}_i - \boldsymbol{V}_i\right)^T \right] \tag{2.18}$$

Puisque le processus de modélisation à long terme vise intrinsèquement à fournir une réduction de dimension de données, nous supposons que $P_i+1<K$. De façon similaire à la Section 1.2.1.2, le vecteur optimal de coefficients est alors donné par :

$$\boldsymbol{C}_i = \boldsymbol{V}_i \boldsymbol{W}_i \boldsymbol{M}_i^T \left(\boldsymbol{M}_i \boldsymbol{W}_i \boldsymbol{M}_i^T\right)^{-1} \tag{2.19}$$

Pour les références concernant ce résultat, on pourra se reporter à la Section 1.2.1.2. Pour en finir avec cette technique, il est utile de préciser qu'on a dans certains cas utilisé une version raffinée de cette méthode, qui a été proposée par [Cappé *et al*., 1995][53] dans

[53] Nous avons déjà mentionné ces travaux, ainsi que ceux de [Galas & Rodet, 1990, 1991], à la Section 2.4.2 de ce document, et nous en reparlerons à la Section 5.1.1 dans le cadre de la modélisation à long terme en deux dimensions.

le cadre de la modélisation d'enveloppe spectrale : l'objectif était dans ce cas de déterminer les coefficients cepstraux conduisant à une enveloppe spectrale passant le plus proche possible des amplitudes des harmoniques tout en restant suffisamment lisse, et ceci sans diminuer l'ordre du modèle. De plus, cette version doit permettre de résoudre d'éventuels problèmes numériques liés au conditionnement de la régression. En effet, la matrice à inverser dans l'équation (2.19) n'est pas nécessairement inversible. En particulier, elle peut être mal conditionnée lorsque P_i est proche de K, ou bien selon la gamme de valeurs qu'elle contient. Pour résoudre ce problème et rendre l'estimation des coefficients du modèle plus robuste, [Cappé *et al.*, 1995] ont introduit dans le cadre du modèle cepstral (c'est-à-dire équivalent au modèle MCD) un terme de régularisation. L'équation (2.17) est alors modifiée selon la forme suivante (avec des notations adaptées à notre problème) :

$$\boldsymbol{\varepsilon}_i = \left(\boldsymbol{C}_i \boldsymbol{M}_i - \boldsymbol{V}_i\right)\boldsymbol{W}_i \left(\boldsymbol{C}_i \boldsymbol{M}_i - \boldsymbol{V}_i\right)^T + \lambda \boldsymbol{C}_i^t \boldsymbol{R} \boldsymbol{C}_i \tag{2.20}$$

Dans cette nouvelle équation, λ est le paramètre de régularisation et $\boldsymbol{R}$ est une matrice carrée donnée par :

$$\boldsymbol{R} = 8\pi^2 \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1^2 & 0 & \cdots & 0 \\ 0 & 0 & 2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & P_i^2 \end{bmatrix} \tag{2.21}$$

Dans le cas où on utilise le modèle MCD, la minimisation de l'erreur définie par (2.20) est alors donnée par :

$$\boldsymbol{C}_i = \boldsymbol{V}_i \boldsymbol{W}_i \boldsymbol{M}_i^T \left(\boldsymbol{M}_i \boldsymbol{W}_i \boldsymbol{M}_i^T + \lambda \boldsymbol{R}\right)^{-1} \tag{2.22}$$

Il est important de noter que le terme de régularisation induit un lissage du modèle, c'est-à-dire que les trajectoires modélisées sont adoucies par rapport aux mêmes trajectoires obtenues sans tenir compte du terme de régularisation. Ceci correspond au fait que le terme de régularisation est un terme correspondant à la courbure du modèle, qui entre ainsi en compte dans la minimisation au détriment du respect des données imposé par le terme quadratique de gauche de l'équation (2.20). Dans la pratique, un choix de valeur pour λ de l'ordre de 10^{-5} à 10^{-6} semble approprié : les éventuels problèmes de conditionnement sont corrigés tout en conservant des trajectoires correctes pour représenter les données.

2.5.5. Algorithme d'estimation de l'ordre optimal

Nous allons maintenant présenter l'algorithme que nous avons élaboré pour estimer automatiquement l'ordre optimal du modèle à long terme pour chaque section de parole voisée, chaque harmonique et chaque type de paramètres, sous contrainte de critères perceptifs. Cet algorithme assure le calcul des poids optimaux associés à ces critères perceptifs, ainsi que les coefficients du modèle à long terme, cela va de soi. On rappelle qu'on donne ici la forme générale de l'algorithme, sachant que le processus de mise à

jour des poids doit être adapté aux deux critères perceptifs différents proposés dans les Chapitre 3 et 4 selon que l'on considère la modélisation des amplitudes ou des phases.

Avant de formuler explicitement l'algorithme, on donne quelques éclaircissements pour mieux comprendre son fonctionnement. Le but de cet algorithme est d'obtenir un compromis optimal entre l'ordre du modèle (pas trop élevé) et le pourcentage R des points de la section modélisée qui vérifient la contrainte perceptive, c'est-à-dire pour laquelle la fonction de l'erreur $f(\boldsymbol{E}_i)$ se situe sous le modèle de seuil perceptif $\boldsymbol{T}_i$ associé à la trajectoire (rappelons que ces éléments seront précisés au Chapitre 3 et 4 dans les cas respectifs des amplitudes et phases). On considère que si un fort pourcentage est atteint, alors l'erreur de modélisation sera globalement inaudible sur la section entière. Au départ de l'algorithme, l'utilisateur fixe ainsi un rapport cible R_{min} qui est la valeur minimale de ce pourcentage R à atteindre. Ensuite, le principe est de mettre en forme le modèle à long terme en jouant à la fois sur l'ordre du modèle et sur la mise à jour des poids dans le processus d'ajustement du modèle au sens des moindres carrés pondérés décrit dans la Section 2.5.4. Pour une phase de l'algorithme où l'ordre du modèle est fixé, la procédure est itérative : l'algorithme alterne entre une phase d'ajustement proprement dit et une phase de mise à jour des poids de ce processus. Cette mise à jour des poids doit permettre de donner plus d'importance aux intervalles temporels sur lesquels l'erreur de modélisation ne remplit pas les contraintes perceptives, c'est-à-dire concrètement les régions où la fonction est au dessus du modèle de seuil[54]. Ceci compose les étapes 4 et 5 de l'algorithme ci-dessous. Ce processus itératif est englobé par une recherche du plus petit ordre pour lequel la contrainte sur R est atteinte après la phase itérative.

Le choix du rapport cible R_{min} affecte la qualité du signal modélisé et il doit rester suffisamment proche de 100% (par exemple, nous prendrons par la suite des valeurs comprises entre 75% et 90% ; voir les Sections 3.3.3 et 4.3.3.3 où ce point est particulièrement discuté). Pour ne pas bloquer l'algorithme dans une recherche d'ajustement impossible car trop contrainte (par exemple si l'ordre du modèle est résolument trop petit par rapport à la complexité de la trajectoire[55]), un nombre d'itération maximum, noté *Itermax*, est fixé pour la mise à jour des poids. Si la condition $R>R_{min}$ est réalisée avant le nombre d'itération maximum, l'ordre du modèle est diminué, pour qu'un modèle d'ordre inférieur soit testé[56], et le processus est réitéré. Si le rapport R n'atteint pas R_{min} en *Itermax* itérations, l'ordre du modèle doit alors être augmenté avant de réitérer le processus d'ajustement. Ce test compose l'étape 6 de l'algorithme. *Itermax* est également défini par l'utilisateur. Une gamme de valeurs approximative de l'ordre de 10-20 est convenable (voir Sections 3.3.2.2 et 4.3.3).

Pour accélérer le processus de recherche de l'ordre du modèle P_i, celui-ci est initialement fixé à la puissance de deux la plus proche de $K/2$. La mise à jour (incrément ou décrément) de cette valeur, notée δP_i, est initialement fixée à $P_i/2$, puis elle est divisée par deux à chaque mise à jour. Ce processus dichotomique permet d'augmenter considérablement la vitesse de l'algorithme tout en garantissant que P_i reste inférieur à

[54] En conséquence, moins de poids relatif est affecté aux indices du temps où la fonction de l'erreur est sous le modèle de seuil, mais on cherche à converger vers un meilleur ajustement global sur toute la section de parole modélisée.
[55] On assiste alors à un comportement « oscillant » de l'algorithme à chaque mise à jour des poids.
[56] Rappelons que l'on cherche l'ordre le plus petit possible réalisant les contraintes perceptives.

$K-1$. L'algorithme s'arrête quand $\delta P_i = 0$ et la dernière valeur de P_i pour laquelle R a atteint R_{min} est la valeur de l'ordre optimal recherchée.

Finissons cette section en présentant formellement l'algorithme. On dénote respectivement par *min* et *max* la fonction minimum et la fonction maximum appliquées sur des vecteurs, et on dénote par *diag* la fonction qui produit une matrice diagonale à partir d'un vecteur, les éléments du vecteur étant mis sur la diagonale. Rappelons enfin que les éléments de $f(\boldsymbol{E}_i)$ seront respectivement précisés dans les Chapitres 3 et 4 pour le cas des amplitudes et le cas des phases.

Algorithme à appliquer sur les jeux de mesure d'amplitude ou de phase (R_{min} est initialisé par l'utilisateur à une valeur comprise entre 75% et 90%, et *Itermax* est initialisé à une valeur entière entre 10 et 20) :

1. Pour chaque indice du temps $k \in [1, K]$ et chaque partiel $i \in [1, I]$, calculer le seuil associé $T_{i,k}$ (voir les Sections 3.1 et 4.3.2). Ensuite, former la trajectoire de seuil $\boldsymbol{T}_i$ correspondante. Puis, pour chaque composante i, $i \in [1, I]$:

2. Initialiser l'ordre P_i à la puissance de deux le plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP_i à $P_i/2$.

3. Initialiser la matrice diagonale de poids $\boldsymbol{W}_i$ de taille K avec tous les éléments sur sa diagonale fixés à un. Itérer alors le processus suivant, de l'étape 4 à l'étape 6 :

4. Calculer le vecteur des coefficients $\boldsymbol{C}_i$ du modèle à long terme avec (2.19) ; calculer l'erreur de modélisation associée $\boldsymbol{E}_i = \boldsymbol{C}_i\boldsymbol{M}_i - \boldsymbol{V}_i$.

5. Augmenter les poids où la fonction d'erreur de modélisation dépasse le seuil perceptif, selon :

   $\Delta \boldsymbol{W} = f(\boldsymbol{E}_i) - \boldsymbol{T}_i$

   $\Delta \boldsymbol{W} \leftarrow \Delta \boldsymbol{W} - \min(\Delta \boldsymbol{W})$ (permet d'assurer que les poids sont positifs)

   $\boldsymbol{W}_i \leftarrow \boldsymbol{W}_i + diag(\Delta \boldsymbol{W} / \max(\Delta \boldsymbol{W}))$ (contrôle la valeur de la mise à jour entre 0 et 1)

6. Calculer le pourcentage R des éléments négatifs de $f(\boldsymbol{E}_i) - \boldsymbol{T}_i$.

   Si $R < R_{min}$ et le nombre d'itérations *Itermax* n'est pas atteint, retourner à l'étape 4.

   Si $R \geq R_{min,}$ diminuer l'ordre du modèle selon $P_i \leftarrow P_i - \delta P_i$, mettre à jour δP_i selon $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

   Sinon, si $R < R_{min}$ et le nombre d'itérations atteint *Itermax*, augmenter l'ordre du modèle selon $P_i \leftarrow P_i + \delta P_i$, faire la mise à jour $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

On stoppe l'algorithme pour la composante i quand P_i se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R \geq R_{min}$. On passe alors à la composante suivante.

2.6. En guise de conclusion pour ce chapitre

Pour terminer ce chapitre de façon un peu originale et inattendue, il est amusant de compléter le bref état de l'art de la Section 2.2 par une étude que nous avons « gardée en réserve » pour cette occasion. Il s'agit de l'étude proposée par Dusan et collègues [Dusan *et al*., 2004] qui a été présentée au congrès ICSLP 2004. Il est en effet intéressant de mentionner que cette étude a été présentée en même lieu et temps qu'une version préliminaire du travail présenté dans cette thèse[57]. Si on cite cette étude ici, c'est parce qu'elle présente un certain nombre de points connexes à la nôtre. En effet, dans [Dusan *et al.*, 2004], les auteurs ont utilisé un polynôme d'ordre P pour modéliser la trajectoire de N valeurs consécutives de paramètres LPC, les LSF mentionnés à la Section 2.2, avec la contrainte que P soit suffisamment inférieur à N pour assurer un gain de codage significatif. Ils ont ensuite mis en œuvre un codeur de parole à très bas débit exploitant cette idée. Cette étude se rapproche donc de la nôtre, dans le principe même de la modélisation à long terme, mais elle diffère cependant de la nôtre sur plusieurs points. Nous finirons ce chapitre par une rapide comparaison qui confirme que la modélisation à long terme est une approche qui semble porteuse de multiples fruits :

- D'abord, nous considérons le cadre du modèle sinusoïdal au lieu de la modélisation par prédiction linéaire. Toutefois, nous avons déjà mentionné la possibilité d'adapter notre approche à long terme à d'autres modèles spectraux et la modélisation LPC fait partie de ceux-ci (ce point sera discuté au Chapitre 6).

- Deuxièmement nous utilisons dans cette thèse différents modèles à long terme, basés sur des fonctions de type cosinus et sinus discrets, des polynômes, des combinaisons entre ces possibilités, tandis que dans [Dusan *et al*., 2004], seul le modèle polynomial est considéré.

- Troisièmement, nous proposons l'utilisation de modèles d'ordre (très) variable pour capturer des trajectoires de paramètre de taille (très) variable, tandis que l'étude de [Dusan *et al*., 2004] considère un ordre fixe destiné à coder un jeu de paramètres de taille fixe et relativement réduite : des résultats sont rapportés pour des polynômes d'ordre 4 appliqués sur dix valeurs de LSP consécutives. Cette dernière approche est ainsi proche dans l'esprit d'une quantification des trajectoires de paramètres LSF par une quantification matricielle de taille fixe [Tsao & Gray, 1985] [Xydeas & Papanastasiou, 1999]. Notre étude est plus proche d'une idée de modèle adaptatif tenant compte de l'évolution de la dynamique des caractéristiques spectrales du signal. Ce caractère adaptatif se traduit en terme de codage par un codage à débit variable plus complexe mais potentiellement plus performant au niveau du débit.

- D'une manière générale, dans l'étude de Dusan et collègues, l'effort est plutôt porté sur l'aspect codage, et à ce titre, une procédure astucieuse de quantification est proposée (nous en reparlerons au Chapitre 6). De notre côté, nous avons

[57] Cette version préliminaire [Girin *et al*., 2004] comporte les grands principes décrits dans cette thèse mais en se restreignant aux paramètres de phase seuls, et sans considérer de contraintes perceptives pour l'ajustement des modèles à long terme (à l'époque, nous avons considéré un critère objectif du type rapport signal sur bruit). Cette étude est présentée à la Section 4.2 de ce document.

plutôt porté nos efforts sur l'aspect modélisation pure, sans tester l'aspect quantification par exemple, à travers le test de différents modèles à long terme, mais aussi avec l'introduction dans le processus de modélisation de critères perceptifs. L'étude de Dusan et collègues ne comporte pas une telle approche perceptive. Cependant, elle fournit une base d'inspiration très intéressante pour appliquer notre propre approche de la modélisation à long terme au codage de la parole à très bas débit. Ce point est aussi discuté au Chapitre 6.

Chapitre 3

3. Application à la modélisation à long terme de l'amplitude

Dans ce chapitre, nous nous intéressons à la modélisation à long terme des paramètres de type amplitudes spectrales du modèle sinusoïdal de la parole et des signaux audio. On rappelle que ces paramètres sont les facteurs multiplicatifs des fonctions cosinus du modèle sinusoïdal (1.1) ou (1.2). Ce sont aussi les « sommets locaux » du spectre d'amplitude à court terme du signal (voir Chapitre 1). Dans cette section, nous précisons d'abord comment l'algorithme d'estimation de l'ordre optimal (Section 2.5.5) est adapté au cas de ces amplitudes, notamment en ce qui concerne le critère perceptif utilisé. Puis nous effectuons une étude expérimentale complète de l'optimisation globale du processus de modélisation à long terme de ces paramètres d'amplitude.

Ce chapitre est donc organisé comme suit. Dans la Section 3.1, nous donnons une présentation détaillée des critères perceptuels utilisés dans la modélisation à long terme des trajectoires des paramètres d'amplitude. Puis nous donnons dans la Section 3.2, la version précise de l'algorithme d'estimation de l'ordre et des coefficients du modèle à long terme, pour traiter ce cas des amplitudes. Ensuite, dans la Section 3.3, nous décrivons le protocole expérimental que nous avons mis en œuvre pour évaluer l'implémentation de l'algorithme, et nous présentons les résultats que nous avons obtenus dans cette série d'expérimentations.

3.1. Critère perceptif : seuil de masquage à long terme

Comme nous l'avons introduit dans la Section 1.5.3.2 du Chapitre 1, le système auditif humain est caractérisé par un seuil auditif et par le phénomène de masquage fréquentiel. On rappelle brièvement que le masquage fréquentiel se définit comme la capacité d'un son à une certaine fréquence donnée à rendre inaudibles les sons moins forts proches de cette fréquence (voir par exemple la Figure 1.12). Dans cette section, nous décrivons précisément quel modèle nous avons utilisé pour tenir compte de ces phénomènes dans notre étude de modélisation à long terme, et comment nous avons adapté ce modèle pour prendre en compte la dimension temporelle des traitements proposés.

Dans notre étude, nous partons du modèle de masquage du standard MPEG [ISO/IEC MPEG, 1992], dont on peut aussi trouver une bonne description dans [Painter &

Spanias, 2000]. Dans ce modèle, le seuil de masquage $T_{i,j,k}$ provoqué à la fréquence $\omega_{i,k}$ par un masqueur tonal (c'est-à-dire une sinusoïde pure) indicé j présent à la fréquence $\omega_{j,k}$, k désignant l'indice de la trame à court terme courante, est donné par :

$$T_{i,j,k} = P_{j,k} - 0.275B_{j,k} + S_{i,j,k} - 6.025 \tag{3.1}$$

Dans cette équation, $P_{j,k}$ désigne la puissance du masqueur tonal à la fréquence $\omega_{j,k}$ (selon le standard MPEG, toutes les puissances sont normalisées en SPL), $B_{j,k}$ est la fréquence $\omega_{j,k}$ passée en échelle Bark (voir la Section 1.5.2), et $S_{i,j,k}$ est une fonction linaire par morceau de $P_{j,k}$ qui est raisonnablement approximée par un triangle dans l'échelle Bark/dB comme suit (voir le détail dans [Painter & Spanias, 2000]) :

$$S_{i,j,k} = \begin{cases} 17\Delta_{i,j,k} - 0.4P_{j,k} + 11 & -3 \le \Delta_{i,j,k} < -1 \\ (0.4P_{j,k} + 6)\Delta_{i,j,k} & -1 \le \Delta_{i,j,k} < 0 \\ -17\Delta_{i,j,k} & 0 \le \Delta_{i,j,k} < 1 \\ (0.15P_{j,k} - 17)\Delta_{i,j,k} - 0.15P_{j,k} & 1 \le \Delta_{i,j,k} < 8 \end{cases} \tag{3.2}$$

où $\Delta_{i,j,k} = B_{i,k} - B_{j,k}$ est la séparation de deux fréquences Barks correspondant à des composantes masquée et masquante.

Le seuil de masquage global à court terme $T_{i,k}$ à la fréquence $\omega_{i,k}$ est obtenu à partir de la sommation (sur une échelle linéaire) de la contribution de tous les masqueurs tonaux individuels $T_{i,j,k}$ dans un voisinage de 10 Bark. De plus, cette sommation inclut aussi la contribution du seuil d'audition absolu $T_A(i)$ (voir Section 1.5.1) à la fréquence $\omega_{i,k}$ [Painter & Spanias, 2000]. On a alors :

$$T_{i,k} = 10\log_{10}\left(10^{0.1T_A(i)} + \sum_{j=1}^{I} 10^{0.1T_{i,j,k}}\right) \tag{3.3}$$

Notons que le modèle de la norme MPEG prévoit aussi d'inclure dans le seuil la contribution des composantes de bruit (à bande étroite). Cependant, dans notre série d'études, les sections de parole modélisées à long terme sont des sections entièrement voisées et considérées comme quasi-harmoniques. C'est pourquoi nous avons employé une version simplifiée du seuil de masquage défini dans le standard MPEG. Nous ne considérons en effet que la contribution additive des masqueurs de type tonal, telle que définie par les équations ci-dessus. De plus, dans notre étude, les fréquences de ces masqueurs tonaux sont les harmoniques du signal $\omega_{j,k} = j\omega_{0,k}$ où $\omega_{0,k}$ est la fréquence fondamentale sur la trame k. Les seuils de masquage globaux sont calculés pour ces mêmes harmoniques $\omega_{i,k} = i\omega_{0,k}$ (on utilise ici l'indice i au lieu de j pour suivre les mêmes notations que dans les équations ci-dessus). En d'autres termes, le seuil de masquage à la fréquence d'une harmonique est déterminé par l'ensemble des harmoniques : le seuil de masquage servant de critère perceptif pour l'erreur de modélisation est déterminé par les composantes du signal lui-même.

Maintenant que nous avons posé les bases de la définition du seuil de masquage à court terme, c'est-à-dire pour chaque trame à court terme de signal k, il s'agit de définir le critère perceptif utilisé dans nos études de modélisation à long terme. Considérons alors une section de parole pseudo-harmonique à modéliser à long terme contenant K trames à court terme, c'est-à-dire K jeux de paramètres d'amplitude successifs $A_{i,k}$, $i \in [1, I]$[58].

La première étape consiste à calculer d'abord le seuil de masquage à court terme pour chaque trame k de la section de parole considérée, $k \in [1, K]$, à partir du spectre d'amplitude. Pour cela, on utilise les équations (3.1) et (3.3), la puissance $P_{j,k}$ du masqueur tonal à la fréquence $j\omega_{0,k}$ étant égale à $(A_{j,k})^2/2$. Une fois que ce seuil global est calculé pour l'ensemble des harmoniques de chaque trame k, $k \in [1, K]$, la trajectoire temporelle $T_{i,k}$, $k \in [1, K]$, du modèle de seuil du masquage à long terme est obtenue pour chaque harmonique i, $i \in [1, I]$, en considérant simplement ces valeurs le long de l'axe du temps. Ceci est conforme au principe général énoncé à la Section 2.5.3. On obtient alors $\boldsymbol{T}_i = [T_{i,1}\ T_{i,2}\ \ldots\ T_{i,K}]$ pour chaque harmonique i, $i \in [1, I]$, ce qui correspond bien à l'équation (2.15). Notons que, par la suite, les amplitudes sont aussi considérées selon l'axe temporel en remplaçant le vecteur générique de données $\boldsymbol{V}_i$ de la Section 2.5.2 par le vecteur des amplitudes $\boldsymbol{A}_i = [A_{i,1}\ A_{i,2}\ \ldots\ A_{i,K}]$ (et de même pour le vecteur des amplitudes modélisées).

On peut compléter ce tableau par deux remarques. La première est que pour tenir compte du fait que le nombre d'harmoniques n'est pas forcément constant d'une trame à l'autre, on peut gérer la « naissance » et la « mort » de trajectoires de seuil de masquage au niveau de la fréquence de Nyquist exactement de la même façon qu'on peut gérer celles d'un partiel (voir la Section 1.3.2.1). Dans la pratique, on verra qu'on restreint nos expérimentations aux harmoniques de rang suffisament bas pour ne pas avoir à considérer ce problème. La deuxième remarque est que la version du modèle à long terme du seuil de masquage tient compte de l'influence dans le masquage de toutes les harmoniques puisque toutes les harmoniques interviennent dans le calcul du seuil à court terme à partir duquel est construit le seuil à long terme. Dans la pratique, les harmoniques sont généralement suffisament espacées pour que seules les harmoniques voisines d'une fréquence donnée interviennent significativement dans le calcul du seuil à cette fréquence.

La Figure 3.1 illustre les principes développés dans cette section sur la portion de signal représentée à la Figure 3.2 (b). Dans cette figure, pour bien illustrer les trajectoires temporelles d'amplitude et de seuil perceptif, nous avons relié les valeurs des trames successives par des lignes pointillées pour la première et quatrième trajectoire. La figure du bas représente la trajectoire d'amplitude de la quatrième harmonique et celle du seuil perceptif correspondant. Dans ce cas, la forme de la trajectoire du seuil suit de près celle de l'amplitude du fait de la contribution majeure de l'harmonique correspondante dans le calcul du seuil de masquage.

[58] En toute rigueur, le nombre d'harmoniques I peut être variable d'une trame à l'autre, c'est-à-dire qu'il peut dépendre de k. On garde cette notation abusive par souci de simplicité. On verra un peu plus tard les conséquences pour la définition du seuil de masquage à long terme. On rappelle par ailleurs que la méthode d'analyse de ces paramètres sera décrite à la Section 3.3.1.2.

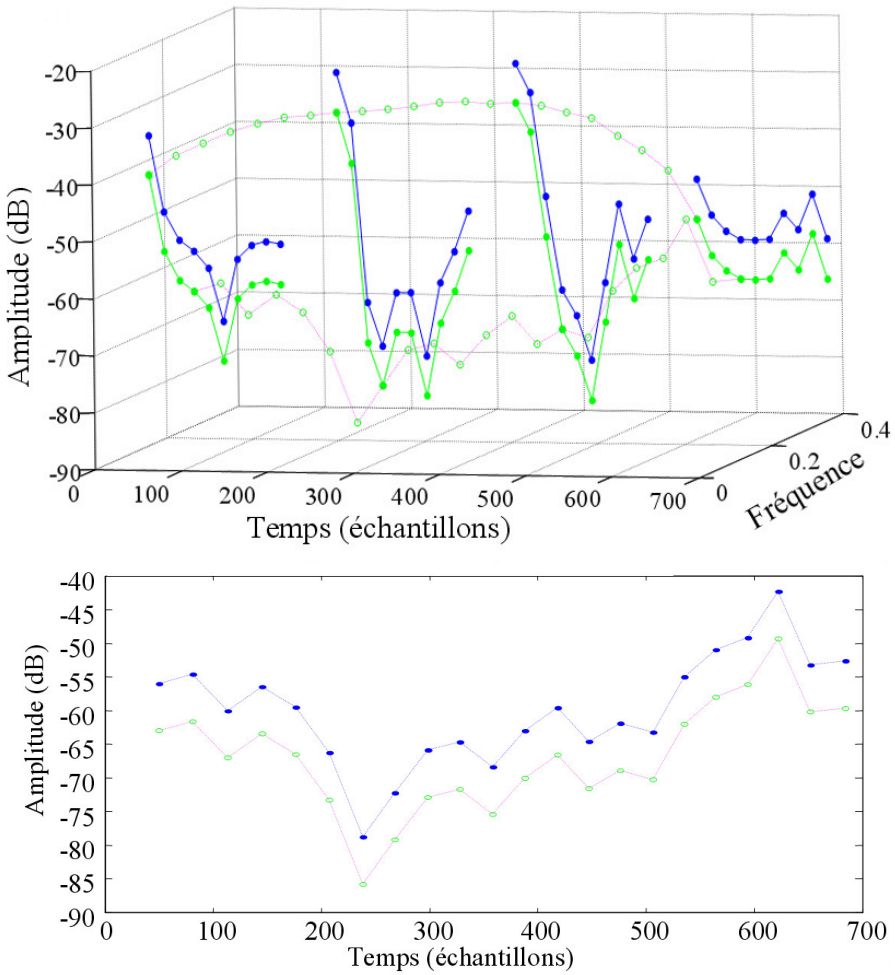

Figure 3.1 : Illustration du seuil de masquage à long terme. Le signal est la voyelle représentée sur la Figure 3.2 (voix de femme, $K = 22$ trames, environ 100 ms de signal, $F_e = 8$ kHz). En haut : représentation fréquentielle du spectre d'amplitude (en bleu, ligne continue), seuil de masquage correspondant (en vert, ligne continue) et version à long terme associée pour les harmoniques 1 et 4 (en magenta, ligne pointillée) ; les valeurs d'amplitude à court terme sont extraites par l'analyse pitch-synchrone décrite à la Section 3.3.1.2. En bas : trajectoire de l'amplitude de la 4ème harmonique (en bleu) avec celle du seuil de masquage correspondant (en magenta).

3.2. *Algorithme de modélisation à long terme des amplitudes spectrales*

Dans cette section, nous présentons de manière précise l'algorithme de modélisation à long terme de la Section 2.5.5 adapté au cas des amplitudes spectrales en utilisant le seuil de masquage à long terme de la section précédente et la fonction d'erreur de modélisation $f(\boldsymbol{E}_i)$ associée. Comme le seuil de masquage est défini pour les puissances des composantes harmoniques, $f(\boldsymbol{E}_i)$ est définie ici comme la puissance de cette erreur de modélisation, soit (*square* dénote la fonction carré élément par élément) :

$$f(\boldsymbol{E}_i) = \frac{1}{2} square(\boldsymbol{A}_i - \hat{\boldsymbol{A}}_i) = \frac{1}{2}\left[\left(A_{i,1} - \hat{A}_{i,1}\right)^2 \quad \left(A_{i,2} - \hat{A}_{i,2}\right)^2 \quad \ldots \quad \left(A_{i,K} - \hat{A}_{i,K}\right)^2 \right] \qquad (3.4)$$

L'algorithme pour la modélisation à long terme des amplitudes spectrales est donné ci-dessous. Rappelons simplement que son principe général est d'ajuster aux données un modèle à long terme aussi simple que possible (c'est-à-dire d'un ordre minimal) tandis que la puissance de l'erreur de modélisation doit rester sous le seuil de masquage afin que l'erreur de modélisation soit inaudible (pour une description détaillée de ce principe, voir la Section 2.5.5). Cette dernière condition doit être vérifiée pour un certain pourcentage R_{min} de l'ensemble de la trajectoire choisi par l'utilisateur.

Algorithme à appliquer sur les jeux des mesures d'amplitude spectrale (dans cet algorithme R_{min} est initialisé par l'utilisateur à une valeur comprise entre 75% et 90% et *Itermax* est initialisé à une valeur entière entre 10 et 20. On rappelle que *max* et *min* dénotent respectivement les fonctions maximum et minimum et *diag* dénote la fonction qui produit une matrice diagonale à partir d'un vecteur, les éléments du vecteur étant mis sur la diagonale) :

1. Pour chaque indice du temps $k \in [1, K]$ et chaque harmonique $i \in [1, I]$, calculer le seuil associé $T_{i,k}$ avec (3.3). Ensuite, former les I trajectoires de seuil $\boldsymbol{T}_i = [T_{i,1}\ T_{i,2}\ \ldots\ T_{i,K}]$ correspondantes pour $i \in [1, I]$. Puis, pour chacune des I harmoniques :
2. Initialiser l'ordre P_i à la puissance de deux le plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP_i à $P_i/2$.
3. Initialiser la matrice diagonale de poids $\boldsymbol{W}_i$ de taille K avec tous les éléments sur sa diagonale fixés à un. Itérer alors le processus suivant, de l'étape 4 à l'étape 6 :
4. Calculer le vecteur des coefficients $\boldsymbol{C}_i$ du modèle à long terme avec (2.19) en remplaçant le vecteur $\boldsymbol{V}_i$ par le vecteur des mesures d'amplitude $\boldsymbol{A}_i$; calculer la fonction d'erreur de modélisation $f(\boldsymbol{E}_i)$ avec (3.4).
5. Augmenter les poids où la fonction d'erreur de modélisation dépasse le seuil perceptif, selon :

   $$\Delta \boldsymbol{W} = f(\boldsymbol{E}_i) - \boldsymbol{T}_i$$
   $$\Delta \boldsymbol{W} \leftarrow \Delta \boldsymbol{W} - \min(\Delta \boldsymbol{W})$$
   $$\boldsymbol{W}_i \leftarrow \boldsymbol{W}_i + diag(\Delta \boldsymbol{W} / \max(\Delta \boldsymbol{W}))$$

6. Calculer le pourcentage R des éléments négatifs de $f(\boldsymbol{E}_i) - \boldsymbol{T}_i$.

   Si $R < R_{min}$ et le nombre d'itérations *Itermax* n'est pas atteint, retourner à l'étape 4.

   Si $R \geq R_{min,}$ diminuer l'ordre du modèle selon $P_i \leftarrow P_i - \delta P_i$, mettre à jour δP_i selon $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

   Sinon, si $R < R_{min}$ et le nombre d'itérations atteint *Itermax*, augmenter l'ordre du modèle selon $P_i \leftarrow P_i + \delta P_i$, faire la mise à jour $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

On stoppe l'algorithme pour chaque composante quand P_i se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R \geq R_{min}$. On passe alors à la composante suivante.

3.3. Expérimentations et résultats

Dans cette section nous décrivons l'ensemble des expériences qui ont été conduites pour évaluer la méthode de modélisation à long terme des trajectoires d'amplitude des sections entièrement voisées de la parole. Nous présentons d'abord le protocole expérimental, puis nous donnons une série de résultats qualitatifs et quantitatifs.

3.3.1. Protocole expérimental

3.3.1.1. Base de données : signaux et harmoniques considérés

Pour réaliser cette campagne d'évaluation, nous avons utilisé des signaux de parole échantillonnés à 8 kHz et produits par 12 locuteurs, six hommes et six femmes. Ces signaux sont du texte lu, et ont été enregistrés dans des conditions très propres (en chambre sourde, sans bruit parasite significatif). Un total d'environ 3500 segments voisés de taille très variée a été extrait de façon semi-automatique. Le logiciel Praat [Boersma & Weenink, 2005] a été utilisé pour mesurer les trajectoires de fréquence fondamentale et les positions des frontières des quasi-périodes de signal (voir la sous-section suivante). A partir des résultats de ces mesures, une segmentation automatique en sections voisées et non voisées a été réalisée, puis le résultat de cette segmentation a été vérifié à la main pour chaque section. Les 3500 segments résultant représentent plus de 13 minutes de parole voisée. Cette quantité de données a été équilibrée entre les voix féminines (environ 1800 segments) et les voix masculines (environ 1700 segments).

Dans la série d'expérimentations présentée dans ce chapitre, pour chaque section, les trajectoires des paramètres d'amplitude des *dix premiers harmoniques* ont été modélisées. On s'est ainsi limité volontairement pour plusieurs raisons :

- La première est que cela permet de s'affranchir de la gestion de la naissance et de la mort des harmoniques dans le processus de suivi temporel au niveau de la fréquence de Nyquist (voir la Section 1.3.2.1 et la Section 1.4.1). En effet, quelle que soit la fréquence fondamentale des sons de parole testés, la zone des dix premières harmoniques est située en dessous de $F_e/2$.
- La seconde raison, la plus importante, est que les trajectoires de paramètres sinusoïdaux pour les harmoniques de rang inférieurs sont plus « lisses » et plus « régulières » que les trajectoires correspondant aux rangs plus élevés. Par conséquent, les premiers harmoniques se prêtent mieux à la modélisation à long terme par les modèles que nous avons proposés au Chapitre 2. Ces observations sont reportées plus en détails et expliquées dans cette section expérimentale. Les conséquences seront prises en compte dans la suite de nos travaux, notamment dans les études exposées au Chapitre 5[59] et dans la discussion du Chapitre 6.
- Enfin, la dernière raison est purement pragmatique : le nombre d'harmoniques pour certaines sections est parfois très grand, de l'ordre de la cinquantaine par exemple, selon la valeur de la fréquence fondamentale. Reporter des résultats quantitatifs pour chaque harmonique, ou à défaut pour chaque région du spectre, s'avérerait alors très lourd. On préfère fournir des résultats plus précis pour une

[59] Rappelons que nous étudions au Chapitre 5 une nouvelle méthode « 2D » qui consiste à modéliser à long terme conjointement toutes les trajectoires d'amplitude de tous les harmoniques.

zone spectrale plus limitée, sachant que la zone en basse fréquence est d'une importance perceptive prédominante dans un traitement harmonique par harmonique du fait de la bonne résolution fréquentielle de l'oreille dans cette zone (voir la Section 1.5.2).

3.3.1.2. Analyse des paramètres

La première étape du processus de modélisation à long terme est en fait d'extraire les valeurs des paramètres d'amplitude devant être modélisés (ainsi que les valeurs des paramètres de phase pour leur modélisation à long terme réalisée dans le Chapitre 4). Nous précisons dans cette sous-section quelle méthode d'analyse nous avons utilisée pour cela. Cette description de la méthode d'analyse n'a pas été faite avant, car bien que nous travaillions sur un processus de modélisation à long terme, l'analyse des valeurs d'amplitude (et de phase) à modéliser est fournie sur une base habituelle à court terme (voir Figure 3.2). Ainsi, on peut aller jusqu'à dire que n'importe laquelle des méthodes d'analyse présentées au Chapitre 1 pourrait être utilisée. C'est pourquoi, on inscrit cette analyse dans la section dédiée au protocole expérimental plutôt que dans la méthode de modélisation à long terme proprement dite.

Les expériences décrites dans cette étude ont été conduites avec une analyse des paramètres par une méthode du type *WMMSE période-synchrone*. L'abbréviation *WMMSE* pour *Weighted Minimum Mean Square Error* fait référence au type de méthode par analyse-par-synthèse présenté à la Section 1.2.1.2. Le terme *période-synchrone* signifie que chaque période de signal est utilisée comme trame d'analyse. Bien que cela complique un peu l'analyse par certains aspects, nous avons fait ce choix pour deux raisons principales qui tendent à privilégier cette approche par rapport aux techniques d'analyse plus habituelles basées sur une fenêtre glissante, par exemple celles basées sur la transformée de Fourier à court terme. D'une part ce choix permet de bien suivre l'évolution du signal période par période, cette évolution étant la base de la non-stationnarité du signal que nous voulons précisément pister dans notre étude à long terme. D'autre part, cela permet d'avoir pour chaque trajectoire de paramètre à modéliser un nombre conséquent de données par rapport aux méthodes avec fenêtre glissante, car en général, la durée d'une période est inférieure au décalage de la fenêtre d'analyse dans ces méthodes. Or, un grand nombre de données représente une information plus riche pour ajuster les modèles à long terme à ces données. Enfin, d'un point de vue calculatoire, cela permet aussi d'éviter certains phénomènes de mauvais conditionnement de la régression linéaire utilisée pour l'ajustement des modèles.

Pour réaliser cette analyse période-synchrone, les signaux sont d'abord « pitch-marqués » en employant le logiciel Praat [Boersma & Weenink, 2005]. Cela signifie que, comme on l'a dit précédemment, le logiciel réalise l'estimation de la fréquence fondamentale du signal et les frontières des quasi-périodes du signal sont automatiquement détectées. Ces quasi-périodes sont utilisées comme trame d'analyse dans nos études. Notons sans la détailler que la méthode employée par le logiciel Praat repose sur une détection du maximum de la fonction d'autocorrélation du signal (voir la note de bas de page n°31). La description de l'algorithme est donnée dans [Boersma, 1993]. On peut voir sur la Figure 3.2 un exemple de détection de *pitch-marks* sur la même portion de signal que celle utilisée à la Section 3.1. Les positions du pitch sont indiquées par les lignes rouges verticales.

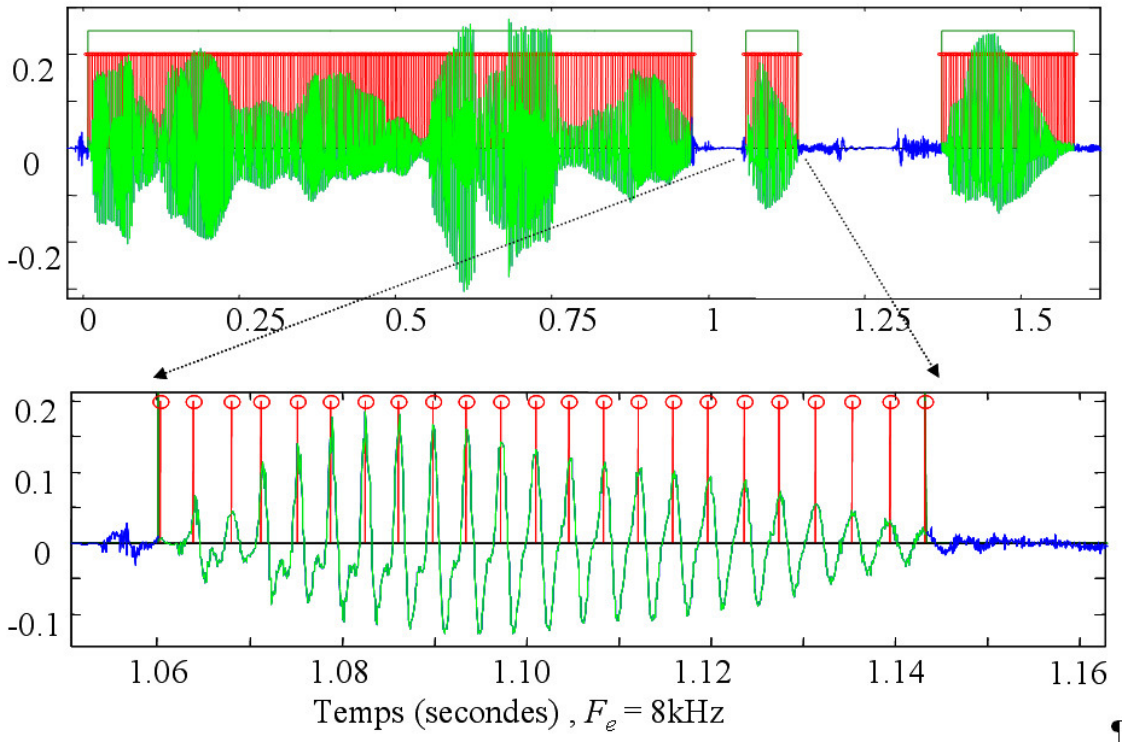

Figure 3.2 : Segmentation d'un signal de parole correspondant à la phrase de notre base de données « Collision de l'avion espion » prononcée par une femme. En haut : représentation des sections voisées (en vert) et non voisées (en bleu). En bas : zoom sur la section voisée centrale (il s'agit du [ε] du mot « espion »), les lignes verticales rouges marquent les positions des frontières de pseudo-période détectées par le logiciel Praat.

Pour chaque pseudo-période, indicée par k, d'une section de parole donnée comprenant K pseudo-périodes, la fréquence fondamentale $\omega_{0,k}$ est alors directement donnée par l'inverse de la période définie comme l'écart entre deux valeurs successives de *pitch-mark*. Etant donnée cette fréquence fondamentale $\omega_{0,k}$, les I_k amplitudes $A_{i,k}$ correspondant à chaque harmonique à la fréquence $i\omega_{0,k}$, et mesurées au centre n_k de chaque période, sont estimées par la procédure *WMMSE* décrite au Chapitre 1, Section 1.2.1.2. On rappelle que cette procédure est basée sur un critère classique de minimisation d'erreur entre le modèle sinusoïdal harmonique et le signal selon un critère des moindres carrés pondérés. Elle fournit une estimation des paramètres très précise avec un coût de calcul très bas. La seule différence entre la méthode effectivement utilisée dans ces expérimentations et celle décrite à la Section 1.2.1.2 est que nous appliquons l'estimation successivement et indépendamment sur chaque période de signal pour obtenir K jeux de paramètres d'amplitude successifs. La fenêtre d'analyse assurant la pondération est donc ici une fenêtre de taille variable exactement égale à la pseudo-période de signal analysée. Cette fenêtre est par ailleurs rectangulaire, ce qui correspond à donner le même poids à chaque échantillon de la période analysée. Notons pour finir que cette procédure fournit également les valeurs des paramètres de phase dont nous réaliserons la modélisation à long terme au Chapitre 4.

3.3.2. Comportement de l'algorithme

3.3.2.1. Quelques exemples

Pour commencer notre campagne d'évaluation, nous proposons d'observer le résultat de la modélisation à long terme des amplitudes spectrales sur quelques exemples typiques de sections de parole traitées avec la méthode proposée. Rappelons que comme annoncé à la Section 2.3.3.2, la longueur et le contenu des sections de parole voisée peuvent être considérablement variables, par exemple selon la séquence de phonèmes. Les sections voisées contiennent de quelques dizaines à quelques centaines de millisecondes, voire plus d'une seconde de parole. Ceci est illustré par exemple sur la Figure 3.2 (a). Dans la phrase en question, on a une longue section continûment voisée d'environ une seconde, depuis le [o] de « collision » jusqu'au « on » de « avion », et on a en même temps une section voisée relativement courte, d'environ 85 ms, limitée au seul phonème [ɛ] de « espion » (il y a une légère pose entre « avion » et « espion »). Pour notre base de données de 3500 sections, nous avons mesuré une longueur moyenne de 230 millisecondes pour les sections voisées extraites. Nous présentons ci-dessous deux exemples de résultat de modélisation, l'un concernant une section voisée plutôt courte (il s'agit de cette voyelle [ɛ] de « espion »), et l'autre concernant une section plutôt longue (il s'agit d'une autre portion de parole de la base de données).

Ainsi, sur la Figure 3.3, nous illustrons le résultat de la modélisation à long terme des amplitudes pour la voyelle [ɛ] de « espion » correspondant au signal de la Figure 3.2. Le modèle utilisé pour la modélisation à long terme est ici le MCD et le résultat présenté sur la figure est celui obtenu pour la quatrième harmonique (déjà représentée sur la Figure 3.1). Cette figure illustre la capacité du modèle à s'adapter globalement à la trajectoire d'amplitude du signal : la courbe correspondant au modèle suit la forme de trajectoire avec peu de coefficients, ici 6, par rapport au nombre de mesures $K = 22$.

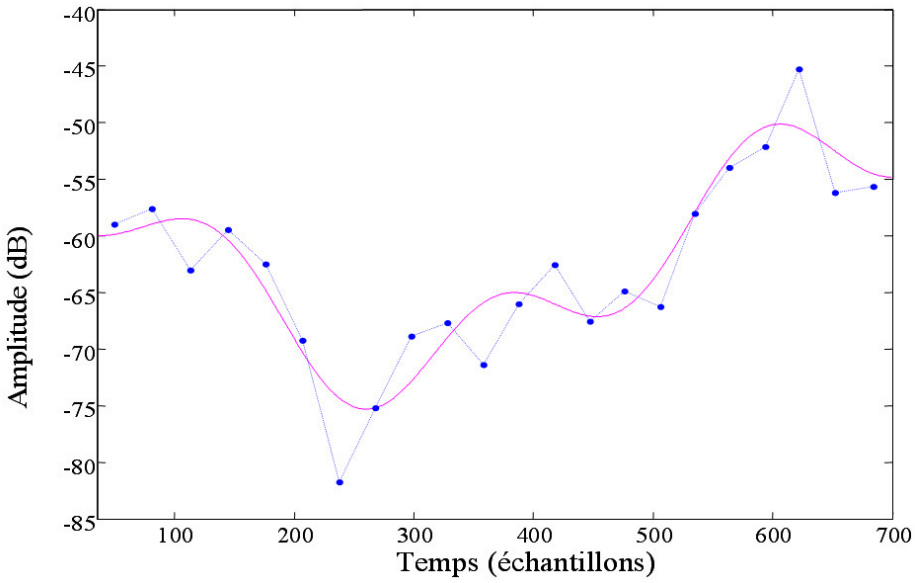

Figure 3.3: Modélisation à long terme de la quatrième trajectoire d'amplitude (en bleu) de la portion de signal correspondant à la Figure 3.2(b) (et aussi à la Figure 3.1). Le modèle à long terme (en magenta) est un MCD d'ordre 6.

La Figure 3.4 concerne une section significativement plus longue de parole voisée, comprenant plusieurs phonèmes, pour une voix féminine : il s'agit de la séquence « maminamidou ». La durée de cette section est d'environ 1,4 seconde, et le nombre de mesures d'amplitude est $K = 408$ (on rappelle qu'il s'agit du nombre de pseudo-périodes de la section considérée). Sur cette figure, nous avons tracé la trajectoire de l'amplitude de la première harmonique de cette longue séquence de parole. Pour cet exemple, dans la figure du haut, nous pouvons voir qu'à nouveau, le modèle MCD se caractérise par une trajectoire lisse autour des mesures d'amplitude. Nous avons tout d'abord voulu illustrer l'effet de la variation de l'ordre du modèle. Pour cela, nous présentons ici les résultats fournis par l'algorithme à la première itération de l'étape 4, c'est-à-dire sans itération sur les poids perceptuels de la matrice $\boldsymbol{W}_i$ (ces poids sont alors tous mis à 1).

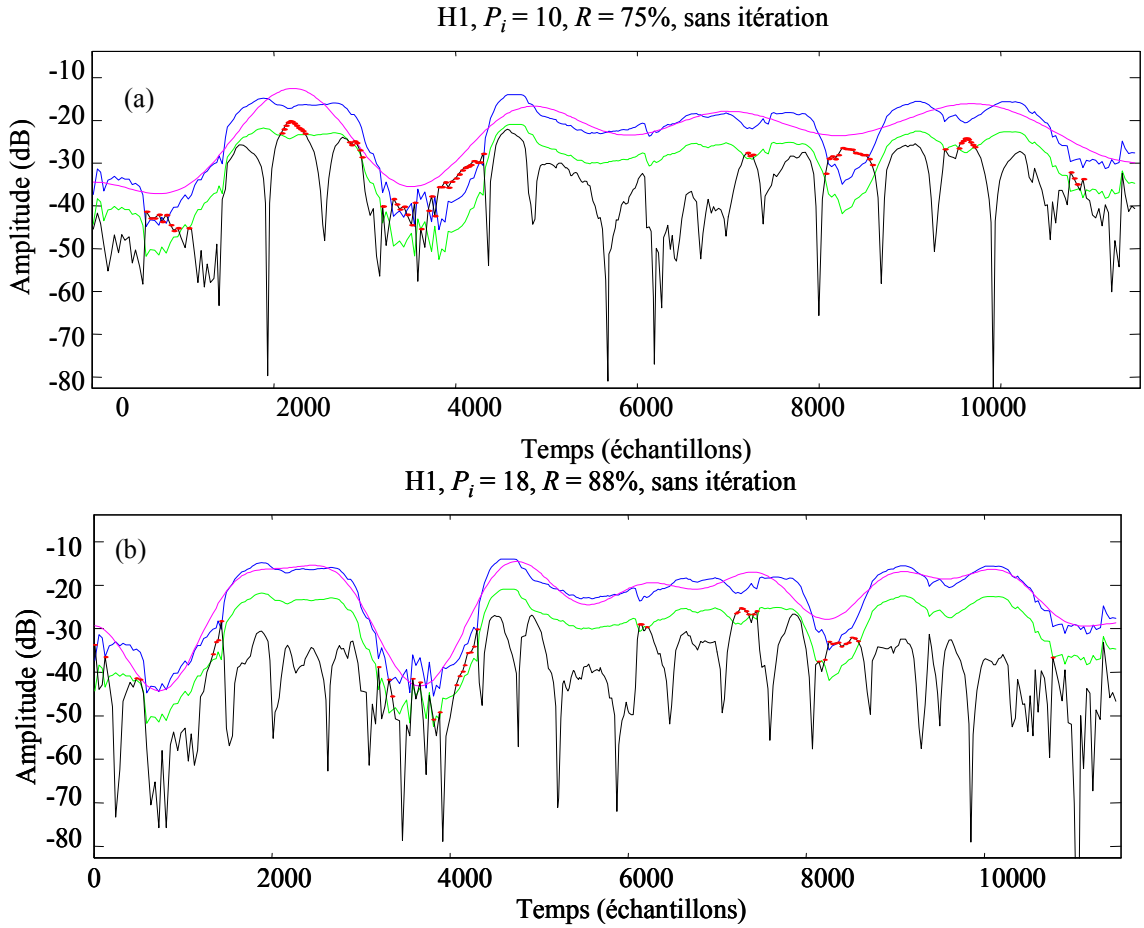

Figure 3.4 : Modélisation à long terme de l'amplitude de la première harmonique d'une longue section entièrement voisée de parole féminine (11630 échantillons à 8 kHz, $K = 408$). Les mesures d'amplitude originales (en bleu) et modélisées à long terme (en magenta), la trajectoire du seuil de masquage (en vert), celle de l'erreur de modélisation (en noir). Les valeurs d'erreur de modélisation dépassant la trajectoire du seuil de masquage sont indiquées en points rouges. Toutes les courbes sont sur une échelle de puissance en dB. (a) : on obtient R = 75% avec le modèle MCD à l'ordre 10. (b) : on a une amélioration du pourcentage R qui passe à 88% en augmentant l'ordre à 18. Dans les deux cas, il s'agit de la première itération de l'étape 4 de l'algorithme.

Pour un ordre du modèle fixé à 10 (figure du haut), le pourcentage R des points de la section modélisée qui vérifient la contrainte perceptive est d'environ 75%. Les points rouges correspondent aux points où l'erreur de modélisation dépasse le seuil de masquage. Dans la Figure 3.4(b), pour atteindre un pourcentage R plus élevé, toujours sans itération sur les poids perceptuels, l'algorithme d'ajustement a augmenté l'ordre du modèle. Sur cet exemple, l'ordre est passé à la valeur 18. On voit qu'avec cette nouvelle valeur, le modèle suit plus fidèlement le contour de la trajectoire d'amplitude. L'erreur de modélisation est alors significativement diminuée, et le pourcentage R est passé d'environ 75% à 88%.

Sur la Figure 3.5, nous avons illustré le comportement global de l'algorithme d'ajustement du modèle à long terme, ce dernier étant toujours le MCD sur cet exemple. Dans ce cas, on effectue les itérations sur la mise à jour des poids perceptuels correspondant à l'étape 5 de l'algorithme d'ajustement, en partant de la situation correspondant à la Figure 3.4(b) (c'est-à-dire un ordre 18 et $R = 88\%$). La Figure 3.5(a) montre que la mise à jour itérative de la pondération perceptuelle permet d'ajuster très correctement le modèle aux données tout en gardant le même ordre. Au bout de 15 itérations, le pourcentage de régions correctement modélisées (au sens du critère perceptif) est passé de 88% à 97%. Par exemple, la majeure partie des dépassements de seuil observés sans les itérations dans les régions situées un peu avant et un peu après l'indice 8000 ont été corrigés. Notons que ce résultat correspond au résultat final fourni par l'algorithme en le laissant s'exécuter sur cette section de signal à partir de son début, et en réglant $R_{min} = 95\%$ (les résultats présentés précédemment sont en fait des résultats intermédiaires de cette exécution). Entre la Figure 3.4(b) et la Figure 3.5(a), on remarquera la mise en forme de l'erreur de modélisation qui explique la réussite de l'ajustement du modèle : sur la première figure, la trajectoire de l'erreur suit les contours de la trajectoire du seuil perceptif de façon relativement grossière, en tout cas de façon beaucoup moins fine que sur la deuxième figure où « l'enveloppe supérieure » de l'erreur colle assez précisément à ce seuil.

Pour compléter l'illustration de ce résultat, et mettre en évidence l'aspect « parcimonieux » de l'algorithme, c'est-à-dire sa capacité à mettre en forme le modèle avec peu de coefficients, la Figure 3.5(b) présente les résultats obtenus avec un ordre du modèle fixé à 28 et à la première itération de l'étape 4 de l'algorithme. Cet ordre de 28 est la première valeur permettant d'obtenir un pourcentage R de 97% dans ces conditions (c'est-à-dire sans itération sur les poids perceptuels). Ainsi, pour une performance de $R = 97\%$, l'ordre optimal du modèle estimé à la fin de l'algorithme d'ajustement est donc de 18 au lieu de 28 grâce au critère perceptuel et à l'ajustement itératif des poids correspondants. Ceci démontre l'efficacité de cette approche et l'intérêt de poser un tel critère. On remarque pour finir que l'ordre optimal est relativement faible devant le nombre de mesures modélisées, qui, on le rappelle, est ici de 408. Nous reviendrons sur ce point plus tard, avec une étude quantitative plus précise sur l'ensemble de la base de données.

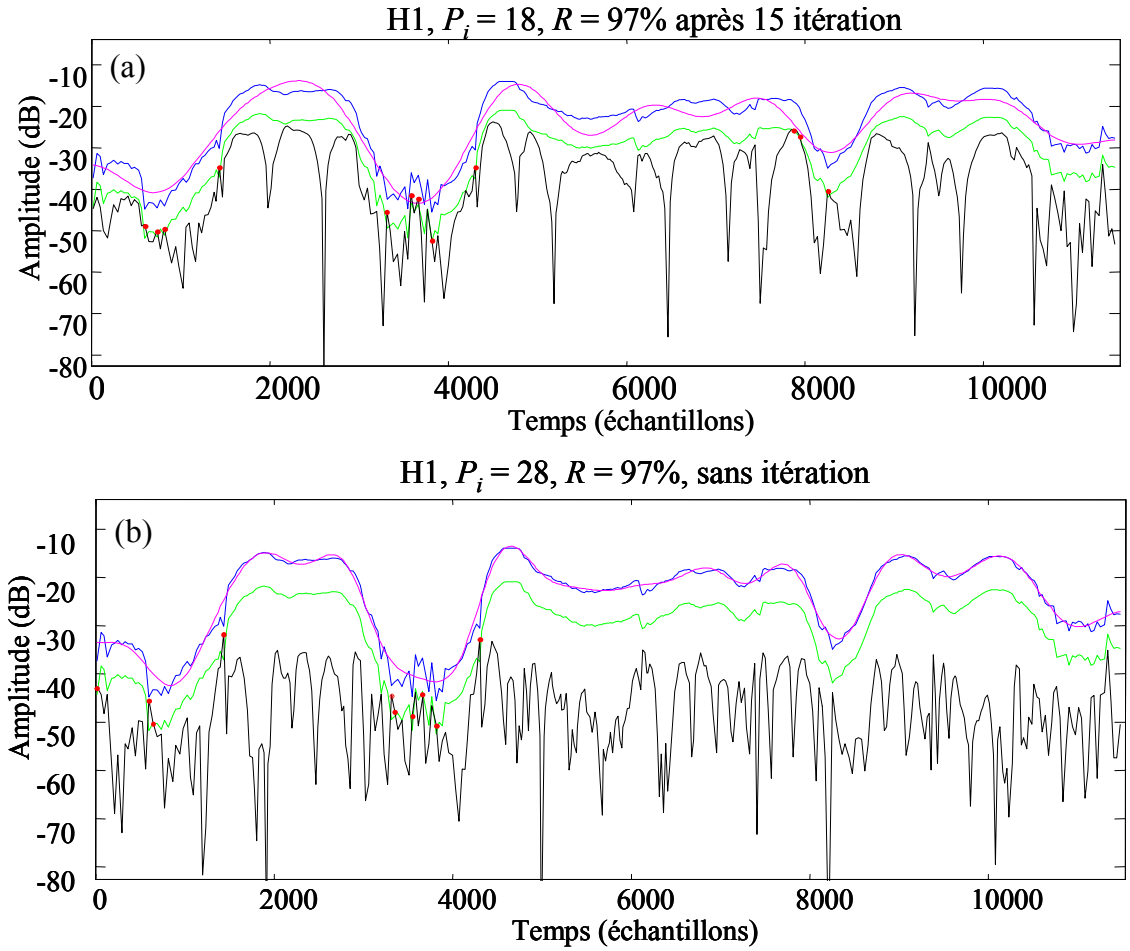

Figure 3.5 : Poursuite de l'expérience de la Figure 3.4 (voir la légende correspondante). (a) : amélioration du modèle MCD en gardant le même ordre qu'à la Figure 3.4(b) (P_i = 18) après 15 itérations de la mise à jour des poids perceptifs. Le pourcentage R est passé de 88% à 97%. (b) : Le même pourcentage R = 97% est obtenu sans effectuer les itérations sur les poids perceptifs, mais il faut alors augmenter l'ordre du modèle à la valeur 28. Les erreurs de modélisation résiduelles dépassant la trajectoire du seuil de masquage sont marquées par des points rouges.

