THÈSE de DOCTORAT de l'UNIVERSITE PARIS 6

Spécialité :

Acoustique Traitement du signal et Informatique Appliqués à la Musique

Présentée par

Geoffroy PEETERS

pour obtenir le grade de DOCTEUR de l'UNIVERSITE PARIS 6

Sujet de la thèse :

Modèles et modification du signal sonore adaptés à ses caractéristiques locales

Soutenue le 11 juillet 2001,
devant le jury composé de :

M.	Ch.	D'ALESSANDRO	Rapporteurs
M.	Th.	DUTOIT	
M.	Ph.	DEPALLE	Examinateurs
M.	St.	MALLAT	
M.	X.	SERRA	
M.	X.	RODET	Directeur de thèse

Travail effectué à l'Ircam - Centre Pompidou
1, Place Igor Stravinsky, 75004 Paris
Rapport version 1.1

Remerciements

Je tiens à remercier

Xavier Rodet, pour m'avoir donné l'opportunité de réaliser ma thèse au sein de l'équipe analyse/synthèse de l'Ircam

Thomas Hélie, pour la relecture de certains chapitres de mon rapport de thèse et les nombreuses discussions fertiles

Marine Oudot, pour m'avoir communiqué l'énergie nécessaire au développement de cette recherche

Rémy Gribonval, pour les discussions fructueuses sur la phase

Philippe Depalle, pour m'avoir donné goût au traitement du signal

Franko Silvio Marzano pour m'avoir donné goût à la recherche

Guillaume Lemaitre et Patrice Tisserand, pour leur bonne humeur contagieuse

Tout les gens présents ou passés dans l'équipe analyse/synthèse en particulier Christophe Vergez, Stéphane Rossignol, Marcelo Wanderley, Diemo Schwarz, Stéphan Tassart

Les membres de l'équipe perception et cognition musicale, pour leur sympathie et nos débats politiques sur les élections parisiennes

Le Ministère de l'Education Nationale, pour avoir financé mes recherches dans le cadre d'une allocation de recherche

Hideki Kawahara et Yegna Yegnanarayana, pour avoir pris le temps de répondre à mes questions

Nortbert Schnell, Serge Lemouton et Philippe Manoury, pour m'avoir permis de participer au projet «K ...»

Hugues Vinet et Vincent Puig, pour la confiance qu'ils m'ont accordée

Laurent Ghys et Julien Boissinot, pour avoir pris soin de mon environnement informatique

Marie, pour sa patience

Ma famille pour son soutient indispensable et, en particulier, mon père pour la relecture (de fichiers TEXsous MS WORD) de mon rapport de thèse et les nombreux commentaires toujours très avisés,

Tous ceux et toutes celles, qui d'une manière ou d'une autre, on rendue cette période de ma vie agréable, en partageant des moments de musique, par leur humour, leur bonne-humeur, leur passion, leur connaissance ou simplement par leur présence

Résumé

L'objectif de cette recherche est l'élaboration d'une méthode de modification du signal sonore utilisant des modèles de représentation du signal adaptés à ses caractéristiques locales.

A cet effet, deux modèles de représentation sont étudiés, correspondant chacun à une interprétation différente d'un signal sonore : la décomposition temporelle d'un signal en formes d'onde élémentaires - interprétation, dans le cas d'un signal périodique, en termes de répétition d'une forme d'onde quasi-identique au cours du temps -, et la représentation d'un signal par une somme de sinusoïdes variables au cours du temps - interprétation en termes de relations harmoniques entre ses composantes fréquentielles -.

L'originalité de notre méthode consiste à tirer parti des avantages de chacun des deux modèles pour la classe de signaux pour laquelle il est le mieux adapté.

Cette méthode hybride nécessite l'estimation des paramètres de chacun des deux modèles (localisation des formes d'ondes, fréquence fondamentale, fréquence, amplitude et phase des composantes fréquentielles) ainsi que d'un ensemble de caractéristiques permettant le choix du modèle le mieux adapté à la représentation d'une région du plan temps/fréquence donnée (singularité des formes d'onde, erreur de modélisation, erreur de spécification, voisement, harmonicité). De nouveaux estimateurs sont proposés et nous montrons la pertinence de l'information de phase pour l'estimation. Dans chaque cas, les estimateurs proposés sont comparés aux estimateurs communément utilisés.

Les méthodes de modification du signal associées à ces deux modèles sont la méthode PSOLA (Pitch Synchronous Overlap-Add) et la synthèse par addition de sinusoïdes. Les possibilités de modification du signal de ces deux méthodes sont étudiées et des améliorations sont proposées, en particulier en ce qui concerne la prise en compte des relations de phase lors de modifications d'un signal.

Les deux modèles sont ensuite utilisés dans une nouvelle méthode hybride de modification du signal (SINOLA), assignant à chaque région du plan temps/fréquence celui des deux modèles le plus adapté à sa représentation et à sa modification.

Mots-clés

informatique musicale, PSOLA, analyse/synthèse sinusoïdale, phase, retard de groupe, estimation, parole

Abstract

The purpose of this research is to develop a method for the modification of the sound signal, using signal models adapted to the local characteristics of the signal.

With this end in view, we study two models of signal representation, each one corresponding to a different interpretation of a sound signal : temporal decomposition of the signal into elementary waveforms - interpretation, in the case of a periodic signal, in terms of temporal repetition of quasi-identical waveforms -, and signal representation by a sum of time-variable sinusoidal components - interpretation in term of harmonic relations between its frequential components -.

The originality of our method consists in taking benefit of each model for the signal class for which the model is best suited.

This hybrid method requires estimating the parameters of each of the two models (waveforms localization, fundamental frequency, frequency, amplitude and phase of the spectral components) as well as a set of characteristics allowing the choice of the best-suited model to represent a given time/frequency region (singularity of the waveforms, modelization error, specification error, harmonicity, voicing). New estimators are proposed and we show the relevance of phase information for the sake of estimation. In each case, the proposed estimators are compared with the commonly used ones.

The signal modification methods associated with these two models are the PSOLA (Pitch Synchronous Overlap-Add) method and the sinusoidal additive synthesis. Signal modification possibilities of each of these methods are studied and improvements are proposed, especially concerning the consideration of phase relations during signal modifications.

The two models are then used in a new hybrid method to modify the signal (SINOLA), which assigns to each time/frequency region the best-suited model for its representation and its modification.

Keywords

computer music, PSOLA, sinusoidal analysis/synthesis, phase, group delay, estimation, speech

Table des matières

Introduction générale

Chapitre 1

Introduction générale

Les méthodes de modification du signal sonore se divisent en deux classes principales, celles qui s'intéressent au modèle de production du son («modèles physiques») et celles qui s'intéressent au signal produit («modèle de signaux»). Dans cette seconde catégorie, nous distinguons les «modèles de signaux paramétriques» et les «modèles de signaux non-paramétriques». Les premiers nécessitent le passage de manière explicite à une modélisation et donc à l'estimation des «paramètres» d'un modèle. Les seconds ne nécessitent pas ce passage et opèrent la transformation de manière directe sur le signal.

Les «modèles de signaux paramétriques» procèdent par modélisation d'un signal. Le son peut être représenté comme la sortie d'un système (cas du modèle source/filtre) ou comme une décomposition sur une base de fonctions (cas de la décomposition en sinusoïdes ou en atomes). La définition du modèle est généralement liée à une catégorie de signaux de caractéristiques communes : ensemble des signaux harmoniques, ensemble des signaux à variation temporelle lente, ensemble des signaux à caractère impulsionnel, ... Le modèle est ensuite utilisé afin de reproduire, de modifier, de compresser, de contrôler, ou de décrire le son. Les «modèles de signaux non-paramétriques», même s'ils ne reposent pas sur un modèle de représentation du son, sont aussi généralement associés à une catégorie de signaux.

De la même manière qu'un système dynamique (modèle physique) n'est apte à représenter qu'une catégorie d'instruments de musique fonctionnant selon un principe physique similaire, un modèle de signal n'est généralement apte qu'à représenter un ensemble de signaux de caractéristiques communes. Dés lors, l'extension de chaque modèle de signal ou l'utilisation simultanée de plusieurs modèles de signaux semblent s'imposer dans le but d'un traitement généralisable à un grand nombre de signaux.

L'objectif de cette recherche est de permettre l'élaboration d'un tel modèle hybride dans le cadre des signaux non-mixtes et monophoniques.

Les deux modèles que nous considérerons sont des modèles de signaux : le modèle PSOLA ou Pitch Synchronous Overlap-Add - modèle non-paramétrique -, et le modèle sinusoïdal - modèle paramétrique -. Ces deux modèles correspondent à deux interprétations différentes des signaux. Dans le cas d'un signal harmonique, le premier considère le signal comme la répétition périodique d'une forme d'onde , le second comme la superposition de composantes de rapport de fréquences harmonique. Chacune de ces méthodes peut être étendue selon sa propre interprétation : signaux pseudo-périodiques (exemple : impulsions d'un signal de parole

de périodicité mal définie ; voir Fig. 1.1 panneau de gauche) et signaux pseudo-harmoniques ou inharmoniques (exemple : partiels de fréquences dilatées d'un son de piano ; voir Fig. 1.2 panneau de droite).

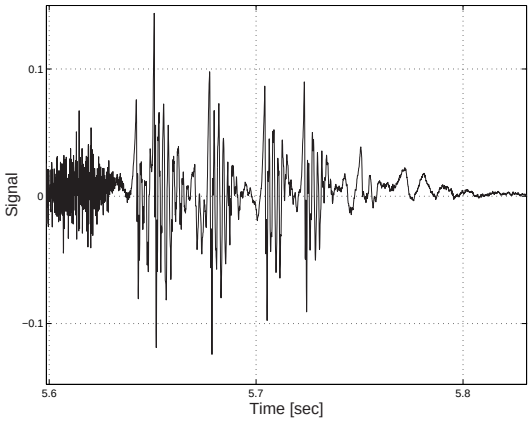


Fig. 1.1 – Signal pseudo-périodique

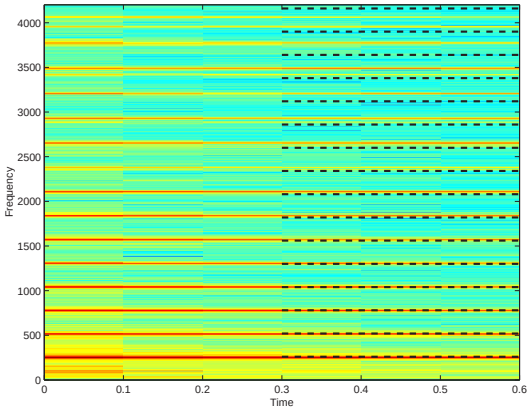


Fig. 1.2 – Signal pseudo-harmonique ou inharmonique

Ces interprétations conduisent à favoriser l'utilisation d'une méthode pour un type particulier de son : le modèle PSOLA pour les signaux pour lesquels l'interprétation en tant que répétition d'une forme d'onde est possible (a priori les signaux pour lesquels une localisation temporelle des formes d'onde est possible), le modèle sinusoïdal pour les signaux pour lesquels l'interprétation en tant que rapport de fréquences est possible (a priori les signaux pour lesquels une localisation fréquentielle est possible, des signaux à variations lentes).

Pour chacun de ces deux modèles, nous devons déterminer la catégorie de signaux auxquels le modèle s'applique, ainsi que définir un ensemble de caractéristiques de cette catégorie permettant, pour un signal donné, de déterminer le modèle le plus approprié. Nous devons également estimer les paramètres de ce modèle lorsque le modèle le requiert. Ceci fait l'objet de la première partie de cette recherche.