3.3.2.2. Observations générales issues des expérimentations

A l'issue de l'observation du processus de modélisation à long terme des amplitudes sur de nombreuses sections de parole de notre base de données, plusieurs remarques peuvent être faites pour caractériser le comportement de notre algorithme, et en tirer des informations pour optimiser son usage.

D'abord, comme on l'a vu sur l'exemple de la sous-section précédente, le principe des itérations fondées sur les poids perceptifs s'avère efficace. Dans l'évolution de l'ajustement du modèle, pour un ordre de modèle fixé, on constate que les régions où l'erreur de modélisation est élevée sont progressivement améliorées d'une itération à l'autre. En contrepartie, la précision de modélisation des régions où cette erreur est faible peut être dégradée. Ceci s'explique par le fait que, pour un ordre fixé, le processus d'ajustement du modèle est un processus d'optimisation sous contrainte fixe.

Il en résulte généralement un bon ajustement « en moyenne » le long de la trajectoire des paramètres, et la précision de cet ajustement dépend de l'ordre du modèle. Si cet ordre est insuffisant, l'algorithme ne peut pas garantir que la contrainte sur R_{min} sera atteinte. Il peut alors en résulter un comportement « oscillant » de l'algorithme, au sens où il y a alternance entre régions bien et moins bien modélisées à chaque itération. Dans ce cas, l'algorithme finira par augmenter l'ordre du modèle. En conséquence, le nombre maximum d'itérations a été fixé arbitrairement à 20. Ce nombre paraît suffisant pour tester si l'ajustement sous la contrainte R_{min} peut être réalisé ou pas. En effet, lorsque l'ordre est suffisamment élevé pour permettre cet ajustement, celui-ci est réalisé plutôt rapidement : nous avons généralement constaté que moins de dix itérations du processus de pondération de l'algorithme d'adaptation étaient suffisantes pour réaliser l'ajustement optimal (généralement après une mise en forme globale du modèle réalisée en quelques itérations, il y a une évolution plutôt faible du modèle d'une itération à la suivante). De même, si l'ordre du modèle est insuffisant, après mise en forme globale rapide, on se dirige assez vite vers le phénomène d'oscillations locales mentionné ci-dessus.

Le deuxième point que nous avons mis en évidence est qu'il n'est généralement pas nécessaire de forcer l'erreur de modélisation à rester entièrement (c'est-à-dire tout le long de la trajectoire) sous le seuil perceptif (en fixant $R_{min} = 100\%$). En effet, dans la pratique, des rapports inférieurs peuvent fournir une synthèse de haute qualité. Autrement dit, le choix du rapport cible R_{min} affecte la qualité du signal modélisé, et il doit rester suffisamment proche de 100% mais pas forcément égal à 100%. On considère que si un pourcentage suffisamment élevé est atteint, alors l'erreur de modélisation sera globalement inaudible sur la section entière. Par exemple, des valeurs de l'ordre de $R_{min} = 90\%$ à 75% selon le rang harmonique sont adéquates. Nous détaillons plus longuement ce point dans la Section 3.3.3 du point de vue de la qualité des signaux synthétisés. Du point de vue du comportement de l'algorithme, ceci s'explique par le fait que, de façon tout à fait cohérente avec le point discuté dans le paragraphe précédent, des « efforts de modélisation » importants à une échelle très locale, c'est-à-dire sur des portions limitées de la trajectoire des paramètres, semblent inutiles. Ils peuvent même produire momentanément (pour un ordre fixé) un ajustement global moins bon, à moins de relâcher les contraintes en augmentant l'ordre du modèle. Or, il apparaît inutile d'augmenter l'ordre du modèle (et d'augmenter conjointement le coût de calcul) pour forcer les petites parties irrégulières de trajectoire à respecter rigoureusement le critère du seuil perceptif. Ceci est particulièrement vrai si on considère que de telles irrégularités locales peuvent résulter en partie des erreurs d'analyse. A titre d'exemple, on peut voir sur la Figure 3.5 que les erreurs résiduelles sont concentrées dans le segment 3000-4000 qui est une zone particulièrement irrégulière. Ces erreurs ne remettent pas en cause le bon ajustement global du modèle. Ainsi, il est judicieux de rappeler ici que le but de l'algorithme est d'obtenir un compromis optimal entre l'ordre du modèle (pas trop élevé) et le pourcentage R des points de la section modélisée qui vérifient la contrainte perceptive. Ce compromis semble donc atteignable, en relâchant un peu la contrainte sur R_{min}.

Le troisième point est qu'avec des valeurs de R_{min} comprises entre 90% et 75%, nous avons trouvé que dans la pratique l'algorithme converge toujours vers une valeur de l'ordre qui est significativement inférieure au nombre K de mesures. Il s'agit d'un résultat important dans l'optique de l'utilisation de la modélisation à long terme pour

des applications de codage, comme déjà mentionné à la Section 2.5.1. C'est pourquoi nous consacrons une sous-section complète à ce point précis, avec une étude quantitative en termes de débit de coefficients à la Section 3.3.4. Avant cela, nous précisons les relations entre modélisation à long terme et qualité des signaux synthétisés.

3.3.3. Tests d'écoute

Pour évaluer qualitativement les effets perceptuels de la modélisation à long terme des trajectoires d'amplitude, des essais d'écoute informelle ont été effectués sur les signaux de la base de données. Comme dans cette étude, le processus d'analyse-modélisation-synthèse concerne seulement les parties voisées de parole, les sections non voisées sont conservées telles quelles et sont concaténées avec les sections voisées modélisées avec une pondération locale pour éviter des artefacts audibles [George & Smith, 1997]. En d'autres termes, les sections voisées modélisées et les sections non voisées non modélisées sont raccordées par *Overlap-Add* local. Nous avons vérifié que ce processus n'entrainait pas de distorsions génantes dans les signaux synthétisés.

Les sections voisées modélisées ont été synthétisées en employant évidemment le modèle à long terme résultant de l'algorithme proposé sur les trajectoires d'amplitudes. Les mesures de phase sont quant à elles interpolées linéairement (après dépliement de ces mesures, comme dans [Girin *et al.,* 2003], voir aussi la Section 1.3.2.2.2). Ceci est justifié par le fait que nous voulons étudier séparément les effets de la modélisation à long terme des amplitudes et les effets de la modélisation à long terme des phases, ce dernier cas étant examiné dans le chapitre suivant. La synthèse à proprement parler de ces sections voisées de signal est ensuite réalisée en appliquant successivement les équations (1.3) et (1.2).

Le modèle à long terme utilisé dans ces tests est le modèle MCD. En outre, nous rappelons que ce modèle à long terme est appliqué seulement sur les dix premières harmoniques de chaque section voisée de parole traitée. Les amplitudes des autres harmoniques sont interpolées linéairement (voir Section 1.3.2.2.1). Pour comparer notre méthode avec une approche à court terme classique, nous avons également synthétisé des signaux de référence avec interpolation linéaire à court terme des mesures d'amplitude de toutes les harmoniques (et naturellement la même interpolation linéaire des mesures de phase).

Deux sujets avec audition normale ont attentivement et extensivement écouté une série de signaux synthétisés. Cette écoute s'est faite par l'intermédiaire d'une carte son de PC de haute qualité, et avec un casque fermé Sennheiser HD280, dans un environnement calme. Les principaux résultats de ces écoutes sont les suivants.

Tout d'abord, on peut dire que la différence perceptive entre chaque signal de test original et le signal de synthèse correspondant est très faible. Les signaux synthétisés avec la méthode à long terme sont donc d'une qualité très proche de celle des originaux (qui était elle-même très bonne, on le rappelle, compte tenu des très bonnes conditions d'enregistrement).

En second lieu, le résultat principal de ces essais est que la modélisation d'amplitude à long terme fournit généralement une qualité de synthèse identique à celle obtenue avec l'interpolation linéaire à court terme des amplitudes mesurées (qui est donc aussi de bonne qualité), pour une large gamme de configurations de test que nous allons préciser par la suite. En d'autres termes, les signaux synthétisés d'une part avec l'interpolation linéaire à court terme des amplitudes et d'autre part avec la modélisation à long terme de ces amplitudes ne peuvent généralement pas être distingués lorsque l'algorithme « a bien fonctionné », c'est-à-dire lorsqu'il a trouvé un bon compromis entre ordre du modèle et respect du critère perceptif. Plus précisément, la transparence perceptuelle entre le modèle long terme et le modèle court terme s'avère être garantie tant que la puissance de l'erreur de modélisation se trouve globalement au-dessous du seuil de masquage, même si elle surmonte « localement » (c'est-à-dire dans des régions relativement isolées de la trajectoire) ce seuil (voir la discussion à ce propos dans la section précédente) : une valeur de R_{min}=75% s'est ainsi avérée raisonnable pour la plupart des harmoniques pour garantir une qualité transparente comparée aux signaux modélisés à court terme. Seules les toutes premières harmoniques (généralement les deux premières) semblent nécessiter une précision un peu plus grande : dans ce cas, on peut préférer fixer R_{min} à 90% par exemple.

Ces observations suggèrent de conserver un peu de « flexibilité » dans l'estimation de l'ordre et l'adaptation du modèle d'amplitude à long terme, notamment en fonction du rang de l'harmonique modélisée. Cette flexibilité dépendante de la perception est illustrée sur la Figure 3.6 : la forme d'onde correspondant à la synthèse de l'harmonique considérée est notablement différente selon que la synthèse a été réalisée à court terme ou à long terme. Pourtant, les deux signaux de parole correspondants, c'est-à-dire intégrant cette harmonique et les neufs autres premières harmoniques modélisées soit à court terme, soit à long terme, sont perceptivement indistinguables. Il en est d'ailleurs de même à l'écoute de la seule harmonique de la Figure 3.6.

Ce potentiel de flexibilité de la modélisation à long terme pourrait être confirmé par une étude formelle complète de l'influence perceptive de la précision de modélisation pour chaque harmonique et chaque type de voix. Cependant, une telle étude semble très lourde à mettre en œuvre, du fait des multiples sources de variabilité des signaux de parole et des configurations possibles pour la modélisation à long terme étudiée. Compte tenu de cette difficulté, nous nous contenterons (au moins dans un premier temps) de ces résultats perceptifs assez généraux exploitables pour une première évaluation quantitative grossière de la méthode proposée en attendant de faire mieux.

Dans la section suivante, nous relions ainsi les observations que nous venons de faire dans cette présente section avec le point de vue de la compression de données : nous étudions l'impact de la modélisation à long terme en termes de débits de coefficients, en tenant compte de nos observations sur l'écoute des signaux.

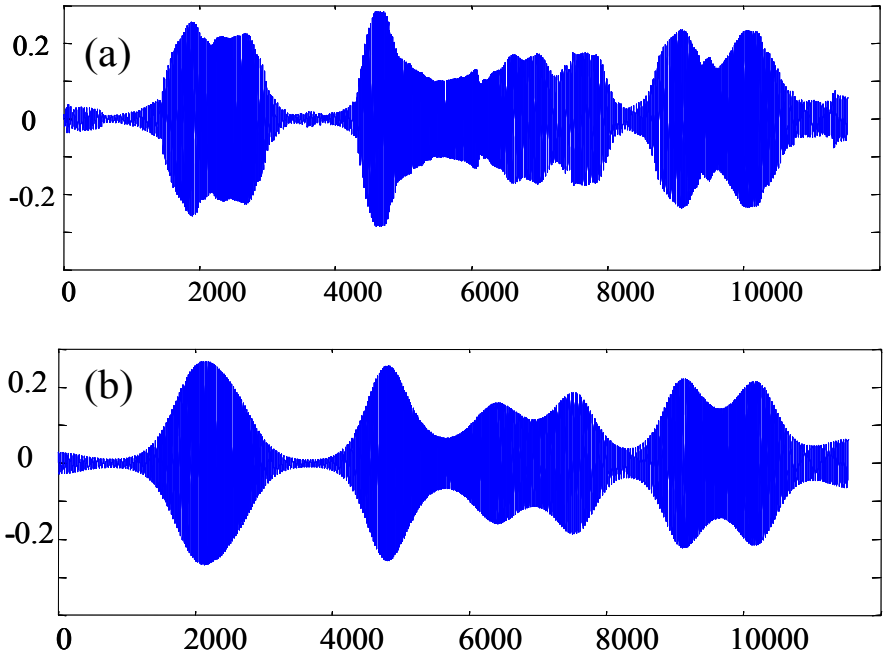

Figure 3.6 : Composante harmonique synthétisée correspondant à la Figure 3.4 (rappel : il s'agit de la première harmonique d'une longue section entièrement voisée de parole féminine ; 11630 échantillons à 8 kHz ; $K = 408$) ; la synthèse est réalisée avec interpolation linéaire à court terme des mesures d'amplitude (en haut) et avec modélisation à long terme de ces même mesures (avec un modèle MCD d'ordre 18) (en bas). Les mesures de phases sont linéairement interpolées dans les deux cas.

3.3.4. Débits de coefficients

Comme nous l'avons déjà mentionné rapidement précédemment, avec des valeurs de R_{min} comprises entre 90% et 75%, nous avons observé que dans la pratique l'algorithme converge généralement vers une valeur d'ordre qui est significativement inférieure au nombre K de mesures. Il s'agit d'un résultat important dans l'optique de l'utilisation de la modélisation à long terme pour des applications de codage, comme déjà mentionné à la Section 2.5.1. Cela illustre la capacité de la méthode proposée à permettre intrinsèquement la compression de données par la réduction efficace de dimension. Dans cette section, nous effectuons une étude quantitative plus précise de ce point.

Pour cela, nous avons d'abord calculé R_{moy} la moyenne du rapport K/P_i entre le nombre de mesures (où le nombre de pseudo-périodes) de la section de parole modélisée et l'ordre du modèle long terme correspondant, pour l'ensemble des 3500 sections de notre base de données, et ceci pour les dix premières harmoniques. Les résultats sont représentés sur la Figure 3.7, avec les écarts-types correspondants. On peut voir que le rapport K/P_i prend des valeurs très importantes pour les harmoniques de rang faible, qui ont une influence perceptive importante : on a un rapport moyen proche de 18 pour la première harmonique, proche de 10 pour la deuxième, et proche de 8 pour la troisième. Ensuite les valeurs diminuent encore un peu pour se stabiliser autour de 5. On constate donc que la modélisation est plus efficace du point de vue de ce rapport pour les harmoniques de rang faible par rapport aux harmoniques plus hautes en fréquence (on peut même souligner le fait que pour les premières harmoniques, elle est très efficace !)

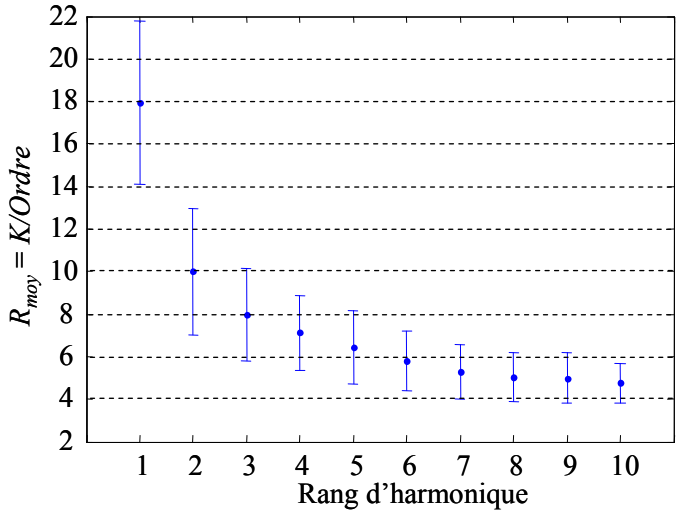

Figure 3.7 : Moyenne et écart-type du rapport K/P_i sur l'ensemble de la base de données (3500 sections voisées de parole) pour les dix premières trajectoires d'amplitude. Le modèle utilisé est le MCD avec $R_{min} = 90\%$.

Ceci peut s'expliquer par le fait que les trajectoires d'amplitudes sont de plus en plus irrégulières, c'est-à-dire bruitées, au fur et à mesure qu'on monte en fréquence. Cette observation est liée à la nature du signal de parole : sur les portions de parole voisées, le spectre est généralement à pente négative, et les composantes de bruit potentielles sont généralement plus fortes en montant dans les fréquences. De plus, l'analyse des paramètres est plus difficile et donc moins précise sur les composantes de faible intensité. Au bruit « naturel » du signal de parole peut donc se superposer un bruit de mesure, les deux sources de bruit étant de plus en plus influentes en montant dans les harmoniques. Le fait que les trajectoires des composantes soient donc de plus en plus irrégulières en montant dans les fréquences explique que les modèles nécessitent des ordres de plus en plus élevés pour respecter leur ajustement aux données.

On peut alors se poser la question de savoir si une telle précision dans la modélisation est nécessaire pour les harmoniques élevés. On donnera quelques éléments de réponse à cette question au Chapitre 6 de ce document. Indépendamment de cette question, une des conséquences pour notre algorithme de cette série d'observations est qu'on pourrait adapter les contraintes sur l'intervalle de recherche de l'ordre optimal pour accélérer le processus, et ceci pour chaque harmonique : on peut prendre par exemple comme valeur initiale de l'ordre la partie entière de K/R_{moy} à la place de celle de $K/2$ (et bien sûr ajuster la valeur initiale de la mise à jour δP_i en conséquence). Alternativement, d'une façon plus générale, la limite supérieure de l'ordre optimal pourrait globalement être choisie à $K/4$. En effet, sur la Figure 3.7, on voit que quasiment toutes les valeurs du ratio K/P_i sont inférieures à 4 à un écart-type près. Ceci est équivalent à dire que quasiment toutes les valeurs finales de l'ordre du modèle sont inférieures à $K/4$. De tels réglages de l'algorithme peuvent contribuer à accélérer significativement sa convergence.

Pour nous rapprocher d'une optique « codage », nous allons maintenant considérer la longueur des sections modélisées en terme d'unité de temps et non plus en terme de nombre de trames. Rappelons en effet qu'une trame est une pseudo-période de signal, de taille variable, et généralement plus courte que les valeurs de décalage des fenêtres

d'analyse-synthèse dans les codeurs usuels : quelques millisecondes dans le premier cas contre environ 20 millisecondes dans le second. Nous allons tester si la modélisation à long terme reste compétitive du point de vue compression de données, en terme de débit de coefficients, et non plus en terme de rapport K/P_i puisque celui-ci dépend de la méthode d'analyse période-synchrone et est en quelque sorte « surestimé » par rapport à l'inverse du débit réel. Toujours dans l'optique d'affiner les résultats, nous avons séparé ici les résultats obtenus pour des voix féminines et ceux obtenus pour des voix masculines, puisque les deux types de voix possèdent des caractéristiques spécifiques, particulièrement des gammes différentes de fréquence fondamentale. Ce dernier facteur a une forte influence sur un processus de type modélisation sinusoïdale.

Ainsi, la Figure 3.8 montre la valeur de l'ordre du modèle à long terme en fonction de la longueur de la section de parole modélisée (en seconde), pour chaque section de notre base de données, pour les harmoniques 2 et 5, et pour $R_{min} = 75\%$. Ces courbes illustrent bien la diversité des configurations de modélisation. Par exemple, des sections longues de parole peuvent être modélisées avec des ordres comparativement très petits (partie inférieure droite des tracés). Ce type de résultat révèle le fort potentiel de la méthode proposée. A l'inverse, des sections courtes peuvent nécessiter des ordres significativement plus élevés (partie supérieure gauche des tracés), ce qui est plus délicat du point de vue du débit. On note par ailleurs que les valeurs d'ordre pour l'harmonique 2 sont plus étendues et décalées vers des ordres inférieurs par rapport aux valeurs pour l'harmonique 5. Ceci confirme les résultats de la Figure 3.7 : la modélisation à long terme semble d'autant plus efficace que le rang de l'harmonique est faible.

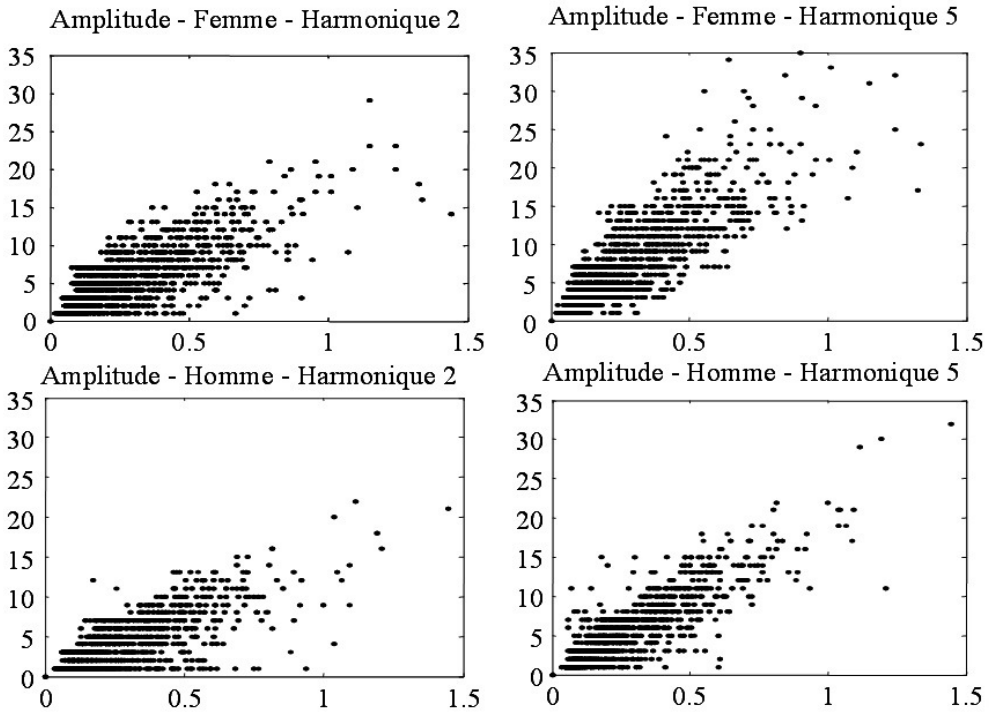

Figure 3.8 : Modélisation à long terme des trajectoires d'amplitude : ordre final du modèle en fonction de la longueur de la section de parole modélisée (en seconde) pour les 3500 sections voisées du corpus. Le modèle est le MCD et on a fixé $R_{min} = 75\%$. En haut : voix de femmes, en bas : voix d'hommes, à gauche : harmonique 2, à droite : harmonique 5.

Afin d'obtenir une vision moyenne des résultats, les valeurs d'ordre (plus un) pour toutes les sections de la base de données ont été additionnées et divisées par la longueur totale cumulée de ces sections. On obtient ainsi un débit moyen de coefficients. Ceci a été fait pour chaque harmonique du rang 1 au rang 10 et pour les deux valeurs préalablement choisies de R_{min} (75% et 90%). De plus, pour ce calcul, nous avons encore séparé les voix de femmes et les voix d'hommes. Les résultats sont recueillis dans le Tableau 3.1. Plusieurs remarques peuvent être faites à partir de l'observation de ces résultats :

- Tout d'abord, la valeur moyenne du débit (ou de l'ordre) augmente avec le rang de l'harmonique. Ceci confirme les résultats des Figures 3.7 et 3.8. Notons que les débits de coefficients sont particulièrement bas pour la première harmonique (moins de 10 coefficients par seconde pour trois conditions sur quatre), qui semble donc particulièrement « facile » à modéliser. Ces débits augmentent ensuite relativement rapidement avec les harmoniques avant de se stabiliser, tout en restant à des valeurs significativement inférieures à 50. Cette dernière valeur peut en effet être utilisée comme référence dans la mesure où elle correspond à un décalage de fenêtre de 20 ms usuellement utilisé dans les systèmes d'analyse-synthèse et de codage à court terme (voir plus loin pour une comparaison en terme de débit moyen entre harmoniques). Ces résultats suggèrent de nouveau, après les remarques de la Section 3.3.5.4, que des contraintes additionnelles sur l'estimation de l'ordre (par exemple, la valeur de R_{min}) devraient être adaptées au rang de chaque harmonique.
- Deuxièmement, les débits moyens de coefficients sont plus hauts et plus étendus pour la parole de femme que pour la parole d'homme. Ceci est cohérent avec la remarque faite plus haut sur la difficulté de modélisation croissante en montant dans les harmoniques, puisque nous devons tenir compte de la différence dans les écarts de fréquence fondamentale pour les voix de femmes et les voix d'hommes. Cette différence explique également que les débits moyens augmentent plus rapidement avec le rang des harmoniques pour la parole féminine que pour la parole masculine. Ainsi, il est plus coûteux de modéliser une harmonique de voix de femme qu'une harmonique de voix d'homme du même rang. Cependant, les voix de femme peuvent requérir globalement moins de coefficients puisqu'elles contiennent moins d'harmoniques. Cette observation est cohérente avec les résultats bien connus sur la dépendance au genre de l'efficacité du codage sinusoïdal de la parole.
- Troisièmement, les taux de coefficients sont plus hauts pour $R_{min} = 90\%$ que pour $R_{min} = 75\%$. Cela est évident puisque l'estimation de l'ordre dépend directement de ce paramètre dans l'algorithme d'ajustement : R_{min} détermine la précision du modèle et donc son nombre de paramètres.

Faisons maintenant la moyenne du débit des coefficients de façon globale sur les dix premières harmoniques. Suite aux observations de la Section 3.3.2.2, on prend ici $R_{min} = 90\%$ pour les harmoniques 1 et 2 et $R_{min} = 75\%$ pour les harmoniques 3 à 10. Nous obtenons alors une valeur moyenne globale de 30,3 coefficients/s par harmonique pour les voix de femmes, et 21,7 coefficients/s par harmonique pour les voix d'hommes. Par conséquent, en considérant une situation générale avec autant de parole féminine que de parole masculine, le débit moyen de coefficients, sans distinction de genre, et par harmonique serait de 26 coefficients/s/harmonique. Comparativement, le nombre

moyen des trames d'analyse à court terme par seconde (qui est également la valeur moyenne de la fréquence fondamentale puisque l'analyse était période-synchrone) est d'environ de 220 pour la parole féminine et de 140 pour la parole masculine. Ainsi, le processus de modélisation à long terme permet de diviser le nombre de paramètres par un facteur minimum de l'ordre de 7 (au moins pour les dix premiers harmoniques), comparé au synthétiseur à court terme en utilisant les amplitudes mesurées de façon période-synchrone, tout en fournissant la même qualité subjective globale (voir la Section 3.3.3). Toutefois, comme déjà mentionné précédemment, il semble plus juste de comparer les débits de coefficients avec ceux des systèmes d'analyse-synthèse et des codeurs de parole à court terme qui utilisent une fenêtre avec un décalage de taille fixe. Cette taille est habituellement dans l'intervalle 10-20 ms (plutôt 20 ms pour les codeurs bas débit), ce qui correspond à 50-100 coefficients/s pour chaque type de coefficient à transmettre. Dans ce cas, *on peut conclure qu'avec la modélisation à long terme des trajectoires d'amplitude, le débit de coefficients est divisé par un facteur de l'ordre de 2 à 4 par rapport à la modélisation à court terme*. Notons de plus que pour une application telle que le codage de parole à très bas débit, voire ultra bas débit, les ordres de modèles peuvent être encore sensiblement diminués pour diminuer significativement le débit de coefficients tout en préservant une qualité de synthèse acceptable. Nous verrons au Chapitre 5 une extension de notre méthode mieux adaptée à la problématique du codage à très bas débit.

Harmonique		1	2	3	4	5	6	7	8	9	10
Voix de femmes	$R_{min} = 75\%$	7,7	20,9	21,7	23,0	27,7	33,0	36,1	37,1	37,6	39,9
Voix d'hommes	$R_{min} = 75\%$	5,4	15,1	19,0	19,9	20,6	27,7	23,5	24,5	25,2	26,0
Voix de femmes	$R_{min} = 90\%$	16,5	29,9	32,9	38,4	43,0	45,1	45,8	45,7	46,0	46,2
Voix d'hommes	$R_{min} = 90\%$	9,8	20,8	23,8	25,2	26,4	27,2	27,6	27,9	28,0	28,3

Tableau 3.1 : Résultats de la modélisation à long terme des trajectoires d'amplitude en termes de débit de coefficients (nombre de coefficients par seconde par harmonique).

3.3.5. Comparaison des différents modèles

Toutes les expérimentations reportées jusqu'ici ont été menées en utilisant le modèle dit MCD de la Section 2.4.2 comme modèle à long terme. Dans la Section 2.4, nous avons cependant proposé plusieurs modèles possibles pour remplir cette tâche. Dans cette section, nous faisons une étude comparative des performances de ces modèles pour déterminer notamment si l'un d'entre eux « sort du lot », ou plus généralement pour mettre en évidence d'éventuelles propriétés spécifiques. Rappelons brièvement que, outre le MCD, ces différents modèles sont le modèle polynomial (MP), le modèle formé de combinaisons de cosinus discrets avec les fonctions sinus correspondantes (MCSD), ou avec des polynômes (MCDP) (voir la Section 2.4). Précisons aussi que dans cette série d'expérimentations, seul le modèle est interchangé dans l'algorithme d'ajustement, c'est-à-dire concrètement la matrice $\boldsymbol{M}_i$ (voir la Section 2.5.2 pour le contenu de chaque matrice en fonction du modèle). Tous les autres réglages de l'algorithme sont effectués de la même façon pour chaque test comparatif (par exemple on teste les différents modèles pour une certaine valeur identique de R_{min}).

3.3.5.1. Configuration des modèles hybrides

Pour les deux derniers modèles, MCDP et MCSD, la configuration utilisée dans ces expérimentations est la suivante. Lorsque l'ordre du modèle est pair, le modèle est composé d'autant de fonction cosinus (de rang strictement supérieur à 0)[60] que de fonctions sinus correspondante (pour le MCSD) ou de termes polynomiaux (pour le MCDP). Lorsque l'ordre est impair, on a une fonction cosinus de plus que les fonctions sinus ou polynomiales. Pour le MCDP, les termes polynomiaux sont ceux d'ordre le plus faible, c'est-à-dire qu'on part de l'ordre 1 (terme en n) jusqu'à $e[P_i/2]$, où $e[.]$ désigne la partie entière. Par exemple, un modèle MCDP d'ordre 4 comporte la constante 1, deux cosinus de rang 1 et 2 et deux termes polynomiaux linéaire et quadratique. A l'ordre 5, ce modèle comporte un cosinus de rang 3 supplémentaire.

3.3.5.2. Critères de performance

L'évaluation comparative des performances nécessite la définition d'un certain nombre de critères. Il s'est avéré en effet qu'un des résultats principaux de cette campagne d'évaluation comparative est que les différents modèles fournissent des résultats similaires en terme de comportement général de l'algorithme (on retrouve les grands traits décrits à la Section 3.3.2 dans le cadre du modèle MCD). De même, on retrouve cette similarité au niveau de la qualité de synthèse : pour une valeur fixée de R_{min}, les différents modèles fournissent une qualité de signal généralement comparable, du moment que l'algorithme a correctement convergé, c'est-à-dire que l'ordre du modèle obtenu est relativement faible devant le nombre K de mesures d'amplitude modélisées[61]. En revanche, d'une part les ordres obtenus pour les différents modèles peuvent être différents, et d'autre part les différents modèles se sont avérés différemment robustes par rapport à des problèmes de type numérique (dus au mauvais conditionnement de la matrice à inverser dans l'équation (2.19)). C'est pourquoi nous axons maintenant cette campagne d'évaluation comparative principalement sur plusieurs critères de type « parcimonie » d'une part et « efficacité numérique » d'autre part.

Ainsi, nous reprenons d'abord le critère de débit de coefficients tel que défini dans la section précédente. Pour que la comparaison entre les différents modèles soit juste, nous devons utiliser le même « matériau » de test pour les quatre modèles. C'est pourquoi nous utilisons pour cette comparaison uniquement les sections de signal telles que le pourcentage cible R_{min} a été atteint pour les 10 premières harmoniques par les quatre modèles mis en jeu, sans occurrence de problème numérique. De plus, nous rajoutons arbitrairement la contrainte suivante : le pourcentage cible R_{min} doit être atteint avec $P_i < K/3$. Ceci permet de tester comparativement les débits dans une gamme où la modélisation à long terme démontre son efficacité (puisque le nombre de coefficients est assuré d'être relativement faible devant celui des données). Ces contraintes sont assez sévères (toutes les harmoniques de chaque section testée doivent avoir été correctement modélisées par tous les modèles, et ceci avec relativement peu de

[60] Ainsi, la constante $1 = \cos(0 \times 2\pi n/N)$ n'est pas vraiment considérée comme faisant partie des fonctions cosinus. Par exemple, pour le modèle MCDP, elle peut être considérée aussi bien comme faisant partie des cosinus que des polynômes. Ceci dit, elle est toujours présente dans le modèle quelle que soit la configuration, car dans la pratique elle est fondamentale pour représenter la moyenne de la trajectoire.

[61] Dans ce cas, la qualité est par conséquent similaire à celle de la synthèse à court terme, puisque c'était le cas pour le modèle MCD.

coefficients). On recense ainsi environ 1550 sections vérifiant ces conditions sur les 3500 de notre base de test pour R_{min} = 75% (soit environ 45% des sections).

En ce qui concerne les critères « numériques », nous voulons comptabiliser pour chaque modèle les occurrences où le processus de modélisation à long terme a été mis en défaut par des problèmes numériques. En effet, dans l'algorithme d'ajustement, il est bon de préciser que la matrice à inverser dans l'équation (2.19) ne doit pas être mal conditionnée. Comme nous gardons dans ces tests la limite à $P_i < K/3$, nous avons trois possibilités de résultats pour chaque harmonique modélisée. En effet, d'une part soit la modélisation est réussie sous cette contrainte, c'est-à-dire qu'on a bien $R \geq R_{min}$ avec $P_i < K/3$, soit elle n'est pas réussie, c'est-à-dire que R n'atteint pas le pourcentage cible R_{min}. D'autre part, dans ce dernier cas, il y a deux possibilités pour expliquer l'échec : soit l'augmentation de l'ordre dans l'algorithme d'ajustement est bloquée par cette limite de $K/3$ (sans connaître de problème numérique : comme avec cette limitation de l'ordre, rien ne garantit que la modélisation permette d'atteindre R_{min} à tout coup, on peut dire que dans ce cas, le modèle n'est simplement pas assez efficace), soit cette augmentation de l'ordre est bloquée par un problème numérique de mauvais conditionnement de l'ajustement du modèle[62] (pour un ordre qui reste inférieur à $K/3$).

Pour quantifier ces occurrences, nous définissons pour chaque modèle trois pourcentages correspondants à ces trois cas de figures : P_{OK} est le pourcentage des harmoniques correctement modélisées avec $P_i < K/3$, P_{NOK} est le pourcentage des harmoniques non correctement modélisées avec $P_i < K/3$ mais sans problème numérique (la limite de $K/3$ est donc atteinte pour ces harmoniques), et P_{PB} est le pourcentage des harmoniques non correctement modélisées avec $P_i < K/3$ à cause d'un problème numérique. Notons qu'on a bien : $P_{OK} + P_{NOK} + P_{PB} = 100\%$. Les valeurs de P_{OK} et P_{NOK} peuvent être considérées comme une mesure comparative (alternative au débit de coefficients) de la capacité des modèles à fournir une représentation parcimonieuse des données, et la valeur P_{PB} peut être plutôt considérée comme une mesure quantitative de leur « souplesse / robustesse » du point de vue du calcul numérique. Notons enfin pour être tout à fait complets que ces calculs de pourcentages sont effectués sur l'ensemble des harmoniques modélisées de tous les signaux de la base de test, et non sur une sélection de sections de parole comme c'est le cas pour les débits. En effet, pour ce calcul de pourcentages, on n'a pas de contrainte de bon comportement des modèles sur un « matériau commun », étant donné qu'on cherche précisément ici à quantifier le comportement de ces modèles[63]. On utilise ainsi pour ces calculs de pourcentages les 35000 (3500×10) harmoniques de la base de test.

[62] Notons que d'une façon générale, dans l'utilisation pratique de l'algorithme, lorsqu'un problème de conditionnement numérique survient, on peut contraindre l'ordre du modèle à être diminué jusqu'à avoir un conditionnement correct, quitte à ce que R_{min} ne puisse jamais être atteint. En d'autre termes, si il s'avère que R_{min} n'est pas atteint du fait de cette contrainte de bon conditionnement numérique, on peut sélectionner la valeur maximum de l'ordre du modèle correspondant au meilleur ratio R possible et garantissant une matrice bien conditionnée pour l'équation (2.19).

[63] En revanche, on peut noter que pour mettre les modèles sur un pied d'égalité dans cette comparaison, on utilise la forme brute du MCD, c'est-à-dire sans la régularisation de l'équation (2.22).

3.3.5.3.Stratégie optimale : approche « multi-modèles » à meilleur choix

Parallèlement au test comparatif des différents modèles, nous proposons de tester aussi une approche optimale en terme de débit de coefficients. Le principe de cette stratégie optimale est très simple : pour chaque section de parole modélisée, nous retenons le modèle qui a nécessité le moins de coefficients en moyenne sur les 10 premières harmoniques (tout en ayant vérifié les contraintes du critère perceptif). Outre les possibilités de l'utiliser dans le cadre du codage à très bas débit[64], l'intérêt est de cette démarche est de nous fournir des résultats de référence pour les autres modèles, en quelque sorte une borne supérieure de performance possible. Comme pour les autres tests, nous moyennons les performances de cette approche sur l'ensemble des sections de la base de données utilisées pour comparer les différents modèles. Nous dénommons par la suite par le terme *multi-modèles* (MM) cette technique.

3.3.5.4. Résultats

Les résultats en terme de débit de cette étude comparative des différents modèles à long terme pour la modélisation des trajectoires d'amplitude sont présentés sur la Figure 3.9 et dans le Tableau 3.2. Sur la Figure 3.9, on a ainsi présenté les débits moyens de coefficients obtenus pour les différents modèles à long terme et pour les dix premières harmoniques des signaux. Le Tableau 3.2 reprend les résultats de la Figure 3.9 en les moyennant sur les dix harmoniques. Notons que dans ces expériences, nous mélangeons les voix de femmes et les voix d'hommes. De plus, le rapport cible R_{min} est le même pour toutes les harmoniques[65], et nous avons sélectionné la valeur « habituelle » $R_{min} = 75\%$. Les débits sont donc moyennés sur les 1550 sections de parole voisées sélectionnées pour cette valeur de R_{min} (voir la Section 3.3.5.2).

Tout d'abord, nous pouvons voir sur la Figure 3.9 que pour l'ensemble des modèles, le débit moyen de coefficients augmente avec le rang des harmoniques. Ceci est cohérent avec les observations que nous avons faites dans les sections précédentes. Rappelons que ce résultat s'explique en grande partie par la complexité croissante des trajectoires d'amplitude lorsqu'on monte dans les fréquences. Ce principe étant acquis, on peut dire que les débits obtenus avec les quatre modèles de base, MP, MCD, MCSD et MCDP, sont assez proches. On retrouve les ordres de grandeur obtenus pour le MCD à la Section 3.3.4, même si les valeurs sont ici un peu plus faibles, de l'ordre de 20 coefficients/s/harmonique en moyenne pour $R_{min} = 75\%$ (ceci s'explique par le fait

[64] Dans le cadre du codage, cette stratégie est séduisante, mais elle requiert cependant de transmettre des bits complémentaires au décodeur pour coder le type de modèle choisi pour chaque section. Cependant, puisque le nombre moyen de sections voisées par seconde est de 4.3 sur notre base de données, et que deux bits sont nécessaires pour coder l'information sur le type de modèle (nous avons quatre modèles différents), le débit binaire additionnel est très bas (moins de 10 bits/s). La question reste de savoir si en moyenne, cet ordre de grandeur est inférieur au nombre de bits sauvés correspondant au gain sur les coefficients fourni par le choix du modèle optimal. Ce point demande à être confirmé par une étude sur la quantification des coefficients du modèle et cette étude dépasse le cadre de cette thèse.

[65] Ceci est justifié par le fait que dans ces expérimentations précises, on cherche surtout à comparer les différents modèles entre eux, avec les mêmes conditions expérimentales, plutôt que de chercher un réglage optimal des modèles pour chaque harmonique, et/ou chaque type de voix. Le fait que R_{min} soit ici le même pour toutes les harmoniques, ajouté au fait qu'on utilise une sélection de sections de la base de test, explique que les résultats concernant le modèle MCD soient différents de la moyenne inter-genres de ceux présentés à la Section 3.3.4 où on avait $R_{min} = 90\%$ pour les deux premières harmoniques et $R_{min} = 75\%$ pour les autres, et où on utilisait les 3500 sections de la base de test.

qu'on utilise ici les sections sélectionnées avec la contrainte $P_i<K/3$; voir aussi la note de bas de page n°65). Il n'y a pas de modèle franchement supérieur aux autres en terme de débit, même si une hiérarchie peut être instaurée selon l'ordre suivant : pour le réglage R_{min} = 75%, le modèle MCD est le plus performant, suivi du MCSD qui est très proche du MCDP, le dernier étant le MP. Cet ordre est confirmé en moyennant les résultats sur les différents harmoniques, comme le montre le Tableau 3.2. Le MCD permet un débit moyen de 18,5 coefficients par seconde et par harmonique, contre 19,9 pour le MCSD, et 20,3 et 20,4 pour respectivement le MCDP et le MP. Par comparaison, le modèle « optimal » MM a un débit de 17,9 coefficients/s/harmonique. D'une manière générale, le MCD semble donc meilleur que ses concurrents en terme de débit, et il semble difficile de discriminer les autres modèles entre eux sur la seule base du débit. Notons que la stratégie du multi-modèle permet d'économiser environ 3% du débit de coefficients, comparé au MCD. Ceci semble *a priori* très faible et met en question l'utilité de cette approche qui complexifie la méthode pour peu de gain. Cependant, encore une fois, il faudrait vérifier si cette stratégie est payante ou non dans le cadre du codage, en prenant en compte le coût additionnel de la transmission de l'information relative au choix du modèle.

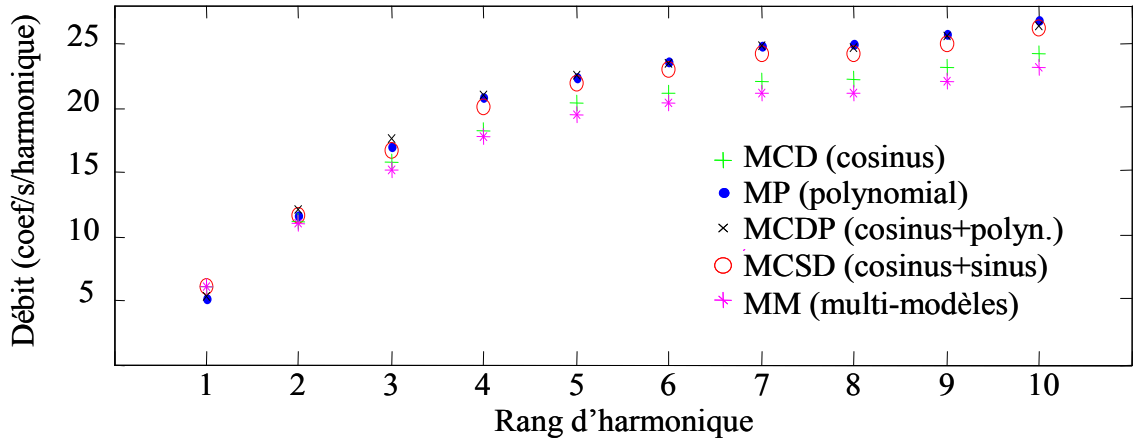

Figure 3.9 : Résultats comparatifs de la modélisation à long terme des trajectoires d'amplitude des dix premières harmoniques par les différents modèles proposés (MCD, MP, MCDP, MCSD et MM) : Débit moyen de coefficients (en nombre de coefficients par seconde) pour chaque harmonique ; on a ici R_{min} = 75%. Les résultats sont moyennés sur l'ensemble des 1550 sections sélectionnées de la base de test.

Modèle	MP	MCD	MCSD	MCPD	MM
Débit	20,4	18,5	19,9	20,3	17,9
P_{OK}	86,5%	94,1%	91,9%	87,9%	*
P_{NOK}	7,3%	5,9%	7,5%	7,3%	*
P_{PB}	6,2%	0%	0,6%	4,8%	*

Tableau 3.2 : Résultats comparatifs de la modélisation à long terme des trajectoires d'amplitude par les différents modèles proposés (MCD, MP, MCDP, MCSD et MM) : Débit moyen de coefficients (en nombre de coefficients par seconde) moyennés sur les dix premières harmoniques des 1550 sections sélectionnées de la base de test. Pourcentages P_{OK}, P_{NOK} et P_{PB} calculés sur l'ensemble des harmoniques des 3500 sections de la base (voir le texte). On a R_{min} = 75%.

En ce qui concerne les critères de robustesse numérique, nous présentons sur la Figure 3.10 les pourcentages moyens P_{OK}, P_{NOK} et P_{PB}, obtenus pour les différents modèles et pour chaque harmonique (pour $R_{min} = 75\%$). D'une façon générale, ces résultats sont plus contrastés que les résultats en terme de débits. Ainsi, comme on peut le voir sur la Figure 3.10(a), les valeurs de P_{OK} pour le modèle MCD surpassent assez nettement les autres modèles, notamment les modèles contenant des termes polynomiaux (MP et MCDP). Partant d'un score comparable et proche de 100% pour les deux premières harmoniques (ce qui confirme l'efficacité de la modélisation à long terme sur ces trajectoires particulièrement régulières), les écarts se creusent en montant vers les harmoniques plus élevées. Ainsi P_{OK} se maintient à environ 90% pour les harmoniques 8, 9 et 10 pour le MCD, alors que le MP chute entre environ 79% et 76%. Comme pour les débits, le Tableau 3.2 résume ces résultats avec un moyennage entre les harmoniques. On retrouve les grandes tendances de la Figure 3.10, lissées par le moyennage. Ainsi, le pourcentage P_{OK} moyen sur les 10 harmoniques obtenu par le MCD pour $R_{min} = 75\%$ est d'environ 94% alors que le pourcentage moyen obtenu par le modèle MP est de 86,5%. D'une façon générale, que ce soit pour chaque harmonique ou en moyenne, le modèle MCSD se situe un peu en dessous du MCD. Après les scores de débit, ceci semble confirmer la moindre efficacité du mélange de cosinus et sinus par rapport aux cosinus seuls (à nombre de termes identiques). Enfin, les termes polynomiaux semblent pénaliser le modèle MCDP dont les valeurs de P_{OK} se situent juste au dessus de celles du MP.

Comme expliqué à la Section 3.3.5.2, les pourcentages P_{NOK} et P_{PB} permettent d'analyser la source des échecs de la modélisation (au sens où R_{min} n'est pas atteint avec les conditions fixées dans ces expériences). Ainsi, comme on peut le voir sur les Figures 3.10(b) et 3.10(c) et dans le Tableau 3.2, d'une façon assez générale, un peu moins de la moitié des échecs des modèles MP et MCDP sont dus à des problèmes purement numériques. Le pourcentage P_{PB} est ainsi de l'ordre de 10% des sections pour le modèle MP pour les harmoniques les plus élevées, ce qui met en évidence la sensibilité numérique de ce modèle. Le modèle MCDP semble aussi efficace que le MP (P_{NOK} est similaire pour ces deux modèles) et légèrement moins sensible du point de vue de P_{PB} (qui reste tout de même autour de 8% pour les harmoniques les plus élevées). A l'inverse, ces pourcentages démontrent la grande robustesse numérique du modèle MCD (et dans une mesure un peu moindre, du modèle MCSD). Ainsi, pour ce test, aucune harmonique n'a révélé de mauvais conditionnement du MCD (et ceci, alors que le MCD est utilisé ici sans la régularisation de (2.22)). Par conséquent, pour ce modèle, toutes les non réalisations de l'objectif R_{min} sont dues à la limitation de l'ordre à $K/3$. Comme les valeurs de P_{NOK} correspondantes sont aussi les plus faibles des quatre modèles, on peut conclure que le modèle MCD est à la fois le plus robuste et le plus efficace de tous les modèles testés, en tout cas pour le réglage $R_{min} = 75\%$.

Ainsi, ces expérimentations ont permis de vérifier que, comme on l'avait brièvement mentionné à la Section 2.4, les modèles à base de terme polynomiaux sont sensibles aux problèmes calculatoires. Ceci provient de la grande gamme des valeurs calculées lorsque la longueur de la section modélisée et/ou l'ordre du modèle est grand. Au contraire, le modèle MCD, qui est très proche d'une Transformée en Cosinus Discrets (comme la MDCT couramment employée en codage, et dont on a déjà parlé plusieurs fois dans ce document), est très robuste pour modéliser des sections de parole assez longues, tout comme la TCD l'est pour coder un grand nombre d'échantillons de signal.

Pour résumer cette série d'expérimentations comparatives, nous pouvons dire que le MCD est le meilleur modèle dans le sens où il donne de bonnes performances générales en terme de débit moyen (c'est même le meilleur pour $R_{min} = 75\%$), tout en garantissant la meilleur robustesse de calcul et le plus fort taux de sections correctement modélisées selon le critère perceptif à long terme. Si on veut raffiner la méthode en adaptant le choix du modèle en fonction de la configuration de la section modélisée, et notamment en fonction de sa taille, les autres modèles, et en particulier le modèle polynomial (MP) peuvent être plus appropriés, notamment sur les sections assez courtes (disons, de l'ordre de 10 trames) où ne se pose pas de problème de conditionnement numérique. Toutefois, le relativement faible gain de débit apporté par l'approche multi-modèles montre qu'une telle stratégie risque d'apporter un gain de performances relativement faible par rapport aux coûts additionnels de calcul et de transmission. Au final, ces expériences semblent confirmer que nous avions raison de « privilégier » le MCD dans nos expérimentations, ce que nous avions fait sur la base d'expérimentations pilotes et sur la base de la littérature (voir par exemple [Li *et al.*, 2001]).

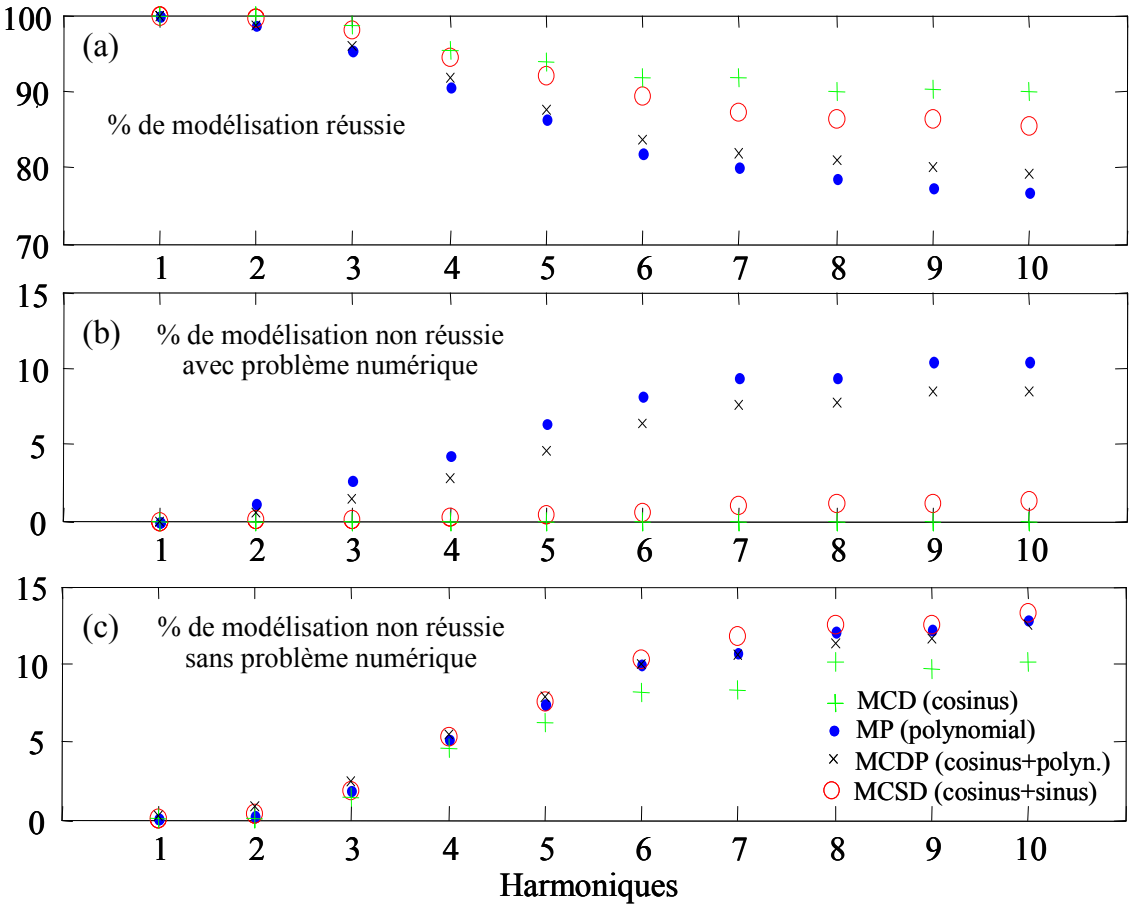

Figure 3.10 : Résultats comparatifs de la modélisation à long terme des dix premières trajectoires d'amplitude par les quatre modèles proposés (MCD, MP, MCDP, MCSD) : (a) pourcentage P_{OK} des harmoniques correctement modélisées ; (b) pourcentage P_{PB} des harmoniques non correctement modélisées à cause d'un problème numérique ; (c) pourcentage P_{NOK} des harmoniques non correctement modélisées sans problème numérique. Dans les trois cas, on a la contrainte $P_i < K/3$. Les pourcentages sont calculés sur les 10 premières harmoniques de l'ensemble des 3500 sections de la base de test.

3.4. Conclusion

Dans ce chapitre nous avons appliqué la méthode de modélisation à long terme présentée au Chapitre 2 sur les paramètres d'amplitude spectrale, en utilisant le critère perceptif de masquage fréquentiel adapté selon l'axe temporel. Les résultats principaux sont les suivants.

Tout d'abord, l'algorithme s'adapte généralement correctement aux différentes configurations des sections modélisées. En général, le modèle à long terme fournit des valeurs d'amplitudes qui sont assez proches des amplitudes mesurées. Ceci est garanti par la contrainte perceptuelle qui guide le comportement de l'algorithme d'ajustement : à la fin de l'algorithme, si la modélisation s'est bien passée, c'est-à-dire si la limite supérieure de l'ordre n'est pas atteinte et si aucun problème numérique n'apparaît, au moins R_{min} pourcent des amplitudes modélisées sont assurées de vérifier cette contrainte.

Les trajectoires modélisées par les différents modèles à long terme ont la propriété intrinsèque d'être lisses, puisque le modèle est composé de fonctions elles-mêmes lisses. La conséquence est que l'ajustement est plus facile pour les harmoniques basses (assez régulières) que pour les harmoniques hautes (plus « bruitées »). Comme pré-senti, l'ordre P varie ainsi beaucoup, selon la longueur et le contenu de la section de parole modélisée. Le réglage de R_{min} à 75% pour la plupart des harmoniques testées (les dix premières) semble être un bon compromis entre précision de la modélisation et respect de la transparence perceptive (avec des valeurs éventuellement plus élevées pour les deux premières harmoniques). Ce réglage permet d'obtenir un gain intéressant sur le nombre de coefficients nécessaire à encoder les trajectoires : un facteur de l'ordre de 7 est obtenu par rapport à la modélisation à court terme utilisant les données issus de l'analyse période-synchrone, et un facteur de l'ordre de 2 à 4 est obtenu par rapport à la modélisation à court terme utilisant une fenêtre glissante avec des valeurs de décalage dans la gamme usuelle (10-20 ms).

Enfin, du point de vue du choix du modèle à long terme, une étude comparative a permis de confirmer l'efficacité du modèle en cosinus discrets, qui se caractérise à la fois par une très bonne robustesse calculatoire et de bonnes performances en terme de débit. Sans toutefois disqualifier complètement ses concurrents, il se dégage donc comme le modèle que nous avons eu raison de privilégier dans nos expérimentations sur la base de sa généralité et de ses connexions avec la DCT.

Chapitre 4

4. Application à la modélisation à long terme de la phase

Dans ce chapitre, nous proposons de modéliser à long terme les paramètres de phase du modèle sinusoïdal. Comme pour les paramètres d'amplitude, ceci signifie que les trajectoires temporelles des paramètres de phase sont modélisées sur de longues sections de parole, au-delà de la longueur de trame usuelle en analyse-synthèse à court terme. Et comme pour les amplitudes, un modèle à long terme unique est employé pour représenter chaque trajectoire de phase sur chaque section de parole entièrement voisée.

Ce chapitre présente deux approches différentes pour le critère d'ajustement du modèle à long terme dans le cas des trajectoires de phase. Dans la première approche, qui sera étudiée à la Section 4.2, nous proposons d'utiliser un critère du type rapport signal à bruit (RSB). Cette première approche assez rudimentaire a permis de poser les bases de la modélisation à long terme des phases. Dans la deuxième approche, que nous développons amplement à la Section 4.3, nous proposons un critère basé sur des considérations perceptives. Cette approche, plus aboutie, permet de fournir une représentation efficace des phases pour les signaux de parole voisés, en terme de débit de coefficients, tout en préservant une bonne qualité de signal. Ce dernier point ouvre de nouvelles perspectives pour le codage de parole avec une bonne qualité à très bas débit.

Ce chapitre est organisé comme suit : d'abord quelques considérations sur la phase du signal en vue de sa modélisation à long terme sont discutées à la Section 4.1. Ensuite, la première étude, basée sur le critère de RSB pour l'ajustement du modèle à long terme, est présentée à la Section 4.2, incluant la définition exacte du critère, la présentation de l'algorithme proposé, et celle des expérimentations et des résultats correspondants. La seconde étude, basée sur un critère perceptif, est décrite dans la Section 4.3, selon la même structure : on décrit successivement le critère perceptif, l'algorithme d'ajustement du modèle à long terme correspondant, les expérimentations et les résultats.

4.1. Quelques considérations sur la phase du signal en vue de sa modélisation à long terme

Dans le Chapitre 1, à la Section 1.1.4.3, nous avons précisé la définition ou plutôt les définitions possibles de la phase d'un signal, et leurs relations avec la notion de fréquence. Nous revenons sur ce point ici de façon plus spécifique pour « préparer le terrain » de la modélisation à long terme de la phase. Nous commençons par une série de rappels rapides sur la définition de phase absolue, puis nous discuterons quelques

considérations sur l'importance de la phase dans la modélisation et le codage de la parole et de la musique. Enfin, nous verrons que la relation entre phase et fréquence prend une importance toute particulière dans le cadre de la modélisation à long terme qui est celui de cette thèse.

4.1.1. Rappels sur la définition de la phase absolue

Rappelons que comme nous l'avons mentionné à Section 1.1.4.3, dans le cadre de signaux comportant plusieurs composantes sinusoïdales potentiellement non stationnaires, la *phase absolue* (ou *phase instantanée*) de chaque composante est définie à chaque échantillon comme l'argument de la fonction cosinus correspondante. Il s'agit d'une fonction positive strictement croissante, résultant de l'intégration des fréquences instantanées au cours du temps. Rappelons que nous nous intéressons particulièrement au cas des signaux quasi-harmoniques avec évolution « lente » des paramètres des composantes, ce qui correspond bien aux portions voisées de la parole ou à de nombreux types d'instruments de musique. Lorsque le signal est (pseudo-)périodique, on peut faire l'hypothèse que les fréquences instantanées sont reliées entre elles par une relation d'harmonicité (*i.e.*, elles sont multiples de la fréquence fondamentale). Notons que les facteurs multiplicatifs associés aux fonctions cosinus de la décomposition sont les amplitudes spectrales, dont nous avons proposé la modélisation à long terme dans le chapitre précédent.

4.1.2. Le problème général de la modélisation/codage de la phase

Si comme on l'a mentionné au chapitre précédent, un encodage relativement précis des amplitudes spectrales du signal semble incontournable pour obtenir un signal codé de haute qualité dans les codeurs de parole, le problème n'est pas aussi tranché en ce qui concerne les phases. On peut par exemple trouver une revue récente de la littérature abordant ce problème de l'influence perceptive de la phase, et comprenant des aspects historiques, dans [Pobloth & Kleijn, 2003]. Les difficultés rencontrées pour traiter ce problème sont d'abord reliées au fait qu'il existe en parole (et aussi en musique) un consensus autour de l'idée d'une moindre importance de la phase (sous-entendu par rapport à l'amplitude) dans la qualité de la perception du signal [Griffin & Lim, 1988] [Lim & Oppenheim, 1979] [Wang & Lim, 1982], en particulier pour les sons voisés relativement réguliers tels que les voyelles par exemple. Cette idée est généralement vraie mais elle mérite d'être précisée sous peine de se révéler fausse ou du moins abusive dans l'absolu, et ceci pour au moins deux raisons.

- D'abord, comme on l'a déjà mentionné à la Section 1.1.4.3, il faut être précis sur la définition de la phase. Cette idée d'une moindre importance perceptive de la phase s'applique en réalité sur les *phases relatives*, c'est-à-dire ce qu'on a défini comme *phase à l'origine* ou *phase de dispersion* dans la Section 1.1.4.3. Rappelons que cette définition de la phase n'est valable que dans un cadre stationnaire, et l'influence du codage de la phase sur la perception n'a été plus ou moins clairement identifiée que dans un tel cadre « idéal ». Par exemple, les auteurs de [Pobloth & Kleijn, 2003] et [Kim, 2001] ont seulement considéré le cas de voyelles périodiques synthétiques. A l'inverse, dans un cadre non stationnaire, il est évident que les trajectoires des fréquences instantanées sont,

elles, perceptivement pertinentes. Or, ces trajectoires sont les dérivées des trajectoires de *phases absolues*, donc une perturbation des phases absolues en tant que fonction du temps, par exemple par une modélisation (interpolation) et/ou une quantification peu performante, peut avoir des répercussions perceptives importantes. En revanche, on peut se permettre (dans une certaine mesure, comme on le verra plus loin) de décaler les composantes sinusoïdales entre elles, à condition que ce décalage soit « entretenu » (c'est-à-dire quasi-constant) tout au long des trajectoires de phase absolue, de façon à ce que l'évolution en fréquence de chaque composante soit respectée. Un exemple courant de cette approche dans les codeurs de parole [Shlomot *et al*., 2001] ou la synthèse musicale [Serra & Smith, 1990] consiste à coder uniquement les trajectoires des fréquences instantanées, sans se soucier des phases à l'origine (de chaque section), et à reconstruire les trajectoires de phases en intégrant les trajectoires de fréquences décodées. On pourrait tenter de coder au moins les relations de phase à l'origine des trajectoires en espérant que les décalages restent fidèles à ceux du signal original au fur et à mesure de l'intégration des fréquences. Mais les erreurs de codage des fréquences ne permettent généralement pas une telle fidélité : sur le long terme, même si elles sont initialisées correctement, les trajectoires se déphasent en général suffisamment pour que le signal reconstruit ne respecte plus la forme d'onde du signal original. Heureusement, l'effet perceptif de ce déphasage est souvent négligeable, et les codeurs ne se privent pas d'exploiter cette propriété.

- Ensuite, tout dépend de ce qu'on entend par qualité, car contrairement à une idée reçue (et à ce qu'on vient de mentionner juste ci-dessus), des changements dans les déphasages relatifs des composantes peuvent être effectivement perçus, même dans un cadre stationnaire. Mais d'une part, ils ne nuisent pas à l'intelligibilité du signal dans le cas de la parole, et d'autre part, les déphasages sensibles ne concernent que des zones limitées du spectre, en moyennes fréquences[66], voir les études de Kim [Kim, 2001, 2003]. Par conséquent, la baisse de qualité du signal est généralement très faible si on ne tient pas compte des relations exactes de phase (relatives) dans le codage, dans un cadre stationnaire ou pseudo-stationnaire.

Cette brève discussion donne un aperçu des difficultés de déterminer un bon degré de précision pour la prise en compte de l'information de phase dans les systèmes de traitement de la parole (et des signaux audio en général). Quoi qu'il en soit, en parole, la différence de qualité selon que l'on prend en compte le phasage exact des composantes ou pas est suffisante pour différencier la classe des codeurs qui respectent cette relation exacte de phase de ceux qui ne la respectent pas[67]. Les premiers peuvent reconstruire

[66] La perception d'un décalage de phase (ou plus précisément d'un changement de décalage de phase) entre des composantes sinusoïdales est liée à la notion de bande critique (voir la Section 1.5.2). En gros, le postulat est que l'on perçoit les variations de phase si les différentes composantes sont dans la même bande critique. Comme ces bandes sont découpées de façon logarithmique, en basses fréquences, les composantes sinusoïdales des sons sont généralement situées dans des bandes critiques différentes. En hautes fréquences, à l'inverse, il y a généralement beaucoup de composantes dans la même bande critique, ce qui fait qu'un changement de décalage de phase « passe inaperçu ».

[67] On peut citer à cette occasion les auteurs de [Pobloth & Kleijn, 2003] qui écrivent ainsi : "*it is fair to state that no sinusoidal-modeling-based speech coders exist that provide transparent speech quality without the use of explicit information about the STFT phase spectrum*".

une onde de forme assez différente de l'onde originale (par exemple voir [Shlomot *et al*., 2001]) alors que les seconds encodent suffisamment précisément la phase pour respecter la structure temporelle fine du signal (la propriété appelée « *shape invariance* » dans [Quatieri & McAulay, 1992]). Les premiers, qui n'encodent généralement que les fréquences, baissent donc légèrement la qualité du signal mais possèdent des gammes de débit inférieures aux seconds qui doivent encoder soit la phase absolue du signal (c'est difficile) soit la phase relative en plus des fréquences (c'est coûteux)[68]. On peut dire que la frontière entre les deux gammes se situe approximativement autour de 4 kbits/s mais tout n'est pas si simple car le débit et la qualité du signal dépendent aussi d'autres facteurs que le codage ou non de la phase (par exemple la façon dont on encode les composantes de bruit).

Pour être relativement complet, il nous reste à mentionner une sous-classe importante des codeurs qui ne respectent pas la relation exacte de phase des composantes et sont basés sur le codage de la fréquence seule : les codeurs harmoniques, qui font l'hypothèse que le signal est suffisamment régulier sur les portions voisées de parole (ou du moins dans la bande harmonique pour les codeurs du type harmonique + bruit, voir la Section 1.4.2.3) pour être considéré localement comme périodique. Dans ce cas, non seulement on n'encode pas la relation de phase des composantes, mais on n'encode pas non plus les fréquences des différentes composantes : on se contente de mesurer et d'encoder la valeur de la fréquence fondamentale F_0, et on fait l'hypothèse que les autres composantes du signal (au moins dans la bande harmonique) sont les multiples de cette fréquence fondamentale. Il s'ensuit un gain considérable au niveau du débit (qui peut alors devenir inférieur à 4 kbits/s). Ce gain se paye généralement par une qualité inférieure aux codeurs plus complets évoqués précédemment, notamment une sonorité de signal parfois trop artificiellement « sinusoïdale », alors que l'original est plus « bruité » et plus « granuleux ».

En résumé, la non importance perceptive de la phase est une notion qui ne s'applique (et encore, avec quelques précautions) que sur les phases relatives (phase à l'origine). Dans tous les cas, il est important perceptivement de coder correctement les trajectoires de fréquence, qui sont implicitement contenues dans les trajectoires de phase absolue. Les codeurs ne codant pas exactement les phases absolues se contentent de coder au moins leur dérivée, et dans le cas le plus économique, au moins la trajectoire du fondamental F_0, tandis que ceux codant la phase doivent permettre de reconstruire les trajectoires de phase absolue, soit directement à partir de celles-ci, soit à partir des fréquences et des phases relatives.

[68] Notons que le codage de la phase (ou autrement dit de la structure temporelle exacte du signal) peut être implicite à certaines méthodes et donc pas forcément explicitement réalisé par la quantification de paramètres de phase. C'est le cas par exemple des codeurs du type analyse-par-synthèse comme les codeurs CELP. Ce type de codeur encode implicitement l'information de phase dans la structure temporelle du signal d'excitation qui modélise le signal résiduel et excite le filtre de synthèse. Le débit typique des codeurs CELP oscille autour de 8 kb/s et de nombreuses déclinaisons du CELP ont été normalisées pour les télécommunications depuis une quinzaine d'années [Atal *et al*., 1993] [Gersho, 1994] [IEEE Com. Mag., 1997].

4.1.3. Conséquences pour la modélisation à long terme des trajectoires de phase

Dans l'approche à long terme de la modélisation et du codage associé de paramètres de phase et/ou de fréquence, on peut aussi distinguer *a priori* les trois cas de figure correspondant aux trois grands types de codeurs à court terme décrits dans la section précédente : on peut modéliser soit l'information complète de phases (en tant qu'information équivalente à celle fournie par l'ensemble phases relatives + fréquences), soit l'ensemble des fréquences, ou bien seulement la fréquence fondamentale. Cependant, dans cette nouvelle approche, comme on s'intéresse aux trajectoires à long terme de ces paramètres, la frontière entre les deux premières catégories est beaucoup plus ténue que dans l'approche à court terme. En effet, les fréquences instantanées étant par définition les dérivées des phases absolues, des modèles à long terme de trajectoires de fréquences peuvent être développés conjointement aux modèles à long terme de trajectoires de phases, en tant que modèles dérivés (au sens de fonction dérivée). Il en est d'ailleurs ainsi dans l'approche à court terme du modèle sinusoïdal de parole dans [McAulay & Quatieri, 1986] : dans cette étude, les modèles de fréquence sont « calés » aux jonctions des trames de synthèse grâce aux mesures de phase en ces points (voir la Section 1.3.2.2.2). Comme on l'a déjà mentionné, dans ce type de codage à court terme « *shape invariant* », une information supplémentaire de phase relative permettant d'estimer la phase absolue doit être régulièrement transmise (c'est-à-dire concrètement pour chaque trame à court terme) en plus des fréquences, pour caler correctement le signal. Par rapport à cette approche, l'approche de type modélisation à long terme permet de préserver la forme de l'onde du signal pour très peu de coût de codage supplémentaire par rapport à des codeurs ne codant que les fréquences : ce coût supplémentaire est celui du passage des modèles à long terme de trajectoires de fréquences aux modèles à long terme de trajectoires de phase absolue, c'est-à-dire l'ajout *d'un seul coefficient pour toute la trajectoire considérée* (pour chaque partiel). Nous verrons plus en détail l'application de ce principe avec les modèles à long terme étudiés. En résumé, dans notre approche à long terme, si on veut coder toutes les fréquences des composantes, on a intérêt à passer directement au codage des phases absolues. C'est ce que nous allons faire par la suite.

4.2. Une première étude

Dans cette section, nous présentons l'implantation et les résultats d'une première étude de modélisation à long terme des trajectoires de phase du modèle sinusoïdal de parole. Nous avons considéré dans cette première étude un critère simple et classique du type rapport signal à bruit (RSB). Le bruit en question est ici identifié à l'erreur de modélisation dans le domaine des échantillons de signal. Par conséquent, on est encore relativement loin d'un critère perceptif puisqu'il est connu que le RSB considéré ainsi sous forme brute n'est pas forcément bien corrélé au jugement perceptif de la qualité des signaux. Ce critère nous a cependant permis de réaliser une première application des principes généraux de la modélisation à long terme à des trajectoires de phase, et de comparer dans ce cadre les résultats de deux modèles à long terme basés sur les fonctions cosinus discrets (MCD) et polynomiales (MP).

4.2.1. Critère de RSB pour l'ajustement du modèle à long terme des trajectoires de phase

Le rapport signal à bruit (RSB) est un terme fréquemment utilisé en traitement du signal ou dans la théorie de l'information pour désigner le rapport entre une quantité d'information utile (le signal) et celle d'information inutile (le bruit). Ce dernier est souvent un bruit parasite venant se greffer sur le signal. Le plus souvent ce rapport est un rapport de puissance, et comme de nombreux signaux ont une échelle dynamique élevée, les rapports signal à bruit sont souvent exprimés en décibels ou dans une échelle logarithmique. Ce concept est utilisé dans de nombreux contextes et applications tels que la compression et la restauration des sons et des images. Dans cette étude, le RSB est défini précisément comme le rapport de la puissance du signal original que l'on veut modéliser à long terme à la puissance de l'erreur de modélisation. Cette erreur est définie ici comme la différence entre ce signal original et le signal synthétisé après modélisation à long terme. On a donc :

$$RSB = 10\log_{10}\left(\frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N}\left(s(n)-\hat{s}(n)\right)^2}\right) \tag{4.1}$$

Cette définition correspond bien à la qualité objective de la modélisation au sens de la fidélité de la forme d'onde synthétisée après modélisation par rapport à la forme d'onde du signal original : un rapport peu élevé signifie que le modèle dénature le signal d'origine au niveau de sa forme d'onde, et à l'inverse, un RSB élevé correspond à un signal bien modélisé au sens de fidèle à la forme d'onde du signal original.

4.2.2. Un premier algorithme de modélisation à long terme des trajectoires de phase

Comme pour la modélisation à long terme des amplitudes abordée au chapitre précédent, nous avons adapté l'algorithme de la Section 2.5.5 au cas des phases. Cet algorithme permet ici de déterminer automatiquement l'ordre optimal (minimal) du modèle à long terme et de réaliser conjointement l'ajustement du modèle aux trajectoires de phase, avec la contrainte d'atteindre un certain RSB minimum. Autrement dit, l'idée de base est un réglage de l'ordre du modèle pour chaque section de parole voisée de sorte qu'un minimum de RSB soit réalisé sur la section modélisée. L'algorithme exact est donné à la fin de cette section.

Notons que dans cette étude préliminaire, l'ordre du modèle à long terme est le même pour chaque partiel, alors qu'il pourrait être différemment ajusté à chaque partiel, comme c'était d'ailleurs le cas pour les amplitudes. Ce choix a été fait ici pour des raisons de simplicité. En effet, le critère d'ajustement est ici un critère global portant sur le signal synthétisé à partir de toutes les harmoniques modélisées, et non sur chaque harmonique indépendamment. Pour calculer ce critère, il faut donc réaliser au préalable la modélisation à long terme de toutes les harmoniques. Pour cette raison, il est difficile d'implanter une méthode réalisant l'ajustement en fonction de ce critère conjoint, tout en considérant toutes les harmoniques de façon indépendante. C'est pourquoi, par souci

de simplicité, on ne réalise l'optimisation de la modélisation qu'en fonction d'un paramètre unique, l'ordre P du modèle qui est identique pour toutes les harmoniques.

Précisons par ailleurs que dans cette étude la synthèse est réalisée en appliquant une interpolation (à court terme) linéaire entre les mesures d'amplitudes (sur une échelle linéaire). Ceci est justifié par le fait que, comme déjà mentionné au chapitre précédent, on veut évaluer séparément la modélisation à long terme des amplitudes et celle des phases. Une fois que les trajectoires d'amplitudes et de phase sont calculées, les équations (1.2) et (1.3) sont utilisées pour produire le signal de synthèse. Au final, l'algorithme est celui présenté ci-dessous.

Algorithme à appliquer sur chaque section de parole voisée (RSB_{min} est choisi par l'utilisateur à une valeur significativement plus grande que zéro ; l'algorithme est terminé pour la plus petite valeur de l'ordre pour laquelle on obtient $RSB \geq RSB_{min}$) :

1. Initialiser l'ordre P à la puissance de deux la plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP à $P/2$.
2. Pour chaque harmonique $i \in [1, I]$, calculer le vecteur des coefficients $\boldsymbol{C}_i$ du modèle à long terme avec (2.19) en remplaçant le vecteur $\boldsymbol{V}_i$ par le vecteur des mesures de phase $\boldsymbol{\varphi}_i = [\varphi_{i,1}\ \varphi_{i,2}\ \ldots\ \varphi_{i,K}]$, et calculer les trajectoires de phases modélisées à long terme par : $\hat{\varphi}_i = \boldsymbol{C}_i \boldsymbol{M}_i$.
3. Calculer le RSB entre le signal original et le signal synthétisé à partir des nouvelles valeurs de phase modélisées à long terme (et à partir des mesures d'amplitude interpolées linéairement).
4. Si $RSB \geq RSB_{min,}$ diminuer l'ordre du modèle selon $P \leftarrow P - \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2.
5. Sinon, augmenter l'ordre du modèle selon $P \leftarrow P + \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2.

On stoppe l'algorithme quand P se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $RSB \geq RSB_{min}$.

4.2.3. Expérimentations

Dans cette section nous décrivons l'ensemble des expériences qui ont été conduites pour évaluer cette première modélisation à long terme des trajectoires de phase basée sur le critère de RSB.

4.2.3.1. Protocole expérimental

Dans cette série d'expérimentations, on reprend les données décrites à la Section 3.3.1.1. Rappelons simplement qu'elles consistent en 3500 segments voisés de taille très variable, pour un total d'environ 13 minutes de parole. Les résultats quantitatifs présentés sont des résultats moyennés sur l'ensemble de ces 3500 segments. De même, le processus d'analyse est le même que celui présenté au Chapitre 3 (Section 3.3.1.2). Rappelons simplement que cette analyse repose sur un ajustement du modèle harmonique sur le signal au sens des moindres carrés, pour chacune des K périodes consécutives de chaque section de parole voisée considérée. A la fin de

l'analyse, chaque section de parole est ainsi représentée par K jeux de paramètres d'amplitude et de phase. Comme précédemment pour les amplitudes, les paramètres de phase sont considérés le long de l'axe du temps comme I jeux de K valeurs consécutives, I étant le nombre d'harmoniques considérées dans les expérimentations. Cependant, par rapport aux amplitudes, les paramètres de phase ainsi organisés nécessitent un pré-traitement supplémentaire : pour assurer la « vraie » trajectoire de phase à partir des mesures effectuées modulo 2π, nous appliquons le processus de dépliement temporel de ces mesures tel que présenté dans le Chapitre 1 Section 1.3.2.2.2. Ainsi, à partir d'ici, nous considérons les ensembles de mesures de phases déroulées :

$$\varphi_i = \left[\varphi_{i,1} \;\; \varphi_{i,2} \cdots \varphi_{i,K}\right] \qquad \text{pour } 1 \leq i \leq I \tag{4.2}$$

Il est aussi important de noter que, comme dans le cas de la modélisation à long terme des amplitudes du Chapitre 3, on modélise seulement les trajectoires de phase des dix premières harmoniques de chaque section de signal considérée. Pour les autres harmoniques, on utilise le procédé d'interpolation linéaire des mesures de phase à court terme décrite à la Section 1.3.2.2.2. Par ailleurs, pour comparer notre méthode avec une approche à court terme classique, nous réutilisons les signaux de référence modélisés à court terme et déjà mentionnés à la Section 3.3.3 : ces signaux sont synthétisés à partir de l'interpolation à court terme de tous les paramètres (amplitudes et phases) pour toutes les harmoniques. Notons que ces signaux peuvent aussi servir de référence au niveau du RSB : pour chaque section de signal à modéliser, on peut ainsi calculer une valeur référence RSB_{ref} correspondant à l'équation (4.1) appliquée avec le signal original et le signal de référence synthétisé à court terme. La valeur cible RSB_{min} pour la modélisation à long terme peut alors être choisie comme un pourcentage donné (proche de 100%) de RSB_{ref}[69].

Enfin, rappelons que lorsque des phrases entières sont synthétisées (avec des sections voisées et non voisées) pour permettre les tests d'écoute, comme expliqué dans la Section 3.3.3, les sections non voisées sont conservées telles quelles et sont raccordées aux sections voisées issues de la modélisation à long terme par un procédé d'*Overlap-Add* local. Notons enfin en ce qui concerne les tests d'écoute que les deux sujets mentionnés dans la Section 3.3.3 ont écouté les signaux nouvellement synthétisés de façon intensive et dans des conditions très propres (voir la Section 3.3.3).

4.2.3.2.Résultats

Le premier résultat que nous rapportons ici est un résultat pratique que nous pouvons tenter d'expliquer en théorie et qui a conduit à modifier légèrement l'algorithme d'une façon que nous allons décrire juste après le report de ces observations. Nous avons en effet constaté le phénomène général suivant (rappelons que nous manipulons 3500 segments de parole de différentes tailles). Pour la plupart de ces segments, et partant

[69] Ceci sous-entend que la modélisation à long terme peut être vue comme une dégradation supplémentaire par rapport à l'interpolation à court terme. Cela n'est pas si évident dans cette étude précise : par exemple, des trajectoires de paramètres lissées par les modèles à long terme peuvent être « meilleures » en terme de RSB que les trajectoires mesurées si les mesures sont particulièrement mauvaises (on a déjà mentionné dans ce document cet effet de filtrage du bruit de mesure). En pratique cependant, prendre RSB_{min} de l'ordre de 80% de RSB_{ref} est un choix qui donne de bons résultats.

d'un ordre assez faible, le RSB augmente généralement assez régulièrement avec l'ordre du modèle, avant de stagner lorsque l'ordre s'approche de la moitié de la taille K des jeux de paramètres modélisés. Lorsque l'ordre augmente encore, le RSB a même tendance à diminuer. Cette diminution est quasiment systématique lorsque l'ordre s'approche de la valeur limite supérieure K–2. Cette observation est assez contre intuitive : on pourrait supposer *a priori* que plus l'ordre augmente, plus le modèle a la possibilité de s'ajuster précisément aux données, et plus le signal synthétisé sera fidèle à l'original. Ce principe est juste, mais le problème est qu'il n'est réalisé par l'algorithme que pour les valeurs mesurées des phases, c'est-à-dire un ensemble relativement restreint de valeurs représentant les trajectoires de phase (une valeur par période de signal). Il se peut (et il est avéré dans ces expériences) qu'un ajustement de plus en plus précis pour cet ensemble de valeurs ne signifie pas pour autant un meilleur ajustement du modèle pour les valeurs en dehors de ces valeurs de référence. Autrement dit, complexifier le modèle pour le contraindre à passer par les bons points peut conduire à complexifier inutilement les trajectoires du modèle entre ces points, et finalement éloigner ces trajectoires des vraies trajectoires de phase (et par conséquent la forme d'onde synthétisée est moins fidèle à l'originale et le RSB diminue).

Ceci peut être typiquement assimilé à un phénomène de sur-apprentissage : on perd en généralité (une trajectoire de phase bien suivie dans son ensemble par un modèle de forme relativement simple, même si les points de référence, c'est-à-dire les mesures, ne sont pas exactement en correspondance) ce qu'on gagne de façon inutile en spécificité (on passe près des points de référence, voire exactement sur les points de référence si le degré de liberté du modèle était égal au nombre de mesures, et ceci au détriment du respect des trajectoires entre les points de mesure). Au final, en moyenne, les trajectoires obtenues en augmentant sensiblement l'ordre du modèle ne sont pas forcément meilleures (c'est-à-dire plus proche des trajectoires « vraies » sous-jacentes aux mesures) que des trajectoires plus simples obtenues avec des ordres inférieurs. Cette observation nous a conduit à limiter la valeur maximale de l'ordre dans la recherche de l'ordre optimal dans l'algorithme proposé. Nous avons ainsi fixé cette valeur maximale à $K/2$ au lieu de la valeur K–2. Conjointement, l'ordre P est initialisé à la puissance de deux la plus proche de $K/4$ au lieu de $K/2$. Nous adoptons cette contrainte dans la série d'expérimentations reportées dans la suite de cette section.

La Figure 4.1 permet de préciser le comportement de l'ordre optimal (c'est-à-dire correspondant au maximum du RSB) en fonction de la taille K de la section modélisée, en tenant compte de cette limitation à $K/2$. Cette figure représente l'ordre optimal obtenu pour chacun des segments voisés du corpus, en fonction de la longueur de segment (cette figure concerne le modèle polynomial mais des graphes très semblables, non représentés ici sont obtenus avec le modèle MCDL). Compte tenu de la discussion ci-dessus, l'ordre du modèle est ici limité par $K/2$ mais il est intéressant de noter que dans la plupart des cas, l'ordre du modèle donnant le RSB maximum est significativement inférieur à cette limite. C'est particulièrement le cas pour les segments plutôt longs qui sont les premiers bénéficiaires de la modélisation à long terme. Pour avoir une vision quantitative d'ensemble, nous avons calculé le nombre moyen de coefficients du modèle à long terme par seconde et par composante harmonique, correspondant à ce maximum de RSB (ainsi ce critère de maximum de RSB remplace dans cette expérience le critère de RSB_{min} dans l'algorithme d'ajustement). Ce débit de coefficients est calculé ici pour le corpus complet de 3500 sections de parole (voix de

femmes et d'hommes ensemble). Il est de 79 coefficients/s/harmonique pour le modèle polynomial et de 91 pour le modèle MCDL. Ces débits fournissent respectivement un RSB moyen de 17,5 dB et 17,9 dB sur l'ensemble du corpus. En comparaison, le nombre moyen de paramètres de phase (et d'amplitude) mesurés est de l'ordre de 200 par seconde et par harmonique[70]. La synthèse à court terme de ces mesures pour générer les signaux de référence modélisés à court terme fournit un RSB moyen de 19,8 dB. Ainsi les modèles à long terme permettent un gain de l'ordre de 55 à 60% sur le nombre de paramètres utilisés pour la synthèse comparés à la synthèse à court terme, en diminuant le RSB d'une valeur de l'ordre de 2 dB seulement.

Pour illustrer ensuite la capacité des modèles à long terme à suivre les trajectoires de phase du signal, nous avons tracé sur la Figure 4.2 un exemple de telles trajectoires pour un segment voisé de voix masculine. Le modèle à long terme est ici le modèle polynomial. Sur cette figure, on a enlevé volontairement le terme linéaire des trajectoires de phase pour une meilleure visualisation des variations locales de celles-ci (le même terme est soustrait aux valeurs des données et du modèle). On peut voir que le modèle affiche une trajectoire lisse autour des mesures de phase, et que cette trajectoire est relativement fidèle à la trajectoire des mesures. Ceci est vérifié même pour une valeur d'ordre très faible devant le nombre de mesures, comme c'est le cas sur cet exemple : sur la figure du haut, on teste ainsi un ordre 8 pour 110 mesures de phases. Une augmentation raisonnable de l'ordre du modèle (c'est-à-dire restant à des valeurs assez faibles par rapport à $K/2$ d'après notre discussion précédente) permet d'améliorer l'ajustement du modèle sur les mesures (et d'améliorer ainsi le RSB ; voir les valeurs ci-après). Ainsi, sur le tracé du bas de la Figure 4.2, la trajectoire du modèle passé à l'ordre 20 devient alors particulièrement fidèle à la trajectoire des mesures.

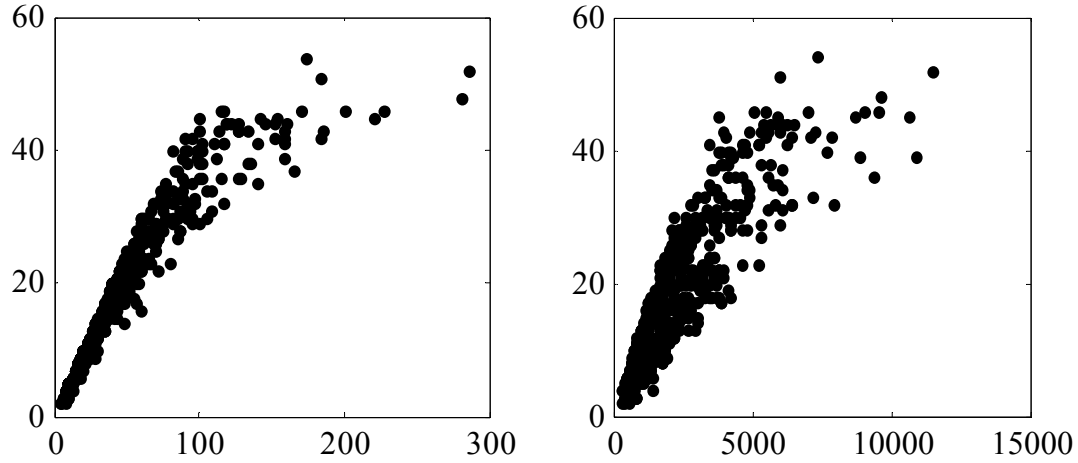

Figure 4.1 : Modélisation à long terme des trajectoires de phase avec un critère de type maximum de RSB : ordre du modèle en fonction de la longueur de la section modélisée (à gauche : en nombre de mesures K ; à droite : en nombre d'échantillons). Le modèle à long terme est ici le modèle polynomial mais des figures très semblables sont obtenues avec le MCDL. L'ordre maximum du modèle est volontairement limité à $K/2$.

[70] Rappel: on a en moyenne 220 trames par seconde pour les voix de femmes, et 140 pour les voix d'hommes. En moyenne inter-genre, compte tenu que l'analyse est pitch-synchrone, que la fréquence fondamentale est plus élevée pour les voix de femmes, et que le corpus est équilibré entre les femmes et les hommes en nombre de sections, on a bien environ 200 trames par seconde.

Ces résultats se transcrivent dans les formes d'onde des signaux synthétisés. Globalement, on peut dire que le signal synthétisé à partir des trajectoires de phase modélisées à long terme (et on le rappelle à partir des trajectoires d'amplitude interpolées à court terme) est relativement bien fidèle à l'original. Ceci est illustré par la Figure 4.3. Sur cette figure, nous avons représenté le signal correspondant au segment 800-1400 de la Figure 4.2. On peut voir que la forme d'onde du signal est globalement préservée même si on peut observer des déphasages locaux lorsque l'ordre du modèle est trop faible. Ainsi, le signal de synthèse issu du modèle à long terme d'ordre 8 (Figure 4.3(c)) est en avance sur le signal original (Figure 4.3(a)), comme cela pouvait être prévu à partir du zoom de la Figure 4.2, où les valeurs de phase modélisées sont supérieures aux valeurs de phase mesurées. Par conséquent le RSB est assez faible, il vaut ici 2.3 dB, même si la forme d'onde du signal est globalement respectée. En revanche, le signal de synthèse issu du même modèle à long terme à l'ordre 20 (Figure 4.3(d)) est synchrone avec le signal original. Le RSB est alors de 13,8 dB (notons que cette nouvelle valeur est obtenue en fixant $RSB_{min} = 13$ dB dans l'algorithme d'ajustement). Pour finir, remarquons que le signal de synthèse est aussi synchrone avec le signal de référence synthétisé avec l'interpolation linéaire à court terme des phases (Figure 4.3(b)).

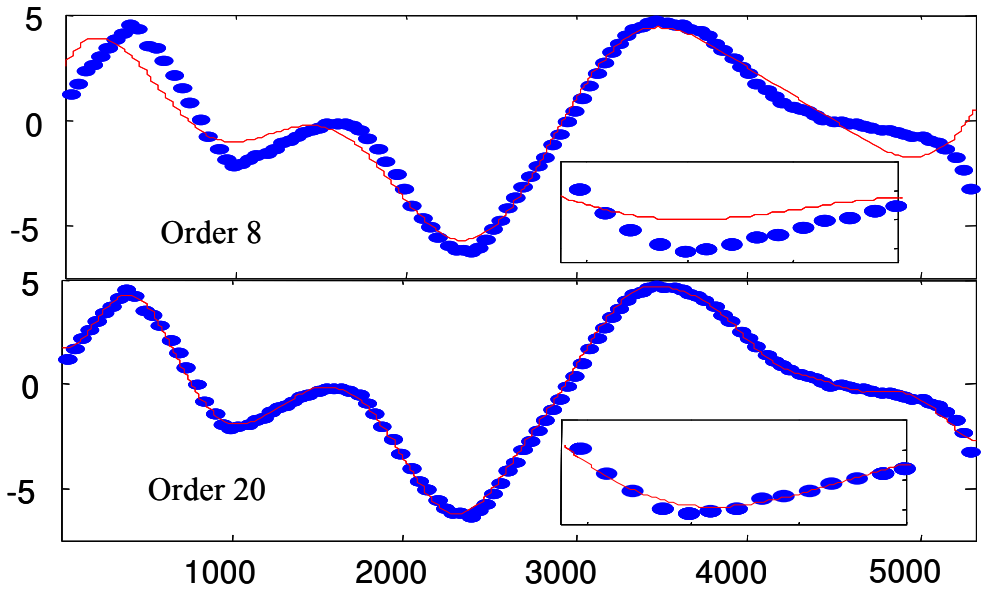

Figure 4.2 : Exemple de modélisation à long terme basée sur un critère de RSB : trajectoires de phase de la première harmonique d'une section de parole voisée d'environ 500 ms (voix d'homme). En abscisse, on a les indices temporels des échantillons ($F_e = 8$ kHz). En bleu : les mesures de phase ($K = 110$). En rouge : le modèle à long terme (modèle polynomial) pour un ordre 8 (en haut) et un ordre 20 (en bas). Le terme linéaire des trajectoires de phase a été retranché pour une meilleure visualisation. Le rectangle inséré est un zoom sur les échantillons 800 à 1400.

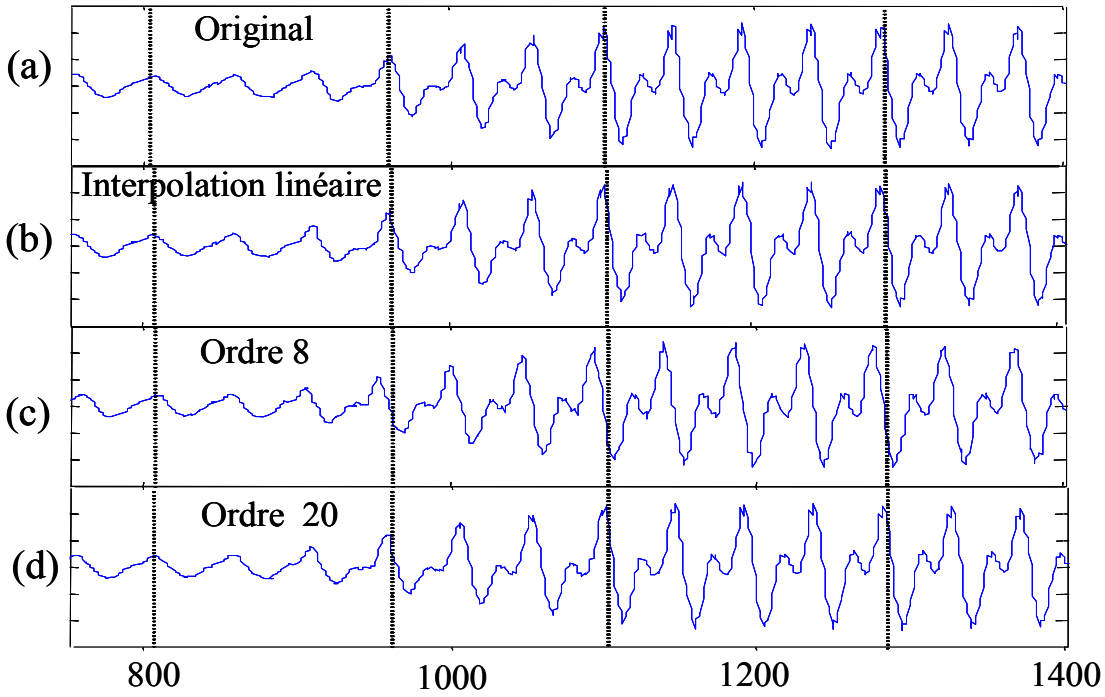

Figure 4.3 : Signaux correspondant au segment 800-1400 de la Figure 4.2 ; (a) signal original ; (b) signal synthétisé avec interpolation linéaire à court terme des mesures de phase dépliées, pour toutes les harmoniques ; (c, d) signal synthétisé avec les dix premières trajectoires de phase modélisées à long terme avec un ordre 8 pour (c) et 20 pour (d) ; les phases de toutes les autres harmoniques sont interpolées linéairement. Pour (b, c, d), toutes les trajectoires d'amplitude sont interpolées à court terme linéairement. Pour (c) et (d), les RSB sont respectivement de 2,3 dB et de 13,8 dB.

Cette série d'observations amène à mettre en regard la qualité du signal vis à vis du respect exact de la forme d'onde. On peut supposer en effet que la qualité du signal doit être préservée en dépit des déphasages observés. Ceci est confirmé par des tests d'écoute informels menés de façon extensive sur un grand nombre d'exemples de la base de données. Le résultat principal de ces tests d'écoute est que les modèles de phase à long terme fournissent une qualité de synthèse semblable à celle obtenue par l'interpolation des mesures de phase à court terme (en tout cas pour les dix premières harmoniques). Il est très important de noter que, même pour des ordres relativement faibles (par exemple 10 coefficients pour modéliser des trajectoires de phase d'un segment voisé comprenant plusieurs phonèmes), la différence perceptuelle entre les signaux de synthèse à long terme et à court terme reste très faible, bien que le RSB baisse de manière très significative. Ceci nous amène à penser que, comme dans d'autres applications telles que le débruitage ou le codage de la parole, le RSB n'est pas le critère le mieux adapté à une approche du problème prenant en compte la perception. Dans notre cas, ce problème de RSB est directement relié à la précision de l'encodage des paramètres de phase, et nous devons en quelque sorte retourner à notre discussion de ce problème abordé à la Section 4.1.2. Pour mieux prendre en compte l'aspect perceptif dans la modélisation à long terme de la phase, nous avons ainsi mené une seconde étude où nous proposons de remplacer le critère de RSB par un critère perceptif. C'est ce que nous allons décrire maintenant dans la section suivante.

4.3. *Une seconde étude*

Dans l'étude préliminaire de la section précédente, nous avons basé la modélisation à long terme des trajectoires de phase sur un critère de type RSB et nous avons donné un aperçu des limites de cette approche. Dans cette section, nous présentons une nouvelle approche où nous tenons compte des propriétés du système auditif en terme de limite de la capacité de l'oreille à percevoir des variations de phase/fréquence des composantes d'un son. Comme pour les études précédentes, nous présentons d'abord le nouveau critère, puis l'adaptation de l'algorithme d'ajustement du modèle à long terme à ce nouveau critère, avant de présenter des résultats d'études expérimentales.

4.3.1. Proposition d'un critère perceptif pour la modélisation à long terme de la phase

Compte tenu de notre discussion de la Section 4.1.2 et des précisions que nous avons apportées sur la définition de la phase et sa relation avec celle de la fréquence, dans cette seconde étude, la distorsion de phase due au processus de modélisation à long terme sera plutôt décrite en terme de modulation de la trajectoire de fréquence associée à la phase modélisée. Les effets perceptuels de cette distorsion seront également analysés de ce point de vue, c'est-à-dire en termes de modulation de fréquence. A cet effet, nous avons choisi de nous inspirer des résultats des études psychoacoustiques de [Zwicker & Feldtkeller, 1981] et de [Demany & Semal, 1989], présentés généralement dans un cadre stationnaire. Nous avons étendu ces résultats à notre approche du problème à long terme, c'est-à-dire dans un cadre non stationnaire où les paramètres de base du modèle sinusoïdal évoluent continûment au cours du temps. Nous décrivons donc d'abord le cadre stationnaire puis l'adaptation des résultats à notre étude.

Les résultats que nous avons exploités sont ceux qui caractérisent les seuils perceptuels dans le cadre d'une modulation de fréquence sinusoïdale (MFS) pour un son tonal de fréquence porteuse fixe. Autrement dit, on part d'une sinusoïde pure de fréquence[71] fixe ω_0, et on module la fréquence de cette sinusoïde par une modulation elle-même sinusoïdale de fréquence ω_M. Ceci est illustré sur la Figure 4.4, sur le tracé du bas. Dans [Zwicker & Feldtkeller, 1981] et [Demany & Semal, 1989], le seuil de MFS est défini comme la déviation maximum $\Delta\omega$ de la fréquence du son pur pour laquelle la modulation reste inaudible. Dans le cas stationnaire, ce seuil s'avère approximativement proportionnel à la fréquence du signal porteur, pour une fréquence de modulation donnée (si cette dernière est sensiblement inférieure à la fréquence porteuse), avec une limite inférieure presque constante. Par exemple, pour une fréquence de modulation de 4 Hz et une fréquence porteuse supérieure à 500 Hz, les auteurs de [Zwicker & Feldtkeller, 1981] proposent le modèle suivant pour le seuil de MFS :

$$\Delta\omega \approx \max(2Hz, 0{,}0035\omega_0) \qquad (4.3)$$

[71] Notons que dans cette section et les suivantes, on garde la notation ω pour désigner une fréquence afin de rester cohérent avec les notations employées jusqu'ici. De plus, ces fréquences sont ici des fréquences analogiques et les valeurs de ces fréquences sont données en Hz.

Intéressons-nous à présent à la possibilité d'élaborer un critère similaire pour la modélisation à long terme de la phase, dans le cas non stationnaire. Dans le cas des signaux de parole (et aussi de certains signaux de musique harmoniques non-stationnaires), la fréquence fondamentale des signaux peut être vue comme une fréquence porteuse de signaux sinusoïdaux variables au cours du temps, dans la marge approximative 100-400 Hz (cet intervalle peut être différent pour les instruments de musique). Évidemment, cette remarque s'applique également sur les différentes harmoniques du signal, avec une gamme de fréquence porteuse proportionnelle en fonction du rang de l'harmonique.

A présent, considérons les deux propriétés suivantes de nos modèles à long terme appliqués aux trajectoires de phase. D'une part, les modèles à long terme proposés sont intrinsèquement « lisses », c'est-à-dire avec des variations régulières au cours du temps. D'autre part, un modèle à long terme dérivé du modèle de phase (au sens de fonction dérivée) est implicitement un modèle à long terme de la trajectoire de fréquence associée à la trajectoire de phase modélisée puisque la fréquence est la dérivée temporelle de la phase. De plus, ce modèle dérivé conserve la propriété de régularité (il est aussi « lisse ») puisque les modèles à long terme choisis (à base de fonctions sinusoïdales ou polynomiales) sont en fait infiniment dérivables. Par conséquent, l'erreur de modélisation des trajectoires de phase se répercute par dérivation en une erreur de modélisation sur les trajectoires de fréquence. Cette dernière erreur peut être vue comme un processus de modulation de fréquence (non sinusoïdale certes) avec une « fréquence »[72] qui dépend de l'ordre du modèle mais qui est toujours petite par rapport à la fréquence porteuse (qui est la fréquence fondamentale ou une de ses harmoniques).

A partir de ces considérations, nous proposons dans cette étude d'adapter de façon très simple le modèle stationnaire de (4.3) au cas non stationnaire de modélisation à long terme, et d'utiliser ce seuil comme critère d'ajustement des trajectoires de phase par l'intermédiaire des trajectoires de fréquences correspondantes (c'est-à-dire les dérivées temporelles). Pour chaque harmonique i, et chaque trame k de la section de parole à modéliser, le modèle de seuil de modulation de fréquence que nous employons dans l'approche à long terme est :

$$T_{i,k} = \Delta\omega_{i,k} \approx \max(2Hz, \alpha\omega_{i,k}) \tag{4.4}$$

Si on fait l'hypothèse d'harmonicité, ce seuil devient :

$$T_{i,k} = \Delta\omega_{i,k} \approx \max(2Hz, \alpha i\omega_{0,k}) \tag{4.5}$$

Ce seuil est donc similaire à celui du cas stationnaire, avec cependant la possibilité de régler le paramètre α de façon adaptée. Ce choix est justifié par le fait qu'à notre connaissance, aucune étude expérimentale exhaustive sur le seuil perceptuel de MFS (ou de modulation de fréquence d'un autre type) n'a été conduite dans le cas non-stationnaire. Compte tenu de la difficulté de mener une telle étude expérimentale complète du fait du grand nombre de paramètres en jeux par rapport au cas stationnaire (deux paramètres supplémentaires, voir la Figure 4.4, tracé du haut), nous avons donc

[72] On utilise abusivement ce terme ici pour représenter l'ordre de grandeur des oscillations temporelles de l'erreur de modélisation, même si ces oscillations ne sont pas sinusoïdales.

opté pour la solution simple proposée ci-dessus tout en ciblant le réglage de α à partir de tests d'écoute pilotes réalisés sur des signaux de parole réels. Ces essais ont révélé que ce paramètre pouvait prendre une valeur significativement plus élevée que dans le cas stationnaire, sans dégradation perceptuelle majeure. Ainsi, dans les expériences de la Section 4.3.3, α est dans la gamme 0,02-0,05, et nous reviendrons sur ce point dans les sections décrivant les résultats expérimentaux.

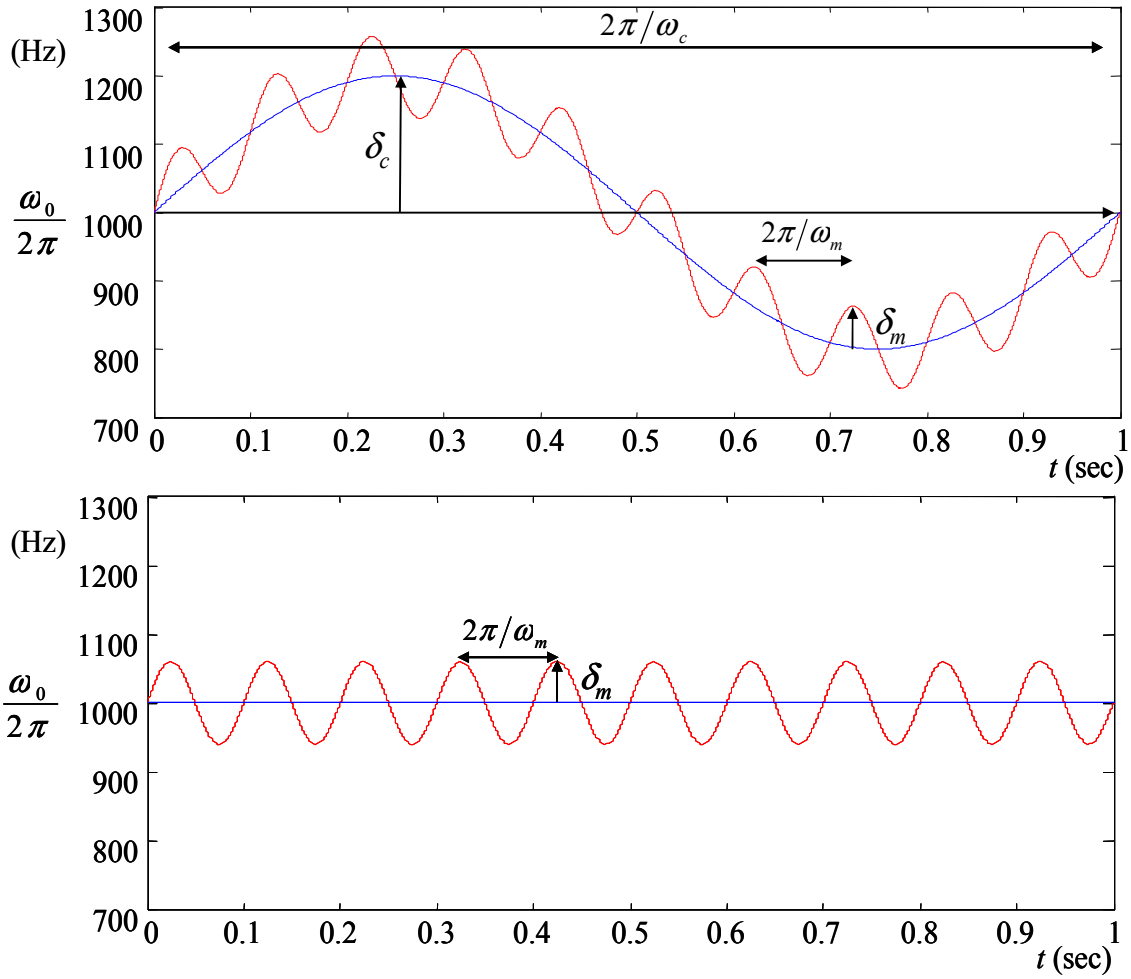

Figure 4.4 : Modulation de fréquence sinusoïdale (MFS) : comparaison entre le cas stationnaire (fréquence porteuse fixe), en bas, et le cas que nous appelons non stationnaire (fréquence porteuse variable, ici sinusoïdale), en haut. Dans le premier cas, il est proposé dans [Zwicker & Feldtkeller, 1981] un modèle de seuil d'excursion en fréquence pour la perception de cette modulation (voir le texte). Dans le deuxième cas, le nombre plus important de paramètres provenant des variations possibles de la porteuse rend très difficile la réalisation d'une étude perceptive complète permettant de réaliser un modèle. En effet, dans ce cas la fréquence du signal s'exprime sous la forme $\omega(n) = \omega_0+\delta_c\cos(\omega_c n)+\delta_m\cos(\omega_m n)$ avec cinq variables : ω_0 correspond à la fréquence fondamentale moyenne, ω_c est la variation globale de fréquence fondamentale au cours du temps modélisant la mélodie du signal, ω_m est la fréquence de la modulation supplémentaire (qui pourrait s'apparenter à l'erreur de modélisation à long terme dans notre étude) ; δ_c et δ_m sont les excursions en fréquence correspondantes.

Une fois le seuil perceptif défini, dans la pratique nous utilisons le formalisme suivant pour comparer ce seuil aux trajectoires de fréquence (nous présentons ce formalisme ici pour simplifier la présentation de l'algorithme d'ajustement qui sera donné à la section suivante). Le modèle à long terme de fréquence associé à un modèle à long terme de phase donné, et défini comme sa fonction dérivée temporelle, peut être évalué aux instants de mesure en utilisant le formalisme d'une matrice de modèle $\boldsymbol{Q}_i$ « dérivée » de la matrice $\boldsymbol{M}_i$ définie à la Section 2.5.2[73]. Par exemple, quand le modèle MCDL défini par (2.8) est utilisé pour modéliser les trajectoires de phase (voir Section 2.4.5), le modèle à long terme de trajectoire de fréquence associé est donné par[74] :

$$\hat{\omega}_i(n) = -\frac{\pi}{N}\sum_{p=0}^{P_i-1} pc_{i,p}\sin\left(p\pi\frac{n}{N}\right) + c_{i,P_i} \tag{4.6}$$

La matrice de modèle dérivé $\boldsymbol{Q}_i$ correspondante est donnée par :

$$\boldsymbol{Q}_i = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -\frac{\pi}{N}\sin\left(\pi\frac{n_1}{N}\right) & -\frac{\pi}{N}\sin\left(\pi\frac{n_2}{N}\right) & \cdots & -\frac{\pi}{N}\sin\left(\pi\frac{n_k}{N}\right) \\ -\frac{2\pi}{N}\sin\left(2\pi\frac{n_1}{N}\right) & -\frac{2\pi}{N}\sin\left(2\pi\frac{n_2}{N}\right) & \cdots & -\frac{2\pi}{N}\sin\left(2\pi\frac{n_k}{N}\right) \\ \vdots & \vdots & \cdots & \vdots \\ -\frac{(P_i-1)\pi}{N}\sin\left((P_i-1)\pi\frac{n_1}{N}\right) & \vdots & \cdots & -\frac{(P_i-1)\pi}{N}\sin\left((P_i-1)\pi\frac{n_k}{N}\right) \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{4.7}$$

et les valeurs de fréquence modélisées aux instants de mesure sont données par :

$$\hat{\boldsymbol{\omega}}_i = \boldsymbol{C}_i\boldsymbol{Q}_i \tag{4.8}$$

A partir de ces valeurs de fréquence, on peut maintenant définir précisément la fonction de l'erreur de modélisation qui sera utilisée dans l'adaptation de l'algorithme d'ajustement de la Section 2.5.5 au cas des trajectoires de phase (donnée à la section suivante). Cette fonction d'erreur est alors ici :

$$f(\boldsymbol{E}_i) = abs(\boldsymbol{\omega}_i - \hat{\boldsymbol{\omega}}_i) = abs(\boldsymbol{\omega}_i - \boldsymbol{C}_i\boldsymbol{Q}_i) \quad \text{avec} \quad \boldsymbol{\omega}_i = \begin{bmatrix} i\omega_{0,1} & i\omega_{0,2} & \ldots & i\omega_{0,K} \end{bmatrix} \tag{4.9}$$

où *abs* dénote la fonction de valeur absolue élément par élément. Dans l'algorithme d'ajustement, cette fonction sera donc comparée au seuil perceptif mis sous forme vectorielle :

$$\boldsymbol{T}_i = \Delta\boldsymbol{\omega}_i = \begin{bmatrix} \Delta\omega_{i,1} & \Delta\omega_{i,2} & \ldots & \Delta\omega_{i,K} \end{bmatrix} \tag{4.10}$$

---

[73] On rappelle que i désigne le rang de l'harmonique considérée. Il n'est pas très important ici pour expliquer la relation entre modèle de phase et modèle de fréquence dérivé, mais on le conserve pour être homogène avec les notations utilisées jusqu'ici, et pour se rappeler que la modélisation est réalisée composante par composante.

[74] On rappelle que pour le modèle MCDL, les termes en cosinus sont indéxés de 0 à $P_i - 1$ de façon à toujours avoir $P_i + 1$ coefficients pour un ordre P_i.

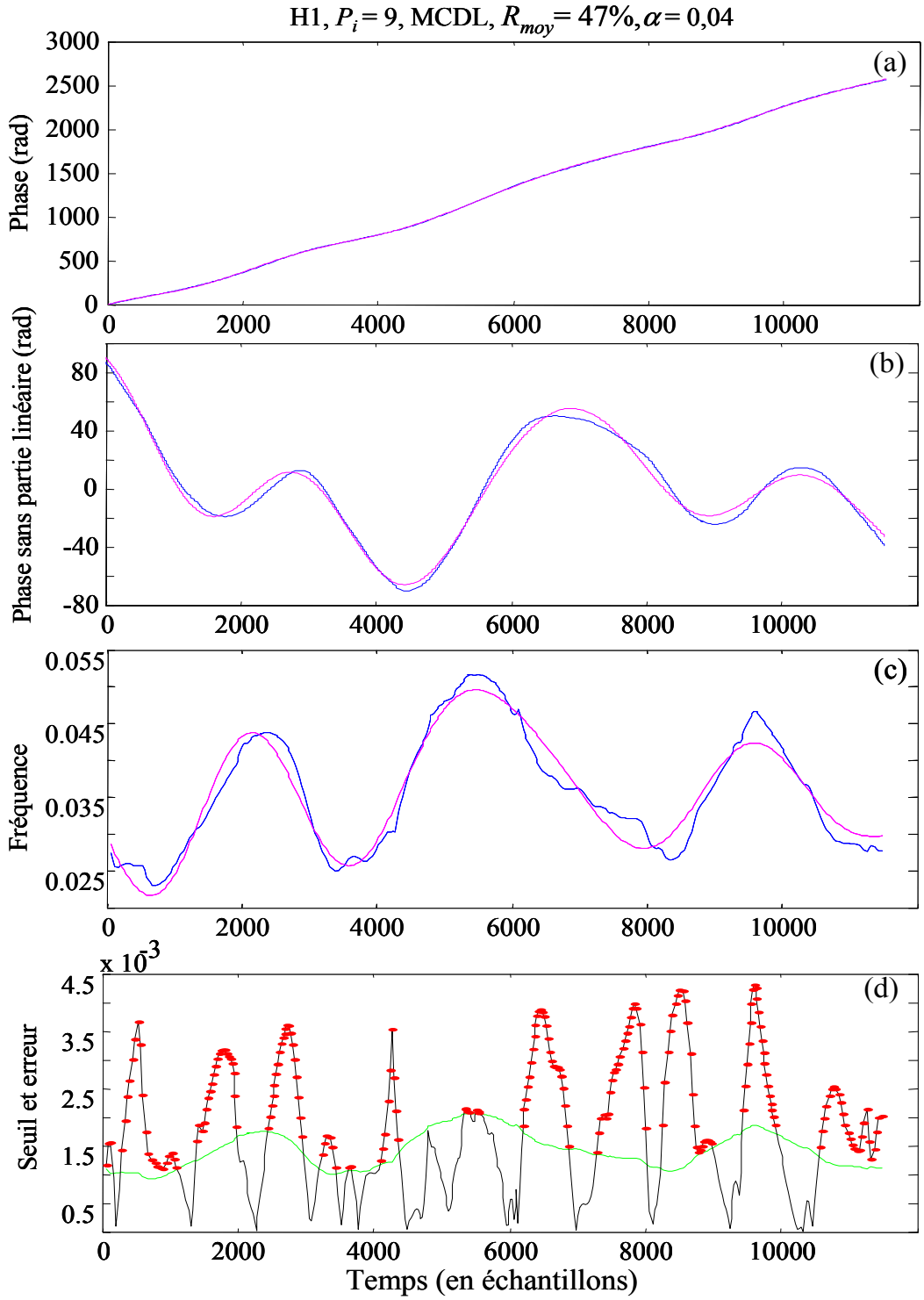

Figure 4.5 : Modélisation à long terme de la trajectoire de phase de la première harmonique de la séquence de parole voisée correspondant à la Figure 3.4 (une longue section entièrement voisée de parole féminine, $K = 408$). Le modèle à long terme est un MCDL d'ordre 9 (soit 10 coefficients). (a) : Trajectoires de phase. (b) : Trajectoires de phase sans le terme linéaire. (c) : Trajectoires de fréquences (numériques). Pour ces trois figures, les mesures sont en bleu et le modèle est en magenta. (d) Seuil perceptif (en vert) pour $\alpha = 0{,}04$ et fonction d'erreur de modélisation (en noir). Les portions de la fonction d'erreur dépassant le seuil perceptif sont marquées par des points rouges.