Dans la deuxième partie, nous étudions les possibilités de modification du signal offertes par chacun des deux modèles, ainsi que les améliorations et extensions qui peuvent y être apportées. Le modèle le plus adapté aux caractéristiques locales du signal, en temps et en fréquence, est alors utilisé pour la modification du signal.

Notre thèse est donc que :
- ▷ l'utilisation conjointe de plusieurs modèles de signaux, par choix d'un modèle adapté aux caractéristiques locales du signal sonore, plutôt que le perfectionnement indéfini d'un même modèle, permet d'atteindre des modifications de grande qualité ;
- ▷ la phase contient une information pertinente pour l'analyse des signaux sonores (localisation temporelle, localisation fréquentielle, synchronie des évènements fréquentiels) ;
- ▷ la prise en compte des relations de phase lors de modifications d'un signal permet d'atteindre un haut niveau de qualité sonore.

Plan du document

Première partie : Caractérisation du signal

Dans la première partie de cette recherche, nous nous intéressons à la caractérisation du signal. Cette caractérisation doit permettre le choix du modèle le plus adapté aux caractéristiques locales du signal (décomposition en formes d'onde élémentaires périodiques ou décomposition en sommes de sinusoïdes) ainsi que la modélisation du signal selon le modèle choisi.

L'organisation de cette partie regroupe les différents types de caractéristiques de manière thématique et hiérarchique.

Chapitre 3 Ce chapitre est consacré à la détection et à la caractérisation des **singularités** d'un signal. Les «singularités» étudiées ici sont définies comme une concentration locale de l'énergie du signal par rapport à une échelle d'observation donnée. L'interprétation de ces singularités dépend de l'échelle considérée.

Pour une échelle d'observation de l'ordre de la période fondamentale du signal, la présence de singularités périodiques dans le signal permettra la décomposition du signal en formes d'onde élémentaires périodiques et la modification du signal à l'aide d'un algorithme de type PSOLA. Dans le cas d'un signal de voix parlée, ces singularités périodiques correspondent aux «instants de fermeture de la glotte» (IFG). Nous rappelons les méthodes existantes habituellement utilisées pour la détection des IFGs : méthodes utilisant la rupture d'un modèle auto-régressif, méthodes utilisant le retard de groupe. Le premier ensemble de méthodes est justifié en raison du système de production de la voix (modèle source/filtre). Le deuxième ensemble de méthodes est justifié en raison des caractéristiques «phase minimale» du signal glottal et plus généralement en raison de la propriété de localisation temporelle du retard de groupe. La méthode que nous proposons fait partie du deuxième ensemble. Nous comparons les différentes méthodes sur des signaux de voix parlée, chantée et d'instruments de musique. Nous introduisons une caractéristique supplémentaire afin de mesurer l'importance des singularités et donc l'applicabilité des algorithmes de type PSOLA.

La présence de singularités non-périodiques dans le signal est interprétée ici comme correspondant aux transitoires (plosives dans le cas de la voix) du signal. De la même manière, leur détection peut s'effectuer par rupture de modèle (modèle de l'enveloppe spectrale) ou sur base de la concentration d'énergie du signal.

Chapitre 4 Ce chapitre est consacré à la détection des **composantes sinusoïdales** et à l'estimation de leurs paramètres. Les deux problèmes sont étroitement liés par la considération d'un modèle sinusoïdal donné (modèle stationnaire ou non, prise en compte du bruit, ...).

Nous étudions les estimateurs des paramètres d'un modèle sinusoïdal les plus communément utilisés : estimateurs morphologiques, estimateurs des moindres carrés dans le cas de modèles sinusoïdaux. Nous introduisons un modèle sinusoïdal non-stationnaire permettant l'estimation des paramètres de fréquence et d'amplitude à l'ordre 1. Ce modèle est comparé aux estimateurs sinusoïdaux communément utilisés sur un ensemble de signaux tests. Les biais, variance et erreur quadratique moyenne de chaque estimateur sont calculés pour différents paramètres d'analyse et différents types de signaux (échelle du spectre, résolution spectrale, finesse spectrale, présence de bruit,

signaux non-stationnaires).
Nous étudions ensuite le problème de la détection des sinusoïdes dans un signal. La détection est étudiée sous l'angle de l'erreur de représentativité et sous l'angle de l'erreur de spécification d'un modèle estimé par rapport à un signal. Nous introduisons un nouvel algorithme de recherche de sinusoïdes dans un signal, basé sur des critères de continuité temporelle de fréquence, amplitude et phase.

Chapitre 5 Ce chapitre est consacré à l'estimation de la **période/fréquence fondamentale** d'un signal et de son caractère périodique/harmonique. Cette estimation peut être vue selon l'angle d'un traitement aval de la détection de singularités dans lequel il s'agit de dériver des propriétés de périodicité de la répétition des formes d'onde élémentaires ; elle peut également être vue sous l'angle de l'estimation sinusoïdale dans lequel il s'agit de dériver des propriétés de rapports harmoniques entre les composantes sinusoïdales estimées. Nous étudions également l'estimation d'un coefficient de voisement et d'inharmonicité dans le plan temps/fréquence.

Chapitre 6 Ce chapitre est consacré aux **marquages des singularités périodiques**. Cette partie est reliée aux algorithmes de type PSOLA dans lequel la décomposition du signal s'effectue en formes d'onde élémentaires périodiques. Ces formes d'onde élémentaires doivent répondre simultanément à une contrainte de périodicité et, du fait du fenêtrage, à une contrainte énergétique. Le problème du «marquage» PSOLA est envisagé sous la forme d'un problème de résolution de contraintes. Deux algorithmes sont proposés : une résolution locale par itération et une résolution globale par minimisation d'un critère d'erreur.

Deuxième partie : Modifications du signal

Dans la deuxième partie de cette recherche, nous nous intéressons à la modification du signal correspondant aux deux types de caractères étudiés : décomposition du signal en formes d'onde élémentaires périodiques, décomposition du signal en somme de sinusoïdes.

Chapitre 7 Ce chapitre est consacré à l'étude de l'algorithme de modification du signal PSOLA-WB. Nous détaillons la méthode PSOLA étudiée ainsi que les améliorations pouvant être apportées au niveau de l'interpolation des formes d'onde élémentaires - interpolation temporelle ou fréquentielle -, ou au niveau du type de transformation autorisée - dilatation et transposition spectrale -, ou enfin au niveau de la modification des sons non périodiques.

Chapitre 8 Ce chapitre est consacré à l'étude des modifications du signal en synthèse par addition de sinusoïdes. Les améliorations de la synthèse sinusoïdale visent à permettre un respect de l'enveloppe spectrale du signal lors de transpositions et à préserver les relations de phase du signal lors de modifications temporelles. En particulier, nous proposons un algorithme fondé sur le retard de groupe relatif. Cet algorithme est comparé aux méthodes de séparation source/filtre, d'analyse/synthèse synchrone et de retard de phase relatif.

Chapitre 9 Ce chapitre est consacré à la mise au point d'une méthode de modification du signal hybride utilisant les avantages de la méthode PSOLA et de la synthèse par addition de sinusoïdes. En fonction des caractéristiques locales de chaque région du signal, la méthode la plus adaptée est utilisée.

Dans le dernier chapitre, nous revenons sur notre thèse et vérifions sa validité au vue des résultats de notre recherche. Nous dégageons également en certain nombre de perspective de

continuité de cette recherche.

Chapitre 2

Présentation des modèles étudiés

2.1 Méthode PSOLA

Depuis 20 ans, de nombreuses méthodes de modification du signal, reposant sur le principe de superposition/addition temporelle ont été proposées [1]. Parmi les plus importantes, citons les méthodes TDHS (Time Domain Harmonic Scaling) [Mal79], LSEE-MSTFTM [GL84], SOLA (Synchronized Overlap-Add) [RW85], WSOLA (Waveform Similarity Overlap-Add) [VR93]). Les modifications du signal permises par ces méthodes sont essentiellement des modifications de l'axe temporel du signal (compression/dilatation temporelle du signal). Pour cela, ces méthodes procèdent de manière tantôt proportionnelle tantôt synchrone à la période fondamentale, tantôt à l'analyse tantôt à la synthèse.

La méthode PSOLA se distingue de ces méthodes par une synchronie à la période fondamentale tant à l'analyse qu'à la synthèse. Ceci permet, à l'inverse des méthodes précédentes, un contrôle à la fois du déroulement de l'axe temporel et de la hauteur du signal.

	Analyse	Synthèse
TDHS	proportionnalité	proportionnalité
LSEE-MSTFTM	-	reconstruction itérative de la phase
SOLA	-	synchronie (par auto-corrélation) avec le signal déjà synthétisé
WSOLA	synchronie (par auto-corrélation) avec le signal déjà utilisé	-
PSOLA	synchronie	synchronie

Nous renvoyons le lecteur intéressé par une comparaison entre la méthode PSOLA et les différentes méthodes de superposition/addition à [MV95] et [Ver98].

La méthode de superposition/addition synchrone à la période fondamentale, PSOLA (Pitch Synchronous Overlap-Add), [CS86] [Cha88] [CM89] [MC90] repose sur une décomposition d'un signal en une série de formes d'onde élémentaires . Ces formes d'onde

[1] Le vocodeur de phase n´est pas considéré ici comme une méthode de superposition/addition temporelle, puisqu'impliquant un traitement fréquentiel.

élémentaires sont obtenues par un fenêtrage exactement centré sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition/addition (Overlap-Add) de ces formes d'onde élémentaires . La modification de la distance relative entre deux formes d'onde élémentaires , ainsi que la modification du nombre de formes d'onde élémentaires , permet de modifier la hauteur et l'axe temporel du signal.

Selon le nombre μ de périodes fondamentales renfermées par une forme d'onde élémentaire , nous parlerons de

- méthode PSOLA à bande large : $\mu = 2$
- méthode PSOLA à bande étroite : $\mu = 4$

La décomposition du signal en formes d'onde élémentaires est effectuée par multiplication du signal $s(t)$ par une fenêtre de pondération $h(t)$ centrée en des temps m_i appelés «marques de lecture». Ces marques de lecture sont positionnées de manière synchrone à la période fondamentale locale du signal. Soient $s_i(t)$ la $i^{\text{ème}}$ forme d'onde élémentaire , et m_i la $i^{\text{ème}}$ marque de lecture :

$$s_i(t) = h_i(t - \mathrm{m}_i)s(t) \tag{2.1}$$

En notant $h(t)$ la fonction de pondération de longueur normalisée à l'unité, la fenêtre servant au découpage du signal en formes d'onde élémentaires s'exprime :

$$h_i(t) = h\left(\frac{t}{\mu T_0(\mathrm{m}_i)}\right) \tag{2.2}$$

dans lequel $T_0(\mathrm{m}_i)$ désigne la période fondamentale autour du temps m_i.

2.1.1 Méthode PSOLA à bande large (PSOLA-WB)

Dans la méthode PSOLA à bande large (PSOLA-WB Wide-Band), chaque forme d'onde élémentaire renferme deux périodes fondamentales ($\mu = 2$).

Du fait de la largeur fréquentielle de la fenêtre de découpage, le spectre de chaque forme d'onde élémentaire $s_i(t)$ de PSOLA-WB peut être considéré comme une approximation de l'enveloppe spectrale (enveloppe spectrale convoluée par la réponse fréquentielle de la fenêtre d'analyse) [MC90]. Cette approximation provoque cependant un étalement des formants (lissage des résonances étroites), même si celui-ci est difficilement perceptible [MC90]. La qualité de cette approximation dépend cependant fortement du positionnement de la fenêtre par rapport au signal ainsi que de la nature du signal (dans le cas d'un modèle source/filtre, l'approximation dépend de la durée effective de la réponse impulsionnelle du filtre, i.e. du facteur d'atténuation, par rapport à la période fondamentale). Cette interprétation des formes d'onde élémentaires de PSOLA-WB en termes d'enveloppe spectrale est illustrée à la Fig. 2.1, où le spectre d'amplitude d'une forme d'onde élémentaire PSOLA-WB est comparé à la réponse fréquentielle du filtre LP du signal.

Cette interprétation des formes d'onde élémentaires de PSOLA-WB en termes d'enveloppe spectrale permet une analogie entre PSOLA-WB et les décompositions du signal sous forme de source/filtre ou encore une analogie avec la synthèse dite par Forme d'Onde Formantique (FOF) (synthèse de type CHANT [dR89], [d'A89]). [2] Les formes d'onde élémentaires de

[2] Dans la synthèse dite par FOF, le signal est obtenu par convolution d'un signal $\tilde{s}(t)$ et d'un train d'impulsions dont la période détermine celle du signal. Le signal $\tilde{s}(t)$ est la somme d'un ensemble de signaux $\tilde{s}_{m,f_c}(t)$, appelés FOFs, représentant chacun la réponse impulsionnelle d'une résonance/anti-résonance (formant/anti-

PSOLA-WB peuvent s'apparenter à la réponse impulsionnelle du filtre d'une décomposition source/filtre ou encore aux FOF de la synthèse CHANT.

Modification du signal

Dans un modèle source/filtre simplifié, dans lequel la source serait représentée par un train d'impulsions périodiques $e(t) = \sum_m \delta(t - mT_0)$ et le filtre serait représenté à chaque instant par sa réponse impulsionnelle $\tilde{s}(t)$, une modification de hauteur correspond à la modification de la périodicité de la source $e(t)$, i.e. une modification de l'inter-distance entre les impulsions. En PSOLA-WB, la décomposition est effectuée de manière implicite : une modification de hauteur est obtenue par modification de l'inter-distance entre formes d'onde élémentaires considérées comme l'approximation des réponses impulsionnelles.

Interprétation fréquentielle

Dans le cas d'un signal stable périodique, une interprétation fréquentielle du changement de hauteur est également possible. En effet, dans ce cas nous pouvons montrer que le spectre du signal de périodicité en 1/f0 correspond à l'échantillonnage aux fréquences multiples hf0($h \in \mathbb{N}$) du spectre d'une forme d'onde élémentaire X_{T_0} de durée T_0.

$$x(t) = \sum_k \frac{1}{T_0} X_{T_0}(h\omega_0) \exp(jh\omega_0 t) \tag{2.3}$$

Une modification de la hauteur du signal correspond à un ré-échantillonnage de X_{T_0} en de nouvelles fréquences $hf(h \in \mathbb{N})$. De ce fait la méthode PSOLA-WB est considérée comme conservant l'enveloppe spectrale («shape invariant»).

Ceci est illustré aux FIG. 2.3 et FIG. 2.4 dans le cas d'un facteur de transposition 1.5 (f0$'$ = 1.5f0). La FIG. 2.3 représente les spectres d'amplitude et de phase du signal original pour une analyse à bande large ($\mu = 2$) et à bande étroite ($\mu = 4$). La FIG. 2.4 représente les mêmes données sur le signal transposé par la méthode PSOLA-WB. Nous voyons que l'enveloppe spectrale du signal transposé possède une enveloppe spectrale proche de celle du signal original.

L'étude approfondie de l'interprétation spectrale, et en particulier l'étude de l'approximation de la réponse du filtre du modèle source/filtre par une forme d'onde élémentaire PSOLA, est faite dans [MC90] et [Mou90].

2.1.2 Méthode PSOLA à bande étroite (PSOLA-NB)

Dans la méthode PSOLA à bande étroite (PSOLA-NB Narrow-Band), chaque forme d'onde élémentaire renferme quatre périodes fondamentales ($\mu = 4$).

Du fait de la durée de la fenêtre de découpage, le spectre de chaque forme d'onde élémentaire $s_i(t)$ présente une structure fine dans laquelle les harmoniques sont résolues. Ceci est illustré à la FIG. 2.2, où le spectre d'amplitude d'une forme d'onde élémentaire PSOLA-NB est comparé à la réponse fréquentielle du filtre LP du signal.

Du fait de la résolution des harmoniques, le PSOLA-NB ne permet pas d'interprétation directe en termes d'approximation de l'enveloppe spectrale, ni en termes de décomposition

formant dans le cas de la voix) d'un filtre à la fréquence f_c. La somme $\tilde{s}(t)$ représente la réponse impulsionnelle d'un filtre ARMA. Par analogie avec la décomposition source/filtre des signaux, la source est ici représentée sous sa forme la plus simple : un train périodique d'impulsions, tandis que le filtre est représenté par sa réponse impulsionnelle.

source/filtre.

Modification du signal

Une modification de hauteur du signal en PSOLA-NB ne peut plus se satisfaire d'une modification de la distance entre deux formes d'onde élémentaires successives. Ceci se comprend en constatant que le ré-échantillonnage du spectre aux multiples d'une fréquence $f \neq$ f0 ne constitue plus un ré-échantillonnage de l'enveloppe spectrale mais un ré-échantillonnage du spectre harmonique [3].

La modification de hauteur en PSOLA-NB nécessite une correction conjointe du spectre d'amplitude et du spectre de phase des formes d'onde élémentaires [Cha88].

Ceci est illustré aux FIG. 2.5 et FIG. 2.6 dans le cas d'une transposition à la fréquence f = 1.5f0. La FIG. 2.5 représente les spectres d'amplitude et de phase du signal à court terme pour une analyse à bande large ($\mu = 2$) et à bande étroite ($\mu = 4$). La FIG. 2.6 représente les mêmes données sur le signal transposé par PSOLA-NB sans correction du spectre. Le spectre du signal transposé présente des distorsions importantes.

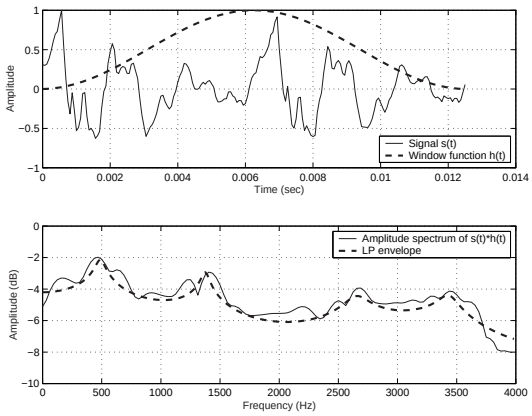


FIG. 2.1 – PSOLA à bande large $\mu = 2$, [H] signal (-) et fenêtrage $h_i(t - \mathrm{m}_i)$ (- -) , [B] spectre d'amplitude de $s_i(t)$ et estimation de l'enveloppe spectrale (- -).

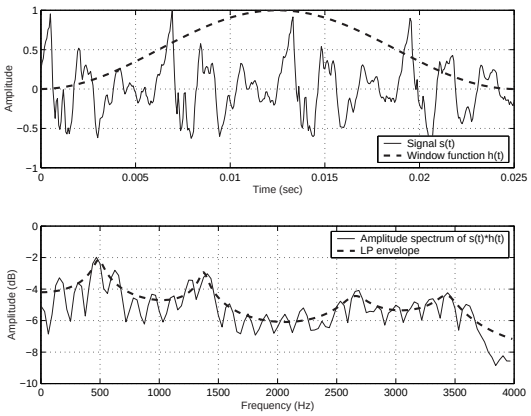


FIG. 2.2 – PSOLA à bande étroite $\mu = 4$, [H] signal (-) et fenêtrage $h_i(t - \mathrm{m}_i)$ (- -) , [B] spectre d'amplitude de $s_i(t)$ et estimation de l'enveloppe spectrale (- -).

2.1.3 Comparaison de PSOLA-WB et PSOLA-NB

La méthode PSOLA est limitée au signaux non-mixtes (une seule source), monophoniques et périodiques.

De plus, la méthode PSOLA-WB, du fait de son fenêtrage étroit dans le domaine temporel, est limitée aux signaux dont la décroissance temporelle de la RI du filtre est rapide face à la période fondamentale. La méthode PSOLA-NB n'est pas limitée par cette contrainte du fait d'un fenêtrage temporel plus large. Une manière de pallier cette contrainte de PSOLA-WB (qui pourrait également s'expliquer en terme de lissage des formants) a été proposée à travers

[3] Pour $f = (2k+1)\frac{\mathrm{f0}}{2}$, le ré-échantillonnage correspond aux creux d'énergie du spectre initial.

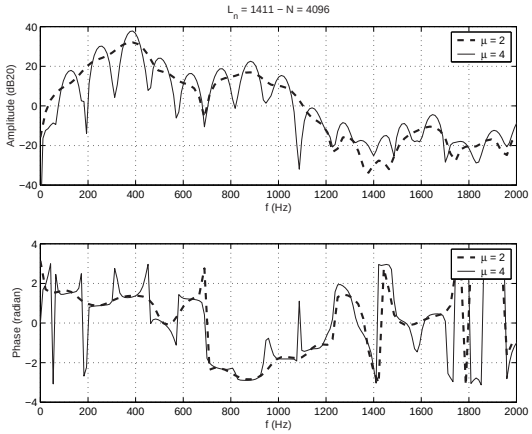


FIG. 2.3 – Spectre à bande large (- -) et à bande étroite (-) du signal original [H] spectre d'amplitude [B] spectre de phase.

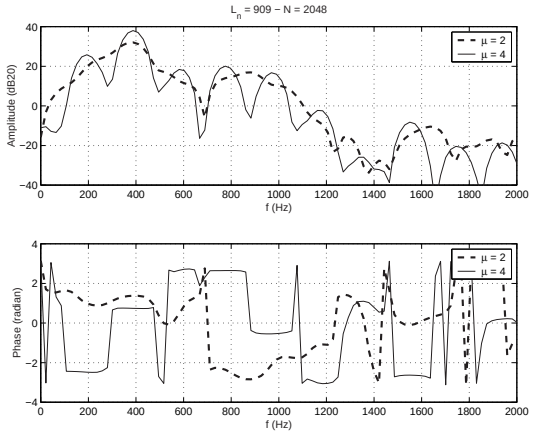


FIG. 2.4 – Spectre à bande large du signal original (- -) et à bande étroite du signal transposé d'un facteur 1.5 vers le haut par PSOLA-WB (-) [H] spectre d'amplitude [B] spectre de phase.

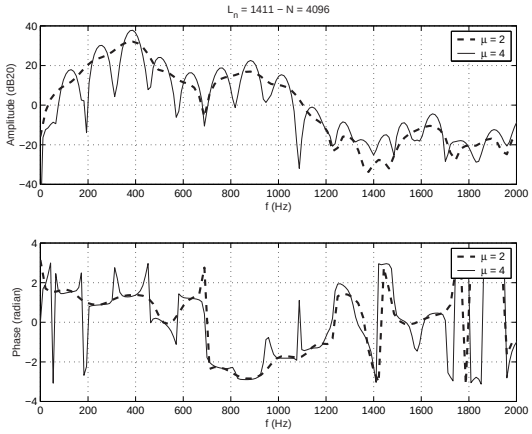


FIG. 2.5 – Spectre à bande large (- -) et à bande étroite (-) du signal original [H] spectre d'amplitude [B] spectre de phase.