Pour bien fixer les idées, nous avons illustré ces principes à la Figure 4.5, en dehors de toute problématique d'optimisation du modèle par l'algorithme d'ajustement (nous verrons bien sûr des résultats qualitatifs et quantitatifs plus poussés dans la partie expérimentale à suivre). Sur cette figure, nous avons ainsi tracé une série de trajectoires de phase et de fréquence pour la première harmonique d'une longue séquence de parole voisée pour une voix féminine. Cette séquence est la même que celle utilisée dans le Chapitre 3, notamment sur les Figures 3.4 et 3.5, pour illustrer la modélisation à long terme des trajectoires d'amplitude. Sur la Figure 4.5(a) nous avons représenté la trajectoire des mesures de phase et celle du modèle à long terme de cette trajectoire. Le modèle est ici un MCDL d'ordre $P_i = 9$ et son réglage est décrit dans la section suivante. La Figure 4.5(b) reprend ces tracés après avoir enlevé le terme linéaire pour une meilleure visualisation des variations non linéaires de phase (le même terme est soustrait aux valeurs des données et du modèle). La Figure 4.5(c) représente les trajectoires de fréquence correspondantes (mesures et modèle dérivé du MCDL). La Figure 4.5(d) représente le seuil perceptif (4.10) réglé avec $\alpha = 0{,}04$, et la fonction d'erreur de modélisation (4.9). On voit que le modèle suit globalement la trajectoire de phase et que le modèle dérivé suit bien celle de la fréquence. Toutefois, la modélisation n'est ici pas assez précise du point de vue du critère perceptif tel qu'il est réglé. En effet, de nombreuses portions de la fonction d'erreur de modélisation dépassent ce seuil (en rouge sur la Figure 4.5(d)) : le pourcentage des régions correctement modélisées, c'est-à-dire en dessous du seuil, est seulement de 47%. Nous allons voir dans les sections suivantes comment le modèle peut être ajusté de façon plus performante.

4.3.2. Algorithme de modélisation à long terme des trajectoires de phase à base de critère perceptif

Nous donnons ci-dessous la version de l'algorithme de modélisation à long terme des trajectoires de phase du signal basée sur le critère perceptif défini dans la section précédente. Cet algorithme est une adaptation directe de l'algorithme générique de la Section 2.5.5, avec ce nouveau critère perceptif et la fonction d'erreur associée.

Algorithme à appliquer sur les jeux des mesures de phase (dans cet algorithme R_{min} et α sont initialisés par l'utilisateur à une valeur comprise respectivement dans les intervalles [75%, 90%] et [0,02, 0,05] et *Itermax* est initialisé à une valeur entière entre 10 et 20 ; ces intervalles sont justifiés à la Section 4.3.3.4) :

1. Pour chaque indice du temps $k \in [1, K]$ et chaque partiel $i \in [1, I]$, calculer le seuil associé de $T_{i,k}$ avec (4.5). Ensuite, former les trajectoires de seuil $\boldsymbol{T}_i$ correspondantes (4.10). Puis, pour chacune des i harmoniques :
2. Initialiser l'ordre P_i à la puissance de deux la plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP_i à $P_i/2$.
3. Initialiser la matrice diagonale de poids $\boldsymbol{W}_i$ de taille $K \times K$ avec tous les éléments sur sa diagonale à un. Itérer alors le processus suivant, de l'étape 4 à l'étape 6 :
4. Calculer le vecteur des coefficients $\boldsymbol{C}_i$ du modèle à long terme avec (2.19) en remplaçant le vecteur $\boldsymbol{V}_i$ par $\boldsymbol{\varphi}_i$; et calculer la fonction d'erreur de modélisation $f(\boldsymbol{E}_i)$ avec (4.9).

5. Augmenter les poids où la fonction d'erreur de modélisation dépasse le seuil perceptif, selon :

   $$\Delta \boldsymbol{W} = f(\boldsymbol{E}_i) - \boldsymbol{T}_i$$
   $$\Delta \boldsymbol{W} \leftarrow \Delta \boldsymbol{W} - \min(\Delta \boldsymbol{W})$$
   $$\boldsymbol{W}_i \leftarrow \boldsymbol{W}_i + diag\,(\Delta \boldsymbol{W} \,/\, \max(\Delta \boldsymbol{W}))$$

   (on rappelle que *max* et *min* dénotent respectivement les fonctions maximum et minimum et *diag* dénote la fonction qui produit une matrice diagonale à partir d'un vecteur, les éléments du vecteur étant mis sur la diagonale).

6. Calculer le pourcentage R des éléments négatifs de $f(\boldsymbol{E}_i) - \boldsymbol{T}_i$.

   Si $R<R_{min}$ et le nombre d'itérations *Itermax* n'est pas atteint, retourner à l'étape 4.

   Si $R \geq R_{min,}$ diminuer l'ordre du modèle selon $P_i \leftarrow P_i - \delta P_i$, mettre à jour δP_i selon $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

   Sinon, si $R<R_{min}$ et le nombre d'itérations atteint *Itermax*, augmenter l'ordre du modèle selon $P_i \leftarrow P_i + \delta P_i$, faire la mise à jour $\delta P_i \leftarrow \delta P_i/2$, et retourner à l'étape 3.

On stoppe l'algorithme pour la composante i quand P_i se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R \geq R_{min}$. On passe ensuite à la composante suivante.

4.3.3. Expérimentations et résultats

Dans cette section, nous présentons un ensemble d'expériences qui ont été effectuées sur les sections entièrement voisées de parole pour évaluer la modélisation à long terme des paramètres de phase avec cette nouvelle approche fondée sur un critère d'ajustement perceptif. Le processus d'analyse utilisé et les données sur lesquelles nous appliquons cette modélisation sont les mêmes que celles présentées à la Section 3.3.1.2. Le protocole expérimental est le même que celui présenté à la Section 4.2.3.1, le seul changement étant évidemment que nous remplaçons les trajectoires de phase modélisées avec l'algorithme basé sur le critère du RSB de la Section 4.2 par les trajectoires modélisées avec l'algorithme basé sur le critère perceptif. Nous renvoyons le lecteur à la Section 4.2.3.1 pour une description de ce protocole et nous enchaînons directement avec la présentation des résultats qualitatifs et quantitatifs obtenus avec la nouvelle version « perceptive » de l'algorithme. Notons que l'ordre de cette présentation est légèrement changé ici : nous présentons d'abord quelques remarques générales, puis un exemple particulier avant de passer à la présentation d'une série de résultats quantitatifs plus complets.

4.3.3.1. Comportement de l'algorithme

Dans cette section, nous illustrons d'abord de façon qualitative le comportement de l'algorithme. Tout d'abord, comme pour la modélisation à long terme des amplitudes du Chapitre 3, nous avons observé qu'il n'est généralement pas nécessaire de forcer l'erreur de modélisation à rester complètement sous le seuil perceptif. En d'autres termes, il n'est pas utile, ni même bénéfique, de fixer R_{min} à 100% (rappelons que R_{min}

est le pourcentage cible des points de la trajectoire modélisée où l'erreur de modélisation est au-dessous du seuil de masquage). En effet, à nouveau, un effort trop important pour ajuster le modèle de façon très précise localement (c'est-à-dire sur les points de mesure) peut entraîner un ajustement global moins bon. On a déjà discuté ce point au Chapitre 3 (rappelons par exemple que les mesures qui servent de référence pour le modèle peuvent être entachées d'erreur et que la modélisation prend implicitement en compte cette imprécision en lissant les trajectoires de mesures). On rejoint par ailleurs la discussion portant sur le phénomène de sur-apprentissage de la Section 4.2.3.2. Mais dans cette nouvelle étude, le critère perceptif améliore notablement le comportement de l'algorithme par rapport au critère RSB : il permet d'obtenir des ordres de modèle significativement plus petits (on verra ce point plus en détail dans la suite de la présentation de cette étude).

Rappelons de plus que le but de l'algorithme est d'obtenir un compromis optimal entre l'ordre du modèle (pas trop élevé) et la précision de la modélisation (déterminée par le pourcentage R). On considère que si un pourcentage suffisant est atteint, l'erreur de modélisation sera alors globalement inaudible sur la section entière. Par ailleurs, il faut garder à l'esprit que dans cette nouvelle approche perceptive de la modélisation de phase à long terme, nous avons à contrôler en plus le paramètre α qui détermine l'excursion de modulation de fréquence (voir Section 4.3). Devant la multiplicité des configurations possibles, nous privilégierons le réglage de R_{min} aux deux valeurs arbitraires $R_{min} = 75\%$ et $R_{min} = 90\%$, comme pour la modélisation des trajectoires d'amplitude décrite au Chapitre 3. Avec ce type de réglage et les valeurs de α testées (voir plus loin), nous avons constaté que, comme dans le cas des amplitudes, l'algorithme converge généralement vers une valeur d'ordre de modèle significativement inférieure au nombre K de mesures. A nouveau ceci justifie qu'on puisse limiter l'intervalle de recherche de l'ordre optimal à une valeur significativement inférieure à K pour accélérer l'algorithme. Nous reviendrons sur ce point un peu plus loin.

4.3.3.2. Un exemple particulier

Pour illustrer le comportement de l'algorithme et la capacité du modèle à long terme à s'ajuster globalement sur les trajectoires de phase du signal, nous reprenons l'exemple de la Figure 4.5. Sur la Figure 4.6, nous avons ainsi tracé les mêmes trajectoires de phase et de fréquence que celles de la Figure 4.5, mais sur cette nouvelle figure, les résultats sont ceux obtenus après convergence de l'algorithme d'ajustement (on rappelle qu'il s'agit de la première harmonique d'une longue séquence d'environ 1,4 s de parole voisée pour une voix féminine). Dans cet exemple, le pourcentage cible R_{min} est fixé à 90%, le facteur de modulation α est égal à 0,04, et le modèle à long terme est le MCDL. En partant de la situation de la Figure 4.5 (c'est-à-dire avec un ordre fixé à 9), l'ajustement des poids par l'algorithme ne permet pas d'augmenter significativement le ratio R, même après *Itermax* itérations. Par conséquent, ce ratio n'atteint pas le pourcentage cible R_{min} : ceci signifie que le modèle n'a pas assez de degrés de liberté pour suivre convenablement toutes les mesures originales de la trajectoire de phase (et donc de fréquence pour le modèle dérivé). L'algorithme doit alors augmenter l'ordre du modèle. Sur la Figure 4.6, l'ordre du modèle est passé à 14. Les modèles de phase et de fréquence sont devenus plus proches des trajectoires des mesures. Par conséquent, l'erreur de modélisation est diminuée et le pourcentage R est passé de 47% à un peu plus de 90%. Le pourcentage cible est donc atteint et l'algorithme est terminé.

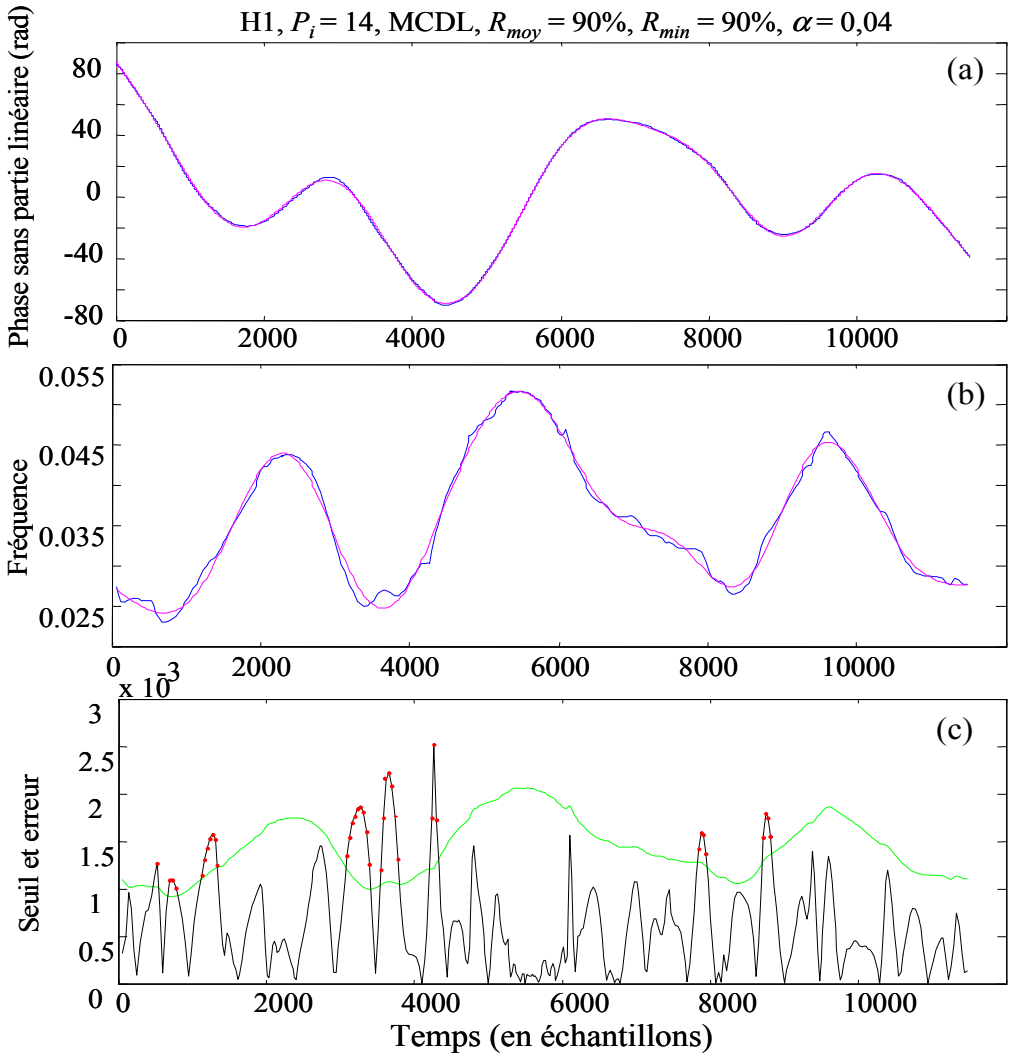

Figure 4.6 : Modélisation à long terme de la même trajectoire de phase que celle de la Figure 4.5 après convergence de l'algorithme d'ajustement du modèle. Pour atteindre la valeur $R_{min} = 90\%$ avec $\alpha = 0{,}04$, l'ordre du modèle MCDL est passé à 14. (a) : Trajectoires de phase sans le terme linéaire. (b) : Trajectoires de fréquences (réduites). Pour ces deux figures, les mesures sont en bleu et le modèle est en magenta. (c) : Seuil perceptif (en vert) et fonction d'erreur de modélisation (en noir). Les portions de la fonction d'erreur dépassant la trajectoire du seuil perceptif sont marquées par des points rouges. On n'a pas représenté la trajectoire de phase avec le terme linéaire, car sur l'échelle correspondante, les deux courbes représentant les mesures et le modèle sont visuellement confondues (comme c'était déjà quasiment le cas sur la Figure 4.5).

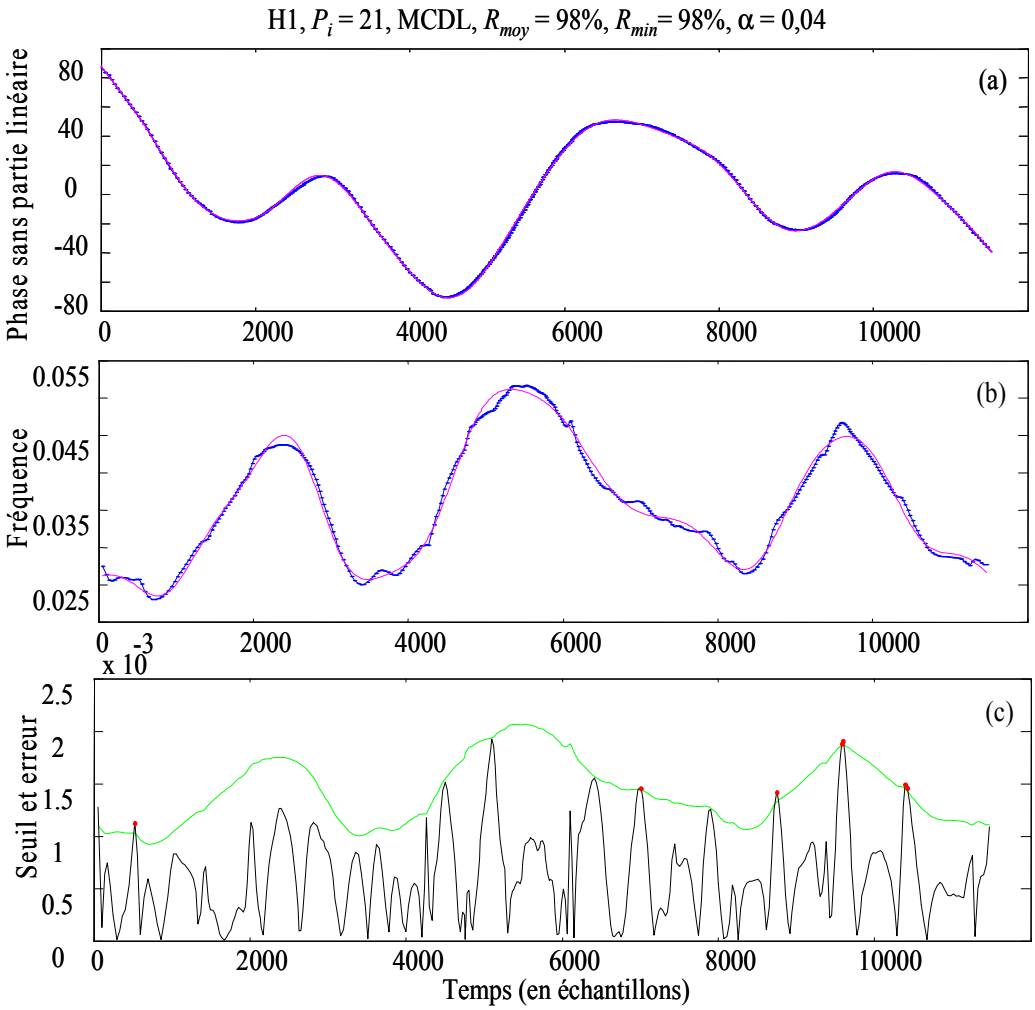

Figure 4.7 : Modélisation à long terme de la même trajectoire de phase que celle de la Figure 4.5 après convergence de l'algorithme d'ajustement du modèle. Pour atteindre la valeur $R_{min} = 98\%$ avec $\alpha = 0{,}04$, l'ordre du modèle MCDL est passé à 21. (a) : Trajectoires de phase sans le terme linéaire. (b) : Trajectoires de fréquences (réduites). Pour ces deux figures, les mesures sont en bleu et le modèle est en magenta. (c) Seuil perceptif (en vert) et fonction d'erreur de modélisation (en noir). Les portions de la fonction d'erreur dépassant le seuil sont marquées par des points rouges.

En complément de cette expérience, à partir de la configuration de la Figure 4.6, nous avons relancé l'algorithme sur les mêmes données en réglant cette fois R_{min} à 98%. On peut voir le résultat sur la Figure 4.7 : les trajectoires des modèles sont encore affinées, du fait de l'augmentation de l'ordre du modèle de phase à 21. On remarque que la forme globale de la fonction d'erreur de modélisation suit particulièrement bien la trajectoire du seuil perceptif. Cependant, comme on l'a déjà mentionné précédemment, une telle précision dans la modélisation n'est pas forcément nécessaire. En effet, l'écoute du signal synthétisé avec les 10 premières harmoniques modélisées à l'ordre 14 et à l'ordre 21 donne des résultats indiscernables perceptuellement. De plus, ces signaux sont indiscernables du signal synthétisé avec la modélisation à court terme de référence (voir plus de détails sur le protocole d'écoute des signaux dans la Section 4.3.3.3). En revanche, le signal synthétisé avec les 10 premières harmoniques modélisées à long

terme à l'ordre 9 est perceptivement assez nettement différent de ce signal de référence. La déformation de la phase est caractérisée par un son « reverbérant » et un peu « métallique » (on donne quelques éléments plus généraux sur l'écoute des signaux dans la sous-section suivante).

Au final, pour cet exemple, nous avons modélisé une longue section de parole comprenant $K = 408$ mesures de phase avec un modèle à long terme à seulement 15 coefficients en respectant les contraintes fixées ($R_{min} = 90\%$, $\alpha = 0{,}04$), ce qui illustre l'efficacité du modèle à long terme et de l'algorithme d'ajustement associé pour la modélisation des trajectoires de phase. Notons pour être tout à fait complets sur cette illustration du comportement de l'algorithme que, comme pour la modélisation à long terme des amplitudes au Chapitre 3, nous fixons à 20 le nombre maximum d'itérations sur les poids perceptifs. Ceci est justifié par le fait que moins de dix itérations du processus de pondération adaptative sont généralement suffisantes pour tester si le modèle peut être ajusté de façon convenable ou non.

En ce qui concerne la forme de l'onde synthétisée à partir des valeurs de phase modélisées à long terme (avec une interpolation à court terme linéaire des mesures d'amplitudes), il faut noter que la propriété de « *shape invariance* » du signal de synthèse est globalement réalisée pour les valeurs de R_{min} et α dans les gammes proposées. Ceci signifie que la forme de l'onde synthétisée est relativement fidèle à celle de l'originale, en dépit du faible nombre de coefficients décrivant la phase. Ceci est dû au fait que, dans cette étude, bien que le critère d'ajustement porte sur les fréquences, ce sont bien les mesures de phase qui sont modélisées par le modèle à long terme. Les relations de phasage entre les différentes composantes du signal sont donc globalement respectées par le modèle à long terme, avec cependant une certaine latitude dépendant de R_{min} et α. Selon les valeurs de R_{min} et α, des déphasages locaux des différentes harmoniques peuvent ainsi apparaître, comme illustré à la Figure 4.8, où $\alpha = 0{,}04$. Cependant, dans la plupart des cas, pour une valeur de α raisonnable, disons inférieure à 3-4%, de tels déphasages sont inaudibles. Ce point est plus amplement discuté à la Section 4.3.3.3.

4.3.3.3. Tests d'écoute

Comme pour la modélisation à long terme des amplitudes, la deuxième phase de résultats concerne les tests d'écoute informels que nous avons effectués pour évaluer qualitativement les effets perceptuels de la modélisation à long terme des trajectoires de phase. Le protocole de test et la technique de génération de signaux complets incluant parties voisées modélisées et parties non-voisées originales sont les mêmes que ceux décrits dans les sections précédentes (voir les Sections 3.3.3 et 4.2.3.1).

Le modèle à long terme utilisé pour modéliser les trajectoires de phase des signaux utilisés dans les tests reportés ici est le modèle MCDL. Des tests supplémentaires ont montré que, pour des paramètres R_{min} et α identiques, la qualité des signaux fournis par la modélisation à long terme avec les différents modèles était indiscernable (cependant les débits de coefficients sont différents, ce sera l'objet de la Section 4.3.3.4).

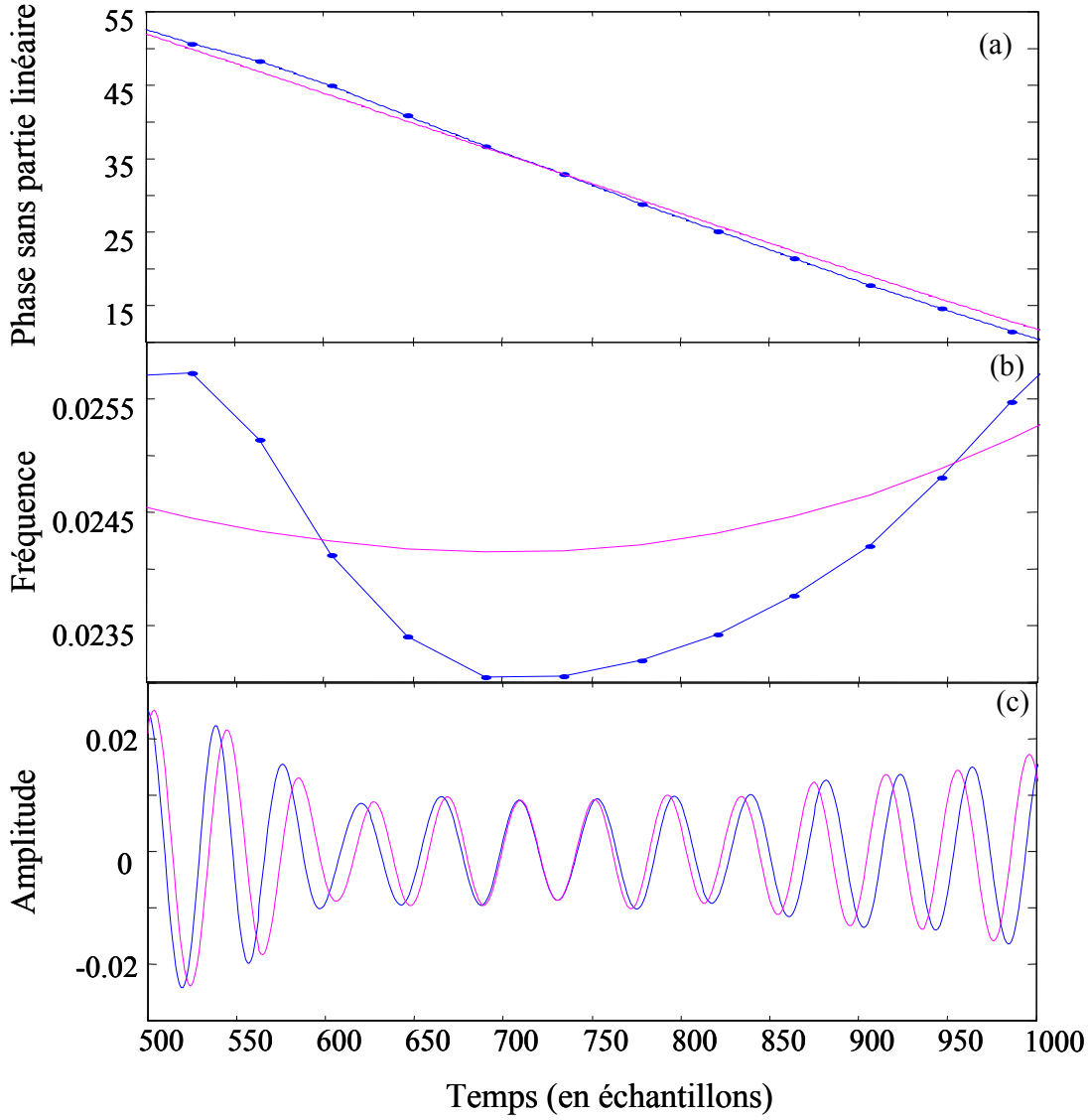

Figure 4.8 : Zoom sur une portion de 500 échantillons des trajectoires de la Figure 4.6, correspondant au segment 500-1000. (a) : Zoom sur les trajectoires de phase sans le terme linéaire. (b) : Zoom sur les trajectoires de fréquence (numérique). Pour ces deux figures, les mesures sont en bleu et le modèle est en magenta. (c) : Zoom sur les portions de signaux de synthèse correspondantes, avec la phase modélisée à long terme (en magenta) et la phase interpolée à court terme linéairement (en bleu) ; dans les deux cas, les mesures d'amplitude sont interpolées à court terme linéairement et on ne représente que la première harmonique synthétisée. Le signal de synthèse à long terme est d'abord en retard par rapport au signal de synthèse à court terme, puis il passe progressivement en avance. Ce phénomène pouvait être prévu à partir des valeurs de phase correspondantes à la figure (a). Un tel déphasage correspond bien à une modulation de fréquence (figure (b)). L'écoute de ce signal révèle que cette modulation est inaudible dans cet exemple, même si dans cette zone zoomée la fonction d'erreur de modélisation dépasse ponctuellement le modèle de seuil perceptif (voir Figure 4.6(c)).

Le résultat principal de ces nouveaux essais est que, comme pour les amplitudes, le modèle à long terme des trajectoires de phase peut fournir une qualité de synthèse semblable à celle obtenue avec interpolation à court terme des mesures de phase originales, selon le réglage des paramètres de la modélisation. Ce résultat a été obtenu d'une façon générale en prenant approximativement $R_{min} \approx 80\%$ et $\alpha \leq 0{,}03$. Par exemple, si le paramètre d'excursion de fréquence α est mis à 0,05, une différence entre les signaux synthétisés par le modèle à long terme et par le modèle court terme peut être audible pour un certain nombre des signaux de parole testés (même si R_{min} est plus grand que 80%). Pour $R_{min} = 80\%$ et $\alpha = 0{,}04$, on est en limite d'audibilité, et pour être plus surs de la qualité des signaux synthétisés, on choisit donc $\alpha \leq 0{,}03$. Il paraît assez difficile d'évaluer plus précisément la valeur de α, et donc du seuil perceptuel, car cette valeur semble dépendre d'autres contraintes de modélisation (par exemple la valeur de R_{min}), et probablement de certaines caractéristiques du signal (par exemple, la forme de la trajectoire d'amplitude de l'harmonique modélisée et/ou sa puissance moyenne). Cependant, il doit être souligné que la gamme des valeurs de α perceptuellement acceptables dans ce contexte de modélisation à long terme intrinsèquement non-stationnaire est significativement plus élevée que les valeurs rapportées dans [Zwicker et Feldtkeller, 1981] et [Demany & Semal, 1989] pour la modulation de fréquence sinusoïdale stationnaire : on passe d'une valeur de 0,0035 dans l'équation (4.3) à environ 0,03 dans notre étude, soit un facteur d'environ 10 entre les deux conditions !

Ce résultat suggère que la non-stationnarité des signaux de parole peut fournir un fort effet de masquage intrinsèque sur la déformation de phase, comparée au cas de signaux stationnaires plus perceptuellement sensibles. Autrement dit, les variations de phase naturelles (au sens de variations des fréquences associées) des signaux multi-composantes complexes comme la parole peuvent cacher des variations continues plus fines mais tout de même significatives (de l'ordre donc de 0,03) de ces paramètres dues à une modélisation (ou à un codage). Il faut cependant garder à l'esprit que l'erreur de modélisation est une modulation irrégulière, dans le sens où elle se caractérise par des ocillations non périodiques (loin s'en faut) autour de la trajectoire originale des fréquences. Elle n'est donc pas de la même nature que les oscillations de fréquence strictement périodiques utilisées en MFS dans le cadre stationnaire. Une large part de la différence de détection perceptive pourrait s'expliquer par cette différence.

Notons à ce propos que les résultats obtenus ici confirment les résultats connexes rapportés précédemment dans [Girin et Marchand, 2004] dans le cadre d'une application de tatouage de signaux audio. Dans cette étude, une modulation porteuse du tatouage était greffée artificiellement sur les trajectoires de fréquence de vrais signaux non stationnaires de parole et d'instruments de musique. Dans cette étude de tatouage, la modulation porteuse du tatouage était à la fois proche d'une sinusoïde et aléatoire ! En effet, les motifs porteurs de l'information binaire à tatouer étaient des périodes de cosinusoïdes surélevées modulés par le message binaire à encoder : on greffe un motif positif pour une valeur 1 d'un bit du message, et on greffe un motif négatif pour une valeur 0 (voir [Girin & Marchand, 2004] pour plus de détails). Comme les messages binaires testés sont aléatoires, on a donc une modulation totale de tatouage aléatoire mais structurée. Il est difficile alors de comparer l'impact perceptif d'une telle modulation par rapport au cas de notre étude à long terme (où l'erreur de modélisation n'est pas du tout périodique) et au cas de l'étude de MFS (où la modulation est

strictement sinusoïdale)[75]. Quoi qu'il en soit, les excursions de fréquence des modulations testées comme inaudibles dans [Girin & Marchand, 2004] étaient déjà très supérieures aux valeurs-référence reportées dans [Zwicker & Feldtkeller, 1981] et [Demany & Semal, 1989]. Elles étaient proches de la valeur de 0,03 constatée dans cette présente étude.

A l'issue de cette discussion, il est évident que des tests formels plus rigoureux doivent être effectués pour préciser de façon quantitative les liens entre l'erreur de modélisation à long terme des trajectoires de phase et la déformation perceptuelle conséquente. L'objectif est de caractériser de façon beaucoup plus fine les gammes acceptables des paramètres contraignants, en vue par exemple de leur exploitation systématique dans des systèmes de codage basés sur la modélisation à long terme (de tels systèmes de codage sont présentés au Chapitre 6). On a vu cependant à la Section 4.3.1 un aperçu des difficultés de ce type de tests pour une composante harmonique seule. Dans le cas de la parole ou d'instruments de musique, cette difficulté est encore démultipliée par le nombre de composantes en jeu et les possibles interactions perceptives entre ces composantes. La réalisation de ces tests dépasse donc le cadre de cette thèse du fait d'un manque de temps pour la mener à bien, mais elle peut tout à fait s'inscrire dans la continuation de ce travail.

4.3.3.4. Débit de coefficients

Comme pour la modélisation à long terme des amplitudes présentée au Chapitre 3, nous nous intéressons maintenant à caractériser les performances de la modélisation à long terme de la phase en terme de « pouvoir de compression ». Et comme pour les amplitudes, nous commençons cette section en donnant une première description quantitative de la relation entre l'ordre optimal P_i du modèle à long terme et le nombre de mesures K de la section de parole considérée.

Nous avons ainsi représenté sur la Figure 4.9, les valeurs moyennes du rapport K/P_i obtenues sur l'ensemble des 3500 sections de parole voisées de notre corpus, pour les dix premières harmoniques. Les écarts types sont aussi représentés sur cette figure. Le modèle est ici le MCDL[76] et on a fixé $R_{min} = 90\%$ et $\alpha = 0{,}02$, soit un réglage assez exigeant pour l'algorithme d'ajustement. On peut voir sur la Figure 4.9 que les valeurs du rapport K/P_i pour les différentes harmoniques sont très homogènes, de l'ordre de 8 à un peu plus de 9, avec un écart-type de l'ordre de 1,5 environ. Ceci reflète la cohérence des trajectoires de phase pour les différentes harmoniques dans le cas de sections de parole voisées, en tout cas pour les harmoniques de rang bas (1 à 10 dans ce cas). En cela, ces trajectoires, et les trajectoires des modèles à long terme associés, s'opposent aux trajectoires des amplitudes qui varient beaucoup en fonction des harmoniques,

[75] On peut difficilement dire que la modulation de tatouage de [Girin & Marchand, 2004] est une configuration intermédiaire entre la modulation résultant de la modélisation à long terme (non périodique) et la modulation de type MFS (périodique) car non seulement la nature de la modulation est différente, mais les ordres de grandeurs des fréquences de ces modulations peuvent être aussi assez différents : l'erreur de modélisation à long terme est une modulation « lente » (de l'ordre de quelques Hz) alors que par comparaison la modulation de tatouage de [Girin & Marchand, 2004] est beaucoup plus rapide (de l'ordre de la centaine de Hz). L'influence de ce dernier facteur reste aussi à approfondir.

[76] On ne donne ici que les valeurs de K/P_i pour le MCDL. La comparaison entre les différents modèles à long terme selon le critère de pouvoir de compression de données est présentée plus loin en termes de débits de coefficients et de robustesse numérique.

comme on l'a vu au Chapitre 3. Les valeurs moyennes du rapport K/P_i étant significativement supérieures à 1, avec un faible écart-type, on pourrait à nouveau poser une limite pour la recherche de l'ordre optimal dans l'algorithme d'ajustement. Une limite de l'ordre de $K/5$ semble raisonnable et cette limite peut ici être posée conjointement pour toutes les harmoniques considérées. Quoi qu'il en soit, ces résultats illustrent à nouveau la capacité de la modélisation à long terme à permettre intrinsèquement la compression de données. D'autant plus que, comme nous l'avons déjà mentionné, nous avons représenté sur la Figure 4.9 les valeurs correspondant à un réglage assez contraignant de l'algorithme d'ajustement ($R_{min} = 90\%$ et $\alpha = 0{,}02$). Des valeurs encore plus élevées sont obtenues en relachant un peu ces contraintes. Par exemple pour $R_{min} = 75\%$ et $\alpha = 0{,}03$, on obtient pour le MCDL des valeurs moyennes de K/P_i de l'ordre de 11 à 12. Plutôt que de donner des valeurs détaillées de ce rapport pour les différentes configurations d'expérimentations, nous allons maintenant à nouveau affiner la caractérisation du pouvoir de compression de la modélisation à long terme en donnant des résultats en termes de débits de coefficients.

Ainsi, de la même façon qu'à la Section 3.3.4 pour la modélisation à long terme des amplitudes, nous avons calculé le débit moyen de coefficients du modèle à long terme de phase. On rappelle que ce débit est défini comme le nombre moyen de coefficients du modèle à long terme par seconde. Il a été calculé sur la base de données complète et pour les 10 premières harmoniques. Les résultats sont donnés dans le Tableau 4.1 pour les voix de femmes et dans le Tableau 4.2 pour les voix d'hommes (dans les deux cas on a fixé $R_{min} = 75\%$ et on fait varier le facteur de modulation α entre 0,02 et 0,05 par pas de 0,01). Le modèle à long terme utilisé pour ces résultats est le MCDL.

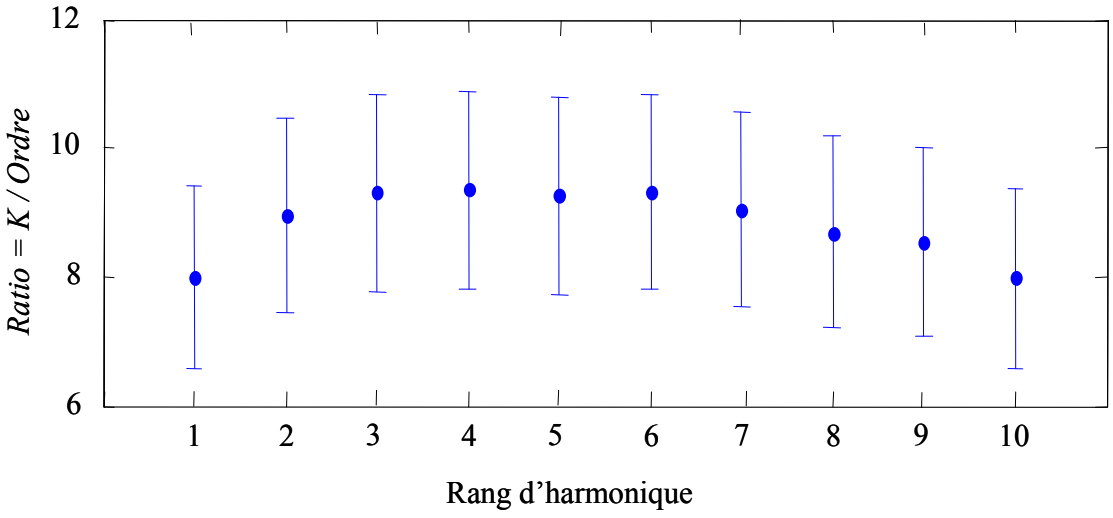

Figure 4.9 : Modélisation à long terme des trajectoires de phase basée sur un critère perceptif : valeurs moyennes et écarts-types du rapport K/P_i calculés sur l'ensemble de la base de données (3500 sections de parole voisées) pour les dix premières hermoniques. Le modèle utilisé pour cette expérience est le MCDL, avec $R_{min} = 90\%$ et $\alpha = 0{,}02$.

Comme dans le cas des amplitudes, on peut voir que le débit est plus grand pour les voix de femmes que pour les voix d'hommes, mais la différence est plus petite que dans le cas des trajectoires d'amplitude (il faut aussi se rappeler que le nombre d'harmoniques pour une voix de femme est généralement significativement inférieur à celui pour une voix d'homme). Par exemple, pour $\alpha = 0{,}03$, le taux moyen de coefficients sur les dix harmoniques est de 21,7 coefficients par seconde et par harmonique pour la voix féminine et de 19,4 coefficients par seconde et par harmonique pour la voix masculine. Le taux global pour l'ensemble des voix de femmes et d'hommes est alors de 20,5 coefficients par seconde et par harmonique. Comme on pouvait le prévoir, le taux de coefficients augmente quand α diminue, et inversement, puisque la précision de la modélisation augmente conjointement. Cependant, pour la gamme des valeurs de α examinées, ce taux reste toujours significativement inférieur au nombre de valeurs de mesure modélisées[77].

Contrairement aux résultats d'amplitude, pour chaque valeur de α, les taux de coefficients obtenus pour les dix harmoniques restent très similaires, ce qui confirme la même observation faite pour le rapport K/P_i. Ceci est confirmé par le calcul de l'écart type de l'ordre optimal calculé à travers les harmoniques pour chaque section de parole modélisée, puis moyenné sur la base de données. Avec $R_{min} = 75\%$ et $\alpha = 0{,}03$, nous avons obtenu un écart-type de 0,6 pour les voix de femmes et de 0,4 pour les voix d'hommes (des valeurs faibles sont également trouvées pour d'autres configurations de la modélisation à long terme). Ceci est probablement dû au fait que, bien que le modèle à long terme s'adapte à des trajectoires de phase, le critère perceptuel qui contrôle le processus est basé sur l'ajustement aux trajectoires de fréquence correspondantes. Or, nous faisons dans cette série d'étude l'hypothèse d'harmonicité. Les trajectoires de fréquences qui servent de référence au critère d'ajustement sont donc les multiples de la fréquence fondamentale. Par conséquent, nous obtenons pour les différents harmoniques, un ensemble de modèles à long terme de trajectoires de phase qui sont presque dans un même rapport harmonique. Toutefois, les différents modèles de phase ne sont pas exactement les multiples du modèle correspondant à la première harmonique (même à une constante près de phase à l'origine, voir la discussion de la Section 4.3.1). En effet, il ne faut pas oublier que ces différents modèles sont ajustés indépendamment aux différentes trajectoires de phase, elles-mêmes mesurées indépendamment les unes des autres (même si l'hypothèse d'harmonicité intervient dans la mesure). En pratique, les différentes harmoniques sont donc modélisées par des modèles « cohérents » et d'ordres très similaires (comme l'indique le faible écart-type à travers les harmoniques). Il est aussi probable que cette cohérence provient du fait qu'on se limite aux premières harmoniques ; en montant plus haut en fréquence, les composantes de bruit peuvent complexifier les trajectoires de fréquence et nécessiter d'augmenter l'ordre des modèles. Ce point sera rediscuté plus longuement dans le Chapitre 6.

[77] Pour la simplicité de la présentation, les résultats détaillés pour $R_{min} = 90\%$ ne sont pas présentés, mais on peut noter que le plus contraignant des réglages testés, soit $R_{min} = 90\%$ et $\alpha = 0{,}02$, donne un taux moyen (voix d'hommes et de femmes confondues) de 31,5 coefficients par seconde par harmonique (toujours avec le MCDL ; voir plus loin pour une comparaison des différents modèles à long terme).

	Harmonique	1	2	3	4	5	6	7	8	9	10
Voix	$\alpha = 0{,}02$	28,3	29,5	31,0	32,4	31,8	31,1	30,7	30,2	30,0	30,0
de	$\alpha = 0{,}03$	20,8	21,3	22,1	22,6	22,3	22,0	21,7	21,5	21,4	21,4
femmes	$\alpha = 0{,}04$	16,7	17,0	17,1	17,3	17,3	17,2	17,0	16,9	16,9	16,9
$R_{min} = 75\%$	$\alpha = 0{,}05$	14,4	14,5	14,6	14,6	14,6	14,6	14,5	14,4	14,5	14,4

Tableau 4.1 : Résultats de la modélisation à long terme des trajectoires de phase en terme de débit moyen de coefficients (nombre de coefficients par seconde) pour les voix de femmes, pour les dix premières harmoniques des signaux de la base de données, et pour les quatre valeurs de α testées (R_{min} = 75%, et le modèle est le MCDL).

	Harmonique	1	2	3	4	5	6	7	8	9	10
Voix	$\alpha = 0{,}02$	24,0	24,0	24,3	24,2	24,2	24,2	24,2	24,4	24,4	24,0
d'hommes	$\alpha = 0{,}03$	19,3	19,4	19,5	19,6	19,6	19,6	19,5	19,3	19,3	19,3
	$\alpha = 0{,}04$	15,9	15,9	15,9	16,1	16,0	16,0	16,0	15,9	15,9	15,9
$R_{min} = 75\%$	$\alpha = 0{,}05$	13,9	13,8	13,8	13,8	13,8	13,8	13,8	13,8	13,8	13,8

Tableau 4.2 : Résultats de la modélisation à long terme des trajectoires de phase en terme de débit moyen de coefficients (nombre de coefficients par seconde) pour les voix d'hommes, pour les dix premières harmoniques des signaux de la base de données, et pour les quatre valeurs de α testées (R_{min} = 75%, et le modèle est le MCDL).

Le taux global (femmes et hommes) moyen de 20,5 coefficients par seconde et par harmonique que nous avons obtenu (avec les valeurs des paramètres $R_{min} = 75\%$ et $\alpha = 0{,}03$) correspond à un facteur de gain d'environ 7 à 11 sur le nombre de paramètres de phase comparés au synthétiseur à court terme utilisant directement les phases mesurées (rappelons que le nombre moyen de paramètres de phase mesurés pour chaque harmonique est de 220 par seconde pour la voix féminine et de 140 par seconde pour la voix masculine). Comparé de façon plus juste aux codeurs utilisant un décalage fixe des trames d'analyse-synthèse de l'ordre de 10 à 20 ms (soit 50 à 100 trames/s), le gain est dans la gamme 2,5–5. Il est donc comparable (et même un peu supérieur) au même gain obtenu pour la modélisation à long terme des amplitudes du Chapitre 3 (rappel : on avait un gain dans la gamme 2–4). Ainsi, comme pour les paramètres d'amplitudes, le modèle à long terme des trajectoires de phase peut permettre un gain de codage significatif dans des applications telles que le codage de parole à (très) bas débit. En outre, pour une telle application, comme pour la modélisation à long terme des amplitudes, les ordres des modèles pourraient être encore diminués tout en préservant une qualité de synthèse acceptable.

Enfin, nous finirons ce tableau en remarquant que les débits atteints avec cette approche perceptive de la modélisation à long terme sont significativement inférieurs aux débits reportés en utilisant le critère de RSB dans la Section 4.2.3.2 (on a ici des débits de l'ordre de 20 coefficients par seconde et par harmonique, contre environ 80 à 90 coefficients par seconde et par harmonique dans la Section 4.2.3.2 pour obtenir un RSB proche de celui obtenu avec la synthèse à court terme classique). D'une part, ceci confirme dans ce cadre de modélisation à long terme que la notion de RSB n'est pas la

plus appropriée pour décrire la qualité perceptive d'un algorithme d'analyse-modélisation-synthèse. D'autre part, cette comparaison met nettement en valeur la contribution apportée par le critère perceptif que nous avons proposé.

4.3.4. Comparaison des différents modèles

Pour comparer les performances des différents modèles à long terme proposés à la Section 2.4 dans le cadre de la modélisation des trajectoires de phase, nous reprenons la méthodologie exposée à la Section 3.3.5 dans le cadre de la modélisation des trajectoires d'amplitude. En particulier, en plus des débits moyens de coefficients, nous reprenons ici les pourcentages P_{OK}, P_{NOK} et P_{PB} définis dans cette section. On rappelle que ces pourcentages sont respectivement les pourcentages des harmoniques de la base de données pour lesquelles le rapport cible R_{min} a été atteint par l'algorithme d'ajustement, non atteint sans problème numérique, et non atteint avec problème numérique, sous la contrainte $P_i < K/3$ (voir Section 3.3.5.2). On rappelle aussi que ces pourcentages sont calculés sur l'ensemble des dix premières harmoniques des 3500 sections de parole de la base de test, alors que les débits sont calculés sur les sections de cette base telles que tous les modèles testés ont vérifié $R \geq R_{min}$ avec $P_i < K/3$ pour les dix premières harmoniques. Ainsi, pour $\alpha = 0{,}03$ et $R_{min} = 0{,}75$, le nombre de sections sélectionnées selon ce critère est de 2330 sur 3500 (soit 67%). Enfin, toujours pour le débit, nous comparons les résultats obtenus par les différents modèles, MP, MCDL, MCSDL, et MCDP, avec l'approche optimale multi-modèles (MM). Rappelons que cette stratégie consiste à choisir le modèle gagnant (défini comme le modèle qui a besoin du nombre minimum de coefficients pour atteindre R_{min}) pour chaque section (voir Section 3.3.5).

Les résultats obtenus dans ces tests comparatifs en terme de débit de coefficients sont présentés sur la Figure 4.10, pour $R_{min} = 75\%$ et $\alpha = 0{,}03$. Les valeurs correspondantes moyennées sur les dix harmoniques sont aussi données dans le Tableau 4.3 qui contient aussi les valeurs du débit pour deux autres valeurs de α testées (0,02 et 0,04), toujours avec $R_{min} = 75\%$. Ces résultats sont assez différents de ceux obtenus dans le cas de la modélisation des amplitudes. On retrouve ainsi sans surprise la relative homogénéïté des valeurs entre les différentes harmoniques, déjà mentionnée dans cette section, alors que les débits augmentaient avec les harmoniques dans le cas des amplitudes. On a déjà expliqué cette homogénéïté de résultat dans le cas des phases : elle provient de l'homogénéïté de la forme des trajectoires des différentes harmoniques, pour les rangs considérés. Mais il ne s'agit pas de la seule différence par rapport aux résultats de la modélisation à long terme des amplitudes. Ainsi, on constate que le modèle polynôme est le plus efficace du point de vue du débit de coefficients par rapport aux autres modèles à long terme (on rappelle que dans le cas des amplitudes, c'était le MCD le plus performant). Pour $R_{min} = 75\%$, $\alpha = 0{,}03$, on peut voir sur la Figure 4.10 que les débits sont de l'ordre de 18 coefficients/s/harmonique pour le MP, alors que de façon assez surprenante, le MCDL est ici le moins performant, avec des débits de l'ordre de 21 coefficients/s/harmonique (les valeurs moyennes exactes, respectivement 18,1 et 21,1 sont données dans le Tableau 4.3). Les autres modèles, MCSDL et MCDP, donnent des performances intermédiaires, autour de 19 coefficients/s/harmonique, pour le même réglage de R_{min} et α. D'une façon générale, les trois valeurs de α testées fournissent des résultats cohérents, avec des valeurs de débit en rapport avec la

flexibilité de modélisation permise par α (les débits diminuent lorsque α augmente et vice-versa). Enfin, on peut noter que les résultats du MCDL sont globalement cohérents avec les résultats présentés dans le Tableau 4.1 et dans le Tableau 4.2 détaillant les débits du MCDL selon le sexe[78] des locuteurs/locutrices.

Les débits concernant la « stratégie optimale » de l'approche multi-modèles sont aussi donnés dans le Tableau 4.3. Les débits moyens de coefficients correspondant à cette stratégie optimale sont respectivement de 19,1, 16,6, et 14,9 coefficients/s/harmonique pour $\alpha = 0{,}02$, 0,03 et 0,04 (et $R_{min} = 75\%$; les valeurs harmonique par harmonique sont aussi représentées sur la Figure 4.10 pour le réglage $R_{min} = 75\%$ et $\alpha = 0{,}03$). Ces valeurs sont à comparer avec celles du MP (qui est ici le meilleur modèle en terme de débit) qui sont respectivement de 21,1, 18,1 et 16,1 coefficients/s/harmonique. Les gains de débit correspondants par rapport au MP sont donc respectivement de 2, 1,5, et 1,2 coefficients/s/harmonique, soit respectivement 9,5%, 8,3% et 7,5%. Bien évidemment, ces gains sont plus élevés si on les prend en référence au modèle MCDL qui est ici moins performant en terme de débit : on a alors respectivement 23,9%, 21,3%, et 22,4%. Rappelons tout de même que la stratégie multi-modèles exige des bits complémentaires pour coder le type de modèle à long terme choisi pour chaque section. Cependant, nous avons évalué au chapitre précédent que le taux additionnel est très faible, de l'ordre de 10 bits/s. Il reste donc à vérifier si ce coût additionnel est plus faible que les bits sauvés correspondant au gain sur les coefficients fourni par le choix du modèle MM. Ce travail serait à inclure dans une procédure de quantification des coefficients des modèles, procédure qui dépasse le cadre de travail de cette thèse.

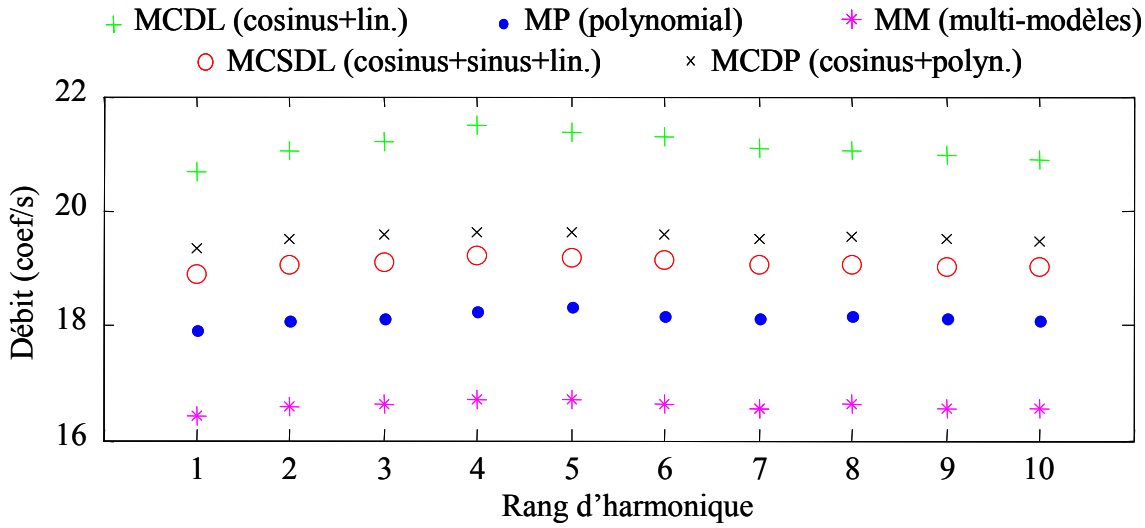

Figure 4.10 : Résultats comparatifs de la modélisation à long terme des trajectoires de phase des dix premières harmoniques par les différents modèles proposés (MCDL, MP, MCDP, MCSDL et MM) : débit moyen de coefficients (en nombre de coefficients par seconde) pour chaque harmonique, et pour $R_{min} = 75\%$ et $\alpha = 0{,}03$. Les résultats sont moyennés sur 2330 sections sélectionnées de la base de test (voir le texte).

[78] Ces nouveaux résultats du Tableau 4.3 ne sont pas exactement les moyennes inter-genre des Tableau 4.1 et 4.2 du fait qu'on utilise dans cette expérience comparative une sélection de sections de la base de données au lieu de l'ensemble des sections.

Modèle	MP	MCDL	MCSDL	MCDP	Multi-M
$\alpha = 0{,}02$	21,1	25,1	21,2	21,5	19,1
$\alpha = 0{,}03$	18,1	21,1	19,1	19,5	16,6
$\alpha = 0{,}04$	16,1	19,2	17,9	18,3	14,9

Tableau 4.3 : Résultats comparatifs de la modélisation à long terme des trajectoires de phase par les différents modèles proposés (MCDL, MP, MCDP, MCSDL et MM) : Débit moyen de coefficients (en nombre de coefficients par seconde par harmonique) moyennés sur les dix premières harmoniques des sections sélectionnées de la base de test, pour $R_{min} = 75\%$ et pour $\alpha = 0{,}02$, 0,03 et 0,04. Pour ces trois valeurs de α, on a respectivement 1133, 2330, et 2876 sections sélectionnées pour ces calculs (voir le texte).

Examinons à présent les pourcentages P_{OK}, P_{NOK} et P_{PB} qui sont reportés sur la Figure 4.11 pour le réglage $R_{min} = 75\%$ et $\alpha = 0{,}03$. Tout d'abord les résultats concernant le critère P_{OK} (le pourcentage des sections de parole pour lesquelles le rapport cible R_{min} est atteint pour les dix premières harmoniques) sont d'une façon générale assez élevés : on est dans un ordre de grandeur de 90-95%, ce qui, après les valeurs de débit, est à nouveau un bon indicateur des performances générales de l'approche à long terme et de l'algorithme d'ajustement, quel que soit le modèle utilisé. Ensuite, les résultats montrent que le modèle MCSDL est globalement le meilleur, suivi de près par le modèle polynôme MP. Le MCDP est le moins bon, mais ce qui est remarquable, ce sont les moindres performances du modèle MCDL par rapport au MP, pour les harmoniques de rang faible. On peut avoir jusqu'à 3% d'écart entre ces deux modèles.

Ceci est confirmé par les valeurs de P_{NOK} (le pourcentage des sections pour lesquelles R_{min} n'est pas atteint pour les dix premières harmoniques à cause de la contrainte $K/3$). Les valeurs pour le MCDL sont autour de 10% pour les harmoniques 2 à 5, contre environ 6% pour le MP. Néanmoins, en montant dans les harmoniques, ces valeurs se ressèrent et chutent en dessous de 5% (même si le MCDL reste un peu moins bon que le MP). Les valeurs de P_{NOK} pour les modèles MCSDL et MCDP se situent *grosso modo* de part et d'autre de celles du modèle MCDL (nous reviendrons sur les performances de ces modèles hybrides un peu plus loin).