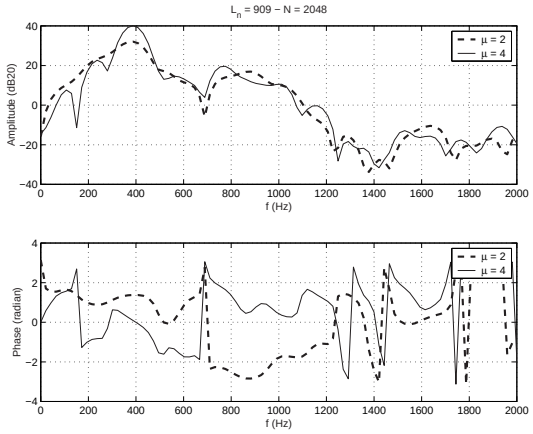


FIG. 2.6 – Spectre à bande large du signal original (- -) et à bande étroite du signal transposé d'un facteur 1.5 vers le haut par PSOLA-NB(-) [H] spectre d'amplitude [B] spectre de phase.

l'utilisation jointe de PSOLA-WB et des méthodes de prédiction linéaire. La méthode LP-PSOLA [MC90] consiste à appliquer le fenêtrage PSOLA-WB sur le signal résiduel (signal d'erreur de prédiction linéaire). Une autre contrainte de PSOLA-WB est de positionner les marqueurs m_i non seulement synchrones à la période fondamentale, mais également proches du maximum local d'énergie à l'intérieur de chaque période. Seule la contrainte de périodicité est nécessitée par la méthode PSOLA-NB.

La méthode PSOLA-NB, ne donne pas lieu à une interprétation en termes de décomposition source/filtre. De ce fait, une modification de la hauteur du signal nécessite une correction des spectres des formes d'onde élémentaires . Cette correction introduit un certain nombre de nouveaux problèmes tels que la transposition des caractéristiques harmoniques/non-harmoniques, la transposition des détails fins du spectre, des relations de phase dans de nouvelles bandes de fréquence, ou l'apparition de trous dans le spectre. Des solutions ont cependant été proposées [MC90] pour ces problèmes. Un autre problème inhérent au fenêtrage large ($\mu = 4$) de PSOLA-NB est l'apparition d'une certaine réverbération dans le signal.

Avantage et Inconvénients des méthodes PSOLA à bande large et étroite :

	Avantages	Inconvénients
PSOLA bande large	coût de calcul faible, signal peu réverbérant	limité à un certain type de signal (atténuation rapide du filtre), nécessité d'un marquage à la fois synchrone à la période fondamentale et contraint par l'énergie locale
PSOLA bande étroite	ne nécessite pas de formes d'onde localisées temporellement	signal plus réverbérant nécessité d'une correction du spectre lors d'une modification de hauteur

2.1.4 Traitement PSOLA sur le signal ou sur le signal résiduel

Le traitement PSOLA sur le signal résiduel (résiduel de prédiction linéaire) a été proposé par [MC90] [ML95] [4]. Le filtre de prédiction linéaire est calculé de manière synchrone à la période fondamentale (autour de chaque marque m_i). Le signal résiduel est calculé par filtrage inverse et l'algorithme PSOLA appliqué au signal résiduel. Le traitement sur le signal résiduel (lorsque celui-ci peut être estimé) présente plusieurs avantages :

- Dans le cas du PSOLA-WB, la distorsion engendrée par le fenêtrage est moindre sur le signal résiduel que sur le signal. Ceci se comprend aisément en considérant la structure impulsionnelle du signal résiduel [5]. Ceci se comprend également en considérant l'interprétation des formes d'onde élémentaires PSOLA-WB en termes d'estimation implicite de l'enveloppe spectrale. Dans le cas de LP-PSOLA, l'approximation de l'enveloppe spectrale est explicite et plus précise. Le traitement sur le signal résiduel permet également de simplifier le problème du positionnement des marqueurs m_i, puisque l'influence des variations d'énergie locale du signal est minimisée par la soustraction de la contribution du filtre.

[4] Des variantes de la méthode LP-PSOLA peuvent être trouvées dans [EL96] REMC-PSOLA (Residual Excited Mel Cepstral PSOLA) [TP99]

[5] Dans le cas du signal vocal, le signal résiduel constitue une approximation de la dérivée du signal glottal

- Dans le cas du PSOLA-NB, un changement de hauteur nécessite une correction du spectre obtenue par un algorithme de compression/dilatation du spectre ou encore par un algorithme de répétition/élimination des harmoniques. Ces traitements ne permettent pas de respecter l'enveloppe spectrale du signal. L'application des algorithmes à un spectre à priori «blanchi» permet de conserver l'enveloppe spectrale.

2.1.5 Choix d'une méthode PSOLA

La méthode PSOLA considérée dans le cadre de cette recherche est une méthode de type PSOLA-WB (appliquée directement sur le signal ou sur le signal résiduel en fonction des propriétés source/filtre du signal) avec traitements dans le domaine fréquentiel. Ce choix a été fait pour les raisons suivantes :

- La méthode PSOLA-NB nécessite un traitement dans le domaine fréquentiel qui l'apparente à l'analyse/synthèse sinusoïdale sans en avoir toute la puissance de modification. La méthode PSOLA-NB, de par son utilisation de fenêtres de durée importante, ne possède pas les avantages de manipulation «microscopique» du son de la méthode PSOLA-WB.
- La méthode PSOLA-WB constitue un complément idéal pour les formes d'onde de caractère pseudo-périodique non modélisable par l'analyse/synthèse sinusoïdale

Remarque à propos de nomenclature: *Les méthodes nommées PSOLA-WB et PSOLA-NB sont généralement désignées dans la littérature par les acronymes TD-PSOLA et FD-PSOLA. Nous avons préféré les dénominations Wide-Band et Narrow-Band, puisque l'utilisation d'un traitement temporel ou fréquentiel est a priori indépendante des conditions d'analyse à bande large ou à bande étroite.*

2.2 Modélisation par addition de sinusoïdes

La synthèse dite «additive» consiste à créer un signal complexe par superposition de formes d'onde simples. Ainsi, dans un orgue, la superposition des registres permet la création de timbres complexes. L'utilisation en électronique de la synthèse additive remonte aux premiers orgues Hammond et au premier synthétiseur par oscillateurs (Telharmonium, 1906).

Un cas particulier de la synthèse additive utilise comme formes d'onde des sinusoïdes. Dans ce cas, les formes d'onde sont orthogonales et peuvent être reliées d'un point de vue perceptif à la hauteur d'un son [HE75]. L'analyse sinusoïdale découle directement de la modélisation implicite à l'analyse de Fourier : modéliser le signal par une somme de sinusoïdes. En ce sens la méthode d'analyse existe depuis les travaux (1822) de Fourier (1768-1830). L'application des techniques de Fourier à l'analyse des signaux a dû attendre le développement des algorithmes de Transformée de Fourier Rapide (FFT) ainsi que l'apparition de calculateurs rapides.

Les premières méthodes d'analyse/synthèse sinusoïdale remontent à [Moo78]. L'estimation est alors effectuée par filtrage hétérodyne. Les premiers modèles d'analyse/synthèse reposant sur la Transformée de Fourier à Court Terme (TFCT) [AR77] [Por80] ont été proposés, quasi-simultanément par [MA86] et [MQ86b] dans le cadre du codage bas-débit de la parole. Le modèle est étendu plus tard aux sons non nécessairement harmoniques, et on lui adjoint une modélisation du bruit [SS90].

Dans la Transformée de Fourier à Court Terme (TFCT), chaque trame du signal est modélisée par une somme de sinusoïdes en nombre égal à celui des points de la TFCT. Dans le modèle sinusoïdal, seules les composantes répondant à certains critères sont modélisées. Ces

critères reposent le plus souvent sur l'observation de l'énergie des composantes, de relations harmoniques entre les fréquences ou encore de régularités temporelles des paramètres.

La formulation proposée par [MQ86b], dans le cadre de la parole, repose sur la contribution de la source glottale $e(t)$ et du filtre $v(t)$ modélisant les propriétés du conduit vocal.

$$\begin{aligned} s(t) &= e(t) \otimes v(t) \\ S(\omega) &= E(\omega) \cdot V(\omega) \end{aligned} \tag{2.4}$$

Dans ce cas le modèle s'exprime :

$$s(t) = \sum_{h=1}^{H(t)} E_h(t) V_h(t) \cos\left(\phi_{E_h}(t) + \phi_{V_h}(t)\right) \tag{2.5}$$

Sous sa forme la plus générale, le modèle sinusoïdal s'exprime :

$$\boxed{s(t) = \sum_{h=1}^{H(t)} A_h(t) \cos\left(\phi_h(t)\right)} \tag{2.6}$$

dans lequel $H(t)$ désigne le nombre de composantes du modèle à l'instant t donné, $A_h(t)$ l'amplitude de la $h^{\text{ème}}$ composante à l'instant t et $\phi_h(t)$ sa phase.

Afin de rendre son estimation plus aisée, deux hypothèses sont généralement faites :

1. les sinusoïdes à un instant donné, $H(t)$, sont en nombre limité,
2. le signal est supposé évoluer lentement dans le temps et donc les paramètres du modèle correspondant sont supposés varier lentement.

Cette deuxième hypothèse permet de sous-échantillonner l'estimation temporelle en un certain nombre d'instants discrets t_m :

$$\begin{aligned} \tilde{s}_m(t) &= \sum_{h=1}^{H(t)} A_h(t_m) cos((t - t_m)\omega_h(t_m) + \phi_h(t_m)) \text{ si } t \in [t_m - L/2, t_m + L/2] \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{sinon} \end{aligned} \tag{2.7}$$

dans lequel L est défini comme la largeur temporelle de l'observation.

Dans ce cas, le signal approximé par le modèle s'exprime :

$$\tilde{s}(t) = \sum_{m} \tilde{s}_m(t) \tag{2.8}$$

Une autre hypothèse souvent utilisée est celle de l'harmonicité des composantes fréquentielles ($\omega_h = h\omega_0$). Cette hypothèse permet de simplifier l'étape d'analyse, puisqu'elle se réduit alors à l'estimation de l'amplitude (voire l'enveloppe spectrale du filtre $v(t)$) et de la phase des composantes harmoniques du signal. Dans le domaine des sciences cognitives, il est souvent admis que l'oreille humaine est insensible aux relations de phase entre les différentes composantes fréquentielles du signal, du moins dans les parties stables (stationnaires) du signal. Dans ce cas, l'estimation des phases n'est pas nécessaire ; les phases seront construites à la re-synthèse par intégration des fréquences. L'analyse se limite alors à l'estimation de la fréquence fondamentale et des amplitudes (voire de l'enveloppe spectrale du filtre $v(t)$). Ceci

permet de représenter le modèle sinusoïdal à l'aide d'un nombre très réduit de paramètres et explique le succès du modèle sinusoïdal dans le domaine du codage. Ces simplifications se font cependant souvent au détriment de la qualité du signal obtenu.

De nombreuses améliorations et extensions ont été apportées au modèle sinusoïdal. La plus importante est sans doute l'ajout d'une composante stochastique $b(t)$ au modèle, permettant de prendre en compte les parties non harmoniques du signal et ainsi d'améliorer le réalisme de la modélisation obtenue (modèle «Deterministic plus Stochastic Decomposition» [SS90], modèle «Harmonic + Noise» [Sty96]).

$$\tilde{s}(t) = \sum_{l=1}^{L(t)} A_l(t) cos(\phi_l(t)) + b(t) \tag{2.9}$$

2.2.1 Choix d'un modèle sinusoïdal

Le modèle sinusoïdal considéré dans le cadre de cette recherche est un modèle sinusoïdal «général», i.e. sans séparation des contributions source/filtre et sans hypothèse d'harmonicité. Ce choix est effectué de manière à permettre la représentation de la plus grande classe de sons. Dans le cas particulier de signaux harmoniques ou de signaux de type source/filtre, le modèle sera cependant simplifié ou adapté.

Première partie

Caractérisation du signal

Chapitre 3

Détection de singularités dans le signal

3.1 Introduction

Dans cette recherche, nous nous intéressons à deux types de modélisation du signal :

1. une modélisation temporelle du signal permettant sa description en tant que répétition d'une même forme d'onde élémentaire au cours du temps, dont le cycle de répétition détermine la période fondamentale du signal,
2. une modélisation spectrale permettant sa description en tant que superposition de sinusoïdes, dont le rapport des fréquences détermine la fréquence fondamentale du signal.

L'objectif de cette double modélisation est de permettre l'utilisation de la modélisation la plus adaptée aux caractéristiques locales (locales en temps et en fréquence) du signal.

Nous utilisons la modélisation du signal afin d'en permettre des modifications. De ce point de vue, il est important de considérer dés l'étape d'analyse la méthode de modification du signal associée au type de modélisation. La méthode de modification du signal correspondant à la modélisation temporelle est la méthode PSOLA à bande large. Dans la méthode PSOLA-bande-large, le signal est décomposé en une succession de formes d'onde élémentaires de manière à ce que chaque forme d'onde élémentaire renferme une période fondamentale et que la superposition/addition de ces formes d'onde élémentaires reconstitue le signal original. Le signal est décomposé en formes d'onde élémentaires par fenêtrage. Du fait de l'étroitesse de ce fenêtrage, chaque fenêtre doit être centrée à proximité du maximum local de l'énergie du signal local, ceci fin de minimiser les détériorations du signal engendrées par le fenêtrage. Il est évident qu'un tel découpage ne peut être appliqué à tous types de signaux et que seuls les signaux présentant une concentration importante de leur énergie locale peuvent faire l'objet d'une décomposition en formes d'onde élémentaires . Il est donc important de caractériser le signal afin de pouvoir déterminer si une telle décomposition est envisageable.

Ceci explique le fait que, dans cette recherche, nous considérons le problème du marquage des formes d'onde élémentaires périodiques en trois étapes distinctes :

1. détection de singularités dans le signal,
2. détection de la fréquence fondamentale,

3. marquage périodique des formes d'onde élémentaires préalablement caractérisées comme singulières.

Dans ce chapitre, nous nous intéressons à la caractérisation du signal en «singularités». Le terme «singularité» est utilisé ici afin de désigner une concentration temporelle importante de l'énergie du signal local. La caractérisation en «singularités» consiste en la localisation des concentrations importantes de l'énergie du signal local ainsi qu'en la mesure de l'importance de cette concentration.

Dans le cas du signal de parole, la décomposition du signal en forme d'onde élémentaire répond au principe de production du signal vocal [1] . Ceci fait du signal vocal une bonne base pour l'étude des méthodes de détection de singularités.

Nous commençons ce chapitre par une description du système de production de la parole et examinons dans quelle mesure les propriétés de ce signal peuvent être exploitées afin de localiser et de caractériser les singularités présentes. Deux approches sont étudiées :

- la première approche repose sur la modélisation du signal de parole comme un modèle source/filtre ; la localisation des singularités peut s'effectuer dans ce cas sur base de la détection des ruptures du modèle auto-régressif du système ;
- la deuxième approche repose sur les caractéristiques phase minimale du signal de parole ; la localisation des singularités s'effectue par observation du retard de groupe du signal ; dans ce cas également il est possible de tirer bénéfice du modèle source/filtre sous-jacent.

Dans cette recherche, nous n'avons pas considéré les méthodes de «marquage périodique du signal» utilisant les caractéristiques à bande étroite du signal (par exemple les caractéristiques résultant d'une modélisation de type superposition de sinusoïdes [MQ86a], [Sty98], [Cha88], [CM89]) même si des rapprochements entre les approches bande large et bande étroite sont possibles. La justification que nous apportons est que, à partir du moment où une analyse à bande étroite du signal est envisageable (et est fiable compte tenu des propriétés de stationnarité du signal), il est préférable d'utiliser une modélisation spectrale plutôt que temporelle, ceci compte tenu des possibilités de modification plus grandes des modélisations spectrales. Nous n'avons pas non plus considéré les méthodes de détection de singularités du signal reposant sur une modélisation dans le plan temps/échelle [Jeh97], [Td99]. Même si celles-ci s'avèrent prometteuses, nous avons préféré rester dans le cadre de l'analyse de Fourier classique.

Nous étudions également dans ce chapitre, la détection des transitoires définis comme des singularités du signal qui ne résultent pas d'une excitation périodique, soit qu'ils s'agissent de plosives dans le cas de la voix ou encore d'attaques d'instruments de musique.

3.2 Production du signal vocal

Le système de production de la voix parlée peut être modélisé par un système source/filtre. Dans le cas de la parole dite «voisée», le cycle d'ouverture/fermeture de la glotte constitue l'action de la source, tandis que le conduit voccal/nasal ainsi que le rayonnement des lèvres constituent un ensemble de filtres en parallèle et en série. La partie filtre du modèle peut être modélisée par un modèle ARMA suivi d'un filtre différentiateur.

- Le filtre AR représente les résonances du conduit buccal ainsi que les contributions de la glotte et des lèvres alors que le filtre MA représente le couplage existant entre le conduit buccal et le conduit nasal produisant des anti-résonances. Malgré cela, un filtre AR remplace le plus souvent le filtre ARMA. Ceci en raison de la plus grande simplicité de son estimation et d'une erreur faible commise dans la majorité des cas. Dans la suite, l'ensemble conduit buccal/nasal sera représenté par un filtre tous-pôles.
- Le filtre différentiateur représente le rayonnement du signal par les lèvres.

Etant donné le caractère linéaire et lentement variable du filtre du conduit buccal et de celui des lèvres (filtres Linéaire et Invariant en Temps sur une durée courte) , ces filtres peuvent être considérés comme constants par rapport à la période de répétition de la source glottale. Nous pouvons donc permuter les filtres tous-pôles et passe-haut (voir Fig. 3.3).

Soit $u_g(t)$ le signal de débit glottal passant dans le filtre buccal/nasal $V(z)$ (filtre ARMA) et le filtre de rayonnement des lèvres $R(z)$ (filtre différentiateur), soit $p(t)$ la pression telle qu'enregistrée par un microphone. Après inversion des filtres, nous considérons la dérivée du signal de débit glottal $p_g(t)$ passant dans le filtre vocal $V(z)$. Le résiduel de prédiction linéaire (modélisation AR) nous fournit donc le signal dérivé de débit glottal.

L'approximation du filtre du système par un filtre tous-pôles n'est vérifiée que pendant la durée où la glotte est fermée. Pendant la durée où la glotte s'ouvre, les conditions du système changent (couplage des cavités supra- et sub-glottales). L'ouverture de la glotte produit une pression sub-glottale et, sous l'action des forces de Bernoulli, la glotte se referme brusquement. Dans le domaine du traitement de la parole, cet instant est désigné par le terme Instant de Fermeture de la Glotte IFG («Glottal Closure Instant», GCI ou encore «epoch») .

L'objectif de cette recherche n'est pas à proprement parler de détecter les IFGs. Cependant l'IFG coïncide avec une impulsion d'énergie dans le conduit buccal/nasal et sa position est donc fortement corrélée avec la position du maximum d'énergie locale du signal qui constitue dans notre cas la localisation de la forme d'onde élémentaire . Qui plus est, l'IFG se produit de manière périodique et constitue dès lors une localisation de l'énergie périodique, localisation recherchée par les décompositions en formes d'onde élémentaires périodiques (traitement de type PSOLA).

Dans la suite de ce chapitre nous considérons les méthodes de détection d'IFGs selon :

- leur propriété de localisation de l'énergie,
- leur potentiel de détection d'une seule impulsion par période
- leur robustesse dans le cas de signaux bruités et seulement presque-périodiques.

Laryngographe : Afin de valider les résultats des méthodes de détection d'IFGs, nous utiliserons dans la suite de ce chapitre la base de données de mesures laryngographiques du «Centre for Speech Technology Research» de l'Université d'Edinburgh [Bag]. Cette base est constituée de 50 phrases prononcées par un locuteur masculin et 50 phrases par une

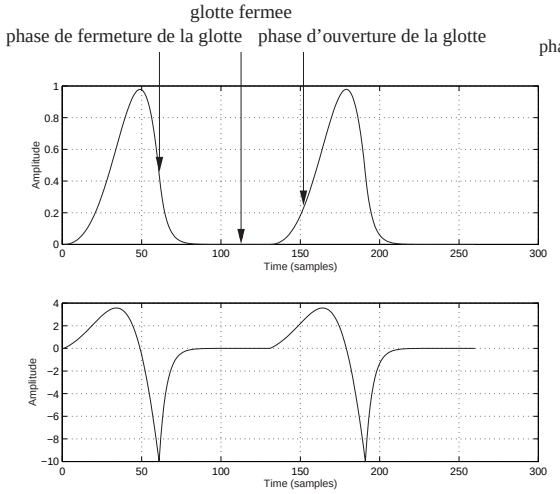


FIG. 3.1 – [H] forme d'onde glottale (modèle de Liljencrants et Fant), [B] dérivée de la forme d'onde glottale

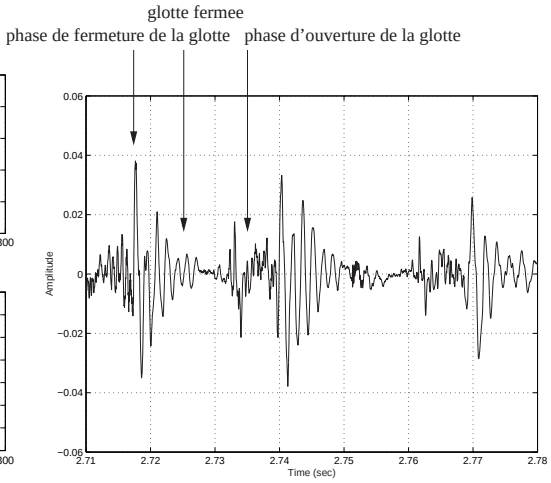


FIG. 3.2 – forme d'onde correspondant à un signal de voix réel (cycle de fermeture/ouverture de la glotte, apparition du bruit) (Signal : voix d'homme, Klaus Maria Brandauer)

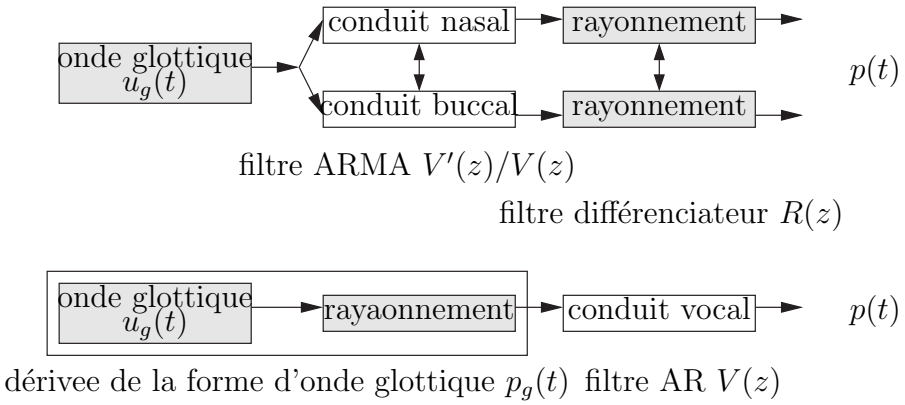


FIG. 3.3 – Système de production du signal voccal

locutrice féminine. Les signaux sont échantillonnés à 20 KHz/12 bits. Chaque phrase a fait l'objet d'une mesure par laryngographie. La laryngographie consiste à mesurer la conductivité électrique du larynx. Celle-ci dépend du coefficient d'ouverture/fermeture de la glotte (circuit électrique ouvert/fermé) (voir FIG. 3.4). Le laryngographe nous fournit une approximation de l'information de fermeture de la glotte.