En revanche, le MCDL (et le MCSDL) reste le(s) meilleur(s) modèle(s) selon le critère P_{PB} (le pourcentage des sections pour lesquelles R_{min} n'est pas atteint pour les dix premières harmoniques à cause d'un problème numérique). A nouveau, comme pour la modélisation des trajectoires d'amplitudes, ce(s) modèle(s) ne connaît (aissent) pas de problème numérique de conditionnement (du moins pour les sections vérifiant $P_i < K/3$). Ce n'est pas le cas des modèles avec des termes polynomiaux : les valeurs de P_{PB} pour le MP oscillent autour de 1,6% (et autour de 0,7% pour le MCDP). Cependant, ces valeurs sont beaucoup plus faibles que dans le cas des amplitudes du Chapitre 3. Ceci confirme le meilleur comportement général du MP sur les trajectoires de phases par rapport aux trajectoires d'amplitudes. Au final, les deux modèles de référence MCDL et MP sont très proches, car les différences opposées sur P_{NOK} et P_{PB} se compensent, en particulier pour les harmoniques élevées, où par conséquent les valeurs de P_{OK} pour ces deux modèles sont très proches (rappelons qu'on a $P_{OK} + P_{NOK} + P_{PB} = 100\%$).

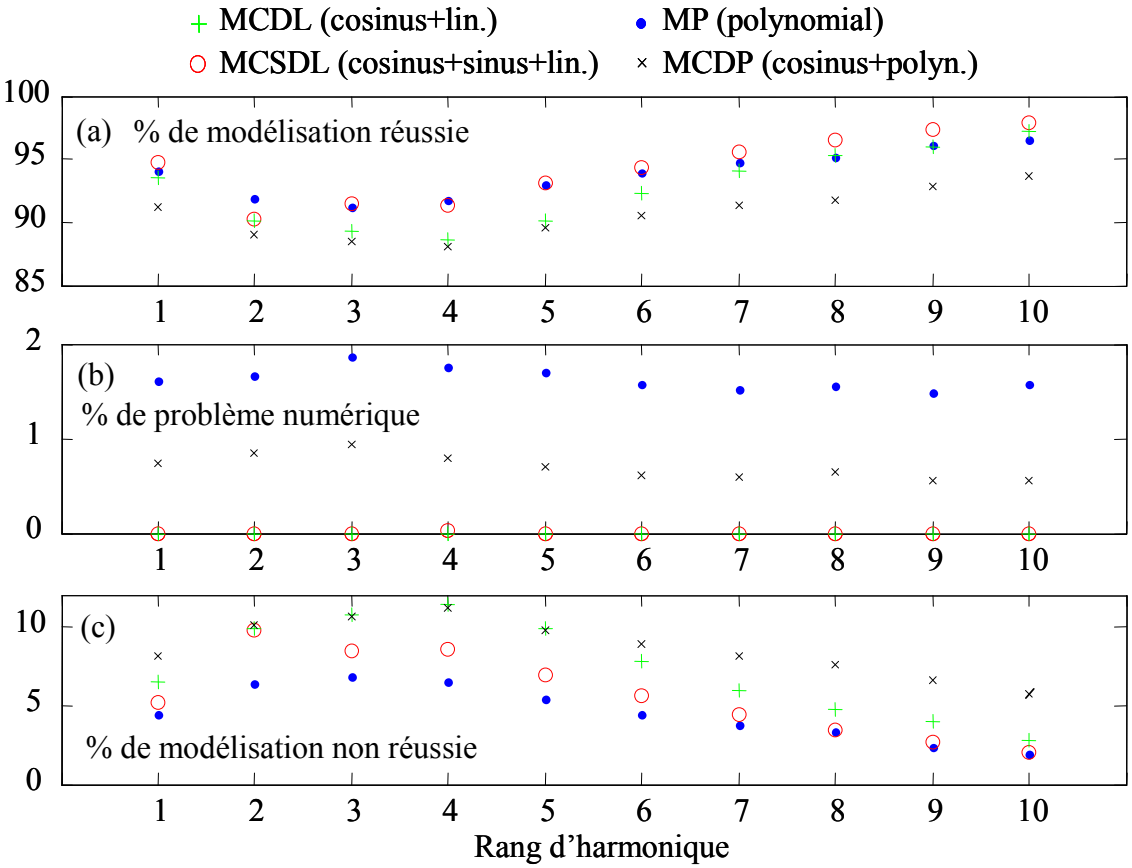

Figure 4.11 : Résultats comparatifs de la modélisation à long terme des dix premières trajectoires de phase par les quatre modèles proposés (MCDL, MP, MCDP, MCSDL), et avec $R_{min} = 75\%$, $\alpha = 0{,}03$, et avec la contrainte $P_i < K/3$: (a) pourcentage P_{OK} des harmoniques correctement modélisées ; (b) pourcentage P_{PB} des harmoniques non correctement modélisées à cause d'un problème numérique ; (c) pourcentage P_{NOK} des harmoniques non correctement modélisées sans problème numérique. Les pourcentages sont calculés sur les dix premières harmoniques des 3500 sections de la base de test.

Ainsi, d'une façon assez générale, le modèle polynomial MP semble avoir un meilleur comportement pour la modélisation à long terme des trajectoires de phase que pour celle des amplitudes. Compte tenu des éléments dont nous avons déjà parlé sur la « robustesse calculatoire » des modèles, ceci peut s'expliquer en partie par le fait que d'une façon générale, les trajectoires de phases sont plus « simples » et plus « régulières » que les trajectoires d'amplitude, comme on peut le voir sur l'exemple de la Figure 2.1. De fait, les débits moyens de coefficients sont plus faibles dans le cas des phases que dans celui des amplitudes, mais il faut être très prudent sur ce point : les débits ne dépendent pas seulement de l'allure des trajectoires mais ils dépendent aussi fortement des critères perceptifs qui sont de nature différente pour les phases et les amplitudes[79]. Simplement, on voit *a posteriori* que les ordres de modèles sont moins

[79] Compte tenu de cette différence assez remarquable des comportements du MP et du MCD(L) selon les paramètres considérés (phase ou amplitude), nous avons tenté de voir si chaque modèle pourrait « être spécialisé » dans un genre spécifique de section voisée, par exemple selon sa longueur. Ainsi, nous avons

élevés dans le cas des phases, ce qui permet éventuellement d'expliquer le meilleur comportement du modèle polynomial dont on a déjà évoqué la sensibilité numérique lorsque l'ordre augmente. Pour finir, on peut dire que le MCDL conserve de bonnes performances générales même si il est un peu décevant au niveau du débit, et que les performances des modèles hybrides MCSDL et MCDP sont assez différentes, le MCDP étant assez décevant, alors que le MCSDL se révèle comme un bon « challenger ».

4.4. Conclusion

Dans ce chapitre, nous avons réalisé la modélisation à long terme des trajectoires de phase de signaux de parole voisée. Nous avons appliqué pour cela les principes développés dans le Chapitre 2, en utilisant un critère perceptif basé sur une limitation de la déviation des trajectoires de fréquence associées aux trajectoires de phase. Cette étude a confirmé le bon comportement de l'algorithme d'ajustement, déjà observé dans le cas de la modélisation à long terme des amplitudes vue au Chapitre 3.

Par rapport à la modélisation à long terme des trajectoires d'amplitude, le cas des trajectoires de phase se caractérise par des trajectoires relativement homogènes entre les différentes harmoniques, tout du moins pour les rangs d'harmonique testés (de 1 à 10). Ceci provient bien sûr du caractère pseudo-périodique des signaux de parole testés (sections voisées). De plus, ces trajectoires qui reflètent la « mélodie de la parole » semblent généralement moins complexes que celles des amplitudes, ce qui a permis de mettre en évidence ici le bon comportement du modèle polynomial. Le modèle à base de cosinus discrets (ici MCDL) est comparativement moins performant que dans le cas des amplitudes, mais il reste un candidat valide compte tenu de sa robustesse calculatoire et de sa généralité.

A nouveau, avec un réglage adéquat des paramètres de l'algorithme d'ajustement, les signaux synthétisés avec l'approche à long terme sont de qualité très bonne, au moins aussi bonne que les signaux synthétisés avec les modèles habituels d'interpolation à court terme. Ceci semble confirmer les bases et l'intérêt du critère perceptif à long terme proposé. Un réglage tel que $R_{min} = 75\%$ et $\alpha = 0{,}03$ semble ici un bon compromis entre qualité perceptive et débit de coefficients. En effet, pour une telle configuration, les débits de coefficients sont de l'ordre de 20 coefficients/s/harmonique (16,6 coefficients/s/harmonique avec la stratégie optimale multi-modèles). Ceci représente une gamme de gain de l'ordre de 7 (pour une voix d'homme avec le MCDL) à 13 (pour une voix de femme avec la stratégie multi-modèles) par rapport à l'interpolation à court terme des paramètres mesurés. Par rapport au débit de coefficient usuel des codeurs à taille et décalage de trame fixe (10 à 20 ms), la gamme de gain est de l'ordre de 2,5 à 6.

---

tracé une série d'histogrammes montrant la répartition du modèle « gagnant » (c'est-à-dire d'ordre minimal) pour toutes les sections voisées du corpus. Même si le modèle polynôme s'avère être généralement sélectionné pour des sections voisées relativement courtes et le modèle MCD(L) s'avère être généralement choisi pour les sections voisées plus longues, les histogrammes sont suffisamment confus pour qu'on ne puisse pas dégager de règles générales pour le choix du modèle. En revanche, il apparaît sur ces histogrammes que les problèmes de conditionnement numérique des modèles avec termes polynomiaux apparaîssent généralement pour des sections de parole « assez longues ». Ces remarques sont communes à la modélisation des trajectoires de phase et d'amplitude.

Enfin, pour finir ce chapitre, on peut rappeler une propriété importante de la modélisation de la phase telle qu'elle a été proposée dans cette étude. Si on veut coder exactement les valeurs de phase du signal, et non seulement les valeurs de fréquence afin de préserver la forme d'onde (la propriété de *shape-invariance*), seul un paramètre additionnel pour chaque harmonique est nécessaire dans ce cadre par rapport à la modélisation des trajectoires de fréquence. Ceci provient du fait que le modèle à long terme des fréquences est la fonction dérivée du modèle à long terme de la phase. Ainsi, la modélisation à long terme de la phase perceptuellement pondérée en fréquence apparaît comme une technique à la fois simple et efficace pour étudier, encoder et préserver efficacement et à un moindre coût la dynamique temporelle de la parole voisée inscrite dans ce paramètre.

Chapitre 5

5. Généralisation de l'approche en 2D : Modélisation à long terme des enveloppes spectrales

Dans les chapitres précédents nous avons étudié la modélisation à long terme des paramètres du modèle sinusoïdal, amplitudes (au Chapitre 3) et phases (au Chapitre 4). Dans ces études, un modèle à long terme unique a été employé pour représenter chacune des trajectoires des paramètres en question sur chaque section de parole entièrement voisée. En d'autres termes, chacune des trajectoires des paramètres d'amplitude (ou bien chacune des trajectoires des paramètres de phase) est modélisée à long terme séparément[80]. Ainsi, un algorithme itératif incluant des contraintes perceptives a permis d'estimer conjointement l'ordre optimal du modèle et ses coefficients indépendamment pour chaque paramètre.

Dans ce chapitre nous étendons et raffinons cette approche en terme de complexité (et de compacité !) de la modélisation, en ajoutant une nouvelle étape de modélisation le long de l'axe des fréquences avant de considérer l'axe du temps : nous modélisons d'abord l'enveloppe des amplitudes par un premier modèle en cosinus discrets (voir Section 2.4.2), puis nous appliquons un deuxième modèle en cosinus discrets selon l'axe du temps pour modéliser à long terme la trajectoire des coefficients résultant de la modélisation d'enveloppe. On voit donc apparaître particulièrement clairement ici le principe d'un modèle de modèle mentionné en introduction de ce document : on modélise à long terme des trajectoires de paramètres eux-mêmes issus d'un modèle. Ce processus aboutit à ce qu'on appelle un modèle à long terme en deux dimensions (2D), ou modèle « spectro-temporel », des enveloppes du signal.

Notons que dans cette approche 2D, nous considérons seulement la modélisation des paramètres d'amplitude, pour des sections de parole voisées, et nous ne traitons pas de la modélisation des phases, ni des sections non voisées. La raison principale est que, comme précisé dans la Section 2.6, un des buts sous-jacents de cette étude est de fournir une représentation particulièrement « compacte » du signal, exploitable par exemple dans un codeur à très bas débit. Dans un tel codeur, l'information de phase est

[80] En réalité, ceci n'est pas tout à fait exact. En effet, il existe un lien entre les différentes harmoniques dans le processus de modélisation à long terme. Dans le cas des amplitudes, le seuil de masquage qui sert de critère perceptif dans l'algorithme d'ajustement du modèle est calculé pour chaque harmonique en tenant compte de l'influence des harmoniques avoisinants (Section 3.1). Dans le cas des phases, le seuil perceptif est calculé en faisant l'hypothèse d'harmonicité. On entend ici par « séparément » le fait qu'un modèle à long terme est utilisé pour chaque paramètre modélisé.

volontairement réduite à la trajectoire de la fréquence fondamentale sur les sections considérées (voir la la Section 4.1 pour une discussion sur la pertinence des divers niveaux d'information de phase). Notons que d'une part, la trajectoire de cette fréquence fondamentale peut être modélisée à long terme en parallèle, d'une façon similaire à celle étudiée au Chapitre 4. Et d'autre part, comme nous le discuterons plus en détails au Chapitre 6, la modélisation en 2D du spectre d'amplitude est un principe qui peut s'appliquer très facilement aux sections non voisées de la parole, plus facilement que la modélisation à long terme en une seule dimension comme vue au Chapitre 3.

Notons enfin pour finir la présentation générale de cette nouvelle approche, que nous proposons comme dans les chapitres précédents un algorithme itératif pour ajuster de façon optimale le modèle aux mesures, toujours en tenant compte d'un critère de type perceptif. Cet algorithme est une adaptation au cas 2D de l'algorithme présenté au Chapitre 3 pour la modélisation à long terme en une dimension des paramètres d'amplitude.

Ce chapitre est organisé comme suit. Dans la Section 5.1, nous présentons le principe de modélisation de l'enveloppe spectrale le long de l'axe des fréquences et nous précisons l'intérêt de l'approche 2D en la reliant aux problèmes plus généraux de la modélisation d'enveloppe spectrale et de la conversion de dimension de données. La Section 5.2 est dédiée à l'étude d'un modèle à long terme pour les trajectoires temporelles des paramètres d'enveloppe obtenues dans la section précédente. L'algorithme d'ajustement et d'estimation des ordres optimaux pour le modèle d'enveloppe spectrale et le modèle à long terme associé sera étudié à la Section 5.3. L'ensemble des expériences que nous avons menées pour évaluer cette approche est donné à la Section 5.4.

5.1. Modélisation de l'enveloppe spectrale

5.1.1. Principe

Considérons une section de parole voisée composée de K pseudo-périodes de signal, sur laquelle nous voulons appliquer cette nouvelle modélisation 2D. Pour cela, nous repartons d'abord du processus d'analyse pitch-synchrone décrit à la Section 3.3.1.2. De façon cohérente aux notations utilisées dans les chapitres précédents, cette analyse fournit K jeux de vecteurs de paramètres d'amplitudes[81] :

$$\boldsymbol{A}^k = \left[A_{1,k} \;\; A_{2,k} \;\; \ldots \;\; A_{I_k,k}\right]^t \qquad \text{pour} \qquad 1 \le k \le K \tag{5.1}$$

Notons que la taille I_k de ces vecteurs peut être variable d'une période à l'autre, selon la fluctuation de la fréquence fondamentale qui détermine directement le nombre d'harmoniques.

[81] On rappelle que les données définies temporellement sont représentées par des vecteurs lignes ou par des lignes de matrices, alors que les données de rang différent à un temps donné (par exemple ici les harmoniques) sont représentées par des vecteurs colonnes ou par des colonnes de matrices. Pour les vecteurs, pour différencier clairement les deux dimensions dans ce chapitre, on met l'index en indice dans le premier cas, et en exposant dans le second cas.

Dans nos études précédentes, nous avons considéré la modélisation des amplitudes directement selon l'axe du temps. Donc, à la fin du processus d'analyse, les amplitudes étaient ré-ordonnées comme I_k jeux de K amplitudes comme suit :

$$\boldsymbol{A}_i = \left[A_{i,1}\ A_{i,2}\ ...A_{i,K}\right] \qquad \text{pour} \qquad 1 \le i \le I_k \tag{5.2}$$

Le problème de la taille variable a été résolu de façon très pragmatique en ne considérant que des trajectoires « complètes », c'est-à-dire ne sortant jamais au-delà de la fréquence de Nyquist (ainsi la modélisation des dix premières trajectoires a été étudiée en détail dans les chapitres précédents).

Alternativement, dans cette nouvelle étude en 2D, les amplitudes sont d'abord considérées avec un ordonnancement selon l'axe des fréquences selon (5.1). Cet ordonnancement est en fait l'ordonnancement habituel dans les études usuelles en analyse spectrale de spectres de raies, et il correspond d'ailleurs au format fournit directement par la procédure d'analyse que nous utilisons ici (voir Section 1.2.1.2). Etant donnés ces K jeux de paramètres d'amplitude, la première étape de la modélisation 2D consiste à remplacer chacun d'entre eux par un modèle d'enveloppe spectrale. Dans cette étude, ce modèle un des modèles utilisés pour la modélisation à long terme « monodimensionnelle » (1D) des chapitres précédents, c'est-à-dire le modèle en cosinus discrets (MCD) présenté à la Section 2.4.2. Ceci n'a rien d'étonnant, bien au contraire, en quelque sorte on retourne à la source : dans la Section 2.4.2, nous avons vu que ce type de modèle avait précisément été proposé dans le cadre de la modélisation d'enveloppe spectrale bien avant notre utilisation dans la modélisation à long terme [Cappé *et al.*, 1995] [Galas & Rodet, 1990, 1991] [Campedel-Oudot, 1998] [Campedel-Oudot *et al*., 2001]. On rappelle que ce modèle est appliqué sur les mesures d'amplitude converties en échelle log, ce qui l'identifie avec celui connu sous l'appellation Cepstre Discret (voir Chapitre 2, Section 2.4.2). On a donc ici, de façon similaire à l'équation (2.2) :

$$\hat{A}^k(f) = \sum_{m=0}^{M} d_{m,k} \cos(m2\pi f) \tag{5.3}$$

où M est un nombre entier positif qui est l'ordre du modèle d'enveloppe spectrale. Bien que le nombre d'harmoniques puisse varier d'une trame à l'autre, dans cette présente étude M a une valeur fixe pour toutes les trames d'une section voisée modélisée à long terme en 2D. Nous justifions ce choix dans la suite et nous verrons comment M est estimé pour chaque section. Pour un M donné, le vecteur des coefficients du modèle est estimé pour chaque trame I de la section par une procédure de type *WMMSE*, similaire à celles que nous avons utilisées dans les études des chapitres précédents. On cherche ainsi à minimiser l'erreur suivante :

$$\varepsilon_k = \sum_{i=1}^{I_k} w_{i,k} \left\| A_{i,k} - \hat{A}^k(i\omega_{0,k}) \right\|^2 \tag{5.4}$$

Les poids de l'équation (5.4) mis au carré sont ici rangés sur la diagonale d'une matrice $I_k \times I_k$ diagonale notée $\boldsymbol{W}_k$. Nous verrons dans l'algorithme de la Section 5.3 que ces poids sont estimés à partir de contraintes perceptives. Ces contraintes sont similaires à

celle utilisées au Chapitre 3 pour la modélisation à long terme « monodimensionnelle » des amplitudes : elles sont basées sur le même seuil de masquage fréquentiel, mais elles sont utilisées ici de façon plus classique le long de la dimension fréquentielle (voir une illustration schématique à la Figure 5.1).

Notons ici $\boldsymbol{H}_k$ la matrice $I_k \times (M+1)$ de terme général $h_{i,m} = \cos(mi\omega_0^k)$,

$$\boldsymbol{H}_k = \begin{bmatrix} 1 & \cos(\omega_{0,k}) & \cos(2\omega_{0,k}) & \cdots & \cos(M\omega_{0,k}) \\ 1 & \cos(2\omega_{0,k}) & \cos(4\omega_{0,k}) & \cdots & \cos(M2\omega_{0,k}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \cos(I_k\omega_{0,k}) & \cos(2I_k\omega_{0,k}) & \cdots & \cos(MI_k\omega_{0,k}) \end{bmatrix} \tag{5.5}$$

Notons par ailleurs $\boldsymbol{D}^k$ le vecteur de coefficients réalisant l'ajustement optimal du MCD pour chaque vecteur d'amplitudes mesurées (en échelle log) :

$$\boldsymbol{D}^k = \begin{bmatrix} d_{0,k} & d_{1,k} & d_{2,k} & \ldots & d_{M,k} \end{bmatrix}^t \tag{5.6}$$

Si on suppose que $M < I_k - 1$, ce vecteur est donné ici par :

$$\boldsymbol{D}^k = (\boldsymbol{H}_k^t \boldsymbol{W}_k \boldsymbol{H}_k)^{-1} \boldsymbol{H}_k^t \boldsymbol{W}_k \boldsymbol{A}^k \qquad \text{pour} \qquad 1 \leq k \leq K \tag{5.7}$$

En effet, les équations (5.7) et (5.4) sont les mêmes que celles utilisées dans les chapitres précédents pour l'analyse des paramètres sinusoïdaux (Section 1.2.1.2) et pour l'ajustement du modèle à long terme (Section 2.5.4), à la transposition près. Cette transposition de tous les éléments en jeu est due au fait qu'on considère ici les données selon l'axe fréquentiel (des vecteurs colonnes) au lieu de l'axe du temps (des vecteurs lignes). De même, la matrice $\boldsymbol{H}_k$ joue le même rôle que la matrice du modèle de l'équation (2.19) : les valeurs d'indice des centres de trame d'analyse sont remplacées ici par les positions des harmoniques. Dans les deux cas, il s'agit des positions (temporelles ou fréquentielles) des mesures.

5.1.2. Intérêt dans le cadre de la modélisation à long terme en 2D

A la fin du processus d'analyse et de modélisation de l'enveloppe spectrale, on obtient donc K vecteurs de paramètres de cepstres discrets de taille $M+1$. Ces vecteurs peuvent être rangés dans une matrice $\boldsymbol{D}$ définie par :

$$\boldsymbol{D} = \begin{bmatrix} d_{0,1} & d_{0,2} & \cdots & d_{0,k} & \cdots & d_{0,K} \\ d_{1,1} & d_{1,2} & \cdots & d_{1,k} & \cdots & d_{1,K} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,k} & \cdots & d_{2,K} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ d_{M,1} & d_{M,2} & \cdots & d_{M,k} & \cdots & d_{M,K} \end{bmatrix} \tag{5.8}$$

Comme la seconde étape de modélisation en 2D est la modélisation de la trajectoire temporelle des coefficients du modèle d'enveloppe, on voit alors tout l'intérêt de cette première étape de modélisation dans notre approche à long terme :

- Elle permet d'abord de ramener le problème de la modélisation à long terme sur des données de taille fixe au cours du temps. En effet, on a déjà mentionné qu'un des problèmes principaux pour encoder les paramètres sinusoïdaux (à long terme aussi bien qu'à court terme d'ailleurs) est la taille variable du jeu de paramètres d'une trame à l'autre en fonction de la variation de la fréquence fondamentale du signal. Il se pose ainsi le problème de gestion des « naissances » et « morts » de ces trajectoires dans la zone des fréquences les plus hautes (voir Section 1.3.2.1). Par conséquent, décrire l'évolution des amplitudes au cours du temps par un nombre de paramètres fixe permet de s'affranchir de ce problème[82].
- Elle permet également une représentation parcimonieuse de l'information. En effet, le nombre de paramètres du modèle sinusoïdal est généralement grand par rapport à l'objectif d'un codage très bas débit. On doit donc chercher non seulement à rendre fixe la taille des données à coder, mais encore à diminuer cette taille autant que possible pour compresser en amont l'information avant de considérer notre problématique de modélisation à long terme (et la problématique sous-jacente de son encodage).

Il est intéressant de mentionner qu'une approche générale à ce problème de changement de la dimension des données sous l'angle des transformations linéaires peut être trouvée dans [Li *et al.*, 2001] sous l'appellation de transformation non carrée (NST pour *Non Square Transform*) des données. Le terme NST dénote en fait tout un ensemble de transformations linéaires de dimension $M \times N$ où N désigne la taille (variable) des données d'origine et M la taille (fixe) des données transformées. Dans la mesure où ces transformations sont linéaires, et où les modèles que nous utilisons dans notre étude sont des combinaisons linéaires de fonctions, il y a identification entre la notion de transformation et celle de modélisation : on peut ainsi retrouver dans les matrices de transformation des valeurs discrétisées des fonctions des modèles. Notons qu'on devrait

[82] On retrouve la même problématique dans le domaine spécifique de la quantification pour les codeurs sinusoïdaux/harmoniques à court terme. En effet, la méthode optimale d'encodage à dimension donnée étant la quantification vectorielle (QV), la méthode optimale globale pour quantifier des données à taille variable serait une QV à dimension variable. Or, celle-ci s'avère coûteuse en données et en complexité d'implémentation car elle réclame l'élaboration de dictionnaires de prototypes pour chaque dimension [Gray & Gersho, 1992]. Là aussi, une solution pourrait consister à se ramener à des données et une QV de taille fixe, en passant par une étape de modélisation. Par ailleurs, ce problème de taille variable des données originelles fait que l'étude en 2D que nous proposons ici se démarque assez nettement de celle de [Farvardin & Laroia, 1989] déjà citée à la Section 2.2. En effet, la DCT-2D proposée par ces auteurs est une transformation bidimensionnelle s'appliquant directement sur des matrices de taille fixe prédéfinie, en l'occurrence des matrices de coefficients LSF qui sont de taille fixe selon les deux dimensions (fréquentielle et temporelle). Par rapport à cette DCT-2D, du fait de la nature des données spectrales que nous manipulons (les amplitudes harmoniques), nous sommes contraints de procéder en deux temps : une première modélisation en fréquence, puis une deuxième modélisation en temps. Remarquons de plus qu'appliquer une DCT-2D similaire à celle de [Farvardin & Laroia, 1989] après régularisation de la dimension fréquentielle de nos données par le premier modèle MCD n'apporterait rien par rapport à notre approche, la corrélation fréquentielle des amplitudes étant précisément déjà extraite par ce premier modèle. Tout cela justifie que nous proposions une approche fonctionnant en deux temps (d'abord fréquentielle puis temporelle) dans le cadre du modèle sinusoïdal.

parler plus rigoureusement d'ensembles de transformations au pluriel plutôt que d'un seul ensemble pour chaque type de données, puisque ces données étant de taille variable N, il faut dans chaque ensemble une transformation $M \times N$ pour chaque taille possible. Les auteurs de [Li *et al.*, 2001] citent différents types de transformations adaptées à différents objectifs (par exemple, la transformée de Karhunen-Loeve, la transformée de Hartley, la DCT,…) et dérivent des résultats intéressants sur l'existence et l'optimalité des transformations inverses associées permettant de revenir à l'espace de départ à partir des données transformées. Ils montrent notamment que la transformation la plus efficace en terme de minimisation de l'erreur de modélisation et de codage est une transformation en cosinus discret de type II (TCD-II) similaire à ce que nous avons appelé dans ce document par modèle en cosinus discrets (MCD)[83]. Il est donc tout à fait cohérent que nous retrouvions ce modèle dans notre approche 2D pour réaliser à la fois la modélisation à long terme selon la dimension temporelle et la modélisation de l'enveloppe spectrale selon la dimension fréquentielle.

En toute logique, l'étape suivante de cette étude en 2D est la modélisation selon l'axe temporel des trajectoires des coefficients du MCD résultant de la première étape de modélisation. C'est ce que nous abordons dans la section suivante.

5.2. Modélisation des trajectoires d'enveloppe

Une fois que la modélisation spectrale a été faite pour toutes les trames de la séquence de parole considérée (section voisée), la seconde étape de modélisation est la modélisation de la trajectoire temporelle des coefficients $d_{m,k}$ du modèle d'enveloppe. Ceci est illustré sur la Figure 5.1 : les coefficients du premier modèle sont maintenant considérés le long de l'axe du temps comme M+1 vecteurs de dimension K (comme on l'a fait directement pour les valeurs d'amplitudes dans le Chapitre 3) :

$$\boldsymbol{D}_m = \left[d_{m,1} \;\; d_{m,2} \;\; \ldots \;\; d_{m,K} \right] \qquad \text{pour } 0 \le m \le M \tag{5.9}$$

On applique alors sur ces vecteurs les principes de la modélisation à long terme en utilisant le MCD, de façon tout à fait analogue à ce qui a été réalisé directement sur les amplitudes spectrales au Chapitre 3. On a donc :

$$\hat{d}_m(n) = \sum_{p=0}^{P_m} c_{m,p} \cos(p\pi \frac{n}{N}) \qquad \text{pour } 0 \le m \le M \tag{5.10}$$

Comme toujours dans nos études, les coefficients du modèle à long terme sont estimés par minimisation des moindres carrés. L'erreur quadratique moyenne pondérée est définie ici par (rappelons que les indices n_k sont les centres des K trames d'analyse) :

$$\varepsilon_m = \sum_{k=1}^{K} w_{m,k} \left| d_{m,k} - \hat{d}_m(n_k) \right|^2 \qquad \text{pour } 0 \le m \le M \tag{5.11}$$

[83] La différence réside dans un décalage de la valeur ½ des indices et un terme multiplicatif constant par ligne ou colonne des matrices en jeu.

Notons à nouveau $\boldsymbol{M}_m$ la matrice $(P_m+1)\times K$ de terme général $m_{p,k} = \cos(p\pi m_k/N)$, comme dans l'équation (2.18), et notons $\boldsymbol{W}_m$ la matrice $K\times K$ diagonale qui contient sur sa diagonale les poids de (5.11) mis au carré[84]. On verra plus tard comment ces poids sont ajustés pour tenir compte du comportement le long de l'axe temporel de l'erreur de modélisation durant l'ajustement du modèle aux données. En supposant que $P_m<K-1$, le vecteur des coefficients du modèle à long terme est donné de façon similaire à (2.19) (voir la Section 2.5.4) par :

$$\boldsymbol{C}_m = \boldsymbol{D}_m\boldsymbol{W}_m\boldsymbol{M}_m^t(\boldsymbol{M}_m\boldsymbol{W}_m\boldsymbol{M}_m^t)^{-1} \qquad \text{pour} \qquad 0 \le m \le M \tag{5.12}$$

Comme dans le cas de la modélisation à long terme 1D, l'ordre optimal du modèle à long terme P_m pour un bon *fitting* des données dépend des caractéristiques de la section de parole considérée. Il dépend aussi potentiellement du rang m du coefficient du modèle d'enveloppe. Par souci de simplicité, dans cette étude, nous choisissons de prendre le même ordre $P_m = P$ pour tous les vecteurs $\boldsymbol{D}_m$, pour $m = 0$ à M. Ce choix correspond aussi à une volonté de garder une représentation de l'information la plus parcimonieuse possible dans l'optique d'une application de la méthode au codage à très bas débit : dans une telle application, l'information encodant les ordres des modèles appliqués à chaque vecteur temporel des coefficients d'enveloppe doit être transmise en plus de l'information encodant la valeur des coefficients eux-mêmes. Prendre le même ordre pour les M composantes des vecteurs d'enveloppe réduit donc très significativement cette information et son coût de transmission. La méthode peut bien sûr être raffinée en gardant la possibilité d'un ordre spécifique pour chaque rang de coefficient.

Ce choix d'un ordre P unique pour toutes les trajectoires des vecteurs de coefficients d'enveloppe permet d'écrire sous une forme matricielle très compacte l'équation d'estimation des coefficients du modèle à long terme. En effet, si on impose de plus que les poids $\boldsymbol{W}_m$ sont appliqués de façon identique sur chacune des M trajectoires, ce que l'on vérifiera par la suite, on a :

$$\boldsymbol{C} = \boldsymbol{D}\boldsymbol{W}\boldsymbol{M}^t(\boldsymbol{M}\boldsymbol{W}\boldsymbol{M}^t)^{-1} \tag{5.13}$$

avec $\boldsymbol{M}_m = \boldsymbol{M}$, une unique matrice de modèle d'ordre P pour $m = 0$ à M, $\boldsymbol{W}_m = \boldsymbol{W}$ une unique matrice de poids correspondante, et la matrice $\boldsymbol{D}$ des coefficients des modèles d'enveloppe est définie en (5.8). La matrice résultante $\boldsymbol{C}$ contient sur chaque ligne m le vecteur de coefficients $\boldsymbol{C}_m$ du modèle à long terme encodant la trajectoire du vecteur ligne de coefficients d'enveloppe $\boldsymbol{D}_m$. Notons qu'à partir de maintenant, on peut considérer $\boldsymbol{C}$ comme l'ensemble des coefficients d'un unique modèle à long terme 2D.

Comme nous l'avons fait dans les chapitres précédents, après avoir présenté le principe de la modélisation à long terme (ici en 2D), nous présentons maintenant l'algorithme que nous utilisons pour estimer conjointement l'ordre P du modèle à long terme, les poids qui sont utilisés dans l'ajustement itératif, la matrice cible $\boldsymbol{C}$ des coefficients du modèle 2D, et également dans cette nouvelle étude l'ordre M du modèle d'enveloppe.

[84] Notons qu'on utilise pour cette matrice de poids la même notation que dans la section précédente pour ne pas multiplier ces notations. Mais cette matrice de poids est ici définie pour une pondération temporelle (pour chaque trajectoire de paramètre) au lieu de fréquentielle. Ceci est en fait intrinsèquement marqué par le fait qu'on utilise ici l'indice m dénotant le rang d'un paramètre au lieu de l'indice temporel k utilisé dans la section précédente.

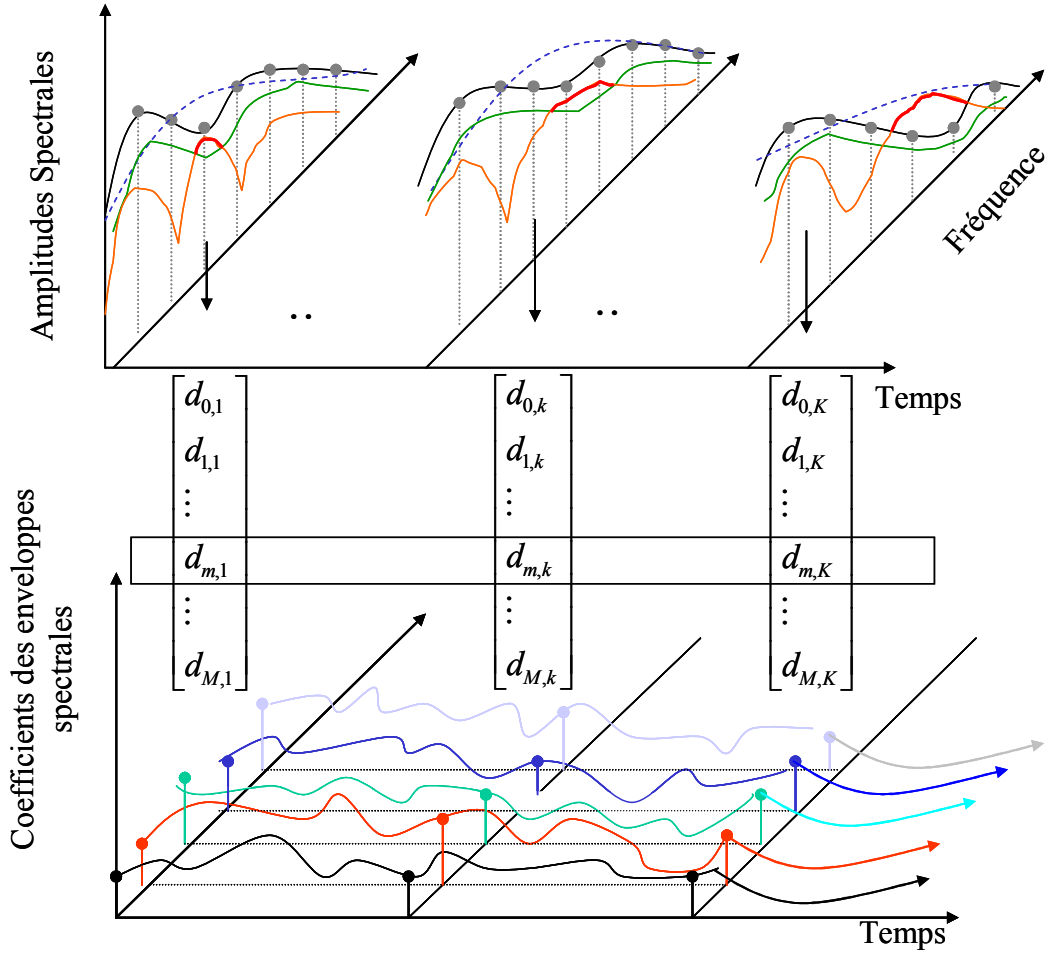

Figure 5.1 : Schéma de principe de la modélisation en 2D des amplitudes spectrales. En haut : première modélisation MCD selon l'axe des fréquences. En gris/noir : points de mesures originaux et spectres originaux ; en vert : le seuil de masquage ; en bleu : le spectre modélisé ; en orange : l'erreur de modélisation ; les zones où l'erreur de modélisation dépasse le seuil de masquage sont accentuées en rouge. En bas : deuxième modélisation à long terme selon l'axe du temps ; les courbes en différents couleurs représentent la modélisation de la trajectoire des différents coefficients MCD des enveloppes spectrales. Pour chaque section de parole modélisée, la taille des vecteurs de paramètres d'amplitude est variable mais la taille des vecteurs de paramètres MCD résultants est fixe (voir Section 5.1.2).

5.3. Estimation des ordres et algorithme d'ajustement

Dans cette section, nous présentons l'algorithme itératif qui permet d'ajuster le modèle à long terme 2D aux données (on considère à nouveau une section de parole voisée de K périodes/trames consécutives). Il s'inspire largement des versions « mono-dimensionnelles » présentées aux chapitres précédents, mais dans cette nouvelle déclinaison, l'algorithme est divisé en deux parties, afin de garder une complexité et un temps de calcul raisonnable pour nos expérimentations. Ainsi, la première partie de l'algorithme consiste à régler M de façon à conjointement adapter K modèles d'enveloppe optimaux aux K jeux d'amplitudes mesurées, et ceci selon un critère

perceptuel moyen. On reprend dans cette partie de l'algorithme le principe de comparaison de l'erreur de modélisation au seuil de masquage fréquentiel développé dans l'approche à long terme du Chapitre 3, mais en revenant à une version plus « conventionnelle » qui considère les données selon l'axe fréquentiel. Les K modèles d'enveloppe d'ordre identique M résultant de cette première étape sont ensuite utilisés comme une référence pour la deuxième partie de l'algorithme qui traite de la dimension temporelle. On réalise alors une estimation itérative de l'ordre P et des coefficients du modèle à long terme 2D de la matrice $\boldsymbol{C}$ de (5.13).

Comme cet algorithme est sensiblement complexifié par rapport à ceux des chapitres précédents, on se permet de le présenter avec quelques explications/commentaires inclus dans la description formelle ci-dessous, pour privilégier la clarté par rapport à la concision. De plus, nous avons choisi de présenter l'algorithme 2D de la façon la plus générale possible. En particulier, le critère de mise à jour des poids temporel de la matrice $\boldsymbol{W}_m$ dans la deuxième partie de l'algorithme n'est pas précisé ici car ce critère a été déterminé à l'issue d'une série intensive d'expérimentations. Il sera donc donné dans la Section 5.4 relative à la description de ces expérimentations, avec toutes les justifications nécessaires. Nous verrons dans cette même section d'expérimentations que ces dernières ont d'ailleurs conduit à d'autres modifications de l'algorithme par rapport à la forme présentée ci-dessous (ces modifications ne remettant pas en cause les principes de cette forme générale ; on verra qu'il s'agit en fait de simplifications).

Algorithme 2D à appliquer sur les jeux des mesures d'amplitude de chaque section de parole voisée. Les variables R_{min1} et R_{min2} sont des pourcentages initialisés par l'utilisateur à une valeur comprise dans l'intervalle [60%, 90%] avec de plus la contrainte $R_{min1} > R_{min2}$ (voir une explication plus loin). La variable *Itermax* dénote un nombre maximum d'itérations et est initialisé à une valeur entière entre 10 et 20.

Première partie de l'algorithme 2D : Modélisation des K enveloppes spectrales de la section de parole considérée (M est initialisé à une valeur arbitraire, typiquement 10) :

1 Pour chaque indice du temps k, k = 1 à K, calculer le seuil de masquage fréquentiel global $\boldsymbol{T}^k$ associé au vecteur d'amplitude $\boldsymbol{A}^k$ en employant le modèle perceptif présenté à la Section 3.1 du Chapitre 3.

2 Initialiser K matrices carrées de poids $\boldsymbol{W}_k$, chacune de taille $I_k \times I_k$ avec tous les éléments de la diagonale à 1. Pour chaque indice k, k = 1 à K, itérer alors le processus suivant de l'étape 3 à l'étape 4, jusqu'à ce que chaque ratio R_k de l'étape 5 soit maximisé.

3 Calculer le vecteur MCD des coefficients d'enveloppe $\boldsymbol{D}^k$ avec (5.7)[85]. Calculer la fonction d'erreur de modélisation (voir la Section 2.5.5, pour la définition des fonctions appliquées) :

$$f(\boldsymbol{E}^k) = \frac{1}{2} square(\boldsymbol{A}^k - \boldsymbol{H}_k \boldsymbol{D}^k)$$

[85] En pratique, on peut utiliser la version régularisée du MCD proposée dans [Cappé *et al.*, 1995] et déjà mentionnée à la Section 2.4.2.

4 Mettre à jour les poids de $\boldsymbol{W}_k$ selon :

$$\Delta \boldsymbol{W} = f(\boldsymbol{E}^k) - \boldsymbol{T}^k$$
$$\Delta \boldsymbol{W} \leftarrow \Delta \boldsymbol{W} - \min(\Delta \boldsymbol{W})$$
$$\boldsymbol{W}_k \leftarrow \boldsymbol{W}_k + diag(\Delta \boldsymbol{W} / \max(\Delta \boldsymbol{W}))$$

Calculer le pourcentage R_k des éléments négatifs de $f(\boldsymbol{E}^k) - \boldsymbol{T}^k$ (c'est-à-dire où la fonction d'erreur de modélisation est en dessous du seuil masquage).

5 Une fois que tous les R_k sont maximisés, calculer la valeur moyenne R_{moy1} de ces rapports sur les K enveloppes :

$$R_{moy1} = \frac{1}{K} \sum_{k=1}^{K} R_k$$

Si R_{moy1} est supérieur au rapport prédéfini R_{min1}, M est diminué de 1, sinon M est augmenté de 1. Retourner ensuite à l'étape 2 (c'est-à-dire qu'on recommence le test complet sur R_{moy1} pour la nouvelle valeur de M). On stoppe cette première partie de l'algorithme quand M se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R_{moy1} \geq R_{min1}$.

Deuxième partie de l'algorithme : Modélisation des coefficients d'enveloppe au cours du temps :

1 Initialiser P à la puissance de deux le plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP à $P/2$.

2 Initialiser la matrice $K \times K$ de poids temporels $\boldsymbol{W}$ avec tous les éléments de la diagonale à 1.

3 Calculer la matrice $\boldsymbol{C}$ des coefficients du modèle à long terme des trajectoires des coefficients d'enveloppe avec (5.13).

4 Calculer les K modèles d'enveloppe modélisés à long terme par $\hat{\boldsymbol{D}} = \boldsymbol{C}\boldsymbol{M}$ (cette équation est équivalente à (5.10) appliquée aux instants n_k et pour toutes les composantes m de 0 à M). On note $\hat{\boldsymbol{D}}^k$ le $k^{\text{ème}}$ vecteur colonne de $\hat{\boldsymbol{D}}$.

5 Restituer les K vecteurs d'amplitudes par : $\hat{\boldsymbol{A}}^k = \boldsymbol{H}_k \hat{\boldsymbol{D}}^k$, pour $k = 1$ à K (cette équation est équivalente à l'équation (5.3) appliquée aux fréquences harmoniques et avec les coefficients d'enveloppe résultant de la modélisation à long terme de l'étape 4).

6 Calculer le rapport R_{moy2} défini de la même façon que R_{moy1} dans l'étape 5 de la première partie de l'algorithme, mais en remplaçant les amplitudes modélisées dans cette première partie par celles issues de l'étape 5 de cette deuxième partie. Pour cela, il faut donc calculer les nouvelles valeurs de la fonction d'erreur de modélisation $f(\boldsymbol{E}^k)$:

$$f(\boldsymbol{E}^k) = \frac{1}{2} square(\boldsymbol{A}^k - \boldsymbol{H}_k \hat{\boldsymbol{D}}^k)$$

puis le nouveau pourcentage R_k des éléments négatifs de $f(\boldsymbol{E}^k) - \boldsymbol{T}^k$, pour $k = 1$ à K, puis moyenner ces valeurs de R_k. Mettre à jour la matrice de poids temporelle selon le procédé qui sera décrit dans la Section 5.4.1.2.

7 Si $R_{moy2} < R_{min2}$, et le nombre d'itérations *Itermax* n'est pas atteint, retourner à l'étape 3 ; (on raffine l'ajustement avec la matrice de poids temporelle mise à jour pour le même ordre de modèle).

Si $R_{moy2} \geq R_{min2}$, diminuer l'ordre du modèle selon $P \leftarrow P - \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2 ; (on recommence complètement l'ajustement en testant un ordre de modèle plus faible avec la matrice de poids temporelle réinitialisée).

Sinon, si $R_{moy2} < R_{min2}$ et le nombre d'itérations atteint *Itermax*, augmenter l'ordre du modèle selon $P \leftarrow P + \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2 ; (on recommence complètement l'ajustement en testant un ordre de modèle plus élevé avec la matrice de poids temporelle réinitialisée).

On stoppe l'algorithme quand P se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R_{moy2} \geq R_{min2}$.

On peut faire ici quelques remarques pour éclaircir certains points soulevés par la lecture de cet algorithme. D'abord, il faut noter que comme la deuxième partie de l'algorithme rajoute une étape de modélisation par-dessus celle de la première partie de l'algorithme, le choix de R_{min2} doit tenir compte de l'erreur additionnelle résultante. C'est pour cela qu'on doit choisir R_{min1} et R_{min2} tels que $R_{min1} > R_{min2}$. En pratique, on peut par exemple choisir $R_{min2} = \gamma \times R_{min1}$, avec γ un nouveau pourcentage, typiquement de l'ordre de 80% (dans la partie expérimentale, on fera varier ces paramètres et on analysera les résultats correspondants).

Ensuite, on peut noter que l'information relative à la trajectoire temporelle de la fréquence fondamentale ω_0 est nécessaire pour permettre la reconstruction des amplitudes modélisées à l'étape 5 de la deuxième partie de l'algorithme (à travers la construction de la matrice $\boldsymbol{H}_k$)[86]. Toutefois, ceci ne pose pas de problème majeur dans l'optique de l'utilisation de cette modélisation 2D dans une application de codage à très bas débit : cette information est toujours transmise sous une forme ou une autre dans les codeurs (voir la discussion de la Section 4.1 du Chapitre 4). Il faut cependant tenir compte du fait que la quantification (et donc la distorsion) de la fréquence fondamentale peut avoir des répercutions significatives sur l'estimation des amplitudes spectrales à partir de la forme d'enveloppe. Cependant, ce dernier point est un problème commun à notre approche et aux approches classiques à court terme utilisant un modèle d'enveloppe pour coder les amplitudes spectrales, et il dépasse le cadre de travail de cette thèse.

De façon similaire, il se pose le même problème pour les instants de mesure n_k dont la connaissance est nécessaire à l'étape 4 de la deuxième partie de l'algorithme (à travers la construction de la matrice $\boldsymbol{M}$). Il faut ici transmettre l'information relative à la position de ces instants de mesure, ce qui peut être très pénalisant dans une application de codage à très bas débit. Toutefois, cette information provient du fait qu'on a arbitrairement utilisé dans nos expérimentations un processus d'analyse période-

[86] Ce point peut être vu comme une illustration du cas général d'une NST avec plusieurs transformations élémentaires de taille $M \times N$ discuté à la Section 5.1.2 : l'information sur la taille des données avant transformation (notée ici N) doit être disponible au décodeur pour appliquer la transformation inverse correspondante (de taille $N \times M$) en vue de revenir à l'espace des données initiales [Li *et al*., 2002].

synchrone (voir Section 3.3.1.2). En fait, dans une application de codage, on peut très bien faire une analyse avec un décalage de fenêtre d'analyse fixe et ainsi se passer de cette information au décodeur (il suffit que le décodeur connaisse la valeur du décalage entre deux mesures, et d'assurer la synchronisation, ce qui est trivial). Pour plus de cohérence dans notre présentation, on a choisi de conserver le même processus d'analyse et les mêmes données que dans les chapitres précédents. On considère ici qu'on connaît donc les instants n_k au décodeur (*i.e.*, dans la partie resynthèse de l'algorithme 2D), ce qui est un peu abusif mais qui ne pose pas de problème en pratique en changeant le processus d'analyse.

Enfin, pour finir cette série de remarques, on peut noter que pour que la modélisation dans chaque dimension (fréquentielle ou temporelle) soit valable, il faut vérifier à chaque appel des étapes 3 de chaque partie de l'algorithme les conditions respectives : $M<I_k-1$ (pour la première partie de l'algorithme) et $P<K-1$ (pour la deuxième partie). La deuxième condition ne pose pas de problème : elle est contrôlée lors de la recherche de l'ordre optimal P et est toujours assurée en pratique. Par contre, dans la première partie de l'algorithme, on rappelle qu'on cherche un ordre M d'enveloppe spectrale « moyen » au sens où il convient à la majorité des K vecteurs d'amplitude modélisés de la section de parole considérée (ce qui est testé par la valeur atteinte par le pourcentage R_{moy1}). Il se peut que pour assurer cette condition, cet ordre moyen doive être ponctuellement égal ou supérieur à la valeur I_k-1 pour certaines trames. On se retrouve alors avec une configuration de type « sur-apprentissage » pour ces trames en question (voir la Section 4.2.3.2 pour une description de ce phénomène de sur-apprentissage lorsque le nombre de coefficients du modèle atteint, voire dépasse, le nombre de données à modéliser). Pour résoudre ce problème, et assurer un modèle convenable pour chaque trame, dans la première partie de l'algorithme, l'ordre « réel » du modèle pour chaque trame k est toujours limité par la condition $M<I_k-1$. Si M doit être augmenté par l'algorithme alors qu'on arrive en limite de cette condition pour une certaine trame (c'est-à-dire qu'on a $M \geq I_k-1$ pour cette trame), on « complète » le dernier vecteur des coefficients du modèle de cette trame, c'est-à-dire celui obtenu pour la valeur limite I_k-1, par des zéros. On a ainsi une augmentation « virtuelle » de l'ordre du modèle pour les trames en question, ce qui permet de garder le même ordre de modèle pour toutes les trames de la section de parole modélisée. Une illustration de ce problème et de cette solution sera donnée dans la section suivante.

5.4. Expérimentations et résultats

Dans cette section nous décrivons un ensemble d'expériences qui ont été conduites pour évaluer la modélisation des paramètres d'amplitudes par le modèle 2D tel que nous l'avons présenté dans les sections précédentes. Dans la suite, nous commentons ainsi le comportement de l'algorithme de modélisation 2D et nous donnons une série de résultats qualitatifs et quantitatifs illustrant les performances de cette méthode. Notons que nous ne nous attardons pas ici à présenter le protocole expérimental en terme de présentation des données utilisées, du processus d'analyse, et du protocole des tests d'écoute. Ces derniers sont en effet les mêmes que dans les chapitres précédents. On pourra se reporter ainsi à la Section 3.3.1 du Chapitre 3 pour une description de ce protocole.

5.4.1. Comportement de l'algorithme et conséquences sur son réglage

5.4.1.1. Première partie de l'algorithme

Les expériences que nous avons menées sur notre base de données ont montré que la première partie de l'algorithme fonctionne tout à fait correctement dans le sens où les modèles d'enveloppe sont bien progressivement ajustés (*i.e.* au cours des itérations sur les poids perceptifs) sur les valeurs d'amplitude, pour chaque trame k. Le comportement de cette partie de l'algorithme sur chaque trame rappelle ainsi le comportement de l'algorithme présenté au Chapitre 3 pour la modélisation à long terme des amplitudes le long de l'axe du temps. Ceci est tout à fait normal puisque cette nouvelle version « fréquentielle » en est directemement inspirée.

En revanche, nous avons observé que la recherche d'un ordre optimal M « moyen » pour l'ensemble des modèles d'enveloppe d'une section de parole voisée (*i.e.* K vecteurs d'amplitudes) pose en général un certain nombre de problèmes. En particulier, comme le rapport R_{min1} qui détermine le critère d'une bonne modélisation des K enveloppes est un critère moyen, pour un ordre M qui est le même pour toutes les enveloppes d'une même section, certains vecteurs d'amplitudes peuvent être modélisés « mieux que d'autres ». Ceci n'est pas surprenant, étant donné qu'on a déjà fait remarquer plusieurs fois que les formes spectrales peuvent varier significativement à l'intérieur d'une même section voisée[87]. Il peut donc apparaître une disparité entre les différentes valeurs du rapport R_k entre les différentes trames k d'une même section, ce qui n'empêche pas la convergence de cette première partie de l'algorithme, pourvu qu'en moyenne, ce rapport atteigne R_{min1}. Cette disparité dans les valeurs de R_k et par conséquent dans la précision et dans la forme même des enveloppes spectrales modélisées pose un problème particulier dans notre approche 2D lorsque cette disparité intervient de façon suffisamment « distribuée » ou « irrégulière » entre trames successives de la section de parole modélisée. En d'autres termes, la première partie de l'algorithme ne garantit pas forcément une grande homogénéité des formes spectrales (et donc des vecteurs de coefficients du premier MCD fréquentiel) entre trames successives. Ceci peut poser un problème important au niveau de la deuxième partie de l'algorithme : la modélisation des trajectoires de coefficients d'enveloppe va être significativement pénalisée par ces « discontinuités » entre valeurs de coefficients successifs de même rang.

Ce problème nous a incité à modifier cette première partie de l'algorithme de façon à ce qu'on améliore l'homogénéïté des formes spectrales successives délivrées. Des observations plus poussées et des expériences pilotes ont alors montré que ces discontinuïtés sont en grande partie une conséquence des itérations sur les poids perceptifs dans le processus d'ajustement de chaque enveloppe au jeu d'amplitude correspondant. En effet, ces itérations visent à améliorer le *fitting* du modèle

[87] Cette observation est d'ailleurs directement reliée à la remarque que nous avons faite à la fin de la Section 5.3 concernant la nécessité de vérifier la condition $M<I_k-1$ pour chaque trame lors du calcul des coefficients du modèle. Rappelons que dans le cas où M doit être augmenté au-delà de cette limite, nous insérons des zéros au niveau des termes supérieurs du dernier vecteur de coefficients du modèle, pour les trames concernées par ce problème.

d'enveloppe à ces données, et le critère est d'abord posé dans l'algorithme en terme « individuel » pour chaque trame k : c'est le rapport R_k. Selon la forme spectrale et l'évolution des itérations sur les poids perceptifs, deux vecteurs d'amplitudes consécutifs relativement proches peuvent ainsi fournir deux modèles d'enveloppes assez différents à la fin des itérations, chacun d'eux vérifiant l'obtention d'une « bonne » valeur de R_k. En d'autres termes, les itérations sur les poids perceptifs permettent d'affiner le fitting des modèles indépendamment sur chaque trame, éventuellement au détriment de l'homogénéïté inter-trame de cet ajustement.

Pour pallier ce problème, nous avons choisi dans la suite de cette étude de simplifier la première partie de l'algorithme en supprimant la phase d'itération sur les poids perceptifs. Dans la boucle de test sur la valeur de l'ordre M du modèle d'enveloppe, le calcul des coefficients de chaque modèle d'enveloppe k n'est ainsi fait qu'une seule fois et on teste le rapport moyen R_{moy1} obtenu avec cet ordre. Il résulte de cette modification que chaque modèle est « un peu moins précisément » ajusté au vecteur d'amplitude correspondante, mais en revanche on gagne en homogénéïté et continuité entre les différents modèles d'enveloppe successifs. Par conséquent les trajectoires de coefficients MCD fréquentiels sont plus homogènes et on va ainsi pouvoir améliorer assez nettement la modélisation à long terme de ces coefficients. Notons que pour contrebalancer cette perte de précision dans la modélisation spectrale, pour une même valeur du ratio cible R_{min1}, l'algorithme appliqué sans les itérations sur les poids perceptifs converge généralement vers une valeur de l'ordre M un peu plus élevée que dans sa version initiale. On a donc généralement des modèles d'enveloppe un peu moins parcimonieux (car moins efficacement ajustés individuellement) mais plus « réguliers » en forme et en trajectoire. On espère ainsi que l'efficacité renforcée de la modélisation le long de l'axe du temps viendra compenser cet effet.

L'ensemble des observations que nous venons de discuter est illustré à la Figure 5.2. Cette figure représente le spectre d'amplitude de trois trames consécutives d'une section de parole voisée pour une voix de femme (il s'agit encore de la longue section de parole déjà utilisée à la Figure 2.1). Cette section est un exemple particulièrement délicat car il s'agit d'une voix de femme avec une grande variation de la fréquence fondamentale. Une portion de cette section comporte ainsi un faible nombre d'harmoniques (de l'ordre de la douzaine pour une fréquence d'échantillonnage de 8 kHz) et l'ajustement du modèle à ce type de spectre relativement « pauvre » est particulièrement sensible.

Ainsi, les Figures 5.2(a), 5.2(c) et 5.2(e) représentent respectivement le spectre d'amplitude des trames 243, 244 et 245 de la section de parole considérée (on rappelle que cette section comporte $K = 408$ trames). Ces spectres d'amplitude sont plutôt homogènes d'une trame à l'autre : les formes sont assez similaires même si le spectre de la première des trois trames est un peu plus lisse que les deux autres qui sont très semblables. Sur chaque sous-figure de droite, on a tracé conjointement le spectre des amplitudes modélisées lors de la première partie de l'algorithme, après convergence des itérations sur les poids perceptifs, et pour un ordre de modèle M fixé à 10. En réalité, la trame 243 ne nécessite pas d'itérations car pour cette trame, le ratio $R_k = 100\%$ est atteint dès le premier calcul du modèle (sans doute du fait du caractère particulièrement lisse de ce spectre). Les trames suivantes demandent respectivement 11 et 2 itérations pour atteindre la valeur maximale de R_k. Cette valeur vaut en l'occurrence 92% pour ces deux trames, soit 11 harmoniques sur 12 « correctement » modélisées. On remarque

alors que l'harmonique qui reste « mal modélisée » au sens du critère perceptif, est la première harmonique pour la trame 244, et que dans ce cas cette harmonique est particulièrement mal modélisée : la valeur résultant du modèle pour cette harmonique est très différente de la valeur obtenue pour la trame précédente et la trame suivante. On pourrait dire que cette harmonique a été « sacrifiée » par le modèle au cours de son ajustement itératif. On a ainsi une illustration des discontinuités d'amplitude entre les trames successives potentiellement engendrées par la modélisation itérative des enveloppes spectrales. Ce type de discontinuité se répercute sur les trajectoires des coefficients de modèle d'enveloppe : les coefficients pour la trame 244 vont être significativement hétérogènes par rapport aux coefficients des trames environnantes. Par conséquent, la modélisation à long terme va être pénalisée par ce « mauvais conditionnement » des coefficients de modèle d'enveloppe.