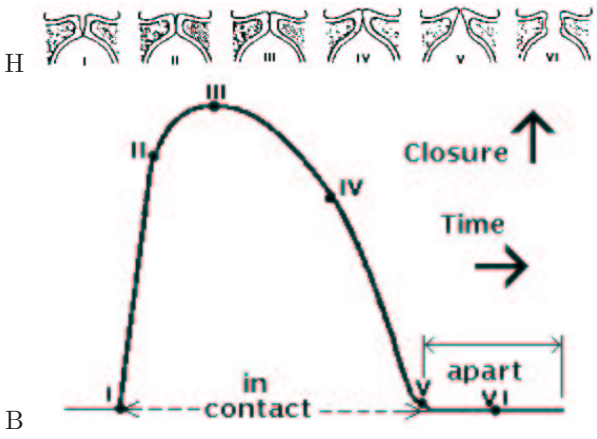


FIG. 3.4 – Cycle d'ouverture/fermeture de la glotte [H] Cycle d'ouverture/fermeture de la glotte en période voisée de parole [B] Signal laryngographique correspondant à ce cycle. (source : `http :// www.linguistics.rdg.ac.uk/ research/ speechlab/ multichannel/ lx/`)

3.3 Détection de singularités par rupture de modèle auto-régressif

La rupture de modèle est une des premières approches envisagées pour la détection des IFGs [Str74]. Le modèle généralement considéré est un modèle AR.

Le signal est considéré comme résultant d'un modèle auto-régressif. Ce modèle auto-régressif n'est en fait vérifié que pendant la durée où la glotte est fermée (voir FIG. 3.1). Dans cet intervalle de temps, l'observation est la réponse libre du système à une impulsion et le modèle AR associé est appelé modèle post-IFG. Suivant cet intervalle, la glotte s'ouvre, introduisant de nouvelles conditions au système. Le modèle auto-régressif associé est le modèle pré-IFG, qui est très difficile à estimer en pratique, du fait du couplage des cavités du système et de bruits de turbulence. La glotte se referme ensuite brusquement, introduisant une discontinuité importante dans le système. Le modèle auto-régressif est considéré comme rompu à cet endroit.

3.3.1 Méthodes existantes

Déterminant de la matrice d'auto-covariance : La rupture d'un modèle auto-régressif est utilisée dans [Str74] pour la localisation des IFGs. Une matrice $\mathbf{S}$ est formée par avancement progressif d'une fenêtre rectangulaire de taille m sur un signal s :

$$\mathbf{S} = \underbrace{\left.\begin{pmatrix} s_{p+1} & s_p & s_{p-1} & \dots & s_1 \\ s_{p+2} & s_{p+1} & s_p & \dots & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+m} & s_{p+m-1} & s_{p+m-2} & \dots & s_m \end{pmatrix}\right\} m}_{p+1} \tag{3.1}$$

Nous supposons $m \geq p+1$, et la matrice de rang p.

En l'absence d'excitation, le modèle auto-régressif d'ordre p impose une dépendance linéaire entre les colonnes de $\mathbf{S}$. Le déterminant de la matrice $\mathbf{S}^T\mathbf{S}$ (matrice d'auto-covariance) théoriquement nul sera donc petit. En présence d'une excitation dans la fenêtre, le déterminant augmentera. [Str74] propose l'utilisation du logarithme du déterminant de la matrice d'auto-covariance comme indication de rupture du modèle.

Erreur de prédiction : [WMG79] propose une méthode proche de celle de [Str74] reposant sur la modélisation LP du conduit vocal par un filtre tous pôles d'ordre p. Soit

$$e_n = s_n + \mathbf{a}_p^T \mathbf{s}_n \tag{3.2}$$

dans lequel s_n désigne l'échantillon n du signal, $\mathbf{s}_n$ le vecteur $(s_{n-1}, s_{n-2}, ..., s_{n-p})^T$ et $\mathbf{a}_p$ le vecteur de coefficients du modèle $(a_1, a_2, ..., a_p)^T$.

Le modèle auto-régressif peut s'écrire pour l'ensemble d'un intervalle $n \in [p+1, p+m]$

$$\mathbf{S} \begin{pmatrix} 1 \\ -a_1 \\ \dots \\ -a_p \end{pmatrix} = \begin{pmatrix} e_{p+1} \\ e_{p+2} \\ \dots \\ e_{p+m} \end{pmatrix} \tag{3.3}$$

La solution des moindres carrés est alors donnée par

$$\mathbf{S}^T\mathbf{S}\begin{pmatrix}1\\-a_1\\\dots\\-a_p\end{pmatrix}=\begin{pmatrix}E_1\\0\\\dots\\0\end{pmatrix} \tag{3.4}$$

dans lequel $E_1=\sum_{i=p+1}^{p+m}e_i^2$ est l'énergie du signal résiduel.
Lorsque la fenêtre d'analyse renferme la fermeture de la glotte, l'énergie du résiduel va croître puis, dés que l'IFG quitte la fenêtre, décroître soudainement. La détection de l'IFG est basée sur la recherche de ce brusque minimum.

Norme de Frobenius : La méthode proposée par [MKW94] consiste à rechercher les maxima de la norme de Frobenius normalisée de $\mathbf{S}$ [2] [3].
La norme de Frobenius normalisée de $\mathbf{S}$ utilisée par [MKW94] s'exprime

$$|\mathbf{S}|_F^2=\frac{1}{p+1}\sum_{i=1}^{m}\sum_{j=1}^{p+1}s_{i,j}^2 \tag{3.7}$$

La justification théorique de son utilisation provient de l'inégalité suivante [MKW94]

$$(\Pi_{i=1}^{p}\sigma_i^2)^{\frac{1}{p+1}}\leq\frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2\leq\frac{\sigma_1^2}{\sigma_{p+1}^2}\frac{1}{p+1}\sum_{i=1}^{p+1}E_i \tag{3.8}$$

Le terme de gauche de l'inégalité est le critère de [Str74], le terme de droite est la moyenne arithmétique de l'énergie des erreurs résiduelles attachées à chaque modèle LLS (Linear Least Square) [4] (modèle de [WMG79]). La norme de Frobenius évolue donc de la même manière que le critère de [Str74] et de [WMG79].
Pour un signal s'étendant de $t=1$ à $t=m+p$, [MKW94] propose d'assigner la valeur de la norme de Frobenius au temps $t=p+1$.

Maximum de vraisemblance : Dans [CO89], il est proposé de modéliser le signal résultant d'un IFG par une ondelette. . L'ondelette $\hat{s}(n)$ utilisée est la réponse impulsionnelle du filtre du modèle AR estimé localement : $\hat{s}(n)=\sum_i a_i\hat{s}(n-i)\ \ \forall n>0$, $\hat{s}(0)=G$. La différence entre le signal autour de n_0 et l'ondelette $x(n)=s(n+n_0)-\hat{s}(n)$ est considérée comme un processus gaussien. De ce fait, la probabilité conditionnelle (ou vraisemblance) de $x(n)$ étant donné le vecteur de paramètre $\underline{a}$, $p(x|G,\underline{a})$ s'écrit comme une loi gaussienne. Le maximum de vraisemblance se produit lorsque les paramètres $\underline{a}$ maximisent $p(x|G,\underline{a})$. Il est montré dans [CO89] que la maximisation de $p(x|G,\underline{a})$ peut se réduire à la maximisation de $\sum s(n+n_0)\hat{s}(n)$ qui est la corrélation entre le signal et l'ondelette. La résolution étant non-linéaire, la solution est obtenue par énumération des valeurs de n_0.
Un traitement aval est appliqué afin d'éviter la détection de pics secondaires à l'intérieur d'une période. Ce traitement repose dans un premier temps sur l'utilisation d'un noyau lissant utilisant la transformée de Hilbert. Dans un second temps, une **fonction structurante** est utilisée : la valeur moyenne locale de la fonction sur un intervalle T est soustraite de la fonction si celle-ci est supérieure à cette moyenne, sinon elle est annulée. Soit $f(x)$ la fonction de soustraction.

$$g(x)=\begin{cases}f(x)-\overline{f(x)} & si \quad f(x)\geq\overline{f_T(x)}\\ 0 & si \quad f(x)<\overline{f_T(x)}\end{cases} \tag{3.9}$$

dans lequel T détermine l'horizon de la moyenne.

Test d'hypothèse : [MF90] propose la détection d'IFGs par tests d'hypothèse. Le signal est considéré comme la concaténation de deux processus AR gaussiens dont deux réalisations sont observées de part et d'autre de l'endroit d'observation. Le test d'hypothèse décide dans quelle mesure l'hypothèse H0 «les deux réalisations sont issues d'un même processus» est rejetée ou conservée faute de preuves [5] . La décision est prise sur base du rapport de vraisemblance des réalisations en considérant l'hypothèse H0 vraie. Si l'hypothèse H0 est rejetée nous sommes à l'endroit de la rupture du modèle. Une autre méthode est proposée dans [MF90] utilisant la divergence entre deux modèles AR gaussiens ; le premier est calculé sur une fenêtre de taille importante, l'autre sur une fenêtre de taille petite.

3.3.2 Observations

Préliminaires : La détection des IFGs par rupture de modèle auto-régressif est dépendante d'un modèle (le modèle auto-régressif) et de son estimation. Le modèle auto-régressif est, pour la voix, un modèle théorique qui n'est pas toujours vérifié. Pour une classe plus large de signaux musicaux, ce modèle n'est même pas du tout justifié. Son estimation dépend de nombreux paramètres : choix de l'ordre du modèle utilisé, positionnement et taille de l'observation. Son estimation dans le cas de signaux de fréquences élevées est particulièrement problématique. Les méthodes basées sur la comparaison de deux modèles AR présentent de plus la difficulté d'estimer deux modèles AR à l'intérieur de chaque période : un modèle AR pré-IFG (en réponse forcée) et un modèle AR post-IFG (en réponse libre) [YV98].

Parmi l'ensemble des méthodes de détection d'IFGs utilisant la rupture de modèle, nous avons uniquement testé la méthode de la norme de Frobenius.