Pour compléter l'illustration du problème, on retrouve cette discontinuité dans le tracé temporel de cette première harmonique (pour cette même section de parole de voix féminine) sur la Figure 5.3(a). Sur cette figure, la discontinuïté au niveau de la trame 244 prend la forme d'un énorme pic (qui sort même de l'intervalle des amplitudes de cette figure) repéré par un cercle rouge. Cette figure permet aussi de se rendre compte que d'une manière générale, la trajectoire d'amplitude reconstruite à partir des valeurs modélisées par la première partie de l'algorithme est relativement « irrégulière » ou « bruitée ». La resynthèse du signal à partir de ces valeurs conduit à un signal assez dégradé.

En revanche, sur les Figures 5.2(b) et 5.2(d), on a représenté les mêmes tracés qu'en 5.2(c) et 5.2(e) respectivement, mais sans effectuer les itérations sur les poids perceptifs. Le modèle est alors le premier modèle calculé avec l'équation (5.7) sans rafinement. On voit que les spectres d'amplitudes modélisées obtenus sont plus homogènes d'une trame à l'autre, même si les valeurs de R_k sont moins bonnes (83% pour les deux trames 244 et 245). En particulier, la valeur modélisée de la première harmonique sur le spectre de la trame 244 est bien plus convenable que précédemment. Par conséquent, la discontinuïté en terme d'amplitude a disparu de la trajectoire, comme on peut le voir sur la Figure 5.3(b). Cette figure montre d'ailleurs que cette amélioration est tout à fait générale : l'ensemble de la trajectoire de l'amplitude modélisée est bien plus lisse et plus fidèle à la trajectoire originale des mesures d'amplitude de cette harmonique.

Parallèlement, et c'est là le point essentiel pour la modélisation à long terme à venir, les coefficients d'enveloppe sont devenus plus homogènes d'une trame à l'autre, et se prêtent donc mieux à cette modélisation à long terme. Notons qu'au final, sans les itérations, l'ordre moyen du modèle MCD d'enveloppe est passé à 11 alors qu'il vaut 10 dans le cas où l'algorithme est appliqué avec itérations sur les poids perceptifs. Ceci illustre la remarque générale mentionnée plus haut sur ce point. En même temps, cela n'est pas en contradiction avec la valeur de 10 annoncée sur la Figure 5.2. En effet, sur cet exemple précis des trames 243, 244 et 245, on a $I_k = 12$ harmoniques dans le spectre. On est donc en situation de limite à la condition $M<I_k-1$. Pour ces trames-ci, lorsque M augmente au cours de l'algorithme, le modèle reste le même : il est complété par un coefficient à zéro (autrement dit, pour ces trois trames, les modèles à l'ordre intermédiaire 10 et à l'ordre final 11 sont identiques).

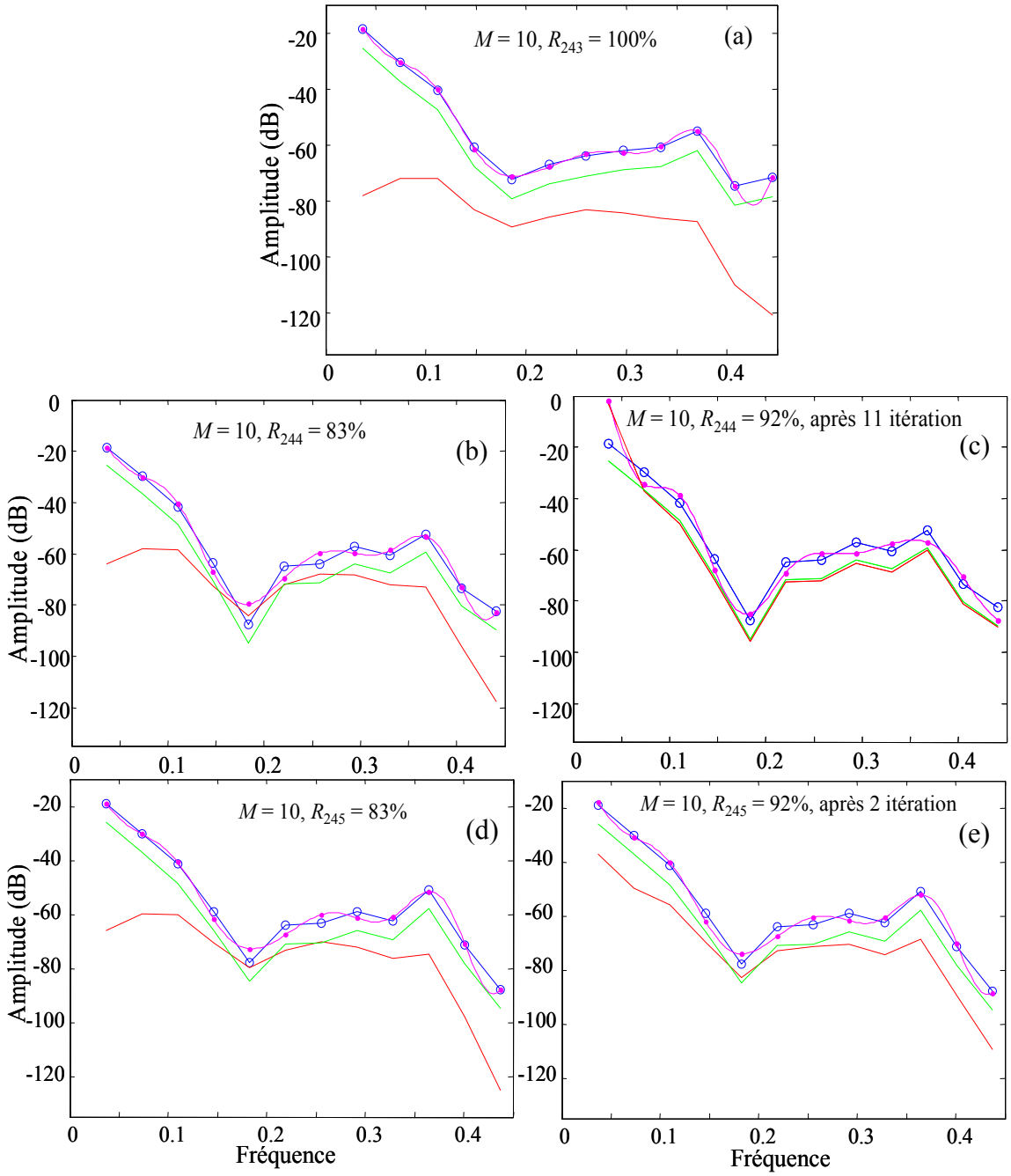

Figure 5.2 : Illustration du fonctionnement de la première partie de l'algorithme d'ajustement 2D, avec (à droite) et sans (à gauche) itérations sur les poids perceptifs, et pour trois trames consécutives d'une longue section de parole voisée de voix de femme (durée 1,4 sec, $F_e = 8$ kHz ; il s'agit du même jeu de données que celui utilisé à la Figure 2.1 ; on représente ici les spectres d'amplitude pour les trames 243, 244 et 245 (les fréquences sur l'axe des X sont des fréquences numériques) ; pour la trame 243, l'ajustement optimal est obtenu au premier calcul de modèle, *i.e.* sans itérations, c'est pourquoi on met la figure correspondante au milieu). En bleu : les mesures d'amplitude ; en magenta : le modèle d'enveloppe MCD ; en vert : le seuil perceptif ; en rouge : l'erreur de modélisation. Dans tous les cas, l'ordre M vaut 10.

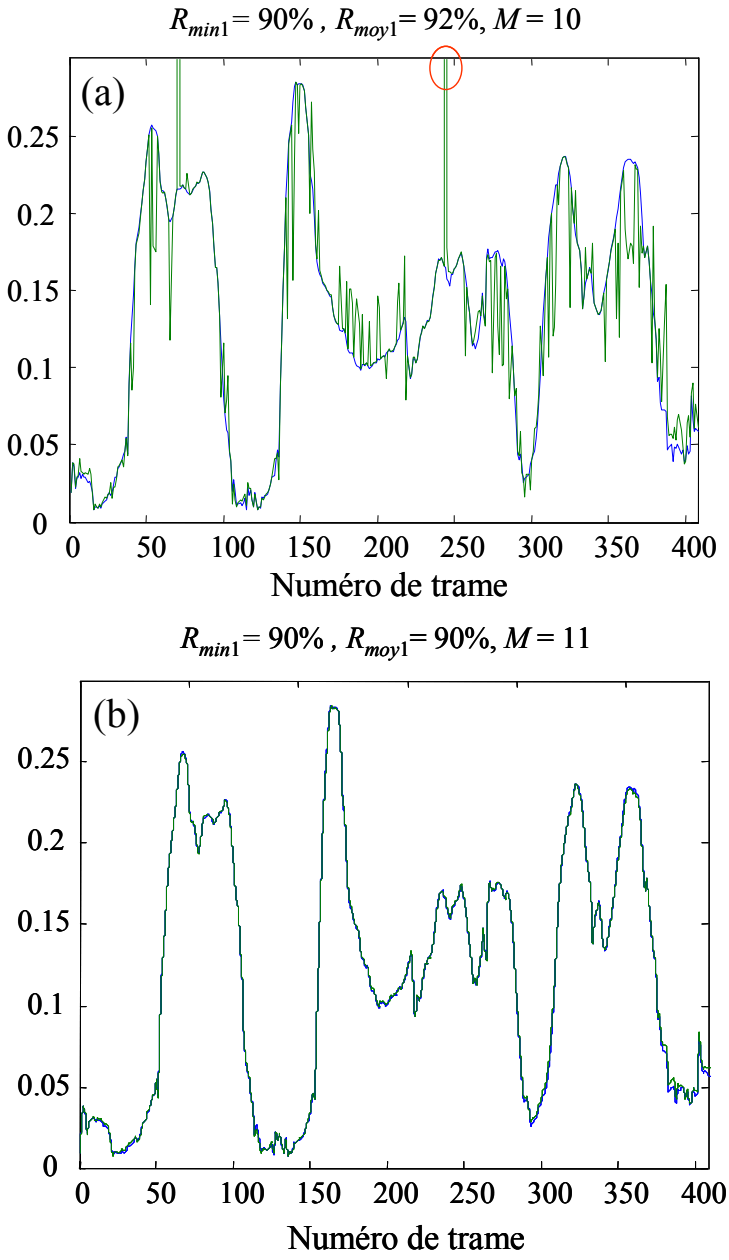

Figure 5.3 : Trajectoires de la première harmonique du signal correspondant au résultat de la Figure 5.2. En bleu : mesures d'amplitude originales ; en vert : amplitudes issues de la modélisation MCD obtenue lors de la première partie de l'algorithme d'ajustement 2D. (a) : résultat obtenu avec la première version avec les itérations sur les poids perceptifs (la trame 244 est marquée par un cercle rouge) ; (b) : résultat obtenu avec la version simplifiée sans les itérations. Notons que dans le premier cas on obtient $M = 10$ (avec $R_{min1} = 90\%$) et dans le second cas, on obtient $M = 11$ (ceci n'est pas en contradiction avec la Figure 5.2, voir le texte).

Ce type de résultat se généralise assez bien à l'ensemble des sections « à problème » que nous avons observées. C'est ce qui nous a conduit à formuler cette version simplifiée de la première partie de l'algorithme sans les itérations sur les poids perceptifs. La version simplifiée de la première partie de l'algorithme est présentée formellement ci-après.

Première partie de l'algorithme 2D en version simplifiée : Modélisation des K enveloppes spectrales de la section de parole considérée (M est initialisé à une valeur arbitraire, typiquement 10) :

1 Pour chaque indice du temps k, k = 1 à K, calculer le seuil de masquage fréquentiel global $\boldsymbol{T}^k$ associé au vecteur d'amplitude $\boldsymbol{A}^k$ en employant le modèle présenté à la Section 3.1 du Chapitre 3.

2 Pour chaque indice du temps k, k = 1 à K, calculer le vecteur MCD d'enveloppe avec (5.7) (les matrices $\boldsymbol{W}_k$ sont toutes la matrice identité $I_k \times I_k$, ce qui veut dire qu'on simplifie cette équation selon $\boldsymbol{D}^k = (\boldsymbol{H}_k^t \boldsymbol{H}_k)^{-1} \boldsymbol{H}_k^t \boldsymbol{A}^k$; on peut cependant toujours utiliser la forme pénalisée de [Cappé *et al.*, 1995]). Calculer la fonction d'erreur de modélisation :

$$f(\boldsymbol{E}^k) = \frac{1}{2} square(\boldsymbol{A}^k - \boldsymbol{H}_k \boldsymbol{D}^k)$$

Calculer le pourcentage R_k des éléments négatifs de $f(\boldsymbol{E}^k) - \boldsymbol{T}^k$.

3 Calculer la valeur moyenne R_{moy1} de ces rapports R_k sur les K enveloppes :

$$R_{moy1} = \frac{1}{K} \sum_{k=1}^{K} R_k$$

Si R_{moy1} est supérieur au rapport prédéfini R_{min1}, M est diminué de 1, sinon M est augmenté de 1. Retourner ensuite à l'étape 2 (c'est-à-dire qu'on recommence l'ajustement complet pour la nouvelle valeur de M). On stoppe cette première partie de l'algorithme quand M se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R_{moy1} \geq R_{min1}$.

Notons que pour que cette version soit efficace, il faut « relacher » quelque peu la contrainte d'ajustement du modèle par rapport à la version de l'algorithme avec itérations sur les poids perceptifs, en ne réglant pas R_{min1} à une valeur trop élevée. Par exemple, sur la Figure 5.3 on obtient de meilleurs résultats sur les trames 244 et 245 avec R_k = 83% plutôt qu'avec R_k = 92%. Ceci devrait inciter à fixer R_{min1} autour de 80% plutôt que autour de 90%. Toutefois, il s'agit ici d'ordres de grandeur « localisés » : ce réglage dépend aussi de toutes les autres trames de la section de parole considérée, et c'est pourquoi la Figure 5.3 correspond en fait à R_{min1} = 90%.

Une fois que les simplifications ont été effectuées sur cette première partie de l'algorithme d'ajustement 2D, on peut donner ici quelques résultats qualitatifs et quantitatifs complémentaires en termes très généraux. Ainsi, même sous forme simplifiée, la première partie de l'algorithme s'adapte généralement correctement aux différentes configurations des sections de parole modélisées. A titre d'illustration, la Figure 5.4 présente les histogrammes décrivant la répartition des valeurs de M pour l'ensemble de la base de données en séparant les voix de femmes (environ 1800 sections voisées) et les voix d'hommes (environ 1700 sections voisées). Pour ces histogrammes, on a fixé R_{min1} = 90%, soit une valeur assez exigeante. On constate que l'ordre M peut varier de façon assez remarquable (n'oublions pas que cet ordre caractérise à chaque fois une section complète de plusieurs trames de signal, le nombre de trames de ces sections pouvant lui-même varier beaucoup). On obtient ainsi des sections avec des

petites valeurs d'ordre d'enveloppe (par exemple 4 ou 5) et donc caractérisées par des spectres assez « pauvres », des sections avec des valeurs d'ordre habituelles pour coder la parole féminine sur la bande 0-4 kHz (en général 10-11) et des sections avec des valeurs d'ordre habituelles pour coder la parole masculine (typiquement 15-16, voir [Cappé *et al.*, 1995] [Campedel-Oudot *et al.*, 2001]). On peut également obtenir des valeurs plus grandes pour des suites de spectres assez « riches » en nombre d'harmoniques et en relief spectral. Au total, on peut affirmer que la contrainte d'un ordre unique pour toute une section continûment voisée ne semble pas globalement géner la variabilité des degrés de libertés sur les formes d'enveloppe spectrales. Ceci est de bon augure pour conserver le maximum de flexibilité dans notre modélisation en 2D. Observons maintenant les résultats portant sur la deuxième partie de l'algorithme.

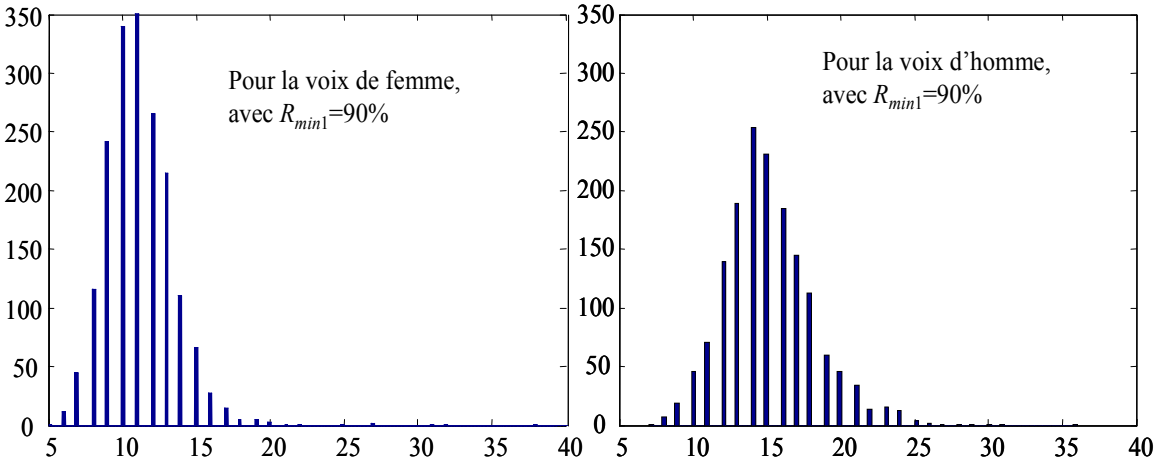

Figure 5.4 : Histogrammes de l'ordre de la modélisation du spectre d'amplitude par un MCD obtenu avec l'algorithme d'ajustement 2D proposé, pour R_{min1} = 90%. A gauche : voix de femmes (1800 sections de parole voisées ; la moyenne de l'ordre vaut approximativement 11). A droite : voix d'hommes (1700 sections ; la moyenne de l'ordre vaut approximativement 15).

5.4.1.2.Deuxième partie de l'algorithme

Comme on l'a vu dans la section précédente, la modification de la première partie de l'algorithme visait à régulariser les trajectoires des coefficients cepstraux du modèle MCD en vue de leur modélisation à long terme. Bien que cette régularisation se soit révélée efficace, nous avons observé que certains spectres d'amplitude pouvaient encore être relativement mal modélisés par le modèle d'enveloppe MCD, même après modification de la première partie de l'algorithme, et ceci de façon relativement isolée le long des trajectoires. Plusieurs raisons peuvent expliquer ces mauvais ajustements résiduels sporadiques : forme particulière des amplitudes des harmoniques, changement « brusque » du nombre d'harmoniques, erreur localisée d'estimation de la fréquence fondamentale (et donc analyse faussée des harmoniques), etc. Comme dans la section précédente, la conséquence de ces irrégularités spectrales locales est de générer des irrégularités tout aussi locales dans les trajectoires des paramètres d'enveloppe issus de la première partie de l'algorithme.

Même si ce type d'erreur est relativement résiduel par rapport aux irrégularités traitées dans la section précédente, nous avons configuré la deuxième partie de l'algorithme de façon à traiter implicitement ce problème. Ainsi, nous proposons de configurer la pondération temporelle de l'ajustement des trajectoires (par la matrice de poids $\boldsymbol{W}$) de façon à donner moins de poids à ces trames isolées posant problème au niveau de la modélisation spectrale. A l'inverse, bien entendu, plus de poids sera donné aux trames ayant un spectre bien modélisé par son enveloppe spectrale issue de la première partie de l'algorithme. Ainsi, on va intrinsèquement lisser les irrégularités dues aux erreurs de modélisation spectrale puisque les trames problématiques auront un poids faible dans le processus d'ajustement WMMSE. Pour cela, avant le calcul de la matrice $\boldsymbol{C}$ des coefficients du modèle à long terme, nous mesurons la précision de l'ajustement du modèle d'enveloppe pour chaque trame k en calculant la distance spectrale suivante :

$$SD_k = \sqrt{\frac{1}{I_k}\sum_{i=1}^{I_k}\left[20\log_{10} A_{i,k} - 20\log_{10} \hat{A}_{i,k}\right]^2} \qquad \text{pour } 1 \leq k \leq K \tag{5.14}$$

Dans cette équation, les $\hat{A}_{i,k}$ désignent les amplitudes modélisées à l'issue de la première partie de l'algorithme, et comme toujours, les $A_{i,k}$ désignent les mesures d'amplitude originales. Il s'agit d'une distance très classique en traitement de parole, largement utilisée dans le cadre du codage notamment[88] [Gray & Markel, 1976] [Gray *et al.*, 1980] [Nocerino *et al.*, 1985] [Paliwal et Atal, 1993]. Il est important d'insister sur le fait que dans cette expression, les valeurs d'amplitudes modélisées sont celles issues de la première partie de l'algorithme (elles resultent exactement de l'échantillonnage du modèle d'enveloppe MCD aux multiples de la fréquence d'échantillonnage pour chaque trame). En effet, une fois que cette distance est calculée, nous avons cherché à l'utiliser pour la mise à jour des poids temporels de $\boldsymbol{W}$ dans l'étape 6 de la deuxième partie de l'algorithme. Cependant, on rappelle que l'algorithme fonctionne en deux temps distincts : modélisation conjointe des enveloppes spectrales des K trames dans la première partie, puis modélisation à long terme des trajectoires des coefficients d'enveloppe résultant dans la deuxième partie. Par conséquent, ce calcul de distance spectrale est figé à la fin de la première partie de l'algorithme, et ne peut pas être « mis à jour » dans la deuxième partie. C'est pourquoi, en pratique, comme pour la première partie de l'algorithme, nous supprimons les phases itératives dans cette deuxième partie. Pour chaque valeur de P testée, nous ne faisons le calcul de $\boldsymbol{C}$ qu'une seule fois, en utilisant les valeurs inverses de la distance spectrale comme poids temporel, soit :

$$W_{k,k} = \frac{1}{SD_k} \qquad \text{pour } 1 \leq k \leq K \tag{5.15}$$

où $W_{k,k}$ est le $k^{\text{ième}}$ terme diagonal de la matrice de poids $\boldsymbol{W}$. En effet, ceci correspond bien à donner plus de poids aux coefficients d'enveloppe correctement ajustée (distance spectrale faible) et moins de poids aux coefficients d'enveloppe mal ajustée (distance spectrale importante). La deuxième partie de l'algorithme, sous forme ainsi simplifiée, devient donc la suivante.

[88] Dans le cadre du codage de type LPC, les valeurs discrètes des amplitudes des harmoniques sont remplacées par les valeurs échantillonnées du spectre LPC sur la même échelle logarithmique.

Deuxième partie de l'algorithme sous forme simplifiée : Modélisation des coefficients d'enveloppe au cours du temps :

1 Initialiser P à la puissance de deux le plus proche de $K/2$, et initialiser la mise à jour de cet ordre δP à $P/2$. Calculer la matrice diagonale $K \times K$ de poids temporels $\boldsymbol{W}$ avec (5.14) et (5.15).

2 Calculer la matrice $\boldsymbol{C}$ des coefficients du modèle à long terme des trajectoires des coefficients d'enveloppe avec (5.13).

3 Calculer les K modèles d'enveloppe modélisés à long terme avec $\hat{\boldsymbol{D}} = \boldsymbol{CM}$. On note $\hat{\boldsymbol{D}}^k$ le $k^{\text{ème}}$ vecteur colonne de $\hat{\boldsymbol{D}}$.

4 Calculer les K vecteurs d'amplitudes spectrales modélisées en 2D avec $\hat{\boldsymbol{A}}^k = \boldsymbol{H}_k \hat{\boldsymbol{D}}^k$, pour $k = 1$ à K.

5 Calculer le rapport R_{moy2} (voir la forme non simplifiée de cette deuxième partie d'algorithme).

Si $R_{moy2} \geq R_{min2}$, diminuer l'ordre du modèle selon $P \leftarrow P - \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2 ; (on recommence l'ajustement en testant un ordre de modèle plus faible).

Si $R_{moy2} < R_{min2}$, augmenter l'ordre du modèle selon $P \leftarrow P + \delta P$, mettre à jour δP selon $\delta P \leftarrow \delta P/2$, et retourner à l'étape 2 ; (on recommence l'ajustement en testant un ordre de modèle plus élevé).

On stoppe l'algorithme quand P se stabilise autour d'une valeur optimale : on conserve alors la valeur minimale pour laquelle on a $R_{moy2} \geq R_{min2}$.

Les expériences pilotes ont montré que cette nouvelle forme de l'algorithme permettait bien d'effectuer une régularisation supplémentaire (par rapport aux résultats de modélisation délivrés par la première partie de l'algorithme). Autrement dit, ce processus de pondération à base de distortion spectrale permet généralement de lisser les pics indésirables résiduels dans les trajectoires des coefficients à modéliser à long terme. Conjointement, ces expériences pilotes ont montré que cette pondération permet généralement de diminuer encore l'ordre P du modèle à long terme par rapport à une version de l'algorithme similaire mais où aucune pondération n'est effectuée avant le calcul de la matrice de coefficients $\boldsymbol{C}$.

5.4.1.3.Résultats qualitatifs sur l'algorithme complet

Nous donnons dans cette sous-section quelques remarques sur le comportement de l'algorithme complet, c'est-à-dire avec les deux parties mises en cascade. Tout d'abord, on peut mentionner que comme dans les versions « 1D-temporelles » des chapitres précédents, c'est-à-dire avec une modélisation séparée de chaque paramètre sinusoïdal le long de l'axe de temps, l'ordre P varie beaucoup selon la longueur et le contenu de la section de parole modélisée. La clé du bon comportement général de cet algorithme 2D est qu'une fois l'ordre P optimal atteint, il délivre des valeurs d'amplitudes qui sont globalement assez proches des amplitudes mesurées. Ceci est en effet garanti par la contrainte perceptuelle « globale » (c'est-à-dire sur le ratio R_{min2}) qui guide le comportement de l'algorithme d'ajustement complet : à la fin de l'algorithme, $R_{min2} = \gamma \times R_{min1}$ pourcent des amplitudes modélisées sont assurées de vérifier cette

contrainte (on rappelle en effet que R_{min2} pourcent de ces amplitudes sont telles que l'erreur de modélisation est au-dessous du modèle de seuil de masquage). Bien entendu, pour une valeur cible globale R_{min2} donnée, l'utilisateur est libre de régler R_{min1} ou ce qui revient au même, le facteur γ. Ceci permet de « doser » la précision de la modélisation entre les dimensions spectrales et temporelles. L'importance de ce dosage s'avère cruciale en pratique. Suivant les valeurs de γ les résultats peuvent en effet être très différents et ce point mérite une attention particulière. Pour l'illustrer, nous allons nous intéresser aux exemples de résultats expérimentaux reportés sur les Figures 5.5 et 5.6.

Ainsi, sur la Figure 5.5 on a représenté les trajectoires des quatre premiers paramètres d'enveloppe issus de la première partie de l'algorithme, pour une longue section de parole voisée (voix d'homme, 1,1 seconde de signal, $K = 159$ trames). On a représenté conjointement les trajectoires du modèle à long terme correspondant. Notons au passage que ces trajectoires de paramètres illustrent bien la corrélation à long terme qui existe entre les valeurs successives de coefficients d'enveloppe, corrélation que l'on cherche précisément à exploiter dans notre approche. On a aussi représenté les trajectoires des (trois premières) amplitudes résultantes, et les trajectoires des amplitudes mesurées en guise de référence. Cette figure correspond au réglage $R_{min1} = 75\%$ et $R_{min2} = 70\%$. Avec ces valeurs, on obtient un ordre $M = 8$, ce qui est trop faible pour caractériser cette section, on le verra par la suite. Cette valeur s'explique par le fait que $R_{min1} = 75\%$ est une contrainte relativement faible pour la modélisation spectrale par le MCD. On obtient parallèlement un ordre à long terme $P = 76$. Cet ordre est par contre assez élevé par rapport au nombre de mesures $K = 159$. Ce dernier point s'explique par le fait que R_{min1} et R_{min2} sont assez proches : pour que la modélisation à long terme effectuée dans la deuxième partie de l'algorithme, et contrainte par R_{min2}, ne dégrade pas beaucoup le score obtenu au cours de la première partie de l'algorithme, déterminé par R_{min1}, il faut que les trajectoires à long terme « collent » suffisamment bien aux valeurs des coefficients d'enveloppe. Pour cela, l'ordre à long terme doit être assez élevé. Notons que bien que la contrainte globale sur les amplitudes dictée par R_{min2} soit atteinte, les trajectoires d'amplitudes modélisées sont assez « décevantes » : elle suivent globalement les trajectoires des amplitudes mesurées mais elles s'écartent parfois localement de façon assez brutale.

Observons à présent les résultats de la Figure 5.6. On a reporté sur cette figure les résultats de la même expérimentation, mais obtenus cette fois avec le réglage $R_{min1} = 90\%$, les autres réglages étant identiques. On voit que bien qu'elles exhibent toujours une évidente corrélation, les trajectoires des paramètres d'enveloppe sont plus irrégulières que dans le cas précédent. Ceci s'explique par le fait qu'avec cette nouvelle valeur plus exigeante de R_{min1}, le modèle d'enveloppe pour chaque trame est plus fin, plus précis que pour $R_{min1} = 75\%$, et par conséquent les variations de ce modèle d'une trame à l'autre sont plus importantes (notons aussi que conjointement l'ordre M du modèle d'enveloppe est passé à 14). En revanche, comme la « marge de manœuvre » de la modélisation à long terme est plus large que dans le cas précédent du fait de la plus grande différence entre R_{min1} et R_{min2}, les trajectoires des coefficients d'enveloppe modélisés sont plus lisses que dans le cas précédent ! En d'autres termes, ces trajectoires ont moins besoin de « coller » fidèlement aux valeurs des coefficients d'enveloppe pour réaliser la contrainte sur R_{min2}. Une conséquence importante est que l'ordre du modèle à long terme est fortement diminué par rapport à l'expérience précédente : on a ici $P = 29$ (contre 76 précédemment).

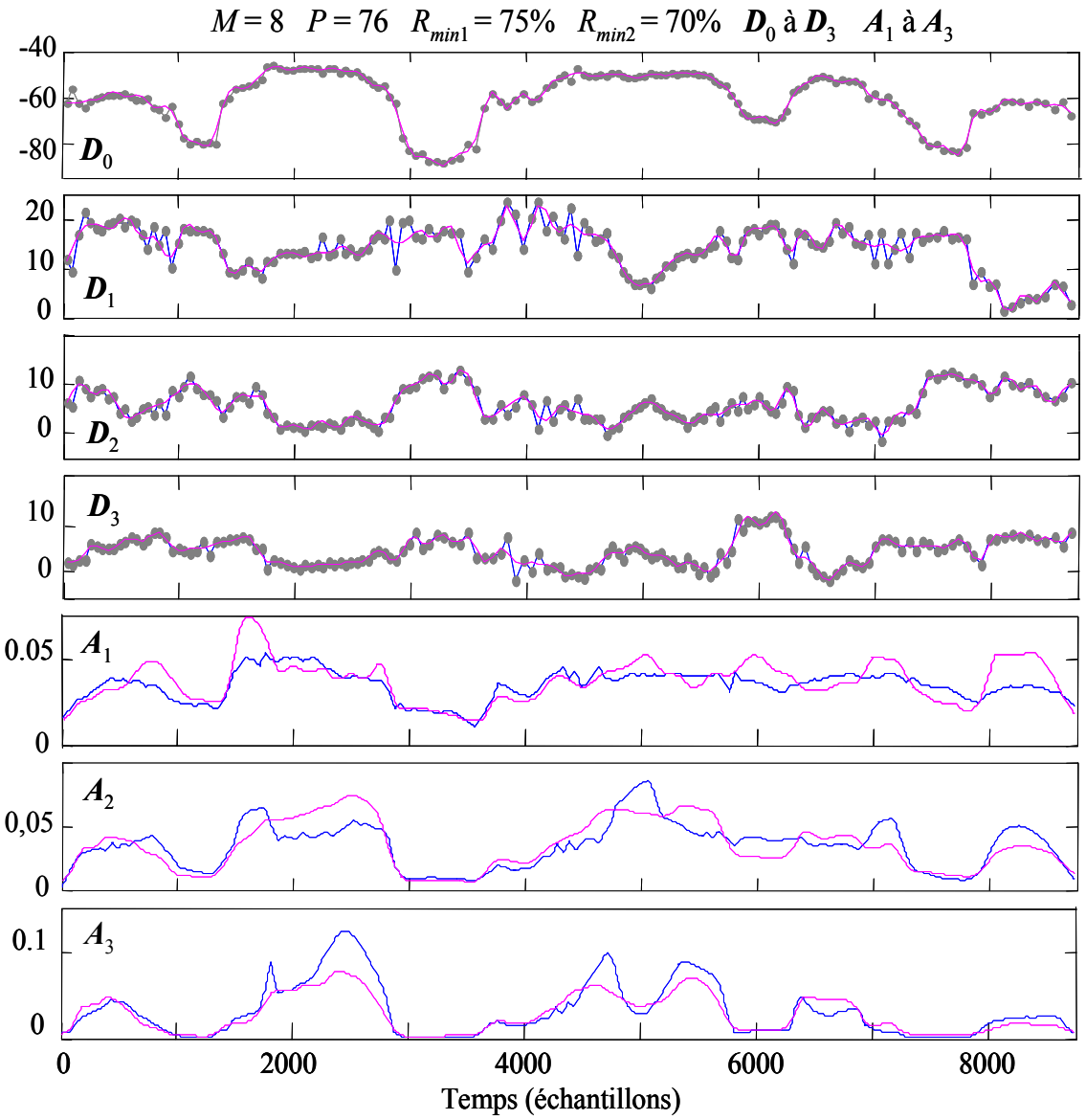

Figure 5.5 : Modélisation à long terme en 2D des amplitudes pour une longue section de parole voisée (voix d'homme, 1,1s de signal, $K = 159$ trames, $F_e = 8$ kHz). Quatre premières figures du haut : trajectoires des quatre premiers coefficients des enveloppes spectrales MCD issues de la première partie de l'algorithme (en points gris / lignes bleues) et modèles MCD à long terme associés (en magenta). Trois figures du bas trajectoires des amplitudes des trois premières harmoniques correspondantes : mesures originales (en bleu) et amplitudes modélisées en 2D (en magenta). On a ici $R_{min2} = 70\%$, $R_{min1} = 75\%$ et on obtient sur cette section $M = 8$ et $P = 76$.

Ainsi, on se rend compte qu'avec deux réglages différents de R_{min1} (et pour le même réglage de R_{min2}), on obtient une répartition très différente de la précision de modélisation entre la dimension fréquentielle et la dimension temporelle. Le deuxième réglage est ici bien meilleur que le premier, on peut s'en rendre compte à travers deux résultats dérivés. D'une part les trajectoires d'amplitudes issues de la modélisation à long terme en 2D avec le nouveau réglage sont plus régulières et plus fidèles aux trajectoires mesurées, comme on peut le voir à la Figure 5.6 pour les trois premières

harmoniques de la section traitée (à comparer avec les mêmes tracé à la Figure 5.5 pour $R_{min1} = 75\%$). Et d'autre part, ce résultat est obtenu pour un débit de coefficients beaucoup plus faible : on a (8+1)×(76+1) = 693 coefficients 2D dans le premier cas, et seulement (14+1)×(29+1) = 450 coefficients dans le second cas. On a donc un gain très important sur le débit de coefficients (ici 243 coefficients sur 693, soit environ 35% de gain) pour une modélisation « quantitativement équivalente » du point de vue du rapport cible global R_{min2} qui nous sert de critère, et vraisemblablement « qualitativement meilleure ».

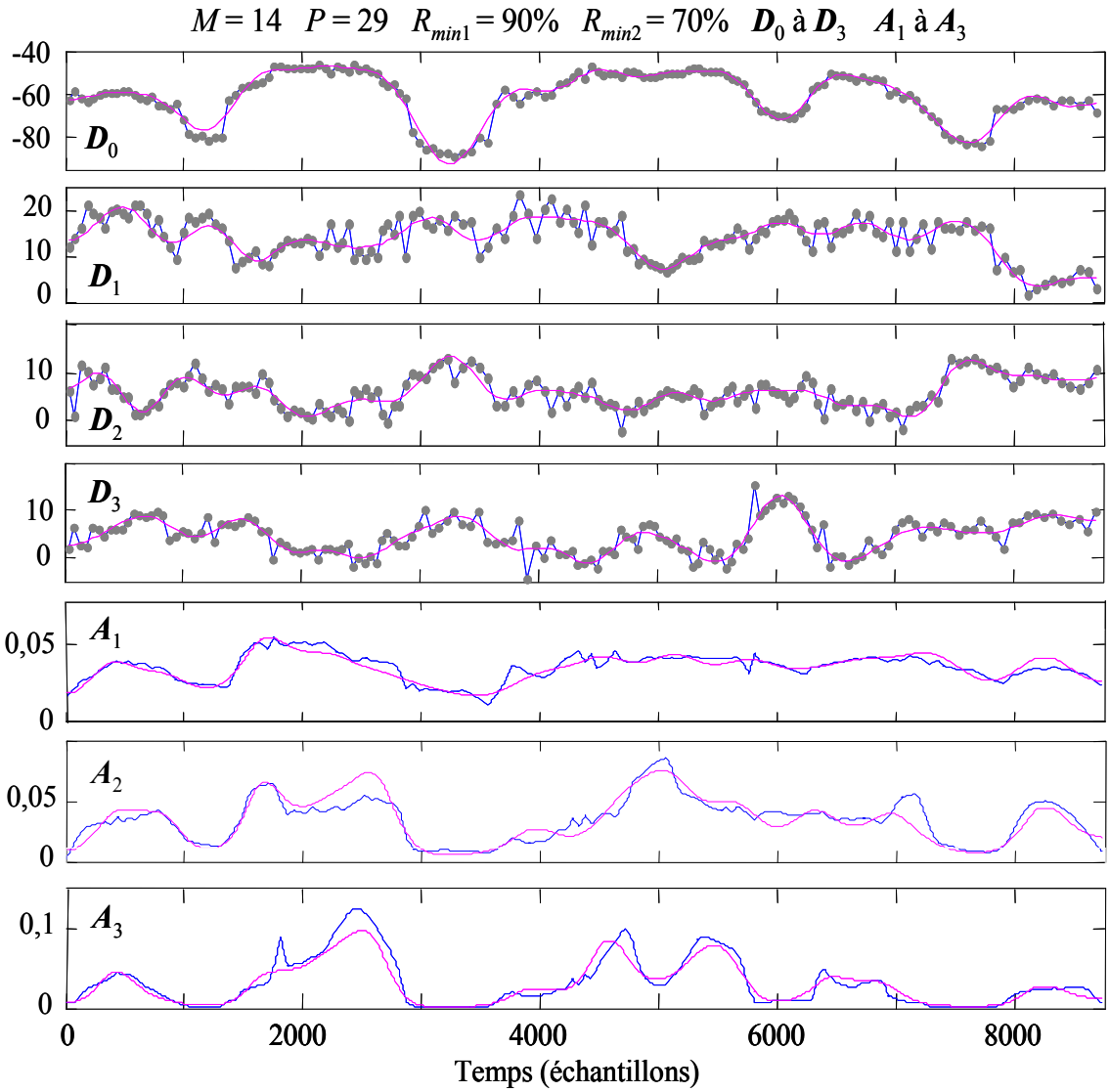

Figure 5.6 : Résultats issus de la même expérience qu'à la Figure 5.5 mais en prenant $R_{min1} = 90\%$. On obtient alors sur cette section $M = 14$ et $P = 29$.

En conclusion de cette expérience, on peut dire qu'il est inutile de faire un gros effort de précision pour la modélisation selon la dimension temporelle, si la modélisation selon la dimension fréquentielle n'est pas adaptée (c'est pourquoi on peut dire *a posteriori* qu'un ordre 14 pour l'enveloppe spectrale est ici sans doute un meilleur choix qu'un ordre 8 pour la section considérée). Avec un tel déséquilibre, on semble perdant à la fois

en terme de qualité des amplitudes synthétisées et de débit. A l'inverse, une modélisation en 2D bien équilibrée selon les deux dimensions semble particulièrement efficace. Nous allons à présent étudier ce point de façon quantitative à la section suivante en testant précisément et systématiquement l'influence du réglage des rapports cibles R_{min1} et R_{min2} sur le débit de coefficients (sur notre base de données complète). Nous analyserons aussi les meilleures valeurs de débit obtenues en regard de nos expérimentations du Chapitre 3 et du contexte du codage sinusoïdal à bas débit.

5.4.2. Débit de coefficients

Dans cette section, nous donnons des résultats de débits moyens pour les coefficients du modèle 2D, c'est-à-dire les coefficients de la matrice $\boldsymbol{C}$ qui représentent l'information à transmettre dans un système de communication pour coder les amplitudes du modèle sinusoïdal avec la technique 2D proposée[89]. Comme mentionné dans la section précédente, la version 2D de l'algorithme d'ajustement (même sous forme simplifiée telle que nous l'utilisons ici) permet de régler indépendamment les valeurs de R_{min1} et R_{min2}. On peut donc allouer arbitrairement plus de précision sur la première modélisation (spectrale) ou sur la deuxième modélisation (temporelle). Parallèlement, le débit de coefficients (et la qualité des signaux, voir la section suivante) doit en principe varier en fonction de ce réglage. Dans cette section, nous cherchons en particulier à caractériser expérimentalement ce point précis de façon quantitative. L'objectif est bien sûr de trouver, si elle existe, une configuration optimale en terme de rapport entre le score de modélisation (*i.e.*, le pourcentage final d'harmoniques vérifiant la contrainte perceptive sur l'erreur de modélisation, déterminé par le seuil R_{min2}) et le débit de coefficients. Pour cela, nous avons appliqué notre algorithme sur l'ensemble de notre base de données (rappel : environ 13 minutes de parole voisée, 3500 sections avec une répartition équilibrée entre voix de femmes et voix d'hommes). Nous avons extrait des débits moyens pour un échantillonnage des valeurs des rapports R_{min1} et R_{min2}, par pas de 5% entre 60% et 90%. De plus, on conserve évidemment la contrainte $R_{min1}>R_{min2}$[90].

Il est important de noter une différence par rapport à la présentation des résultats de débits de coefficients effectuée dans les Chapitres 3 et 4 aux Sections 3.3.4 et 4.3.3.4 : nous complétons ici les résultats obtenus avec la méthode d'analyse pitch-synchrone utilisée dans toutes les expériences présentées jusqu'ici par des résultats obtenus avec une méthode d'analyse à fenêtrage de taille fixe (25 ms) et décalage fixe (20 ms). Cette dernière configuration est une configuration typique des codeurs usuels à court terme. Elle nous permettra donc de comparer plus équitablement les résultats obtenus avec notre méthode avec les résultats d'un codage à court terme (en terme de débits de coefficients). Cette comparaison servira de base à la mise en perspective de nos travaux dans des applications de codage à très bas débit, mise en perspective qui sera plus largement discutée au Chapitre 6.

[89] Comme on l'a déjà mentionné avant, on suppose que dans un tel codeur la fréquence fondamentale est transmise séparément. On ne s'intéresse ici qu'au débit relatif aux coefficients codant les amplitudes, comme on l'a fait au Chapitre 3.

[90] En fait, on teste aussi la condition $R_{min1} = R_{min2}$ dans cette expérience pour vérifier dans quelle mesure cette contrainte est appropriée ou pas (d'après les résultats de la section précédente, elle ne devrait pas l'être !)

5.4.2.1. Débits de coefficients avec analyse pitch-synchrone :

Les débits étant en principe sensibles au genre de la voix du fait des différences entre les caractéristiques spectrales, nous avons ici encore séparé les résultats obtenus pour les voix de femmes et les voix d'hommes. Les résultats sont reportés dans le Tableau 5.1 pour les voix de femmes et dans le Tableau 5.2 pour les voix d'hommes. En toute logique, on peut voir que la gamme des débits possibles selon les différentes configurations varie beaucoup (comme on pouvait s'y attendre à partir de la variation observée dans nos expériences des ordres M et P des modèles d'enveloppe et des modèles à long terme, en fonction de cette configuration). De plus, pour une valeur de R_{min2} donnée, on peut voir que la valeur du débit varie beaucoup en fonction de R_{min1} (lorsqu'on a suffisamment de valeurs de R_{min1} possibles bien sûr). Dans les tableaux, on a ainsi mis en bleu la valeur optimale du débit obtenue pour chaque valeur de R_{min2} testée. On peut constater que cette valeur est distribuée assez régulièrement dans les tableaux, c'est-à-dire approximativement en « diagonale centrale » parmi l'ensemble des couples de valeurs testées : elle est ainsi obtenue pour un écart suffisant entre R_{min1} et R_{min2}, ce qui confirme les résultats présentés dans la section précédente (voir une explication dans cette même section précédente), et à l'inverse, il semble inutile et même pénalisant de fixer une valeur de R_{min1} trop grande lorsque R_{min2} est inférieur à 75%. Dans ce dernier cas, le fameux déséquilibre dont on a parlé dans la section précédente penche du côté de la modélisation spectrale. Au total, le débit optimal semble atteint pour un rapport $\gamma = R_{min2}/R_{min1}$ de l'ordre de 0,80. Par exemple, si on fixe le rapport cible global R_{min2} égal à 65%, il faudrait régler le rapport R_{min1} aux alentours de 80%. Notons que la valeur optimale de R_{min1} n'est pas forcément exactement la même pour les voix d'hommes et les voix de femmes. Nous allons plus largement discuter des différences inter-genres dans la suite de cette section[91].

F R_{min2} / R_{min1}	90%	85%	80%	75%	70%	65%	60%
90%	2533	2083	1510	994	635	433	321
85%	*	2052	1549	1010	620	399	285
80%	*	*	1641	1119	663	395	262
75%	*	*	*	1298	784	433	258
70%	*	*	*	*	989	527	278
65%	*	*	*	*	*	763	352
60%	*	*	*	*	*	*	544

Tableau 5.1 : Débits de coefficients du modèle 2D des amplitudes spectrales obtenus pour différentes valeurs du couple $\{R_{min1}\,;\,R_{min2}\}$. Ces débits sont des débits moyens calculés sur l'ensemble des sections de parole voisée de la base de test pour les voix de femmes (environ 1800 sections) et en utilisant la méthode d'analyse *pitch-synchrone*. En bleu, on a marqué les débits optimaux pour une valeur de R_{min2} donnée.

[91] Notons que outre les différences de caractéristiques spectrales inter-genres, il faut aussi tenir compte de l'échantillonnage assez grossier de R_{min1} et R_{min2}. Cet échantillonnage est en effet limité par le coût de calcul de ces expérimentations : un calcul de débit moyen pour une configuration donnée sur l'ensemble des 3500 sections de la base de donnée peut prendre plusieurs heures. Cependant, on cherche ici à dégager un résultat assez général et pour cela cet échantillon de configurations est suffisant.

H R_{min2} / R_{min1}	90%	85%	80%	75%	70%	65%	60%
90%	2152	1778	1300	896	643	486	388
85%	*	1661	1365	846	570	411	320
80%	*	*	1303	895	562	376	279
75%	*	*	*	998	622	379	257
70%	*	*	*	*	770	436	258
65%	*	*	*	*	*	596	297
60%	*	*	*	*	*	*	417

Tableau 5.2 : Mêmes résultats que le Tableau 5.1 mais pour les voix d'hommes (environ 1700 sections de parole).

Si on s'intéresse maintenant aux valeurs optimales obtenues en tant que telles, on peut dégager les ordres de grandeur suivants : pour R_{min2} égal à respectivement 80%, 75%, 70%, 65% et 60%, on obtient pour les voix de femmes, un débit de approximativement 1500, 1000, 600, 400 et 250 coefficients/s respectivement, et on obtient pour les voix d'hommes 1250, 850, 550, 375 et 250 coefficients/s respectivement. Comme on a déjà dit qu'une valeur de R_{min2} de l'ordre de 75% assure une bonne qualité du signal de synthèse (voir la section suivante pour plus de détails sur ce point), on peut noter que cette qualité est atteinte pour un débit de l'ordre de moins de 900 coefficients/s en moyenne inter-genre. Notons que par la suite, on ne considérera donc pas les valeurs de débit pour R_{min2} = 90% et 85%, qui sont trop élevées à notre goût, d'autant plus que ces débits chutent relativement vite en fonction de la diminution de R_{min2}. On peut imaginer qu'on va ainsi conserver une qualité assez bonne en regard de la baisse du débit. Nous reviendrons sur ce point à la Section 5.4.3.

On peut constater par ailleurs que les débits optimaux pour les voix de femmes sont un peu supérieurs à ceux pour les voix d'hommes. Ceci est assez surprenant en regard des remarques déjà faites sur les différences de caractéristiques spectrales inter-genre, et s'oppose aux résultats connus dans la littérature de la modélisation et du codage de la parole (voir par exemple [Cappé *et al.*, 1995] déjà cité sur ce point, ou [McAulay & Quatieri, 1995]). Avant de tenter d'expliquer ce phénomène, il faut noter que ce résultat n'est pas vérifié pour toutes les valeurs de débits. En particulier, on observe le résultat inverse (et donc plus prédictible) pour les valeurs « supérieures hautes » au débit optimal dans le Tableau 5.1, c'est-à-dire pour les valeurs de R_{min1} plus grandes que celles correspondant au débit optimal. De fait, on dirait que la valeur optimale du débit correspond à une valeur de « basculement » où les valeurs de débit pour les voix d'hommes deviennent inférieures à celles pour les voix de femmes. L'explication que nous tentons d'apporter à ce phénomène consiste en deux points principaux. D'une part, il semble donc que par rapport au voix d'hommes, les spectres de voix de femmes doivent être suffisamment bien modélisés selon la dimension fréquentielle avant d'être modélisés selon la dimension temporelle. Ceci peut s'expliquer par la relative sensibilité de ces spectres à une modélisation d'enveloppe du fait qu'ils comportent généralement moins d'harmoniques que les spectres de voix d'hommes. Il est donc important que la première modélisation fréquentielle soit de bonne qualité avant d'aborder la deuxième modélisation temporelle, ce qui peut expliquer que les débits augmentent comparati-

vement plus vite pour les voix de femmes que pour les voix d'hommes lorsque R_{min1} décroit à partir de la valeur optimale. D'autre part, une part importante de cette différence inter-genre entre les débits pourrait provenir… de l'analyse ! Rappelons en effet que celle-ci est effectuée de façon pitch-synchrone dans notre étude. Comme la fréquence fondamentale des voix de femmes est généralement significativement plus élevée que celle des voix d'hommes, on a un échantillonnage temporel des trames d'analyse, et donc des vecteurs de paramètres spectraux, significativement plus dense pour les voix de femmes que pour les voix d'hommes[92]. Or, il est tout à fait possible, et même fortement probable que la modélisation à long terme selon l'axe temporel soit sensible à cet échantillonnage. En effet, comme le critère à vérifier par cette modélisation est une contrainte moyenne sur l'ensemble des amplitudes (en terme de pourcentage R_{min2}), plus d'amplitudes signifie plus de contraintes à vérifier pour le modèle à long terme et donc potentiellement la nécessité d'avoir des ordres P plus élevés que dans un cas où on a moins de mesures. Ainsi, la lecture comparative des résultats hommes/femmes semble indiquer qu'en termes de débit de coefficients, le choix d'une procédure d'analyse pitch-synchrone que nous avions fait pour volontairement augmenter le nombre de données disponibles pour la modélisation à long terme (voir la Section 3.3.1.2) se retourne en quelque sorte contre nous ! Bien entendu, ce constat n'est pas dramatique en soi. N'oublions pas que nous cherchons à capturer la dynamique temporelle des caractéristiques spectrales de nos signaux de parole, certes efficacement, mais aussi suffisamment précisément, ce qui justifie un bon échantillonnage temporel dans notre série d'études. Dans une optique de codage à bas débit, ce point doit être reconsidéré, avec un échantillonnage des modèles d'enveloppe moins dense et représentatif de celui utilisé dans les codeurs usuels (par exemple avec des fenêtres d'analyse de taille fixe décalées de 20 ms)[93] : cette observation est précisément à l'origine des expérimentations complémentaires que nous donnons dans la sous-section suivante.

Avant de donner les résultats pour ce mode « codage », nous allons comparer les débits obtenus par la modélisation 2D avec des débits obtenus en configuration « 1D-fréquentielle » (c'est-à-dire avec la seule modélisation d'enveloppe spectrale issue de la première partie de l'algorithme sans modélisation à long terme) *pour le même échantillonnage des trames d'analyse* (c'est-à-dire ici avec l'échantillonnage pitch-synchrone). Les résultats de ces diverses configurations sont donnés dans les Tableaux 5.3 et 5.4. Notons que pour que la comparaison entre modélisation 2D et modélisation 1D-fréquentielle soit juste, la valeur de R_{min1} dans le second cas est fixée à la même valeur que R_{min2} dans le premier cas. Ainsi, on compare les débits de ces deux approches pour les mêmes pourcentages cibles finaux sur les amplitudes modélisées. Dans les Tableaux 5.3 et 5.4, on a reporté l'ordre moyen M_{moy} des modèles d'enveloppe obtenus pour la modélisation 1D-fréquentielle. Cet ordre moyen est calculé comme la moyenne de l'ordre M obtenu pour chaque trame, sur l'ensemble de la base de données.

[92] Rappel : environ 220 mesures par seconde pour les voix de femmes contre environ 140 mesures par seconde pour les voix d'hommes pour notre base de données ; on a déjà donné ces valeurs dans les Chapitres 3 et 4.

[93] Un tel échantillonnage, identique pour les voix de femmes et les voix d'hommes, devrait vraisemblablement rééquilibrer les débits de coefficients 2D selon l'ordre « habituel », avec des débits inférieurs pour les voix de femmes par rapport à ceux pour les voix d'hommes.

Les valeurs des Tableaux 5.3 et 5.4 permettent d'affirmer que la stratégie de modélisation à long terme des coefficients d'enveloppe spectrale permet de diminuer de façon significative le débit des coefficients par rapport à une approche purement 1D selon la dimension spectrale : en effet, pour les voix de femmes, on passe respectivement d'un débit de 1918, 1640, 1425, 1224 et 1068 coefficients/s en 1D à respectivement 1510, 994, 620, 395 et 258 coefficients/s en 2D pour les valeurs de pourcentage cible R_{min} allant de 80% à 60% par pas de 5%. Les gains respectifs en débit de coefficients sont donc de 21,3%, 39,4%, 56,5%, 67,7% et 75,8%. Pour les voix d'hommes, on passe de respectivement 1481, 1240, 1049, 870 et 733 coefficients/s en 1D à 1300, 846, 562, 376 et 257 coefficients/s en 2D pour le même échantillonnage du pourcentage cible. Les gains respectifs sont alors de 12,2%, 31,8%, 46,4%, 56,8% et 64,9%. On remarque donc que les gains de débit augmentent significativement lorsqu'on tend vers les bas débits, c'est-à-dire lorsqu'on relâche la contrainte sur le pourcentage cible des amplitudes modélisées vérifiant la contrainte perceptive. Ceci est dû au fait que les débits 2D diminuent plus vite avec la baisse du pourcentage cible que les débits 1D. Ceci confirme l'efficacité de l'approche 2D en terme de pouvoir de compression des données lorsque les contraintes de la modélisation ne sont pas trop rigoureuses. Pour des valeurs de R_{min2} qui restent raisonnables en regard de la bonne qualité des signaux de synthèse (voir la section suivante), la modélisation à long terme joue pleinement son rôle de représentation efficace des données. Par exemple, pour $R_{min2} = 70\%$, on a un gain de coefficients moyenné entre voix de femmes et voix d'hommes supérieur à 50%. Pour des valeurs de R_{min2} inférieures (par exemple 60%), on atteint des gains moyens inter-genres impressionnants : de l'ordre de 70%, mais dans ce cas, la qualité des signaux est aussi dégradée (voir la section suivante).

R_{min1}	R_{min2}	M_{moy}	Modèle 1D	Modèle 2D
90%	80%	7,6	1918	1510
90%	75%	6,3	1640	994
85%	70%	5,3	1425	620
80%	65%	4,4	1224	395
75%	60%	3,7	1068	258

Tableau 5.3 : Débits moyens de coefficients obtenus pour les voix de femmes (environ 1800 sections de parole voisées) pour la modélisation 2D et pour la modélisation 1D-fréquentielle avec analyse des paramètres pitch-synchrone (Modèle 1D). M_{moy} est l'ordre moyen du modèle d'enveloppe issu de la modélisation 1D-fréquentielle, moyenné sur l'ensemble des sections. La valeur de R_{min1} utilisée pour cette modélisation 1D-fréquentielle est celle de R_{min2} affichée dans le tableau (voir le texte).

R_{min1}	R_{min2}	M_{moy}	Modèle 1D	Modèle 2D
90%	80%	9,3	1481	1300
85%	75%	7,5	1240	846
80%	70%	6,1	1049	562
80%	65%	4,9	870	376
75%	60%	4,0	733	257

Tableau 5.4 : Mêmes résultats que le Tableau 5.3 mais pour les voix d'hommes (environ 1700 sections de parole voisées).

On peut noter enfin que les gains de débits sont supérieurs pour les voix de femmes par rapport aux voix d'hommes. Cela n'a rien de surprenant vu notre discussion menée plus haut sur l'influence de l'échantillonnage temporel des enveloppes spectrales qui est plus élevé pour les voix de femmes que pour les voix d'hommes. Certes, on a déjà dit que cet échantillonnage plus élevé pouvait vraisemblablement expliquer le besoin de débits 2D plus élevés par rapport aux voix d'hommes. Mais parallèlement, il conduit aussi à des valeurs de débits de modélisation 1D-fréquentielle plus élevés (on peut dire que par rapport aux voix d'hommes, l'échantillonnage temporel des spectres plus élevé a un effet dominant sur le débit par rapport à leur échantillonnage fréquentiel, c'est-à-dire le nombre d'harmoniques, lui-même plus faible). Finalement le rapport des débits entre modélisation 2D et modélisation 1D-fréquentielle est favorable ici aux voix de femmes (on voit que l'écart relatif entre débits 2D pour les voix de femmes et d'hommes se resserre au fur et à mesure que le pourcentage cible diminue, alors que ce n'est pas le cas pour les débits de la configuration 1D-fréquentielle).