Normalisation de la norme de Frobenius Malgré son appellation, la méthode de la norme de Frobenius «normalisée» proposée par [MKW94] n'est pas normalisée en énergie. La valeur de (3.7) dépend de l'énergie locale du signal observé. Une normalisation de cette norme par l'énergie locale du signal est envisageable. Soit en notant $x(n)$ le signal, m la taille de la fenêtre, p l'ordre du modèle, et n_0 la position du début de la fenêtre :

$$|\mathbf{S}|_F^2(n_0+p) = \frac{1}{\sum_{i=n_0}^{n_0+p+m-1} x^2(i)} \sum_{i=n_0}^{n_0+m-1} \sum_{j=i}^{i+p} x^2(j) \tag{3.10}$$

Cette normalisation, effectuée par calcul de l'énergie sur une durée courte, introduit cependant des discontinuités dans la fonction norme de Frobenius. Ceci est illustré aux figures Fig. 3.5 et Fig. 3.6. Chaque figure représente, du haut vers le bas, (1) le signal, (2) la norme de Frobenius et (3) le résultat de la fonction structurante $g(n)$ dénommé **AVS** pour "Averaged Value Substracted". Nous avons superposé à cette dernière, en traits légers, la partie positive du signal laryngographique. Les fonctions $|\mathbf{S}|_F^2(n)$ et $g(n)$ sont indiquées sous forme normalisée entre 0 et 1. La figure Fig. 3.5 utilise la norme de Frobenius telle que proposée par [MKW94], la figure Fig. 3.6 utilise la norme de Frobenius normalisée par l'énergie. Du fait des discontinuités introduites par la normalisation en énergie, nous n'avons pas poussé plus en avant son utilisation. Les résultats suivants sont donc indiqués pour la norme de Frobenius telle que proposée par [MKW94].

Test de la méthode de la norme de Frobenius : Les résultats de l'application de la norme de Frobenius «normalisée» sont indiqués pour les signaux suivants :

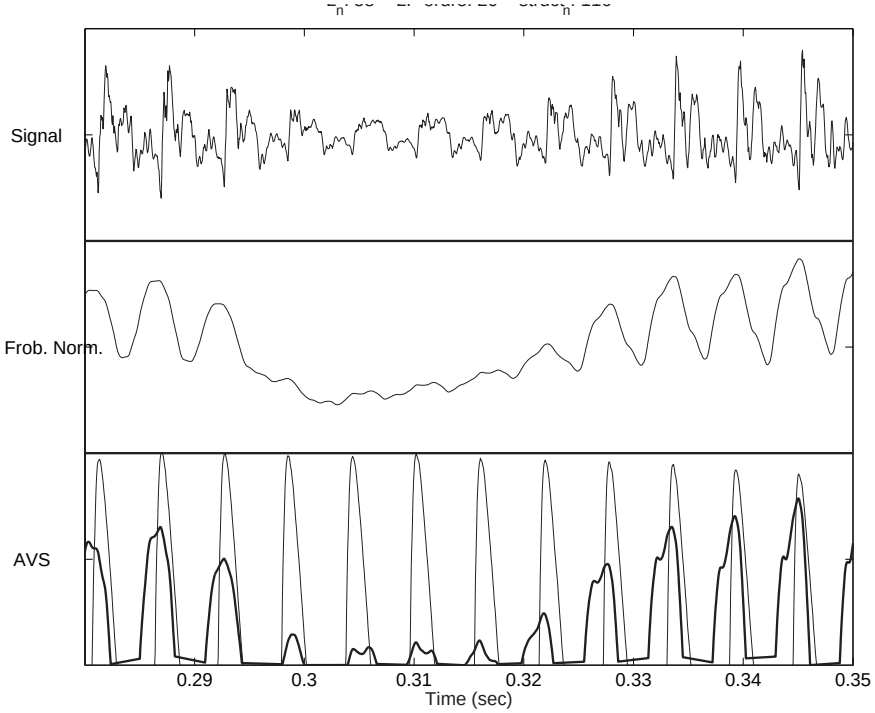


FIG. 3.5 – Méthode de la norme de Frobenius, pas de normalisation en énergie, Signal= rl001-2

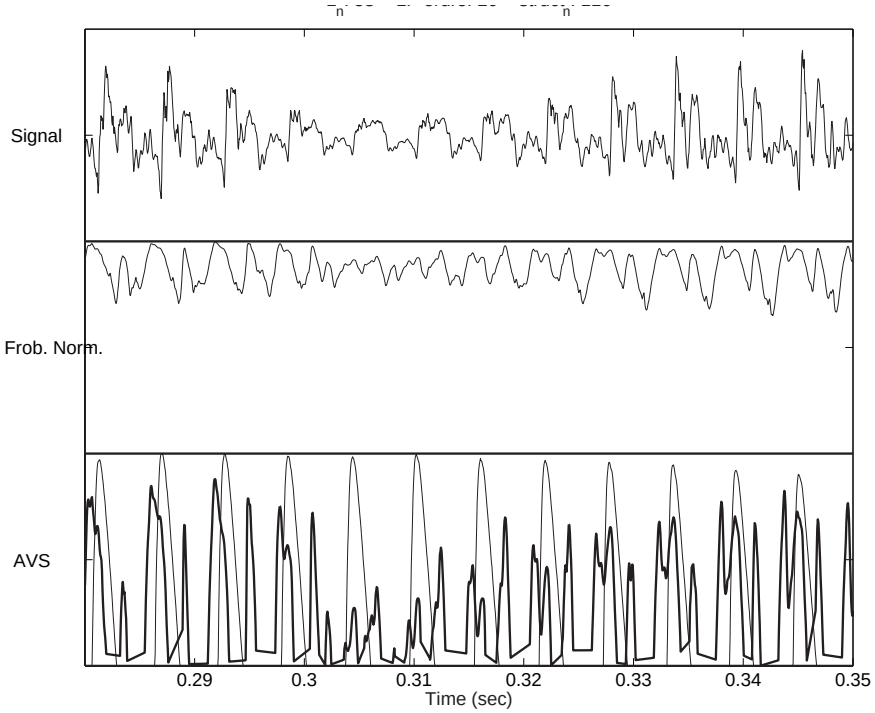


FIG. 3.6 – Méthode de la norme de Frobenius, normalisation en énergie, Signal= rl001-2

Numéro de figure	Description	$\overline{\text{f0}}$	Laryngographe
FIG. 3.17 page 57	voix d'homme	172 Hz	oui
FIG. 3.19 page 58	voix d'homme	138 Hz	oui
FIG. 3.21 page 59	voix d'homme	121 Hz	oui
FIG. 3.23 page 60	voix de femme	317 Hz	oui
FIG. 3.25 page 61	voix de femme	307 Hz	oui
FIG. 3.27 page 62	voix parlée, signal périodique	111 Hz	non
FIG. 3.29 page 63	vocal fry, signal composé de pulses peu périodiques	40 Hz	non
FIG. 3.31 page 64	sinusoïde, premier formant à f0, forme d'onde symétrique	281 Hz	non
FIG. 3.33 page 65	voix chantée de femme	510 Hz	non
FIG. 3.35 page 66	sons de trompette	350 Hz	non
FIG. 3.37 page 67	sons de violon	260 Hz	non
FIG. 3.39 page 68	sons de piano	260 Hz	non

Les cinq premiers signaux sont issus de la base de donnée de l'Université d'Edinburgh [Bag] et autorisent la comparaison des résultats avec le signal laryngographique. Les signaux suivants sont, soit des signaux de parole considérés comme des cas difficiles, soit des signaux d'instrument de musique. Dans ce dernier cas, nous cherchons à tester l'applicabilité de la méthode de la norme de Frobenius en dehors du contexte de la voix.

La taille de la fenêtre est choisie égale à la moitié de la période fondamentale moyenne sur l'ensemble du segment considéré, notée $\overline{\text{T0}}$. La taille choisie pour l'intervalle T pour le calcul de $g(x)$ est de $\overline{\text{T0}}$. L'ordre du modèle auto-régressif est choisi de manière à obtenir un formant (une paire de pôles) par bande de fréquence de 1000Hz, excepté dans le cas de signaux de fréquence fondamentale élevée où l'ordre est choisi de manière à garantir la condition $m \geq p + 1$.

Les critères de qualité choisis sont les suivants :

- pour les signaux accompagnés d'une mesure laryngographique, nous comparons la position des maxima de $g(x)$ (fonction structurante "Averaged Value Substracted") par rapport aux maxima du signal laryngographique.
- pour les signaux non accompagnés d'une mesure laryngographique, nous observons la périodicité des maxima de $g(x)$ ainsi que le nombre de maxima par période.
- nous regardons la variance locale de la fonction $|\mathbf{S}|_F^2(n)$.

Sur cette base de données (réduite), la combinaison de l'utilisation de la norme de Frobenius passée dans une fonction structurante donne de bons résultats, excepté dans le cas de signaux

de fréquence fondamentale élevée [6] (voir FIG. 3.33). Cependant, la variance locale de $|\mathbf{S}|_F^2(n)$ est relativement faible (voir FIG. 3.17 autour du temps 0.31 sec, FIG. 3.19 autour du temps 0.68 sec, FIG. 3.21 autour du temps 0.24 sec, FIG. 3.23 autour du temps 0.585 sec), ce qui rend le résultat obtenu après fonction structurante sensible à la taille T de $g(x)$. D'une manière générale le maximum de $g(x)$ tend à précéder le maximum de la fonction laryngographe.

De manière fort surprenante, les résultats obtenus sur des instruments de musique ne répondant pas à un modèle source/filtre (violon FIG. 3.37 et piano FIG. 3.39) sont acceptables. Ces résultats s'explique en partie par la dépendance envers l'énergie de la norme de Frobenius telle que proposée par [MKW94].

3.4 Détection de singularités par utilisation de l'information du spectre de phase de la transformée de Fourier

3.4.1 Introduction

Afin d'illustrer nos développements dans cette partie, nous considérons le signal constitué d'une sinusoïde amortie. Selon le facteur d'amortissement, ce signal permet la jonction entre le paradigme de la forme d'onde élémentaire et celui de la composante sinusoïdale.

Soit la sinusoïde amortie d'amplitude A, de taux d'amortissement α, de fréquence ω_0 et de phase initiale ϕ_0.

$$x(n) = A\exp(-\alpha n)\cos(\omega_0 n + \phi_0) \tag{3.11}$$

Sa transformée en Z est un filtre AR réel du 2ème ordre

$$X(z) = \frac{A}{2}\left[\frac{\exp(j\phi)}{1 - z_0 z^{-1}} + \frac{\exp(-j\phi)}{1 - z_0^* z^{-1}}\right] \tag{3.12}$$

dans lequel z_0 est un pôle d'amplitude $\exp(-\alpha)$ et de fréquence ω_0.

$$z_0 = \exp(-\alpha)\cdot\exp(j\omega_0) \tag{3.13}$$

$x(n)$ peut être considéré comme une approximation du signal d'un formant du signal vocal (réponse d'un conduit vocal à résonance unique à une excitation glottale impulsionnelle).

$\exp(-\alpha)$ étant le module du pôle dans le plan complexe, α détermine la résonance (le taux de décroissance temporel) ainsi que la largeur de bande en fréquence du système.

- pour une valeur α importante, le pôle se trouve proche de l'origine, la résonance est faible (forme d'onde d´énergie concentrée temporellement) et la bande passante large,
- pour une valeur α tendant vers zéro, le pôle se rapproche du cercle unité, la résonance devient infinie (forme d'onde de type sinusoïde non amortie) et la bande passante nulle.