5.4.2.2. Débits de coefficients avec analyse avec fenêtrage classique

Comme mentionné plus haut, pour faire une comparaison plus juste de notre méthode 2D avec les codeurs usuels, nous avons appliqué notre algorithme sur l'ensemble de notre base de données avec cette fois une analyse des paramètres avec fenêtre fixe (25 ms) et décalage fixe (20 ms, soit un recouvrement de 5 ms). Ceci correspond à un échantillonnage à 50 Hz, à comparer avec les 140 et 220 mesures/s en moyenne de l'analyse pitch-synchrone (respectivement pour les voix d'hommes et de femmes). Les nouveaux résultats sont reportés dans le Tableau 5.5 pour les voix de femmes et dans le Tableau 5.6 pour les voix d'hommes, pour les mêmes réglages que nous avons appliqués pour les Tableaux 5.1 et 5.2. Dans ces nouveaux tableaux, nous avons à nouveau mis en bleu la valeur optimale du débit obtenue pour chaque valeur de R_{min2} testée. Ces valeurs sont distribuées assez régulièrement, ce qui confirme les résultats présentés dans la section précédente. Les valeurs optimales obtenues pour R_{min2} égal à respectivement 80%, 75%, 70%, 65%, 60%, sont respectivement 345, 266, 205, 158 et 120 coefficients/s pour les voix de femmes, et respectivement 422, 321, 246, 189 et 141 coefficients/s pour les voix d'hommes. La première conclusion qui s'impose est donc une baisse très significative des débits de coefficients dans cette nouvelle configuration d'échantillonnage des paramètres modélisés. Comme on pouvait s'y attendre, la parcimonie de la procédure d'analyse comparée à l'analyse pitch-synchrone se répercute complètement sur la parcimonie du modèle 2D résultant. A titre d'exemple, pour R_{min2} = 75% (on verra à la section suivante que cette valeur continue d'assurer une qualité assez bonne du signal de synthèse malgré l'échantillonnage temporel plus faible), le débit passe de 994 coefficient/s pour l'analyse pitch-synchrone à 266 coefficients/s pour l'analyse à fenêtre fixe, pour les voix de femmes. Pour les voix d'hommes, la comparaison donne 321 vs. 846 coefficients/s. L'influence de l'analyse est donc tout simplement énorme : les débits sont divisés par un facteur allant de 2 à 4 environ selon les configurations. Ceci permet d'envisager sérieusement l'utilisation de la modélisation 2D dans une optique de codage de parole à très bas débit, ce que nous rediscuterons dans le Chapitre 6.

Notons pour finir cette comparaison que dans la nouvelle configuration d'analyse, et donc avec le même échantillonnage des données pour les voix d'hommes et les voix de femmes, la hiérarchie habituelle de coût de codage supérieur pour les voix d'homme par

rapport aux voix de femmes est cette fois respectée, ce qui confirme la discussion que nous avons donnée sur ce point dans la sous-section précédente.

F R_{min2} / R_{min1}	90%	85%	80%	75%	70%	65%	60%
90%	*	442	345	296	248	207	175
85%	*	*	361	266	222	183	153
80%	*	*	*	273	205	167	137
75%	*	*	*	*	208	158	126
70%	*	*	*	*	*	159	120
65%	*	*	*	*	*	*	121
60%	*	*	*	*	*	*	*

Tableau 5.5 : Débits de coefficients du modèle 2D des amplitudes spectrales obtenus pour différentes valeurs du couple $\{R_{min1} ; R_{min2}\}$. Ces débits sont des débits moyens calculés sur l'ensemble des sections de parole voisée de la base de test pour les voix de femmes (environ 1800 sections) pour la méthode d'analyse avec une fenêtre fixe (25 ms). En bleu, on a marqué les débits optimaux pour une valeur de R_{min2} donnée.

H R_{min2} / R_{min1}	90%	85%	80%	75%	70%	65%	60%
90%	*	566	422	384	323	279	244
85%	*	*	462	321	278	234	203
80%	*	*	*	338	246	205	174
75%	*	*	*	*	254	189	157
70%	*	*	*	*	*	192	141
65%	*	*	*	*	*	*	146
60%	*	*	*	*	*	*	*

Tableau 5.6 : Mêmes résultats que le Tableau 5.5 mais pour les voix d'hommes (environ 1700 sections de parole).

Pour finir cette étude, et pour avoir une bonne appréciation quantitative des valeurs de débit obtenues, nous reprenons le principe de comparer les résultats de l'approche « 2D » avec l'approche « 1D », c'est à dire avec une modélisation uniquement fréquentielle, sans modélisation temporelle. Dans ce cas, comme dans la sous-section précédente, le ratio cible global de ces deux approches est identique. Dans le Tableau 5.7, nous avons reporté les débits en terme de nombre total de coefficients/s utilisé par chaque modélisation pour trois configuration des ratios cibles. En comparant ces débits, on obtient un gain allant de 10% à 50% selon les conditions en faveur de la modélisation 2D. Ces résultats confirment ceux présentés dans les Tableaux 5.3 et 5.4 et montrent à nouveau l'apport de la modélisation temporelle des enveloppes spectrales, dans cette nouvelle configuration « analyse classique des paramètres ».

*Rmin*1	*Rmin2*	Voix de Femmes		Voix d'hommes	
		Modèle 1D	Modèle 2D	Modèle 1D	Modèle 2D
90%	80%	395	345	462	422
80%	70%	297	205	334	246
70%	60%	230	120	243	141

Tableau 5.7 : Débits moyens de coefficients obtenus pour les voix de femmes (environ 1800 sections de parole voisées) et les voix d'hommes (environ 1700 sections de parole voisées) pour la modélisation 2D et pour la modélisation 1D-fréquentielle avec analyse des paramètres avec fenêtre fixe (25 ms) et décalage fixe (20 ms). La valeur de R_{min1} utilisée pour la modélisation 1D-fréquentielle est celle de R_{min2} affichée dans le tableau (voir le texte).

5.4.3. Tests d'écoute

Dans cette section, nous reportons quelques résultats concernant l'évaluation de la qualité des signaux traités avec l'approche 2D. Avant cela, il est nécessaire de donner brièvement quelques précisions sur le processus de synthèse dans ce cadre de modélisation 2D. Ainsi, la synthèse est réalisée en appliquant d'abord une interpolation linéaire entre les amplitudes résultant de l'étape 4 de la seconde partie de l'algorithme 2D dans sa version simplifiée décrite à la Section 5.4.1.2. Cette interpolation inclut le processus habituel de « naissance et de mort » pour les harmoniques qui dépassent la fréquence de Nyquist (voir Section 1.3.2). Rappelons de plus qu'un codeur de parole à bas débit utilisant la méthode proposée doit coder au moins la trajectoire de fréquence fondamentale, à défaut de coder toutes les trajectoires de fréquence ou de phase de toutes les harmoniques. Dans cette étude 2D, nous ne nous intéressons qu'à la modélisation des amplitudes et les trajectoires de phase sont synthétisées par interpolation linéaire des mesures dépliées, comme dans le Chapitre 3. Les équations 1.2 et 1.3 sont ensuite utilisées pour produire le signal de synthèse. Comme mentionné précédemment, dans cette étude le processus d'analyse-modélisation-synthèse concerne uniquement les parties voisées. Ainsi, pour la synthèse de phrases complètes, les sections non voisées sont concaténées avec les sections voisées modélisées avec une pondération locale, comme on l'a fait dans les chapitres précédents.

Pour comparer les signaux originaux et les signaux de synthèse 2D (*i.e.* issus de la modélisation 2D des amplitudes), nous avons à nouveau fait des essais d'écoute informelle sur les signaux de la base de données (voir les Chapitres 3 et 4 pour le protocole des tests d'écoute). Sauf quand mentionné explicitement, les résultats présentés sont ceux obtenus avec les paramètres fournis par l'analyse pitch-synchrone. Les résultats principaux de ces tests sont les suivants. Pour un réglage du rapport-cible R_{min2} minimum d'environ 70%, et un réglage de R_{min1} tel qu'on ait approximativement $R_{min2} \approx 0{,}80 \times R_{min1}$,[94] on obtient un signal de synthèse de bonne qualité. Ce réglage des rapports-cibles est bien sûr inspiré par les résultats des sections précédentes, notamment

[94] Un réglage typique peut ainsi être $R_{min1} = 90\%$, $\gamma = 80\%$, d'où $R_{min2} = 72\%$.

en terme de débit optimal à R_{min2} donné. On entend par « bonne qualité » une qualité assez proche de celle du signal original, même si les deux signaux ne sont pas complètement indiscernables. Bien que la forme d'onde d'un signal de parole ne soit pas en relation directe avec sa qualité sonore, on illustre la qualité de la synthèse sur la Figure 5.7. Cette figure reporte la forme d'onde du signal correspondant aux Figures 5.5 et 5.6 : on peut voir que le signal de synthèse est relativement fidèle à l'original. Ceci est particulièrement vrai pour le cas $R_{min2} = 70\%$, $R_{min1} = 75\%$, $M = 8$ et $P = 76$ que nous avons critiqué dans la section précédente. C'est précisément l'effort (mal proportionné) porté sur la dimension temporelle qui assure ici cette fidélité de la forme d'onde. Paradoxalement, la forme d'onde associée à la « bonne » configuration $R_{min2} = 70\%$, $R_{min1} = 90\%$, $M = 14$ et $P = 29$ est moins fidèle au signal original, du fait de la parcimonie de la modélisation selon la dimension temporelle. En particulier, l'enveloppe temporelle de ce signal semble quelque peu lissée par rapport à celle du signal original. Cependant, les deux signaux sont de qualité identiques : ils sont indiscernables à l'écoute (et on le confirme, proches de l'original). Ainsi, l'efficacité de la modélisation 2D, notamment selon la dimension temporelle, ne nuit pas à cette qualité. Ce résultat est un pilier de notre approche, qu'on avait déjà pu mettre en évidence auparavant (voir la Figure 3.6 par exemple).

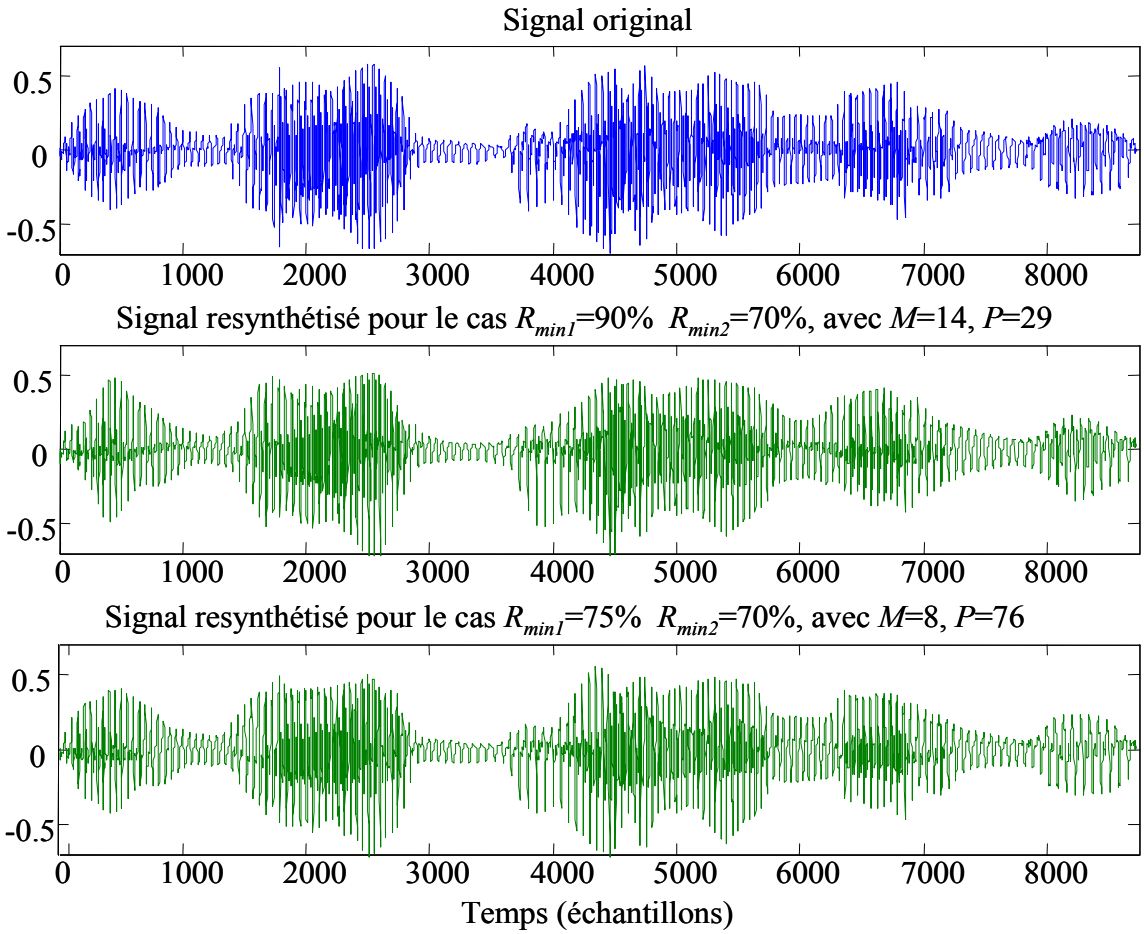

Figure 5.7 : Signal de synthèse généré avec les amplitudes issues de la modélisation à long terme 2D ($F_e = 8$ kHz). Au milieu : $R_{min2} = 70\%$, $R_{min1} = 90\%$, $M = 14$ et $P = 29$ correspondant à la Figure 5.6 ; en bas : cas $R_{min2} = 70\%$, $R_{min1} = 75\%$, $M = 8$ et $P = 76$ correspondant à la Figure 5.5 ; en haut : signal original.

Par ailleurs, la qualité de synthèse des signaux issus de la modélisation 2D est aussi comparable à celle des signaux générés au Chapitre 3 avec la modélisation à long terme des 10 premières trajectoires d'amplitudes et avec le même pourcentage-cible sur le critère perceptif[95]. Ce dernier point est très important. En effet, par rapport à ces résultats du Chapitre 3, un point fort majeur de l'approche en 2D est qu'elle englobe l'ensemble des harmoniques sur toute la bande du signal (0-4 kHz dans cette étude) et pas seulement les 10 premières harmoniques comme dans nos études précédentes de modélisation à long terme 1D. Par conséquent, le pourcentage cible du critère perceptif s'applique ici conjointement à toutes les harmoniques en même temps, alors qu'il ne caractérisait que les 10 premières harmoniques dans le Chapitre 3. Le fait qu'on obtient une qualité similaire à celles des signaux du Chapitre 3 montre qu'avec l'approche 2D, même les harmoniques de rang plus élevés (que dix) sont relativement bien représentées. Ceci est confirmé par le fait que (pour R_{min2} et R_{min1} bien dimensionnés) la qualité des signaux synthétisés à partir de la modélisation 2D est aussi très proche de celle des signaux de références synthétisés avec une interpolation à court terme des mesures de toutes les harmoniques.

Compte tenu de ces résultats de test d'écoute, il est intéressant de revenir sur les débits et de comparer les débits 2D donnés dans la section précédente (avec l'analyse pitch-synchrone[96]) avec les débits de la modélisation 1D à long terme (*i.e.* dans la configuration de modélisation des trajectoires d'amplitudes du Chapitre 3) « extrapolés » à l'ensemble des harmoniques. On a vu au Chapitre 3 que les débits moyens obtenus avec un pourcentage-cible de 75% étaient de l'ordre de 25 coefficients/s/harmonique en moyenne sur les dix premières harmoniques. En prenant respectivement une voix de femme avec une fréquence fondamentale autour de 220 Hz en moyenne et une voix d'homme avec une fréquence fondamentale de l'ordre de 140 Hz en moyenne (on rappelle que ce sont les valeurs moyennes de notre base de données), on obtient sur la bande 0-4 kHz respectivement une moyenne d'environ 18 et 29 harmoniques respectivement. Ceci correspondrait à des débits respectifs de l'ordre de 450 et 725 coefficients/s pour l'ensemble des harmoniques, soit des débits compétitifs par rapport aux débits 2D reportés dans le Tableau 5.2 (pour $R_{min2} = 75\%$). Toutefois, cette analyse est biaisée par le fait qu'on ne tient pas compte d'une observation importante effectuée au Chapitre 3 : pour le même pourcentage-cible sur le critère perceptif, les débits de coefficients augmentaient significativement avec le rang des harmoniques du fait de la complexité croissante des trajectoires d'amplitudes. Les débits de la modélisation à long terme 1D extrapolés ci-dessus sont donc sans doute significativement sous-estimés par rapport à des débits réellement équivalents aux débits 2D avec le même pourcentage-cible sur toutes les harmoniques du signal. De plus, la qualité des signaux resynthétisés avec l'approche 2D laisse à penser que la

[95] On rappelle qu'un score de 75% sur le critère perceptif pour ces dix premières harmoniques permettait alors une bonne qualité de synthèse, même s'il fallait augmenter ce seuil, notamment sur les toutes premières harmoniques, pour atteindre la transparence ou quasi-transparence par rapport aux signaux de référence synthétisés à court terme.

[96] Avec une analyse sur fenêtre fixe et décalage de l'ordre de 20 ms, on a vu qu'on obtient des débits 2D très inférieurs à ceux obtenus avec l'analyse pitch-synchrone, mais si on veut les comparer avec l'approche 1D temporelle, il faudrait effectuer aussi celle-ci avec la même analyse sur fenêtre fixe, ce que nous n'avons pas réalisé dans ces travaux.

modélisation 2D est efficace pour capter intrinsèquement cette complexité des formes spectrales contenue dans les hautes harmoniques.

Il reste à mentionner brièvement l'influence de la méthode d'analyse (ou plutôt de l'échantillonnage temporel des paramètres modélisés. Nous avons procédé à l'écoute comparative des signaux avec les amplitudes modélisés en 2D à partir des mesures obtenues avec fenétrage fixe à 50 Hz (voir la Section 5.4.2.2). Il apparaît qu'en dépit d'un échantillonnage temporel plus faible, les signaux restent de qualité tout à fait correcte. Même si dans certains cas, des différences peuvent apparaître entre les signaux modélisés en 2D avec les deux méthodes d'analyse différentes, les signaux modélisés en 2D à partir des mesures à 50 Hz sont très semblables (*i.e.* indiscernables) des signaux synthétisés avec interpolation linéaire des amplitudes reconstruites à partir des modèles cepstraux 1D tirés des mêmes mesures à 50 Hz (ce qui est la condition de référence pour les codeurs usuels, sans quantification). Compte tenu des valeurs de débits (2D vs. 1D) obtenus dans cette configuration d'analyse (voir la Section 5.4.2.2), ces tests confirment donc le fort potentiel de la méthode 2D proposée dans le cadre du codage de la parole à très bas débit.

Au final, on peut donc conclure sur l'efficacité de l'approche 2D à la fois en terme de qualité de synthèse, de débit de coefficients, et de « généralité » de la modélisation réalisée de façon globale sur l'ensemble de la bande du signal. Pour finir cette section de façon un peu « ludique », nous reportons une dernière observation amusante. Si le rapport R_{min2} diminue significativement, les trajectoires des coefficients du MCD à long terme deviennent notablement « simplifiées » (puisque la contrainte sur l'ordre P est relâchée). Il peut en être de même pour la forme de l'enveloppe spectrale si R_{min1} est aussi relaché, mais on peut aussi choisir de conserver une bonne modélisation spectrale en gardant une valeur suffisamment élevée de R_{min1} tout en diminuant R_{min2}. Dans ce dernier cas, le signal de synthèse restitué, bien qu'ayant une bonne sonorité, s'éloigne alors du signal original, et ceci d'une façon assez particulière : il tend vers une version hypo-articulée de ce dernier, tout en conservant un aspect relativement naturel. En d'autres termes, on a vraiment l'impression que le locuteur fait moins d'effort d'articulation, plutôt qu'une dégradation de type bruit, artefacts, ou perte de naturel, comme il en arrive dans les codeurs à très bas débit. Comme on a vu que les débits de coefficients de la modélisation à long terme 2D chutent assez vite en fonction du rapport cible R_{min2}, un des challenges ouverts par cette technique est de trouver le bon compromis qualité/débit dans une application de codage à très bas débit, en tenant compte de cette particularité.

5.5. Conclusion

Les résultats de ce chapitre portent deux points : le premier point, la modélisation de l'enveloppe spectrale qui précède la modélisation à long terme proprement dite, fournit plusieurs avantages :

- Elle permet de résoudre le problème de « taille variable » des jeux de paramètres d'amplitude d'une trame d'analyse à l'autre, dû aux variations de la fréquence fondamentale (et à la présence de composantes de bruit), car l'enveloppe est modélisée en utilisant un ordre fixe sur la section de parole considérée à long terme.

- Elle permet de réduire la taille de ces jeux de paramètres avant la modélisation à long terme, puisque l'ordre du modèle d'enveloppe est généralement très inférieur au nombre d'amplitudes mesurées, c'est un point important dans l'optique de l'utilisation de cette modélisation 2D pour le codage de parole à très bas débit.
- Elle permet également d'avoir une représentation paramétrique du spectre d'amplitude adapté à de nombreuses méthodes de transformation des signaux sonores, telles que l'étirement temporel ou bien la conversion de voix par exemple. En effet, la plupart des travaux menés dans ce domaine traitent essentiellement de la transformation de l'enveloppe spectrale. Cette représentation est particulièrement adaptée aux transformations du signal en fréquence, telle que le *pitch-scaling* par exemple.

Le deuxième point est que ce travail a confirmé la robustesse et la généralité du Modèle en Cosinus Discrets qui est adéquat pour modéliser à la fois l'enveloppe spectrale (comme déjà montré dans [Cappé *et al.,* 1995]) et la trajectoire temporelle de paramètres (comme déjà montré aux Chapitres 3 et 4). Tenir compte de ces deux aspects réunis dans un seul modèle 2D a abouti à de nouvelles avancées. Dans cette nouvelle étude, la raison principale de l'efficacité de la modélisation 2D et de l'algorithme d'ajustement associé est la « double » variabilité intrinsèque du débit : l'ordre M du modèle d'enveloppe et l'ordre P du modèle temporel sont tous les deux ajustés sur les caractéristiques locales du signal. L'approche 2D et l'algorithme associé peuvent ainsi permettre de réduire significativement le nombre de coefficients pour la représentation des amplitudes spectrales : pour des valeurs du pourcentage-cible de l'algorithme assurant une bonne qualité du signal de synthèse (soit $R_{min2} \approx 70\%$), un gain d'environ 50% est obtenu par rapport au débit de coefficients issu d'une modélisation par enveloppe spectrale 1D réalisée avec le même échantillonnage temporel des données.

Dans les études présentées, cet échantillonnage est généralement assez riche du fait qu'il est pitch-synchrone : on a un jeu d'amplitude et donc une enveloppe spectrale à chaque période de signal. Dans une optique de codage à bas débit, il est plus juste de comparer les débits issus des deux types de modélisation (1D-fréquentielle et 2D) appliquées sur des données obtenues avec un échantillonnage plus représentatif de celui utilisé généralement dans ce type d'application, soit 100 ou 50 Hz (c'est-à-dire des trames d'analyse-synthèse décalées de 10 ou 20 ms). Dans cette optique, nous avons mené une série d'expérimentations complémentaires reprenant le calcul des débits sur les mêmes données échantillonnées par un fenêtrage fixe à 50 trame/s (décalage de 20 ms). Les résultats, avec des débits particulièrement parcimonieux, ont confirmé le potentiel de l'approche 2D pour le codage sinusoïdal (il ne manque plus que la quantification, si l'on ose dire…), un point qui sera largement rediscuté au Chapitre 6.

En conclusion de ce chapitre, on peut dire que les résultats obtenus, bien que très prometteurs, notamment en terme de débits de coefficients, peuvent encore être améliorés. En particulier, dans l'algorithme proposé, pour des raisons de simplicité et de coût calculatoire, nous avons séparé en deux temps l'estimation de l'ordre M du modèle d'enveloppe et l'estimation de l'ordre P du modèle à long terme. Cette démarche est théoriquement sous-optimale par rapport à une estimation conjointe de M et P pour chaque section de parole modélisée. Ainsi, un algorithme plus performant, mais plus

coûteux, incluant la réalisation de cette estimation conjointe reste à implémenter et à évaluer. Ce point représente à notre sens une étape à explorer dans la perspective de l'utilisation de notre approche 2D pour le codage *half-duplex* (sans interactivité temps-réel) de la parole (et potentiellement d'autres types de sons) à très bas débit.

Chapitre 6

6. Bilan de l'étude, discussion autour des perspectives, et conclusion

Dans cette thèse nous avons successivement posé le contexte (Chapitre 1) et les grands principes (Chapitre 2) de la modélisation à long terme. Ensuite, nous avons réalisé la modélisation à long terme des paramètres d'amplitude du modèle sinusoïdal (Chapitre 3) puis celle des paramètres de phase (Chapitre 4). Cette modélisation a été effectuée « en une dimension », selon l'axe temporel. Dans le Chapitre 5, nous avons généralisé l'approche à une modélisation « en deux dimensions », d'abord selon l'axe fréquentiel puis selon la dimension temporelle. Dans ce chapitre final, nous dressons d'abord à la Section 6.1 un bilan des points qui ressortent de ce cheminement. Ce bilan va du général (par rapport à l'objectif général de la thèse) au particulier (ce qui ressort de chaque étape) même s'il n'a pas pour ambition d'être complètement exhaustif, loin de là. On dresse aussi à la fin de cette section un bilan rapide des publications qui ont jalonné ce travail.

En nous appuyant sur ce bilan, nous nous tournons ensuite à partir de la Section 6.2 vers les perspectives plus générales offertes par ce travail. L'objectif est de présenter un éventail (sans doute assez incomplet) de portes ouvertes pour des améliorations des méthodes proposées, pour de nouveaux développements et pour des applications à venir inspirées de l'approche à long terme et exploitant ses propriétés. Ainsi, nous décrivons d'abord à la Section 6.2 deux applications immédiates de notre travail : tout d'abord la possibilité d'appliquer la modélisation à long terme pour effectuer des transformations de base sur les signaux, notamment des étirements/compressions temporelles et des changements de fréquence fondamentale. Puis nous présentons l'utilisation possible des modèles à long terme dans le cadre d'une technique de tatouage audio élaborée récemment en coopération entre l'ICP et le LaBRI. Enfin, *last but not least*, nous décrivons à la Section 6.3 l'extension de la modélisation à long terme à d'autres représentations spectrales. Le modèle sinusoïdal/harmonique + bruit (voir Section 1.4) et le modèle LPC sont spécifiquement concernés par cette extension et une attention particulière leur est portée. Comme ces modèles sont particulièrement performants dans le cadre du codage de la parole, cette section traite conjointement de l'application de la modélisation à long terme au codage à très/ultra bas débit, une application de toute première importance qui nous servira à dessiner une conclusion générale à ce document.

6.1. Bilan du travail réalisé

6.1.1. Un rapide bilan général

Le principal objectif de ce travail était de chercher une représentation à la fois efficace, parcimonieuse, et flexible (c'est-à-dire adaptée à de possibles transformations à effectuer sur le signal) de la dynamique temporelle du signal de parole. En particulier, la dynamique temporelle de la structure spectrale des sections voisées de la parole est considérée. Ceci a été réalisé en deux temps : tout d'abord nous avons procédé à une extraction de cette structure spectrale en employant le modèle sinusoïdal (en version harmonique), puis nous avons appliqué un modèle paramétrique pour chaque trajectoire de paramètre spectral, phase que nous avons dénommé « Modélisation à Long-Terme » du fait que nous avons considéré ces trajectoires sur de longues sections de parole (sections de parole continûment voisées). Ce dernier point constitue le cœur de notre contribution originale, qui outre le principe même de cette étude, est composé de deux points principaux :

- Un point en relation avec les connaissances sur la perception du signal de parole et plus largement des sons : l'adaptation de critère psycho-acoustiques existants au cas spécifique du traitement à long terme qui nous intéresse ;
- Un point technique algorithmique : la proposition, l'implémentation et le test d'une série d'algorithmes d'optimisation (basés sur la même structure générale) réalisant la mise en forme des modèles à long terme.

Les paramètres modélisés à long terme sont soit des amplitudes spectrales (Chapitre 3), soit des phases (Chapitre 4), soit des coefficients d'enveloppe spectrale (Chapitre 5). Dans la suite de cette section nous dressons un bilan plus détaillé de ce travail (bien que certainement non exhaustif) en listant les points principaux positifs qui ressortent à chaque étape de notre étude, et aussi les points plus discutables ou plus problématiques qui restent en suspens ou pour lesquels nous proposons des pistes de travail en vue de leur amélioration.

6.1.2. Bilan de l'application de la modélisation à long terme sur les trajectoires d'amplitude (Chapitre 3)

Les points positifs principaux qui se dégagent de notre série d'études sur la modélisation à long terme des paramètres d'amplitude spectrale sont les suivants :

- L'algorithme proposé s'est montré efficace ; le calcul optimal des modèles est obtenu en un nombre raisonnable d'itérations et les calculs mis en jeu (ajustement au sens des moindres carrés) sont très classiques. Une optimisation de ces calculs pourrait sans doute être mise en œuvre, mais cet aspect ne faisait pas partie des priorités du travail ;
- L'adaptation du modèle de masquage fréquentiel défini généralement dans un cadre stationnaire a été validée par l'expérimentation : la version à long terme de ce modèle que nous avons proposée s'est montrée efficace pour guider perceptivement la précision de la modélisation. De plus cette adaptation est très

directe : il s'agit essentiellement d'un suivi des valeurs du seuil de masquage calculé à court terme le long de l'axe fréquentiel, une tâche relativement simple.

- La modélisation à long terme des trajectoires d'amplitude nous apporte une bonne compression des données. Ceci a été obtenu à la fois grâce à l'exploitation du critère perceptif, et la propriété intrinsèquement lisse des modèles à long terme, combinée à celle des harmoniques, en tout cas pour celles de rang relativement faible. Un facteur de gain de l'ordre de 7 est obtenu par rapport à la modélisation à court terme (c'est-à-dire par rapport au nombre de mesure), dans le cas d'une analyse des paramètres pitch-synchrone.

Parmi les points plus discutables, on peut noter les suivants :

- Dans cette étude du Chapitre 3, à cause de la variation de la fréquence fondamentale qui implique une taille variable pour les vecteurs d'amplitude successifs, nous avons considéré seulement les dix premières composantes sinusoïdales. Il s'agit d'une restriction assez sévère mais qui a tout de même permis de mettre en évidence quelques points intéressants. Par exemple, la difficulté croissante de la modélisation en montant dans les harmoniques du fait de leur caractère de plus en plus bruité. Ou encore un ordre de grandeur fiable pour le débit de coefficients à long terme a aussi été mis en évidence. Ce problème a été en fin de compte résolu implicitement au Chapitre 5, on considérant la modélisation à long terme de coefficients d'enveloppe qui représentent la totalité des harmoniques. Bien sûr, le passage à l'information d'enveloppe ne résout pas la nécessité *in fine* de caractériser le caractère harmonique ou bruité des composantes sous l'enveloppe, mais on peut dire que ce problème est aussi vrai dans les modèles spectraux d'enveloppe à court terme.
- La méthode d'analyse des paramètres sinusoïdaux, basée sur un découpage du signal période par période (et donc une estimation de la fréquence fondamentale) a montré ses limites. Cette méthode dite *pitch-synchrone* a été privilégiée tout au long de ce travail car l'objectif était de suivre au plus près l'évolution temporelle des paramètres spectraux pour « ne rien perdre d'important » dans cette évolution lors de la modélisation à long terme. En fin de compte cet objectif est peut-être atteint mais au prix de limitations notables : on a ainsi montré au Chapitre 5 que l'utilisation (tardive il est vrai) d'une méthode d'analyse plus classique (avec fenêtrage fixe) est finalement plus efficace en terme de parcimonie de l'information (avant, et surtout après modélisation à long terme) avec des différences perceptives non significatives. Cette nouvelle configuration d'analyse a certes été testée uniquement dans le cadre de la modélisation à long terme des coefficients d'enveloppe cepstrale, mais il est vraisemblable qu'une observation similaire apparaisse pour les amplitudes spectrales. De plus, il est aussi vraisemblable qu'une partie du bruit observé sur les trajectoires d'amplitude à modéliser en montant dans les différentes harmoniques soit précisément due à la méthode d'analyse pitch-synchrone. Ainsi, « si c'était à refaire », au vu des observations du Chapitre 5, les expérimentations du Chapitre 3 seraient menées avec l'analyse classique (au moins en parallèle avec l'analyse pitch-synchrone, pour confirmer cette discussion). Bien entendu, ces critiques s'appliquent au choix de la méthode d'analyse des paramètres de phase.

6.1.3. Bilan de l'application de la modélisation à long terme sur les trajectoires de phase (Chapitre 4)

En ce qui concerne la modélisation à long terme des trajectoires de phase, les points « positif » du bilan sont les suivants :

- On peut reprendre les mêmes remarques que pour les amplitudes en ce qui concerne la pertinence et le bon fonctionnement de l'algorithme proposé, ainsi que les bons résultats en terme de compression de l'information. Sur ce dernier point, on peut insister sur le fait déjà mentionné que dans le cadre à long terme, pour réaliser un codeur « *shape invariant* », le coût de codage des trajectoires de phase est égal au coût de codage des trajectoires de fréquence plus un seul paramètre additionnel pour chaque harmonique.
- En ce qui concerne le critère perceptif, l'apport proposé dans cette thèse est particulièrement original : il a consisté à proposer de traiter la précision de la modélisation temporelle de trajectoires de phase en terme de modulation de la fréquence associée à ces trajectoires. A notre connaissance, le problème du suivi fin et de la représentation optimale des trajectoires de phase des composantes d'un signal de parole non-stationnaire est un problème peu abordé dans la littérature. A fortiori, en ce qui concerne l'introduction de critères perceptifs dans cette représentation. Nous avons considéré ce point de près, comme l'illustre notre discussion de fond de la Section 4.1. A nouveau, le critère que nous avons proposé semble validé par les expérimentations.

Au niveau des points à améliorer ou bien, formulons-le de façon plus positive, qui mériteraient un approfondissement, on peut citer :

- La spécificité des trajectoires de phase a été expérimentalement mise en évidence : on a mentionné l'homogénéité des trajectoires des différentes composantes, ce qui est la moindre des choses pour nos signaux voisés et donc pseudo-harmoniques. En revanche cette spécificité n'a pas été exploitée par notre modélisation qui est restée au stade développé dans le Chapitre 4 : une modélisation séparée composante par composante. A l'image du modèle d'enveloppe pour les amplitudes, on pourrait chercher une représentation temporelle conjointe optimale des différentes trajectoires de phases des différentes composantes. Dans la suite de ce chapitre, au niveau des perspectives en codage à bas débit données à la Section 6.4, on donne la version « minimaliste » de la représentation conjointe : on peut se contenter dans ce cas de modéliser à long terme la trajectoire de la fréquence fondamentale. Mais si le débit n'est pas une contrainte, la recherche de la qualité maximale des signaux modélisés passe peut-être par la capture fine des déphasages possibles entre les différentes composantes, dues par exemple au couplage source-conduit vocal et à diverses sources de non-linéarités. On pourrait chercher une modélisation à long terme conjointe des phases des différentes composantes vérifiant cet objectif. Ce point reste à explorer.

6.1.4. Bilan de l'application de la modélisation à long terme sur les trajectoires d'enveloppe (modélisation 2D ; Chapitre 5)

Cette nouvelle approche est certainement une avancée notable par rapport à la modélisation des amplitudes spectrale. Parallèlement à la nouveauté du modèle 2D lui-même, du point de vue algorithmique :

- Un nouvel algorithme de mise en forme de ce modèle 2D a été proposé, plus général que les versions « mono-composantes » données aux Chapitres 3 et 4. Il s'est révélé lui-aussi performant pour un coût de calcul qui reste raisonnable (surtout dans sa version simplifiée présentée dans un deuxième temps) ;

En conséquence, un pas particulièrement important a été franchi au niveau de la compression de données :

- Les débits obtenus montrent que cette représentation est particulièrement efficace en terme de parcimonie. Par définition même, cette modélisation encode de façon très flexible (voir la discussion sur le contrôle des paramètres M et P dans le Chapitre 5) l'ensemble de l'information représentant la totalité des harmoniques pour un coût comparativement bien plus faible que l'approche composante par composante (rappelons au passage que cette nouvelle approche lève le problème de la taille variable des vecteurs d'amplitude et lève ainsi la limitation aux dix premières harmoniques considérée au Chapitre 3).
- Le passage à une analyse des paramètres sur une fenêtre de taille et décalage fixe dimensionnée aux valeurs typiques des codeurs à court terme a permis d'obtenir des débits très encourageants dans l'optique d'un codage à très bas débit de la parole avec une bonne qualité de signal : par exemple, on obtient des débits moyens inter-genre de l'ordre de 300 coefficients/s pour un ratio cible de 75% (rappel : c'est le pourcentage d'amplitudes correctement modélisées au sens du critère perceptif à long terme). Il est raisonnable d'estimer pouvoir réaliser une quantification correcte des paramètres 2D pour un coût moyen de l'ordre de 3 à 5 coefficients par paramètres (c'est un ordre de grandeur courant en codage par transformée à bas débit basée sur le même type de coefficients DCT). Ainsi, le débit binaire serait de l'ordre de 900 à 1500 bits/s pour le codage de tous les paramètres d'amplitude. De plus, on peut envisager sérieusement la réalisation d'un codeur sinusoïdal de type « ultra-bas débit », c'est-à-dire avec un *débit total* de moins de 1kbits/s. Pour cela, on peut diminuer R_{min2} et la quantification des paramètres, tout en conservant une qualité de parole raisonnable (et potentiellement meilleure que les codeurs ultra-bas débits n'exploitant pas la dimension temporelle à long terme). Ces points sont repris dans les perspectives présentées plus loin.

Parmi les améliorations possibles, on a déjà mentionné dans la conclusion du Chapitre 5 la possibilité d'améliorer la procédure d'estimation des ordres spectral M et temporel P du modèle 2D par une estimation conjointe plutôt que séparée (telle qu'elle est faite en deux temps dans la version actuelle de l'algorithme). Ceci serait possible au prix d'une complexité plus grande.

6.1.5. Publications réalisées sur ce travail

Nous dressons ici rapidement un petit bilan en terme de publications. Le travail réalisé dans cette thèse a donné lieu à un article de revue internationale (*IEEE Transactions on Audio speech and Language Processing*) et plusieurs articles de conférences de premier plan. Dans l'ordre chronologique :

- L'article (Girin, Firouzmand & Marchand, 2004) pose les bases de l'approche à long terme proposée dans le cas de la modélisation des trajectoires de phase. Comme il s'agissait de nos premiers travaux, l'étude présentée dans cet article est l'étude préliminaire réalisée en considérant le critère de RSB pour l'ajustement du modèle de phase, étude que nous avons présentée en détails à la Section 4.2 de ce document.
- Les bases de la modélisation à long terme des amplitudes, incluant le critère perceptif basé sur le seuil de masquage fréquentiel à long terme, ont été présentées dans (Firouzmand & Girin, 2005).
- L'étude (Firouzmand, Girin & Marchand, 2005) focalise sur la comparaison des différents modèles que nous avons utilisés dans ce travail pour les tâches de modélisation à long terme à la fois des amplitudes (voir la Section 3.3.5) et des phases (voir la Section 4.3.4).
- L'article de revue (Girin, Firouzmand & Marchand, 2007) est un article de synthèse, qui reprend l'essentiel des résultats des Chapitres 3 et 4 les plus aboutis, concernant à la fois la modélisation à long terme des amplitudes à base de seuil de masquage fréquentiel, et la modélisation à long terme des phases à base de seuil de modulation de fréquence. Une partie de la discussion que nous avons menée à la Section 4.1 est notamment reprise (de façon beaucoup plus compacte !) dans cet article.
- Enfin, l'approche 2D développée au Chapitre 5 a été soumise à publication : c'est l'article (Firouzmand & Girin, 2007).

6.2. Deux exemples d'application directe de la modélisation à long terme

Dans cette section, nous présentons deux applications immédiates pour la modélisation à long terme telle que nous l'avons considérée dans ce document. Il s'agit de la transformation des signaux et du watermarking. Comme on l'a déjà mentionné, l'application au codage est aussi extrêmement intéressante. Nous la réservons pour la section suivante où on verra qu'elle peut se marier avec toute une série d'extensions plus générales à nos travaux, ce qui constitue un vivier de travaux possibles dans ce domaine pour le futur.

6.2.1. Application à la transformation des signaux

Comme on l'a déjà mentionné au Chapitre 1, le modèle sinusoïdal et ses dérivés sont des modèles bien adaptés pour transformer le signal de multiples façons. Ceci a contribué largement à leur popularité dans les communautés de chercheurs en traitement

des signaux de musique et de parole, à partir, disons, de la fin des années 1980, lorsque la puissance des machines a commencé à permettre d'effectuer ces transformations avec des temps de calculs raisonnables. Dans le cadre de la modélisation à long terme, ce besoin de pouvoir transformer les signaux voit naître de nouvelles possibilités.

Ainsi, on peut appliquer ces modèles à long terme à la problématique du changement d'échelle temporelle, que ce soit pour un étirement ou une compression de la durée du signal. Avant de préciser cette idée, donnons rapidement quelques éléments sur ce type de transformation. Le changement d'échelle temporelle est généralement caractérisé par un facteur d'étirement/de compression qui est supérieur à 1 (pour l'étirement) ou compris entre 0 et 1 (pour la compression) (voir la Figure 6.1) : ce facteur donne la durée relative du signal transformé par rapport au signal original. Notons que ce facteur peut varier au cours du temps. On a alors un étirement ou une compression temporel(le) non linéaire. Par exemple, en utilisant un facteur d'étirement passant progressivement d'une valeur inférieure à 1 à une autre valeur supérieure à 1, on peut ralentir la « vitesse d'articulation » (*i.e.* de production) de la parole (on peut bien sûr aussi l'accélérer en permutant les valeurs du facteur d'étirement). Cette adaptation peut être réalisée en fonction de descripteurs extraits du son à traiter. Par exemple, on peut utiliser un indice de voisement pour créer un ralenti sélectif, qui ne ralentit que les voyelles, et laisse les consonnes telles quelles, pour conserver plus de naturel au signal transformé.

Dans le cadre de la modélisation à long terme, le changement d'échelle temporelle est particulièrement aisé à effectuer : en effet, il revient tout simplement à ré-échantillonner les trajectoires de paramètres générées par les modèles à long terme. De plus, *ce ré-échantillonnage peut être généré directement à partir des coefficients des modèles à long terme* : il suffit d'appliquer les équations de génération des trajectoires de paramètres à partir des coefficients des modèles à long terme, par exemple l'équation (2.1), avec un nouvel échantillonnage cible des indices temporels. On évite ainsi les calculs assez lourds des procédures de ré-échantillonnage traditionnelles, par exemple celles basées sur une interpolation-décimation du signal [Crochiere & Rabiner, 1975]. Notons que ce ré-échantillonnage du modèle à long terme peut être régulier ou irrégulier, c'est-à-dire soit régulier par portion, soit adaptatif (ajustement progressif des valeurs au cours des échantillons de paramètres), pour correspondre à une évolution du facteur de compression.

Pour les paramètres d'amplitudes, ce ré-échantillonnage ne pose pas de problème particulier : on a vu dans le Chapitre 1 notamment que ce type de paramètres est particulièrement robuste à des interpolations temporelles très simples (par exemple linéaire), et cette robustesse se généralise à ce problème de ré-échantillonnage : les jeux de paramètres d'amplitude ré-échantillonnés à partir des modèles à long terme sont de bons candidats pour la resynthèse. Pour les phases, le problème est « comme d'habitude » plus difficile (cf. Sections 1.3.2 et 4.1.3 par exemple) : dans le cadre à long terme comme dans le cadre à court terme, un ré-échantillonnage des trajectoires de phase permettant de conserver la forme de la forme d'onde du signal est un problème délicat (on a déjà mentionné cette contrainte dite de « *shape-invariance* », voir par exemple [Quatieri & McAulay, 1992]). Par contre, dans une version simplifiée de ce problème, on peut tout aussi facilement que les amplitudes se contenter de ré-échantillonner les trajectoires à long terme des *fréquences*. Comme pour l'interpolation simple sans changement d'échelle, ce ré-échantillonnage des trajectoires de fréquence

ne garantit pas la préservation de la forme d'onde exacte du signal. Par contre, combiné avec le ré-échantillonnage des amplitudes, ce ré-échantillonnage aboutit à des résultats très satisfaisants du point de vue de la qualité des signaux générés, en regard de la simplicité du traitement. On peut voir sur la Figure 6.1 un exemple simple d'un tel traitement (avec un facteur d'étirement ou de compression constant) permettant de changer le débit de parole.

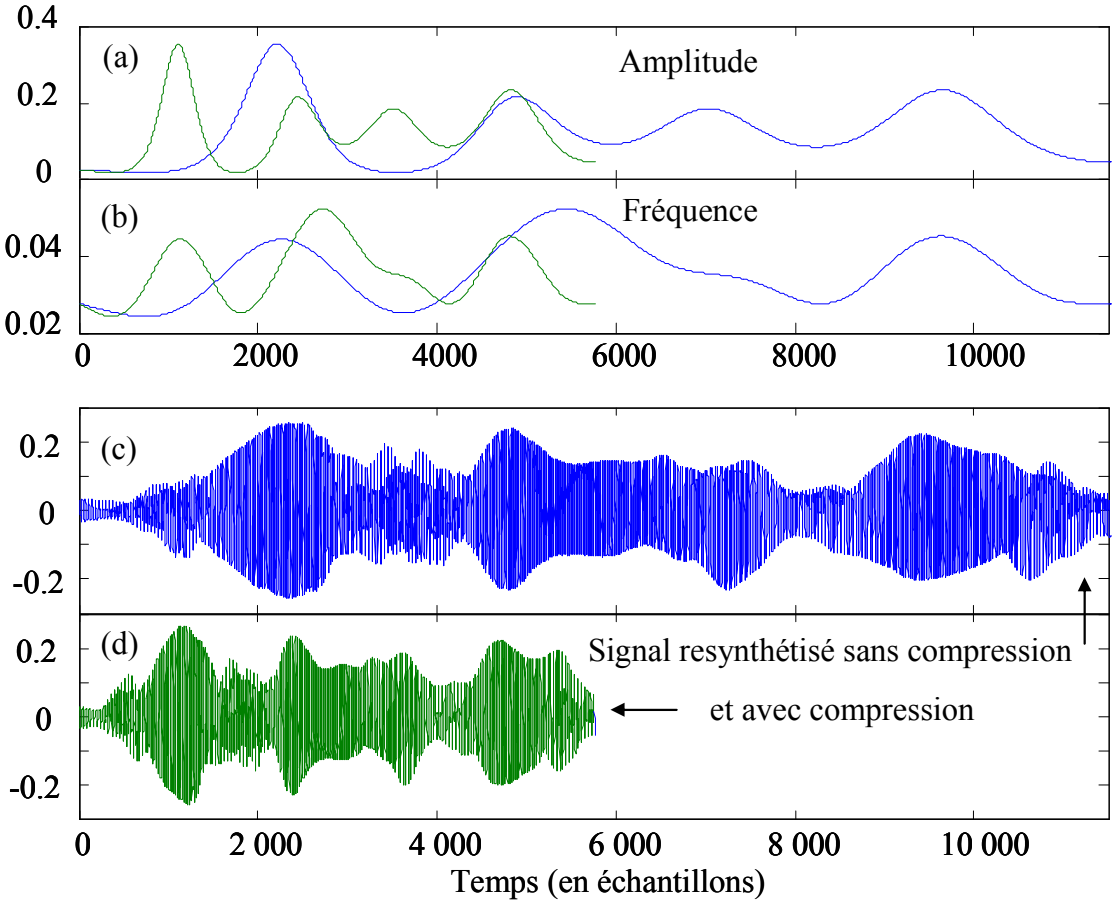

Figure 6.1 : Exemple d'une compression temporelle de facteur 0,5 pour les trajectoires d'amplitude (a) et de fréquence (b) de la première harmonique d'une section de parole voisée de voix de femme (durée 1,4 s, F_e = 8 kHz). Les trajectoires des paramètres avant compression sont en bleu et les trajectoires compressées sont en vert. En bas, on a le signal resynthétisé sans compression (c) et avec compression (d). Un étirement temporel d'un facteur 2 fournit des résultats visuellement très similaires (identique pour les paramètres), avec un rôle « inversé » entre les courbes avant et après changement d'échelle et un changement d'échelle correspondant pour l'axe des X.

Pour finir cette illustration du potentiel de l'approche à long terme dans le cadre de la transformation des signaux, on peut mentionner assez rapidement une autre manipulation classique : le changement de pitch. Un changement d'échelle fréquentielle peut être effectué très facilement par une manipulation très simple sur les trajectoires à long terme des fréquences des différentes harmoniques du signal : il suffit d'effectuer une multiplication par un facteur d'échelle avant la resynthèse. Notons que comme dans le cadre à court-terme, pour respecter le caractère naturel du signal, cette manipulation doit s'affranchir de l'influence du conduit vocal par une décomposition de type source-

filtre (par exemple par un modèle LPC). Le modèle sinusoïdal à long terme peut alors être appliqué sur le signal source, qui est ensuite transformé en fréquence, puis utilisé pour exciter le filtre du conduit vocal pour resynthétiser le signal. Dans ce cas, on peut bénéficier de la combinaison de l'approche à long terme dans le cadre sinusoïdal avec celle généralisée à des modèles de type source-filtre comme la LPC, telle que présentée à la Section 6.3 qui clôt ce document.

6.2.2. Application au tatouage des signaux

Comme déjà mentionné dans l'introduction de ce document, le tatouage (ou *watermarking* en anglais) des signaux consiste à insérer des données supplémentaires dans un signal de façon imperceptible (*i.e*. inaudible pour les signaux de parole et pour les signaux audio en général) et si possible robuste aux attaques éventuelles et aux traitements effectués sur le signal porteur. Un premier prototype de tatouage original basé sur le modèle sinusoïdal harmonique a été développé récemment à l'ICP [Girin & Marchand, 2004]. Cette technique repose sur une modulation particulière des trajectoires de fréquence du signal. Le principe général est le suivant : on commence par analyser les paramètres du modèle sur une portion de signal à tatouer. Dans l'étude citée, on considère seulement le cas de sections voisées de parole car on verra qu'on a besoin de l'hypothèse d'harmonicité[97]. Puis on réalise le suivi de partiel (qui se réduit au suivi des harmoniques dans le cas voisé ; voir Section 1.3.2.1). Le processus de tatouage s'applique indépendamment sur chaque harmonique et pour simplifier la présentation, considérons qu'on veuille tatouer une harmonique donnée de rang p. La base de la technique consiste à moduler la trajectoire fréquentielle de cette harmonique avec une modulation particulière : avant de resynthétiser le signal, on somme à cette trajectoire de fréquence un signal de tatouage (le *watermark*) composé d'une suite de motifs élémentaires eux-mêmes modulés par le message binaire à transmettre. Dans l'étude en question, le motif est une fenêtre en cosinus surélevé et ce motif est de signe positif pour un 1 et de signe négatif pour un 0 (voir la Figure 6.2).

Une fois qu'on a défini ce principe de modulation de trajectoire fréquentielle de l'harmonique tatouée, la difficulté se trouve au niveau du décodeur du système de tatouage : il s'agit d'extraire correctement la suite de motifs greffée, à partir du signal resynthétisé, pour retrouver le message binaire. En effet, l'excursion en fréquence de la modulation doit rester relativement faible pour être inaudible. Or, les performances du détecteur de *watermark*, basé sur un estimateur de fréquence instantanée, dépendent fortement de cette valeur d'excursion : plus elle est élevée, meilleure est la détection[98] (voir la Figure 6.2 pour un exemple de bonne estimation de la fréquence tatouée). Ces performances dépendent aussi fortement de la régularité de la trajectoire de fond de la fréquence, sur laquelle est greffée la modulation : des irrégularités de cette trajectoire peuvent parasiter significativement la détection des motifs de *watermark*. Pour obtenir des scores de détection performants en dépit de la contrainte sur l'excursion de

[97] Cette approche peut ainsi s'appliquer sur des sections pseudo-périodiques de signaux de musique.

[98] Le terme « faible » dans la phrase précédente est ambigu. En effet, en réalité, l'excursion en fréquence de la modulation de tatouage peut être relativement élevée par rapport aux seuils de perception usuellement adoptés dans le cas stationnaire, on a largement discuté ce point au Chapitre 4. Mais on utilise le terme « faible » du fait qu'on se place ici du point de vue du détecteur de *watermark* pour lequel cette modulation est toujours trop faible ! Toute la difficulté et la finesse de cette technique de tatouage consiste à trouver le bon compromis entre ces deux contraintes.

modulation, une solution consiste à s'assurer que les trajectoires porteuses de la modulation sont suffisamment lisses. Dans [Girin & Marchand, 2004], pour assurer ce caractère lisse, les auteurs remplacent la trajectoire de fréquence à tatouer par une interpolation linéaire des valeurs de la fréquence fondamentale (elle-même jugée suffisamment lisse) multipliée par le rang *p* de l'harmonique considérée. Ce point peut être amélioré, et en particulier, il doit pouvoir largement bénéficier des résultats apportés par l'approche à long terme présentée dans cette thèse : on a déjà dit plusieurs fois que les modèles à long terme proposés se caractérisent justement par un aspect lisse et régulier (du fait qu'on utilise des modèles à base de fonction infiniment dérivables). Ces modèles doivent donc pouvoir servir de support particulièrement bien adapté à la modulation de tatouage, et permettre d'augmenter significativement les performances du détecteur de *watermark*. D'autant plus que ce détecteur pourrait tenir compte de contrainte *a priori* résultant de l'utilisation des modèles à long terme. Ce point est actuellement en cours d'étude à l'ICP.

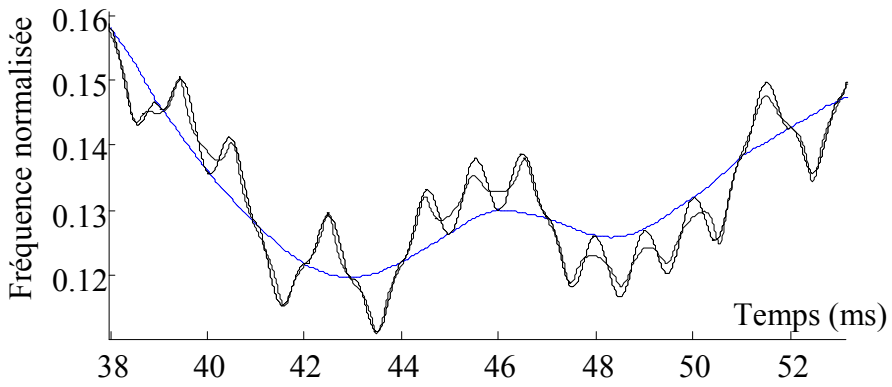

Figure 6.2 : Exemple de tatouage par modulation de trajectoire de fréquence (6ème harmonique d'une section voisée de voix de femme) ; en tiret-point : trajectoire de fréquence avant tatouage : il s'agit de la fréquence fondamentale multipliée par le rang de l'harmonique ; en pointillés : trajectoire de fréquence tatouée à l'encodeur (les motifs au-dessus de la trajectoire de fond encodent une valeur binaire 1, ceux au-dessous encodent une valeur binaire 0 ; en continu : trajectoire tatouée estimée au décodeur. La suite de motifs est estimée par soustraction entre cette trajectoire et la trajectoire de fond, et les valeurs binaires sont déduites par sommation et seuillage du résultat (d'après [Girin & Marchand, 2004]).

6.3. Extension de la modélisation à long terme à d'autres modèles spectraux et application au codage

Dans cette thèse, nous avons étudié spécifiquement et en détail la modélisation à long terme des paramètres du modèle sinusoïdal. Comme nous l'avons déjà mentionné, la modélisation à long terme peut s'appliquer sur diverses autres représentations du signal. Dans cette section, nous illustrons ce principe sur deux exemples majeurs : la modélisation sinusoïdal/harmonique + bruit d'une part, et le modèle de prédiction linéaire (LPC) d'autre part. Comme ces modèles sont particulièrement adaptés et exploités dans le cadre du codage à bas débit de la parole, nous finirons ce document en revenant conjointement sur les applications de l'approche à long terme au codage.

6.3.1. Extension vers un modèle sinusoïdal + bruit ou harmonique + bruit à long-terme

Le modèle sinusoïdal + bruit et sa version simplifiée harmonique + bruit ont été présentés dans le Chapitre 1 à la Section 1.4.2. On rappelle que dans ces modèles, le signal est décomposé en une partie sinusoïdale composée de sinusoïdes « régulières » (en relation harmonique les unes par rapport aux autres dans le cas du modèle harmonique + bruit) et une partie bruitée, représentant les composantes aléatoires du signal. Rappelons que, tout comme les composantes sinusoïdales/harmoniques, les composantes bruitées évoluent aussi continûment dans le temps. On a aussi vu au Chapitre 1 que dans le domaine de l'analyse-synthèse des signaux audio, ces deux types de composantes peuvent être analysées et séparées [Yegnanarayana *et al*., 1998] [Jackson & Shadle, 2000] [Girin, 2006], afin d'être modélisées et/ou modifiées séparément. Une attention spécifique peut ainsi être portée aux composantes bruitées, en particulier pour la synthèse [Richard & d'Alessandro, 1996] [Stylianou, 1996], pour le codage [Macon & Clements, 1997] et pour l'étude fondamentale de la production de parole [Jackson & Shadle, 2001]. Dans cette section, nous allons voir comment ces principes peuvent être étendus au cas de l'approche à long terme qui est celui de cette thèse. Pour simplifier la présentation de cette section, nous nous placerons dans le cadre du modèle harmonique + bruit mais la plupart des notions discutées peuvent être étendues au cas plus général du modèle sinusoïdal + bruit.

On s'intéresse donc ici à la modélisation à long terme d'une section de signal continûment voisée (comme dans les chapitres précédents) mais avec une composante bruitée significative sur l'ensemble de cette section. On suppose que les composantes harmoniques et les composantes bruitées ont été analysées et séparées sur l'ensemble de la section de signal. Cette séparation peut être faite par une des méthodes citées dans le paragraphe précédent[99]. Ainsi, il est tout d'abord évident qu'à l'issue de cette décomposition harmonique + bruit, les trajectoires des paramètres de la partie harmonique peuvent être modélisées à long terme par la nouvelle approche que nous avons étudiée telle qu'elle a été présentée au Chapitre 3 et au Chapitre 4 (voir même au Chapitre 5 si la bande harmonique est suffisamment compacte, voir plus loin). Nous nous intéressons donc ici particulièrement à la modélisation à long terme conjointe des composantes de bruit (rappelons qu'à la fin du processus de modélisation séparée, le signal synthétisé est une sommation du signal de synthèse sinusoïdal et du signal de synthèse bruité). Nous proposons deux grandes voies pour la modélisation à long terme de ces composantes bruitées selon que le bruit est considéré large bande ou à bande limitée.

Dans le premier cas, nous proposons de réutiliser directement l'approche 2D développée au Chapitre 5 pour modéliser à long terme l'enveloppe pleine bande du bruit. Celle-ci est généralement décrite à court terme par un jeu de paramètres d'amplitudes. Par exemple, on rappelle que dans [Serra & Smith, 1990], ces amplitudes sont les valeurs résiduelles des pics du spectre après soustraction des composantes harmoniques

[99] On rappelle cependant que sans hypothèses particulières sur le signal, ce problème de séparation n'est pas un problème facile car les composantes sinusoïdales et bruitées peuvent être très mélangées sur toute la bande du signal. Si la séparation s'avère relativement infructueuse, la modélisation à long terme n'apporte rien pour la manipulation séparée des composantes, mais elle peut être quand même utilisée pour encoder l'enveloppe spectrale globale du signal.

estimées ; ces valeurs sont ensuite interpolées linéairement pour avoir l'enveloppe et « boucher les trous » laissés par la soustraction des harmoniques. De la même façon que nous l'avons fait pour les jeux d'amplitudes harmoniques au Chapitre 5, ce modèle d'enveloppe spectrale du bruit pleine bande, qui est caractérisé à court terme par un nombre très variable de paramètres (encore plus que dans le cas harmonique) peut être d'abord remplacé par un modèle d'enveloppe de dimension fixe (et petite par rapport à la dimension de départ si on veut faire un codage efficace du signal). Puis, pour modéliser à long terme les trajectoires des coefficients d'enveloppe résultant, on peut appliquer un deuxième modèle sur chacune de ces trajectoires selon l'axe du temps exactement comme on l'a fait au Chapitre 5. A nouveau, dans les deux phases de cette modélisation 2D (fréquentielle et temporelle), l'ajustement des modèles peut reposer sur une procédure de type WMMSE (voir Section 2.5.4). Au niveau de la synthèse de cette partie bruitée du signal, il faut bien entendu d'abord décoder la suite d'enveloppes spectrales à court terme à partir des coefficients modélisés à long terme, comme on l'a fait au Chapitre 5. Puis on peut appliquer deux grandes possibilités utilisées dans l'approche classique à court terme. La première consiste à générer un filtre de synthèse à partir de chaque enveloppe[100] puis à filtrer un bruit blanc à travers ce filtre en assurant l'évolution du processus au cours du temps. La deuxième possibilité consiste à réaliser un échantillonnage régulier et bien dimensionné sur le modèle d'enveloppe spectrale déduite du modèle « 2D », pour en tirer un jeu de valeurs d'amplitude et ensuite faire la synthèse du signal trame à trame avec les équations du modèle sinusoïdal. Dans ces équations, on utilise alors une phase relative aléatoire : nous avons déjà mentionné que le caractère aléatoire du paramètre de phase, par exemple avec une répartition uniforme entre $-\pi$ et π, permettait d'assurer la génération d'un bruit (voir la Section 1.4 et [McAulay & Quatieri, 1995] [Macon & Clements, 1997]).

La deuxième façon de considérer les composantes de bruit, valable en particulier pour la parole et certains instruments de musique, est l'approche à bande limitée haute fréquence qu'on a décrite à la Section 1.4.2.3, issue principalement en parole des travaux de [Stylianou, 1996]. Dans ce cas, pour chaque trame à court terme, la bande harmonique (basse fréquence) et la bande bruitée (haute fréquence) sont séparées par la fréquence de voisement (voir la Section 1.4.2.3). On peut tout à fait envisager une version à long terme de cette approche. En effet, comme la fréquence de voisement est susceptible d'évoluer au cours du temps, on propose alors de modéliser la trajectoire de cette fréquence de voisement par un modèle à long terme. Puis, on pourrait tout simplement appliquer séparément sur la bande harmonique et sur la bande bruitée la modélisation à long terme telle qu'elle a été proposée précédemment dans ce document. En particulier, pour la partie bruitée, on pourrait adapter les principes donnés ci-dessus dans le cas pleine bande au cas à bande limitée (en adaptant bien sûr de façon parcimonieuse la quantité de coefficients mis en jeu pour une bonne efficacité de la

[100] Pour cela, on peut par exemple calculer des coefficients d'autocorrélation par transformée de Fourier inverse du spectre d'amplitude converti en puissance, puis déduire de ces coefficients d'autocorrélation des coefficients de filtre d'un modèle auto-régressif (identifié à un modèle LPC) par l'algorithme de Levinson, classique en modélisation LPC [Levinson, 1947] [Markel & Gray, 1976]. Notons par ailleurs, qu'au cas où cette approche source-filtre est utilisée pour les composantes de bruit, on peut alternativement utiliser directement les développements à long terme réalisés sur les paramètres LPC et présentés à la Section 6.3.2. En d'autres termes, on peut encoder à long terme une enveloppe spectrale LPC plutôt qu'une enveloppe MCD ou autre, et profiter des avantages de ce modèle LPC en terme de synthèse (filtrage d'un bruit blanc à travers le filtre LPC de synthèse).

représentation), et ceci aussi bien à l'analyse, à la modélisation à long terme, et à la synthèse. On peut remarquer ici que comme la fréquence de voisement évolue au cours du temps, le nombre d'harmoniques n'est pas nécessairement constant et il faut en tenir compte dans la modélisation à long terme de la partie harmonique, comme on l'a fait dans le cas pleine bande[101].

Cette remarque nous amène à finalement proposer une version simplifiée de cette décomposition en deux bandes en version à long terme : on peut imaginer que la trajectoire de la fréquence de voisement peut être identifiée à une certaine harmonique. On prendrait l'harmonique ayant la trajectoire de fréquence la plus proche de celle de la fréquence de voisement sur l'ensemble de la trajectoire (au sens d'un certain critère très simple, par exemple on peut prendre à nouveau un critère du type moindre carrés moyens). Cette simplification permet bien de s'affranchir du problème de gestion du nombre d'harmoniques modélisées à long terme si on utilise l'approche 1D des Chapitres 3 et 4. Elle a aussi un autre avantage au niveau du codage du signal : comme l'information de trajectoire de la fréquence fondamentale est forcément encodée[102], on restreint le codage de la fréquence de voisement à la seule transmission de la valeur entière représentant le rang de l'harmonique jouant le rôle de fréquence de voisement. Ceci étant valable pour toute la section de signal modélisée avec cette approche, ce coût de transmission est vraisemblablement négligeable.

6.3.2. Extension vers un modèle LPC à long-terme

Comme on l'a déjà mentionné auparavant plusieurs fois dans ce document, le principe de la modélisation à long terme peut être adapté aux paramètres du modèle LPC. En réalité, c'est l'idée de base des travaux de Dusan et collègues [Dusan *et al*., 2004] que nous avons déjà mentionnés pour conclure le Chapitre 2. Rappelons que ces travaux ont été proposés simultanément à nos travaux préliminaires sur la modélisation à long terme de la phase ([Girin *et al*., 2004] et Section 4.2 de ce document), et qu'ils consistaient à modéliser la trajectoire temporelle de K vecteurs de paramètres LSF par des polynômes d'ordre P. Dans cette étude[103], les auteurs donnaient des résultats pour $K = 10$ et $P = 4$.

Dans la conclusion du Chapitre 2, nous avons présenté les points communs et les points « parallèles » entre notre approche et celle de Dusan et collègues. Nous n'allons pas redétailler ces points ici mais disons simplement que si on remplace les paramètres du modèle sinusoïdal par des paramètres LSF (et en se munissant également d'une mesure de distance adaptée à ces paramètres), notre approche à long terme peut être vue comme

[101] En toute honnêteté, on ne l'a pas vraiment fait dans l'approche 1D des Chapitres 3 et 4, puisqu'on a limité le nombre des harmoniques modélisés pour ne pas rencontrer ce problème. Par contre, dans l'approche 2D du Chapitre 5, on a vu que ce problème est intrinsèquement résolu en considérant implicitement toutes les harmoniques dans la bande modélisée.

[102] Toujours dans une optique de codage à très/ultra bas débit, on peut estimer et encoder la fréquence de voisement tout en s'affranchissant de la séparation effective des modèles spectraux dans la bande harmonique et dans la bande bruitée : on peut en effet ne considérer qu'un seul modèle d'enveloppe globale sur toute la bande du signal. Ce modèle sera peut-être moins précis, mais vraisemblablement plus facile à coder, à la fois à court terme et à long terme. Pour la resynthèse du signal, on utilise alors les valeurs de ce modèle global échantillonnées aux fréquences harmoniques dans la bande harmonique, et échantillonnées arbitrairement dans la bande bruitée.

[103] Notons qu'un article de Dusan et collègues présentant plus en détails ces travaux vient tout juste d'être publié dans une revue [Dusan *et al*., 2007].

une généralisation des travaux de Dusan et collègues[104]. Cette généralisation porte sur deux points essentiels : d'une part la possibilité de coder des trajectoires de LSF de taille temporelle[105] variable par un modèle lui-même de taille variable grâce à l'adaptation de l'algorithme d'ajustement du modèle à long terme développé dans le cadre sinusoïdal, et d'autre part la possibilité d'utiliser d'autres modèles à long terme que le modèle polynomial (nous avons ainsi tendance à privilégier notre favori, le modèle en cosinus discrets). Dans les travaux que nous avons présentés dans cette thèse, la taille variable dépend d'une segmentation préalable du signal en sections voisées et non voisées. Compte tenu que, même dans une approche à court terme, c'est-à-dire sur des trames de 20 ms environ, le comportement des LSF est assez différent sur les trames voisées de signal et sur les trames non voisées [Hagen *et al.*, 1999], nous pouvons conserver ce type de segmentation pour l'étude à long terme de la modélisation et du codage des trajectoires des paramètres LSF.

De fait, tous ces points ont fait l'objet d'une étude réalisée récemment à l'ICP en parallèle des derniers développements de cette thèse [Girin, 2007][106]. Dans cette étude, le traitement va jusqu'au codage (c'est-à-dire la quantification) à long terme des paramètres LSF (et pas seulement la modélisation de leur trajectoire). Nous ne détaillerons pas les points techniques de cette étude qui peuvent être trouvés dans la référence ci-dessus. Mentionnons juste que cette quantification vient après l'étape de modélisation à long terme des LSF par un MCD, et que pour la réaliser, l'auteur inspire d'une astuce utilisée dans l'étude de [Dusan *et al*., 2004]. Cette astuce consiste à exploiter une transformation linéaire bi-univoque entre les P+1 coefficients du modèle à long terme et un jeu réduit de P+1 vecteurs de LSF[107]. C'est ce jeu réduit de LSF qui est quantifié (par une quantification vectorielle [Gray and Gersho, 1992]) et transmis au décodeur. La transformation bi-univoque permet de retrouver le modèle à long terme au décodeur, ce qui permet de déduire ensuite la valeur quantifiée des K vecteurs LSF originaux. Cette procédure est inscrite au cœur d'un algorithme d'ajustement procédant selon le même principe d'analyse par la synthèse que celui développé dans cette thèse dans le cadre sinusoïdal. Le critère d'ajustement à long terme des LSF quantifiés (c'est-à-dire ceux issus de la double tâche de modélisation et de quantification) est bien sûr adapté à ce nouveau cadre d'étude : il s'agit ici de la moyenne temporelle *ASD* sur les K

[104] Notons que dans un cadre de codage LPC, l'approche à long terme peut aussi être appliqué sur d'autres types de paramètres des codeurs LPC tels que ceux caractérisant l'excitation. Nous reviendrons sur ce point par la suite.

[105] Rappelons que la taille fréquentielle des vecteurs de paramètres LSF est fixe (elle vaut typiquement 10 dans le cadre du codage à bas débit de la parole en bande téléphonique). Ceci permet de ne pas avoir de problème de régularisation de taille des données spectrales avant la modélisation à long terme, problème qui nous avait conduit au Chapitre 5 à modéliser l'enveloppe spectrale du signal avant d'effectuer la modélisation à long terme des paramètres résultant. Dans le cadre du modèle LPC, l'enveloppe spectrale LPC, codée par un vecteur de LSF de taille fixe, remplace directement l'enveloppe spectrale MCD du Chapitre 5. Ainsi, la modélisation LPC à long terme telle qu'on l'introduit dans cette section peut être vue comme une variante de notre approche 2D du Chapitre 5.

[106] L'auteur de cette présente thèse n'a pas participé formellement à ces travaux, mais seulement par le biais de discussions informelles. C'est pourquoi il n'apparaît pas en tant que co-auteur de cette étude, bien qu'il soit partie prenante de ces développements.

[107] Rappelons qu'on a K vecteurs LSF sur chaque section et que la modélisation à long terme assure P+1<K. C'est précisément cette dernière inégalité qui assure le gain de codage de la méthode proposée par rapport à une approche classique à court terme.

trames de chaque section traitée de la mesure de distorsion spectrale[108] [Atal & Paliwal, 1993] [Gray & Markel, 1976] [Gray *et al*., 1980] [Nocerino *et al*., 1985], soit :

$$ASD = \sqrt{\frac{1}{K}\sum_{k=1}^{K} SD_k}$$

avec

$$SD_k = \frac{1}{2\pi}\int_0^{2\pi}\left[10\log_{10} P_k\left(e^{j\omega}\right) - 10\log_{10}\hat{P}_k\left(e^{j\omega}\right)\right]^2 d\omega$$

où les arguments de l'intégrale sont les spectres de puissance LPC associés aux vecteurs de LSF originaux et quantifiés. Ainsi, ce critère permet de contrôler les performances de la quantification à long terme en terme de distorsion spectrale moyenne.

Dans [Girin, 2007], on exécute l'algorithme sur une grande base de données en faisant varier les valeurs de distorsion moyenne cibles (qui remplace ainsi le critère du ratio R_{min} utilisé dans le cadre sinusoïdal) et on calcule les débits moyens (ici en bits/s) obtenus dans chaque cas. Les résultats, tels que ceux donnés aux Figures 6.3 et 6.4, montrent qu'un gain de débit important peut être réalisé par rapport à la quantification vectorielle classique, réalisée sur une base à court terme, c'est-à-dire trame par trame. On obtient par exemple un gain relatif pouvant dépasser les 50% (selon les régions du plan débit/distorsion) pour les sections voisées de parole, et un gain relatif maximal de l'ordre de 30% pour les sections non-voisées. Ces résultats montrent à nouveau d'une façon générale le potentiel de compression impressionnant qui caractérise la modélisation à long terme, et ils montrent d'une façon particulière qu'après le cadre sinusoïdal, le cadre du modèle LPC peut bénéficier très largement de cette approche.

Nous avons insisté ici particulièrement sur le codage à long terme des paramètres LSF représentant la contribution du filtre d'analyse-synthèse LPC. Nous pouvons terminer cette section avec quelques considérations sur les autres paramètres représentant le signal résiduel utilisé comme excitation du filtre de synthèse.

Ainsi on peut commencer par prendre l'exemple très simple du vocodeur LPC type FS 10-15 où ce signal d'excitation est modélisé de façon très simple, soit par un peigne de Dirac dans le cas d'un son voisé, soit par un bruit blanc dans le cas des sons non-voisés. Dans tous les cas, on ne transmet qu'une valeur de fréquence fondamentale (avec un indice particulier pour encoder le cas non-voisé) et une valeur de gain pour représenter la variation d'énergie du signal. A l'issue de nos travaux, il est évident que ces deux paramètres peuvent très bien être modélisés et encodés avec une approche à long terme. En effet, le cas de la fréquence fondamentale est un cas particulier des trajectoires de fréquences, déjà maintes fois mentionné dans ce document. Quant au gain, il se caractérise aussi généralement par une évolution relativement lente et régulière au cours du temps, en tout cas sur les portions de signal qui se prêtent bien à la modélisation à long terme (voir le Chapitre 2). Une modélisation à long terme de ce paramètre devrait donc se révéler tout aussi efficace.

[108] Nous avons déjà utilisé une variante de cette distance spectrale à la Section 5.4.2.1 pour pondérer la modélisation à long terme en 2D.

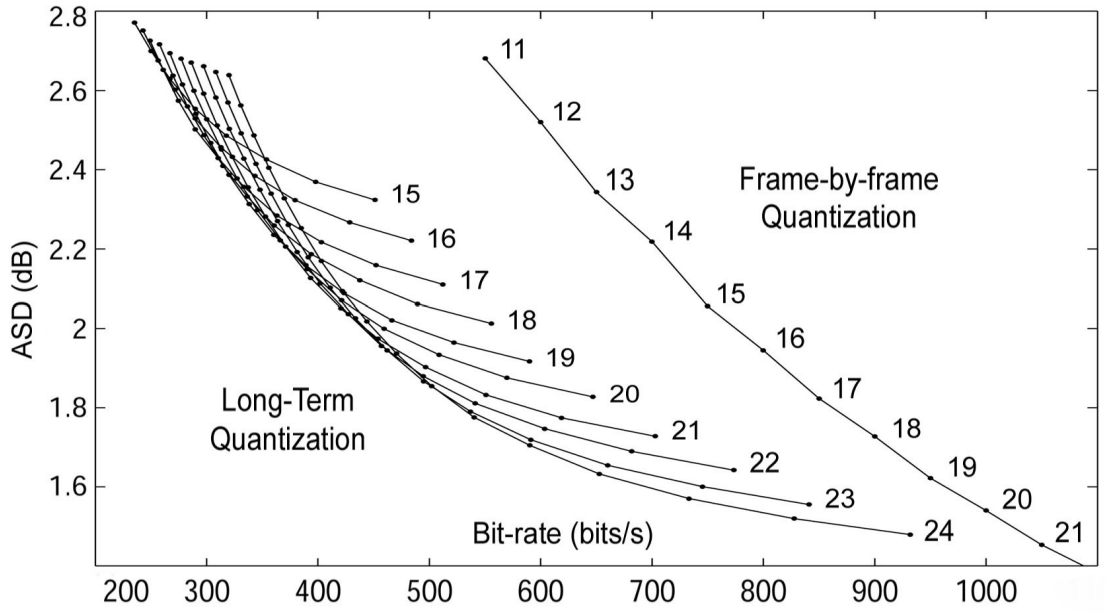

Figure 6.3 : Résultats de la quantification à long terme proposée dans [Girin, 2007] : distorsion spectrale moyenne (ASD) en fonction du débit moyen, calculés sur une large base de données (4656 sections voisées de parole, soit 67080 vecteurs, 88 locuteurs différents). Courbes de gauche : résultats de la quantification à long terme pour différents valeurs cibles de ASD et différentes résolutions du quantificateur vectoriel utilisé dans l'algorithme à long terme (ces résolutions, en bits par vecteur, sont indiquées sur la figure). Courbe de droite : résultat de la quantification vectorielle à court terme (trame par trame) utilisant les mêmes dictionnaires que dans le cas à long terme, pour différentes résolutions (indiquées sur la figure).

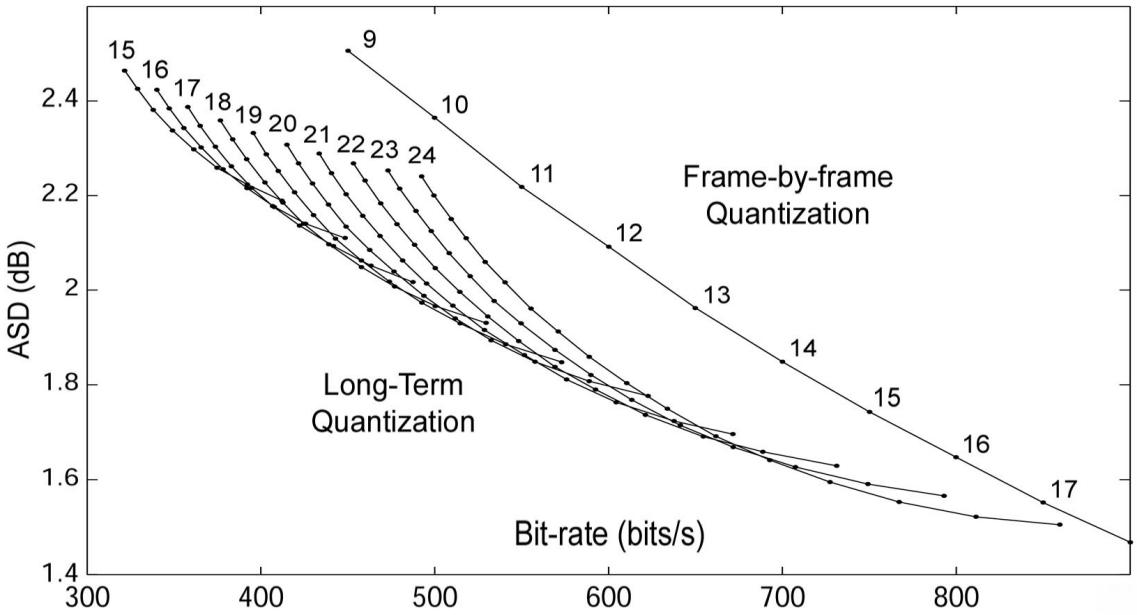

Figure 6.4 : Même chose que Figure 6.3 mais pour les sections non-voisées de la base de données (4427 sections, 22101 vecteurs, 88 locuteurs) (d'après [Girin, 2007]).

Au-delà du vocodeur LPC de base, les codeurs basés sur le modèle LPC se caractérisent par des modèles de signal d'excitation plus ou moins raffinés et complexes à mettre en œuvre. L'objectif est d'enrichir le modèle d'excitation pour améliorer notablement la qualité des signaux codés par rapport à la qualité très basique du vocodeur LPC de base, tout en essayant de garder un débit raisonnable. Sans entrer ici dans les détails, pour généraliser l'application de l'approche à long terme au codage de parole et faire la jonction entre la Section 6.3.1 et cette présente section, il est intéressant de mentionner le fait que l'approche harmonique + bruit peut en effet être appliquée de différentes façons au signal d'excitation des codeurs LPC. Cette approche permet d'une part de représenter la partie bruitée naturelle du signal d'excitation, et d'autre part de rattraper les « imperfections » du spectre LPC pour représenter finement le relief spectral du signal : on retrouve l'information précise manquante dans le codage des amplitudes des harmoniques du signal résiduel [Shlomot *et al*., 2001]. Ainsi, le modèle harmonique + bruit est un bon candidat pour représenter le signal d'excitation (même si bien sûr il n'est pas le seul, voir la note de bas de page 68). Il est donc légitime de penser que l'adaptation des principes de la modélisation à long terme à ce modèle harmonique + bruit, telle que présentée à la Section 6.3.1, pour le codage à long terme de l'excitation est une voie prometteuse. Elle permettrait de compléter le codage à long terme des paramètres LSF dans l'optique de la réalisation complète d'un codeur à long terme de la parole avec un débit performant et une bonne qualité de signal.

6.3.3. Conclusion sur cette section et sur le document

L'application de la modélisation à long terme au codage de la parole (et on l'espère à celui d'autres types de signaux audio) semble particulièrement prometteuse, et c'est la raison pour laquelle nous finirons ce document en complétant et commentant ce point. Un point majeur est qu'à partir des travaux présentés dans cette thèse, des développements peuvent être poursuivis à la fois en codage de type harmonique et en codage de type LPC, avec, comme on vient de le voir ci-dessus, des combinaisons fructueuses possibles entre ces deux grandes approches. En codage sinusoïdal/harmonique, nous devons poursuivre l'effort que nous avons porté en modélisation vers l'étape de quantification des coefficients à long terme, c'est-à-dire la représentation binaire de cette information. En particulier, l'approche 2D étant particulièrement prometteuse pour représenter les amplitudes spectrales, nous devons à présent nous intéresser à la quantification des coefficients DCM du modèle 2D à long terme. Dans ce cadre sinusoïdal, il semble inutile d'essayer d'éviter la quantification directe des coefficients du modèle par une astuce tel que celle mentionnée dans le cadre du modèle LPC (rappel : on applique une transformation bi-univoque entre les coefficients DCM et un jeu réduit de vecteurs LSF pour quantifier ces derniers au lieu des paramètres DCM en utilisant des techniques bien maîtrisées de quantification vectorielle de LFS). En effet, une telle transformation des paramètres DCM temporels ramènerait ici à… des paramètres DCM ! (du premier modèle d'enveloppe). Il faut donc d'abord tenter d'appliquer sur ces paramètres DCM en dimension 2 des méthodes assez traditionnelle de quantification non uniforme (d'abord scalaire puis vectorielle si nécessaire) en espérant profiter de la robustesse générale de ces coefficients à la quantification.

En ce qui concerne l'approche LPC, il faut poursuivre les efforts sur la quantification des paramètres du signal d'excitation, comme on l'a mentionné à la fin de la sous-section précédente. Dans tous les cas, codeur LPC ou harmonique, il faut réaliser la quantification de la trajectoire de fréquence fondamentale. La combinaison de cette information très simple avec l'information d'enveloppe à long terme (enveloppe DCM ou LPC+gain) devrait permettre de réaliser un codeur de parole à ultra bas débit (inférieur à 1kb/s) avec une qualité tout à fait intéressante en regard de ce faible débit (un qualité proche du vocodeur LPC de base, voire supérieure si on arrive à appliquer efficacement la modélisation à long terme dans un cadre harmonique + bruit, voir les discussions déjà données plus haut dans cette Section 6.3).

D'une façon plus générale, les résultats présentés aux Chapitres 3, 4 et 5, notamment en ce qui concerne les débits de coefficients des modèles à long terme, ainsi que les extensions présentées dans cette section ont fourni des pistes pour l'élaboration de codeurs à long terme de diverses qualités et divers débits. On voit par exemple que ce nouveau cadre de modélisation à long terme permet un nouveau degré de liberté dans le dosage entre qualité et débit (cf. les Figures 6.3 et 6.4). *Dans tous les cas, ces codeurs à long terme doivent être caractérisés par un débit significativement plus faible que les codeurs « classiques » (à court terme) de même catégorie (c'est-à-dire basés sur les mêmes techniques dans un cadre à court terme), pour la même qualité de signal*. En contrepartie, la modélisation et le codage à long terme imposeront un délai de codage élevé, pouvant rendre inadéquates les applications de communication interactive et limiter l'utilisation de ces nouveaux codeurs au stockage de la parole et aux applications *offline*. Toutefois, ces dernières sont de plus en plus nombreuses et gourmandes en ressource mémoire car on cherche à stocker de plus en plus d'informations : boîtes de messagerie vocales, transmission de données vocales, « boîtes noires » pour tout type d'application, archivage de grandes bases de données, serveurs de synthèse de parole sur systèmes fixes ou embarqués, etc. On peut donc légitimement penser qu'une approche spécifique plus efficace du codage prenant pleinement en compte la redondance temporelle des signaux doit continuer d'être développée.

Enfin, pour finir cette discussion sur les applications de la modélisation à long terme au codage, et pour clore en même temps ce document sur une note originale, on peut revenir sur les dernières observations reportées dans la Section 5.4.3. Rappelons que dans cette section on avait observé un effet assez amusant sur la synthèse du signal lorsqu'on abaissait progressivement les valeurs du rapport cible R_{min} qui contrôle la qualité des amplitudes spectrales modélisées (par l'intermédiaire de la quantité des amplitudes vérifiant la contrainte du critère perceptif) : le signal paraissait progressivement de plus en plus hypo-articulé. Cet aspect provenait bien sûr du lissage trop accentué des trajectoires de paramètres utilisés à la synthèse. Ce type de dégradation du signal est plutôt original dans le cadre du codage et il s'oppose même à certaines dégradations plus usuelles telles qu'un aspect bruité, saccadé, voire relativement « incohérent » au cours du temps. La « perte de naturel » qui est le terme usuellement consacré pour qualifier la qualité d'un signal codé à très bas débit prend donc ici un sens tout particulier. Au final, les codeurs à long terme proposés ici pourraient ainsi être les premiers codeurs permettant à l'utilisateur de régler systématiquement le compromis entre degré d'articulation de la parole codée et débit, un choix potentiellement cornélien mais pour le moins original !

Bibliographie

[Ahlbom *et al*., 1987]
G. Ahlbom, F. Bimbot and G. Chollet, Modeling spectral speech transitions using temporal decomposition techniques, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'87)*, Dallas, Texas, USA, pp. 13-16, 1987.

[Ahmadi & Spanias, 1999]
S. Ahmadi and A. S. Spanias, Cepstrum-based pitch detection using a new statistical V/UV classification algorithm, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 3, pp. 727–730, May 1999.

[Almeida & Silva, 1984]
L. B. Almeida and F. M. Silva, Variable-frequency synthesis: an improved harmonic coding scheme, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'84)*, San Diego, California, USA, pp. 27.5.1-27.5.4, 1984.

[Atal, 1983]
B. S. Atal, Efficient coding of LPC parameters by temporal decomposition, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'83)*, Boston, Massachusets, USA, pp. 81-84, 1983.

[Atal *et al.*, 1993]
B. S. Atal, V. Cuperman and A. Gersho, *Speech and Audio Coding for Wireless and Network Applications*, Springer-Verlag, Boston, 1993.

[Auger & Flandrin, 1995]
F. Auger and P. Flandrin, Improving the readability of time-frequency and time-scale representation by the reassignment method, *IEEE Transactions on Signal Processing*, Vol. 49, No. 5, pp. 1068-1089, May 1995.

[Bailly & Benoit, 1992]
G. Bailly and C. Benoit (Editors.), *Talking Machines: Theories, Models, and Designs*, Elsevier, North Holland, Amsterdam, 1992.

[Bimbot & Atal, 1991]
F. Bimbot and B. S. Atal, An evaluation of temporal decomposition, *Proceedings of the International Conference on Speech Technology (Eurospeech'91)*, Genova, Italy, pp. 1089-1092, 1991.

[Boëffard & d'Alessandro, 2002]
O. Boëffard et C. d'Alessandro, *Synthèse de la parole,* dans J. Mariani (éditeur), *Traitement Automatique du Langage Parlé, tome 1: Analyse, codage et synthèse*, collection Information-Commande-Communication, Hermès, pp. 115-154, 2002.

[Boersma & Weenink, 2005]
P. Boersma and D. Weenink, *Praat: doing phonetics by computer* (Version 4.3.14) [logiciel de traitement de la parole et des sons], disponible à http://www.praat.org/, 2005.

[Boersma, 1993]
P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *Proceedings of the Institute of Phonetic Sciences of Amsterdam*, Holland, Vol 17, pp. 97-110, 1993.

[Boyer & Abed-Meraim, 2002]
R. Boyer and K. Abed-Meraim, Audio transients modeling by damped and delayed sinusoids (DDS), *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, Orlando, Florida, USA, 2002.

[Campedel-Oudot, 1998]
M. Campedel-Oudot, *Application du modèle sinusoïdes et bruit au codage, au débruitage, et à la modélisation des sons de parole*, Thèse de Doctorat de l'Ecole Nationale Supérieure des Télécommunications, Paris, France, 1998.

[Campedel-Oudot *et al*., 2001]
M. Campedel-Oudot, O. Cappé and E. Moulines, Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 5, pp. 469-481, July 2001.

[Cappé *et al.,* 1995]
O. Cappé, J. Laroche and E. Moulines, Regularized estimation of cepstrum envelope from discrete frequency points, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'95)*, New Paltz, NY, USA, 1995.

[Cermak, 2002]
G. W. Cermak, Subjective quality of speech over packet networks as a function of packet loss, delay, and delay variation, *International Journal of Speech Technology*, Vol. 5, pp. 65-84, 2002.

[de Cheveigné & Kawahara, 2002]
A. de Cheveigné and H. Kawahara, YIN: A fundamental frequency estimator for speech and music, *Journal of the Acoustical Society of America*, Vol. 111, No. 4, pp. 1917-1930, 2002.

[Cohen *et al.,* 2001]
E. Cohen, R.F. Riesenfeld and G. Elber, *Geometric Modeling with Splines*, A K Peters, 2001.

[Collen, 2002]
P. Collen, *Techniques d'enrichissement de Spectre des Signaux Audionumériques*, Thèse de Doctorat de l'Ecole Nationale Supérieure des Télécommunications, Paris, France, 2002.

[Crochiere & Rabiner, 1975]
R. E. Crochiere & L. R. Rabiner, Optimum FIR digital filters implementation for decimation, interpolation and narrow-band filtering, *IEEE. Transactions on Acoustics, Speech and Signal Processing*, Vol. 23, pp. 444-456, October 1975.

[Demany & Semal, 1989]
L. Demany and C. Semal, Detection thresholds for sinusoidal frequency modulation, *Journal of the Acoustical Society of America*, Vol. 85, No. 3, pp. 1295-1301, 1989.

[Depalle *et al.*, 1993]
P. Depalle, G. Garcia and X. Rodet, Analysis of sound for additive synthesis: Tracking of partials using hidden markov models, *Proceedings of the International Computer Music Conference (ICMC'93),* San Francisco, California, USA, 1993.

[Depalle, 1991]
P. Depalle, *Analyse, modélisation et synthèse des sons basées sur le modèle source-filtre*, Thèse de Doctorat de l'Université du Maine, Le Mans, France, 1991.

[Desainte-Catherine & Marchand, 2000]
M. Desainte-Catherine and S. Marchand, High precision fourier analysis of sounds using signal derivatives, *Journal of the Audio Engineering Society*, Vol. 48, No. 48(7/8) , pp. 654-667, July/August 2000

[Ding & Qian, 1997]
Y. Ding and X. Qian, Processing of musical tones using a combined quadratic polynomial phase sinusoid and residual signal model, *Journal of the Audio Engineering Society,* Vol. 45, No. 7/8, pp. 571-585, 1997.

[Dolson, 1986]
M. Dolson, The phase vocoder: A tutorial*, Computer Music Journal*, Vol. 10, No. 4, pp. 14–27, 1986.

[Dusan *et al*., 2004]
S. Dusan, J. Flanagan, A. Karve and M. Balaraman, Speech coding using trajectory compression and multiple sensors, *Proceedings of the International Conference on Speech and Language Processing (ICSLP'2004)*, Jeju, South Korea, 2004.

[Dusan *et al.*, 2007]
S. Dusan, J. Flanagan, A. Karve and M. Balaraman, Speech compression by polynomial approximation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 2, pp. 387-395, February 2007.

[Dusan & Flanagan, 2002]
S. Dusan and J. Flanagan, Low bit rate speech coding using trajectory modeling, *Journal of the Acoustical Society of America*, Vol. 112, No. 5, Pt. 2, 3pSC6(A), 2002.

[En-Najjary, 2005]
T. En-Najjary, *Conversion de voix pour la synthèse de la parole*, Thèse de Doctorat de l'Université de Rennes I, France, Avril 2005.

[Farvardin & Laroia, 1989]
N. Farvardin & R. Laroia, Efficient encoding of speech LSP parameters using the discrete cosine transformation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'89)*, Glasgow, Scotland, pp. 168-171, 1989.

[Firouzmand & Girin, 2005]
M. Firouzmand & L. Girin, Perceptually weighted long term modeling of sinusoidal speech amplitude trajectories, *Int. Conf. on Acoustics, Speech & Signal Processing (ICASSP 2005),* Philadelphia, USA.

[Firouzmand, Girin & Marchand, 2005]
M. Firouzmand, L. Girin & S. Marchand, Comparing several models for perceptual long-term modeling of amplitude and phase trajectories of sinusoidal speech, *Euro. Conf. on Speech Communication and Technology (Interspeech 2005),* Lisboa, Portugal.

[Firouzmand & Girin, 2007]
M. Firouzmand & L. Girin, Long-term flexible 2D modeling of speech spectral amplitudes, Submitted to the *Int. Conf. on Speech Science and Technology (Interspeech 2007)*, Antwerp, Belgium.

[Flanagan & Golden, 1966]
J. L. Flanagan and R. M. Golden, The phase vocoder, *Bell System Technical Journal*, pp. 1493–1509, November 1966.

[Freed *et al*., 1993]
A. Freed, X. Rodet and P. Depalle, Performance, synthesis and control of additive synthesis on a desktop computer using *FFT*[1], *Proceedings of the International Computer Music Conference (ICMC'93)*, Tokyo, Japan, 1993.

[Galas & Rodet, 1991]
T. Galas and X. Rodet, Generalized functional approximation for source-filter system modelling, *Proceedings of the International Conference on Speech Technology (Eurospeech'91)*, Genova, Italy, pp. 1085-1088, 1991.

[Galas & Rodet, 1990]
T. Galas and X. Rodet, An improved cepstral methode for deconvolution of source-filter system with discret spectra : application to musical sound signals, *Proceedings of the International Computer Music Conference (ICMC'90)*, Glasgow, Scotland, pp. 82-84, 1990.

[Garcia & Pampin, 1999]
G. Garcia and J. Pampin, Data compression of sinusoidal modeling parameters based on psychoacoustic masking, *Proceedings of the International Computer Music Conference (ICMC'99)*, Bejing, China, pp. 40-43, 1999.

[George & Smith, 1997]
E. B. George and M. J. T. Smith, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 5, pp. 389-406, September 1997.

[Gerhard, 2003]
D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*, Technical Report TR-CS 203-06, Department of Computer Science, University of Regina, Saskatchewan, Canada, November. 2003.

[Gersho, 1994]
A. Gersho, Advances in speech and audio compression, *Proceedings of the IEEE*, Vol. 82, No. 6, 1994

[Gersho & Gray, 1992]
A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston, 1992.

[Girin, Firouzmand & Marchand, 2004]
L. Girin, M. Firouzmand & S. Marchand, Long term modeling of phase trajectories within the speech sinusoidal model framework, *Int. Conf. on Speech & Language Proc. (ICSLP 2004),* Jeju, South Korea.

[Girin, Firouzmand & Marchand, 2007]
L. Girin, M. Firouzmand & S. Marchand, Perceptual long-term variable-rate sinusoidal modeling of speech, *IEEE Transactions on Audio, Speech, and Language Processing,* **15**(2), 2007.

[Girin *et al.,* 2003]
L. Girin, S. Marchand, J. di Martino, A. Röbel, and G. Peeters, Comparing the order of a polynomial phase model for the synthesis of quasi-harmonic audio signals, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, New Paltz, NY, USA, 2003.

[Girin & Marchand, 2004]
L. Girin, and S. Marchand, Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, Montréal, Canada, 2004.

[Girin, 2006]
L. Girin, Theoretical and experimental bases of a new method for accurate separation of harmonic and noise components of speech signals, Proceedings of the *European Signal Processing Conference (EUSIPCO'2006)*, Florence, Italy, 2006.

[Girin, 2007]
L. Girin, Long-term quantization of LSF parameters, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP'2007*), Honolulu, Hawaï, 2007.

[Ghaemmaghami & Deriche, 1996]
S. Ghaemmaghami and M. Deriche, A new approach to very low-rate speech coding using temporal decomposition, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'96)*, Atlanta, Georgia, USA, pp. 224-227, 1996.

[Ghaemmaghami *et al.*, 1997]
S. Ghaemmaghami, M. Deriche, and B. Boashash, Comparative study of different parameters for temporal decomposition based speech coding, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, Munich, Germany, 1997.

[Golub & van Loan, 1983]
G. H. Golub and C. F. van Loan, *Matrix computations*, North Oxford Academic, Oxford, 1983.

[Goodwin, 1997]
M. Goodwin, Matching pursuit with damped sinusoids, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, Munich, Germany, pp. 2037-2040, 1997.

[Gottesman, 1999]
O. Gottesman, Dispersion phase vector quantization for the enhancement of waveform interpolative coder, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'99)*, Phoenix, Arizona, USA, pp. 269-272, 1999.

[Gray & Markel, 1976]
A. H. Gray and J. D. Markel, Distance measures for speech processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 24, No. 5, pp. 380-391, 1976.

[Gray *et al*., 1980]
R. M. Gray, A. Buzo, A. H. Gray and Y. Matsuyama, Distortion measures for speech processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 367-376, 1980.

[Griffin & Lim, 1988]
D. W. Griffin et J. S. Lim, Multiband-excitation vocoder, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 2, pp. 236-243, 1988.

[Guéguin, 2006]
M. Guéguin, *Evaluation objective de la qualité vocale en contexte de conversation*, Thèse de Doctorat de l'Université de Rennes 1, France, 2006.

[Hagen *et al.*, 1999]
R. Hagen, E. Paksoy and A. Gersho, Voicing-specific LPC quantization for variable-rate speech coding, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 5, 1999, pp. 485-494.

[Hammer *et al*., 2005]
F. Hammer, P. Reichl and A. Raake, The well-tempered conversation: interactivity, delay and perceptual VoIP quality, *Proceedings of the IEEE International Conference on Communications (ICC'2005)*, Seoul, Korea, 2005.

[Hammer *et al*., 2004]
F. Hammer, P. Reichl and A. Raake, Elements of interactivity in telephone conversations, *Proceedings of the International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004)*, Jeju, Korea, pp. 1741-1744, 2004.

[Hanna, 2003]
P. Hanna, *Modélisation statistique de sons bruités*, Thèse de Doctorat de l'Université de Bordeaux 1, France, Décembre 2003.

[Hanna & Desainte-Catherine, 2005]
P. Hanna & M. Desainte-Catherine, CNSS model: a statistical and spectral model for representing noisy sounds with short-time sinusoids, *EURASIP Journal of Applied Signal Processing*, Vol. 2005, No. 12, pp. 1794-1806, 2005.

[Harris, 1978]
J. F. Harris, On the use of windows for harmonic analysis with the discrete fourier transform, *Proceedings of the IEEE*, Vol. 66, pp. 51-83, 1978.

[Helmholtz, 1877]
H. L. F. von Helmholtz, *Die Lehre von den Tonempfindungen, als physiologische Grundlage für die Theorie der Musik*, Braunschweig, Verlag Vieweg, 1877.

[IEEE Com. Mag., 1997]
IEEE Communications Magazine, Special Issue: Standardization and characterization of G.729, Vol. 35, No. 9, September 1997

[ISO/IEC MPEG, 1992]
ISO JTC1/SC29/WG11 MPEG, *IS11172-3 Information technology – Coding of moving Pictures and associated audio for digital storage media at up to about 1.5 Mbit/s*, Part 3: Audio, 1992.

[Jackson & Shadle, 2000]
P. Jackson & C. Shadle, Fricative noise modulated by voicing, as revealed by pitch-scaled decomposition, *Journal of the Acoustical Society of America*, Vol. 108, No. 4, pp. 1421-1434, 2000.

[Jackson & Shadle, 2001]
P. Jackson and C. Shadle, Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 7, pp. 713-726, 2001.

[Jayant *et al.*, 1993]
N. S. Jayant, J. D. Johnston and R. Safranek, Signal compression based on models of human perception, *Proceedings of the IEEE*, Vol. 81, pp. 1385-1422, October, 1993.

[Jayant & Noll, 1984]
N. S. Jayant and P. Noll, *Digital coding of waveforms – Principles and applications to speech and video*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1984.

[Kay, 1988]
S. M. Kay, *Modern spectral estimation*, Signal Processing Series, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[Kay, 1993]
S. M. Kay, *Fundamentals of statistical signal processing – Estimation theory*, Signal Processing Series, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[Kim, 2001]
D. S. Kim, On the perceptually irrelevant phase information in sinusoidal representation of speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 8, pp. 900-905, 2001.

[Kim, 2003]
D. S. Kim, Perceptual phase quantization of speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 4, pp. 355-364, 2003.

[Kitawaki & Itoh, 1991]
N. Kitawaki and K. Itoh, Pure delay effects on speech quality in telecommunications, *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 4, pp. 586-593, 1991.

[Lagrange, 2004]
M. Lagrange, *Modélisation sinusoïdale des sons polyphoniques*, Thèse de Doctorat de l'Université Bordeaux 1, France, Décembre 2004.

[Lagrange *et al.,* 2002].
M. Lagrange, S. Marchand and J-B. Rault, Sinusoidal parameter extraction and component selection in a non stationary model, *Proceedings of the International Conference Digital Audio Effects (DAFx'2002)*, Hamburg, Germany, pp. 59-64, 2002.

[Lim & Oppenheim, 1979]
J. S. Lim & A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech, *Proceedings of the IEEE*, Vol. 67, No. 12, pp. 1586-1604, 1979.

[Li *et al.,* 2001]
C. Li, P. Lupini, E. Shlomot and V. Cuperman, Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 6, pp. 622-631, 2001.

[Levine, 1998]
S. N. Levine, *Audio representations for data compression and compressed domain processing*, Ph.D. Thesis of the Stanford University, Department of Electrical Engineering, Palo Alto, California, USA, December 1998.

[Levinson, 1947]
N. Levinson, The Wiener RMS (root mean square) error criterion in filter design and prediction, *Journal of Mathematics and Physics*, Vol. 125, pp. 261-268, 1947.

[Liu & Lin, 2001]
D. J. Liu and C. T. Lin, Fundamental frequency estimation on the joint time-frequency analysis of harmonic spectral structure, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, pp. 609–621, September 2001.

[Macon, 1996]
M. Macon, *Speech synthesis based on sinusoidal modeling*, Ph.D. Thesis of the Georgia Institute of Technology, Atlanta, Georgia, USA, October 1996.

[Macon & Clements, 1997]
M. W. Macon and M. A. Clements, Sinusoïdal modeling and modification of unvoiced speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 5, No. 6, pp. 557-560, 1997.

[Makhoul, 1975]
J. Makhoul, Linear prediction: A tutorial review, *Proceedings of the IEEE*, Vol. 63, No. 5, pp. 561–580, April 1975.

[Manfredi *et al.*, 2000]
C. Manfredi, M. d'Aniello, P. Bruscaglioni and A. Ismaelli, A comparative analysis of fundamental frequency estimation methods with application to pathological voices, *Medical Engineering & Physics*, Vol. 22, No.^{o}2, pp. 135-147, 2000.

[Markel & Gray, 1976]
J. D. Markel and A. H. J. Gray, *Linear prediction of speech*, Springer-Verlag, New-York, 1976.

[Marchand, 2000]
S. Marchand, *Modélisation informatique du son musical*, Thèse de Doctorat de l'Université de Bordeaux 1, France, décembre 2000

[Marchand & Raspaud, 2004]
S. Marchand and M. Raspaud, Enhanced time-stretching using order-2 sinusoidal modeling, *Proceedings of the International Conference on Digital Audio Effects (DAFx'2004),* Napoli, Italy, 2004.

[Marques & Almeida, 1986]
J. Marques and L. Almeida, A background for sinusoid based representation of the voiced speech, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'86)*, pp. 1233-1236, Tokyo, Japan, 1986.

[Master & Liu, 2003]
A. S. Master & Y. W. Liu, Robust chirp parameter estimation for Hann windowed signals, *Proceedings of IEEE International Conference on Multimedia and Exposition (ICME'2003)*, Baltimore, MD, USA, 2003.

[McAulay & Quatieri, 1986]
R. J. McAulay and T. F. Quatieri, Speech analysis/synthesis based on a sinusoidal representation, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 34, No. 4, pp. 744-754, August 1986.

[McAulay & Quatieri, 1995]
R. J. McAulay and T. F. Quatieri, *Sinusoidal coding*, Chapter 4 from *Speech Coding and Synthesis*, Edited by W. B. Kleijn and K. K. Paliwal, Elsevier Science, 1995.

[Möller, 2000]
S. Möller, *Assessment and prediction of speech quality in telecommunications*, Kluwer Academic Publishers, Boston, USA, 2000.

[Moorer, 1977]
J. A. Moorer, Signal processing aspects of computer music – A survey, *Computer Music Journal*, Vol. 1, No. 1, pp. 4-37, 1977.

[Moorer, 1978]
J. A. Moorer, The use of the phase vocoder in computer music applications, *Journal of the Audio Engineering Society*, Vol. 26, No. 1, pp. 42-45, 1978.

[Moore & Glasberg, 1983]
B. C. J. Moore and B. R. Glasberg, Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *Journal of the Acoustical Society of America*, Vol. 74, pp. 750–753, 1983.

[Moulines & Charpentier, 1990]
E. Moulines and F. Charpentier, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication*, Vol. 9, pp. 453-467, 1990.

[Nieuwenhuijse *et al.,* 1998]
J. Nieuwenhuijse, R. Heusdens and F. Deprettere, Robust exponential modeling of audio signals, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, USA, pp. 3581-3584, 1998.

[Nocerino *et al*., 1985]
N. Nocerino, F.K. Soong, L.R. Rabiner and D.H. Klatt, Comparative study of several distortion measures for speech recognition, *Speech Communication*, Vol. 4, pp. 317-331, 1985.

[Oppenheim & Schaffer, 1989]
A. V. Oppenheim and R. W. Schaffer, *Discrete time signal processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[Painter & Spanias, 2000]
T. Painter and A. Spanias, Perceptual coding of digital audio, *Proceedings of the IEEE*, Vol. 88, No. 4, pp. 451-513, 2000.

[Paliwal, 1983]
K. K. Paliwal, Comparative performance evaluation of different pitch estimation methods for noisy speech, *Acoustics letters*, Vol. 6, No 11, pp. 164-166, 1983.

[Paliwal & Atal, 1993]
K. K. Paliwal and B. S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Transactions on Speech and Audio Processing*, Vol. 1, pp. 3-14, January 1993.

[Papoulis, 1977]
A. Papoulis, *Signal Analysis*, McGraw-Hill, New-York, USA, 1977.

[Pobloth & Kleijn, 2003]
H. Pobloth and W. B. Kleijn, Squared error as a measure of perceived phase distortion, *Journal of the Acoustical Society of America*, Vol. 114, No. 2, pp. 1081-1094, 2003.

[Poli *et al.*, 1991]
G. De Poli, A. Piccialli and C. Roads (Editors), *Representation of musical signals*, MIT Press, Cambridge, Massachusetts, USA, 1991.

[Portnoff, 1976]
M. R. Portnoff, Implementation of the digital phase vocoder using the fast Fourier transform, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 24, No. 3, pp. 243–248, 1976.

[Prandom *et al*. 1997]
P. Prandom, M. Goodwin et M. Vetterli, Optimal time segmentation for signal modeling and compression, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97)*, Munich, Germany, pp. 2029-2032, 1997.

[Press *et al.* 1992]
W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C (The Art of Scientific Computing)*, Chapter 10: Minimization or Maximization of Functions, pp. 402–405, Cambridge University Press, Cambridge, MA, USA, 1992

[Princen & Bradley, 1986]
J. P. Princen and A. B. Bradley, Analysis/synthesis filter bank design based on time domain aliasing cancellation, *In IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. 34, No. 5, pp. 1153-1161, October. 1986.

[Quatieri & McAulay, 1992]
T. F. Quatieri and R. J. McAulay, Shape invariant time-scale and pitch modification of speech, *IEEE Transactions on Signal Processing*, Vol. 40, No. 3, pp. 497-510, 1992.

[Rabiner & Schaffer, 1978]
L. R. Rabiner and R. W. Schaffer, *Digital processing of speech signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[Raspaud *et al*., 2005]
S. Marchand, M. Raspaud, and L. Girin, A generalized polynomial and sinusoidal model for partial tracking and time stretching, *Proceedings of the International Conference on Digital Audio Effects (DAFx'2005),* Madrid, Spain, 2005.

[Richard & d'Alessandro, 1996]
G. Richard and d'Alessandro, Analysis/synthesis and modification of the speech aperiodic component*, Speech Communication*, Vol. 19, pp. 221-244, 1996.

[Roads, 1996]
C. Roads, *The Computer Music Tutorial*, MIT Press, Cambridge, Massachusetts, USA, 1996.

[Roads *et al.,* 1997]
C. Roads, S. T. Pope, A. Piccialli and G. De Poli (Editors), *Musical signal processing*, Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[Rodet, 1997]
X. Rodet, Sinusoidal + residual models for musical sound signals analysis/synthesis, *Journal of Applied Signal Processing*, Vol. 4, No. 3, pp 131-141, 1997.

[Rodet & Depalle, 1993]
X. Rodet and P. Depalle, Spectral envelope and inverse FFT synthesis*,* Proceedings of the *93rd Convention of the Audio Engineering Society*, San Francisco, 1993.

[Schroeder *et al.*, 1979]
M. Schroeder, B. S. Atal and J. L. Hall, Optimizing digital coders by exploiting masking properties of the human ear, *Journal of the Acoustical Society of America*, Vol. 66, No. 6, pp. 1647-1652, December 1979.

[M. H. Serra, 1997]
M. H. Serra, Introducing the phase vocoder, In *Musical Signal Processing*, C. Roads, S. T. Pope, A. Piccialli and G. De Poli (Editors), Swets & Zeitlinger, Lisse, the Netherlands, pp. 31–90, 1997.

[Serra & Smith, 1990]
X. Serra and J.O. Smith, Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition, *Computer Music Journal*, Vol. 14, No. 4, pp. 12-24, 1990.

[Serra, 1997]
X. Serra, Musical sound modeling with sinusoids plus noise, In *Musical Signal Processing*, C. Roads, S. T. Pope, A. Piccialli and G. De Poli (Editors), Swets & Zeitlinger, Lisse, the Netherlands, pp. 91-122, 1997.

[Shimamura & Kobayashi, 2001]
T. Shimamura and H. Kobayashi, Weighted autocorrelation for pitch extraction of noisy speech, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 7, pp. 727–730, 2001.

[Shlomot *et al*., 2001]
E. Shlomot, V. Cuperman and A. Gersho, Combined harmonic and waveform coding of speech at 4 kb/s, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 6, pp. 632-646, 2001.

[Smith & Serra, 1987]
J. O. Smith and X. Serra, An analysis/synthesis program for non-harmonic sounds based on sinusoidal representation, *Proceedings of the International Computer Music Conference (ICMC'87)*, San Francisco, California, USA, 1987.

[Sugamura & Itakura, 1986]
N. Sugamura & F. Itakura, Speech Analysis and synthesis method developed at ACL in NTT - From LPC to LSP, *Speech Communication,* Vol. 5, pp. 199-215, 1986.

[Stylianou, 1996]
Y. Stylianou, *Modèles harmoniques plus bruit combinés avec des méthodes statistiques pour la modification de la parole et du locuteur*, Thèse de Doctorat de l'Ecole Nationale Supérieure des Télécommunications, Paris, 1996.

[Stylianou, 1996]
Y.Stylianou, Applying the harmonic plus noise model in concatenative speech synthesis, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 4, pp. 21-29, 2001.

[Terhardt, 1979]
E. Terhardt, Calculating virtual pitch, *Hearing Research*, Vol. 1, pp. 155-182, 1979.

[Tsao & Gray, 1985]
C. Tsao and R. M. Gray, Matrix quantizer design for LPC speech using the generalized Lloyd algorithm, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 33, No. 3, pp. 537-545, 1985.

[Unser et al., 1993]
M. Unser, A. Aldroubi, and M. Eden, B-Spline signal processing : Part I-Theory, *IEEE Transactions on Signal Processing*, Vol. 41, No. 2, pp. 821-833, 1993.

[Van Dijk-Kappers *et al.*, 1989]
A. M. L. Van Dijk-Kappers and S. M. Marcus, Temporal decomposition of speech, *Speech Communication*, Vol. 8, pp. 125-135, 1989.

[Van Dijk-Kappers, 1989]
A. M. L. Van Dijk-Kappers, Comparison of parameter sets for temporal decomposition, *Speech Communication*, Vol. 8, pp. 203-220, 1989.

[Wang & Lim, 1982]
D. L. Wang & J. S. Lim, The unimportance of phase in speech enhancement, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 30, No. 4, pp. 679-681, 1982.

[Wiener, 1949]
N. Wiener, *Extrapolation, interpolation and smoothing of stationnary time series, with engineering applications*, Wiley, New-York, 1949.

[Xydeas & Papanastasiou, 1999]
C. S. Xydeas and C. Papanastasiou, Split matrix quantization of LPC parameters, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 2, pp. 113-125, 1999.

[Yegnanarayana *et al.,* 1998]
B. Yegnanarayana, C. d'Alessandro & V. Darsinos, An iterative algorithm for decomposition of speech signals into periodic and aperiodic components, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, pp. 1-11, 1998.

[Yong *et al.*, 1988]
M. Yong, G. Davidson and A. Gersho, Encoding of LPC spectral parameters using switched-adaptive interframe vector prediction, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'88)*, New-York, USA, pp. 402-405, 1988.

[Zölder, 2002]
U. Zölder, *DAFx'2002: Digital Audio Effects*, John Wiley & Sons, 2002

[Zwicker & Feldtkeller, 1981]
E. Zwicker et R. Feldtkeller, *Psychoacoustique: l'oreille récepteur d'information,* (version française de *Das oh als nachrichtenmpfänger*), Masson, Paris, 1981.

[Zwicker & Fastl, 1990]
E. Zwicker & H. Fastl, *Psychoacoustics: facts and models*, Springer-Verlag, Heidelberg, Germany, 1990.

Résumé

La modélisation sinusoïdale du signal de parole est usuellement définie à « court terme », c'est-à-dire sur des trames successives de signal d'une durée de l'ordre de 10 à 30 ms. Cette thèse apporte une contribution nouvelle à ce domaine en ajoutant à ce niveau traditionnel de modélisation spectrale un niveau supplémentaire le long de l'axe temporel : on cherche à modéliser les trajectoires de paramètres sinusoïdaux (amplitudes et phases) sur des durées significativement plus longues que celles des trames à court terme (typiquement plusieurs centaines de ms ; on considère dans cette thèse des sections de parole continûment voisées). Nous proposons pour cela d'utiliser différents modèles à long terme à base de fonctions en cosinus discrets et de fonctions polynomiales. L'ajustement des trajectoires est réalisé par une régression au sens des moindres carrés pondérés, les poids de la régression étant déterminés par des critères perceptifs adaptés au traitement à long terme. Pour cette tâche, une série d'algorithmes itératifs est proposée et testée. L'approche à long terme se révèle à la fois efficace et parcimonieuse pour décrire la dynamique des signaux de parole voisés.

Mots-clés : Modèle sinusoïdal de la parole, amplitudes, phases, modélisation temporelle, long terme, modèles psycho-acoustiques, compression de parole, transformation de parole.

Abstract

The sinusoidal model of speech signals is usually defined on a "short-term" basis, *i.e*. on successive frames of about 10–30 ms. In this thesis, we add to this usual spectral modeling a new level of modeling along the temporal axis: the goal is to model the temporal trajectories of the sinusoidal parameters (amplitudes and phases) over durations which are significantly longer than the short-term frames (typically several hundreds of ms; continuously voiced sections of speech are considered in this study). For this, we propose to use different long-term models based on discrete cosine and polynomial functions. The fitting of these models with the parameters trajectories is achieved by a weighted least square minimisation technique, the weights being derived from perceptual criteria which are adapted to the long-term processing. For this task, a series of iterative algorithms is proposed and tested. The proposed long-term approach is shown to provide an efficent and sparse representation of the dynamics of voiced speech signals.

Keywords : Sinusoidal model of speech, amplitudes, phases, temporal modeling, long term, psychoacoustic models, speech compression, speech transformation.