La position du pôle influence également le comportement local du spectre de phase, et ce d'autant plus que le pôle se rapproche du cercle unité (voir FIG. 3.7 page 33). Le cas limite du pôle sur le cercle unité produit, théoriquement et dans le cas d'un signal s'étendant de $-\infty$ à $+\infty$, une discontinuité locale du spectre de phase (saut de phase de π). Dans le cas pratique d'une analyse à court terme (TFDCT), cette discontinuité donne place à un saut de phase de grandeur déterminée par les paramètres de la fenêtre de pondération utilisée

.

3.4.2 Caractérisation en localisation temporelle

3.4.2.1 Utilisation du retard de groupe pour la localisation temporelle

Retard de groupe : L'observation des variations du spectre de phase se fait généralement au travers du retard de groupe : $\tau_g(\omega) = -\frac{\partial\phi(\omega)}{\partial\omega}$ (voir annexe D). $\tau_g(\omega)$ donne le décalage temporel (par rapport au centre de la fenêtre d'analyse t_m) de chaque composante fréquentielle ω. Nous l'estimons par utilisation du ré-assignement temporel [AF95] (voir annexe G).

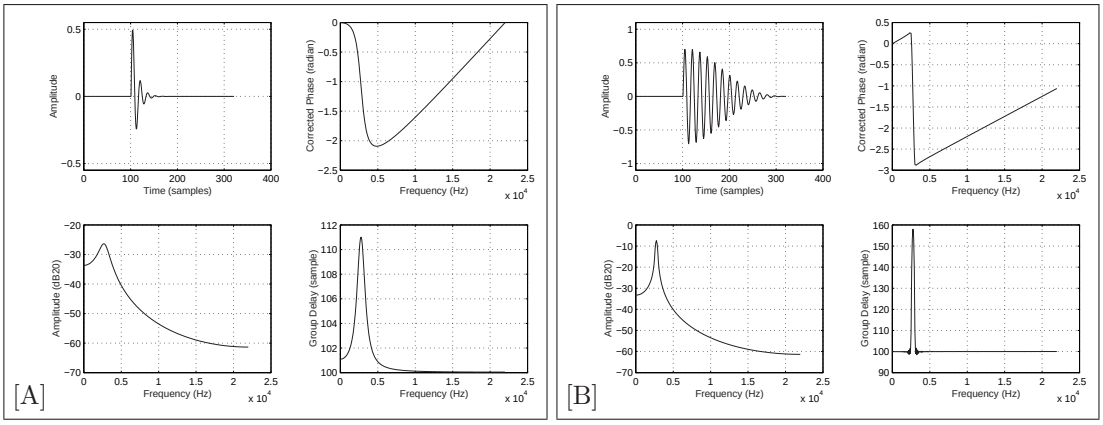


FIG. 3.7 – Sur chaque panneau [A] et [B] : [HD] signal temporel, [BG] spectre d'amplitude, [HD] spectre de phase corrigé par rapport au début du signal, [BD] retard de groupe. Signal : sinusoïde de taux d'amortissement [A] $\alpha = 0.1$, [B] $\alpha = 0.01$

Pour un signal $s(t)$, une fenêtre d'observation $h(t)$ de taille L centrée en un instant t_m, et en utilisant la convention passe-bas (BP) de la TFCT, nous pouvons écrire

$$\tau_g(s, t_m, L, \omega) = \Re\left\{ \frac{STFT^{BP}_{(t-t_m)h}(x, t_m, \omega)}{STFT^{BP}_{h}(x, t_m, \omega)} \right\} \tag{3.14}$$

$\tau_g(s, t_m, L, \omega)$ prend ses valeurs dans l'intervalle $[0, \frac{L}{2}]$. Une valeur de $|\tau_g| > \frac{L}{2}$ consiste à assigner temporellement la composante ω à une position extérieure à la fenêtre d'analyse, c'est-à-dire à une position pour laquelle nous n'avons pas observé le signal. Ces valeurs seront dés lors considérées comme aberrantes et exclues de nos calculs. Soit Ω l'ensemble des valeurs ω_k pour lesquelles $|\tau_g(s, t_m, L, \omega)| < \frac{L}{2}$.

Lissage de la fonction τ_g : En présence de bruit, des oscillations de τ_g sont observées. Diverses techniques de lissage du retard de groupe ont été proposées : moyenne du retard de groupe sur trois signaux transposés $x(n), x(n)e^{-jt\Delta\omega}, e^{jt\Delta\omega}$ [MMY89], filtrage médian du retard de groupe [YV98], pondération du retard de groupe par une fonction dérivée de l'amplitude [YMR91]. La solution retenue ici est le lissage par filtrage médian. La taille du filtre est choisie égale à 4. Une valeur plus faible conduit à l'élimination des τ_g dus à des résonances fortes, alors qu'une grande valeur ne produit pas de lissage significatif.

Localisation temporelle des singularités : Le retard de groupe indiquant la localisation temporelle de l´énergie, il semble intuitif de l'utiliser pour la localisation des singularités. Il semble naturel d'utiliser pour cela sa moyenne fréquentielle

$$\boxed{\mu_{min}(s, t_m, L) = \int_{\omega} \tau_g(s, t_m, L, \omega) d\omega} \tag{3.15}$$

Il semble également naturel de pondérer le retard de groupe par l'énergie locale du spectre, afin de minimiser l'influence de régions fréquentielles dont l'énergie est faible et donc poten-

tiellement perturbées par du bruit.

$$\boxed{\mu_{ener}(s,t_m,L)=\int_w \tau_g(s,t_m,L,\omega)|S(t_m,\omega)|^2 d\omega} \tag{3.16}$$

Les deux formulations (3.15) et (3.16) ont une signification très différente. Comme nous allons le voir dans le paragraphe suivant, (3.15) trouve sa justification à travers la théorie des signaux à phase minimale et peut être rattaché au mode de production de la voix ; à l'inverse (3.16) ne se justifie qu'à travers ses propriétés de localisation de l'énergie.

3.4.2.2 Utilisation des propriétés des signaux à phase minimale pour la localisation temporelle

La séquence dite «à phase minimale» est la séquence qui, parmi l'ensemble des séquences ayant une réponse en amplitude $|S(\omega)|$ donnée, possède la concentration d'énergie la plus importante.

Parmi les propriétés des signaux à phase minimale (voir annexe E pour un résumé de ces propriétés), une propriété intéressante est celle de *«spectre de phase de pente moyenne nulle»*. La valeur $\mu_{min}(s,t_m,L)=0$ de (3.15) peut donc s'interpréter comme le début d'une séquence à phase minimale en t_m.

L'utilisation des propriétés des signaux à phase minimale pour la détection des IFGs part de la constatation que le signal dérivé du débit glottal $p_g(t)$ peut être considéré comme un signal à phase minimale [SY95] (ceci est vrai du moins au-dessus de 600 Hz [7]). Le signal en sortie du système de production vocale $p(t)$ peut donc être considéré comme le résultat de la convolution d'un signal à phase minimale $p_g(t)$ par la réponse impulsionnelle du filtre du conduit vocal $v(t)$.

Dans [SY95] et [YV98], l'estimation des IFGs s'effectue par utilisation des propriétés des signaux à phase minimale. Une analyse à fenêtre glissante est effectuée. La durée de la fenêtre est choisie de manière à observer, à un instant t_m donné, au maximum un seul signal à phase minimale à l'intérieur de la fenêtre. Pour cela, le choix d'une durée inférieure à la période fondamentale locale est effectué. Deux estimateurs sont proposés pour la détection des IFGs :

1. la **pente moyenne du spectre de phase** obtenue par régression linéaire [SY95].
   Nous notons sa valeur $\mu_{pente}(s,t_m,L)$,
2. la **moyenne du retard de groupe** [YV98].
   Nous notons sa valeur $\mu_{min}(s,t_m,L)$.

Ces estimateurs sont calculés à chaque instant t_m. La localisation du début des signaux à phase minimale se fait alors par détection des passages par zéro de la fonction $\mu(s,t_m,L)$. Un passage par zéro de la fonction $\mu(s,t_m)$ à l'instant $t_m=t_0$ correspond au démarrage d'une séquence à phase minimale en t_0, donc à un IFG en t_0.

3.4.2.3 Influence des paramètres du signal et d'analyse

μ_{min} est un estimateur du début t_0 d'une séquence à phase minimale alors que, d'après (D.3), μ_{ener} est un estimateur du centre de gravité t_{ener} de la séquence à phase minimale. Dans le cas d'un signal à phase minimale non-fenêtré, la différence entre μ_{min} et μ_{ener} dépend uniquement de la valeur du taux d'amortissement α.

Dans le cas général d'un signal à phase minimale non-symétrique fenêtré [8], tant l'estimation de t_0 par μ_{min} que l'estimation de t_{ener} par μ_{ener}, et que le décalage de μ_{ener} par rapport à t_0 dépendent

- du taux d'amortissement α,
- de la position du signal t_0 dans la fenêtre d'observation,
- de la longueur $2L$ et du type de la fenêtre d'observation $h(t)$.

Un autre paramètre déterminant la qualité de l'estimation de μ_{min} et μ_{ener} est la présence de bruit additif dans le signal. Dans le cas général, (3.11) se réécrit

$$x(t) = \begin{cases} [A\exp(-\alpha(t-t_0))\cos(\omega_0(t-t_0)+\phi_0)+b(t)]\cdot h_{2L}(t) & \forall t \geq t_0 \\ b(t)\cdot h_{2L}(t) & \forall t < t0 \end{cases} \tag{3.18}$$

dans lequel $b(t)$ est un bruit gaussien additif non nécessairement blanc. Il s'agit donc de comparer

- $\mu_{min}(x,t_m,L) = \int_\omega \tau_g(x,t_m,L,\omega)d\omega$ [9] à t_0
- $\mu_{ener} = \frac{\int_{t=-L}^{L} t|s(t_m,t-t_m)|^2 dt}{\int_{t=-L}^{L}|s(t_m,t-t_m)|^2 dt}$ à t_{ener} [10] .
- μ_{ener} à μ_{min}

μ_{ener} et μ_{min} conduisent l'un comme l'autre à des développements analytiques compliqués [11] ; aussi proposons-nous une étude comportementale.

◊ *Etude comportementale*

Aux FIG. 3.8, nous comparons le comportement des estimateurs μ_{pente}, μ_{min} et μ_{ener} pour différentes conditions de signaux, et différentes conditions d'analyse. Nous étudions l'influence

- du facteur d'amortissement de la sinusoïde [G] $\alpha = 0.05$ [D] $\alpha = 0.01$;
- de la position $t_0 \in [0, L]$ du début du signal dans la fenêtre ; l'abscisse des figures représente cette position ; elle simule les conditions d'une analyse à fenêtre glissante ;
- du rapport signal/ bruit ; pour une valeur de bruit fixée, le SNR dépend de la position du signal dans la fenêtre ; aussi exprimons-nous le niveau de bruit en valeur absolue : pour une amplitude $A = 1$ de la sinusoïde [H] $\sigma_b^2 = 0$, [B] $\sigma_b^2 = 0.1$.

En présence de bruit, les valeurs indiquées sont celles obtenues en moyenne pour un ensemble de 100 réalisations de bruit.

Dans chaque panneau de FIG. 3.8, l'ordonnée représente la moyenne (trait gras) et la moyenne ± l'écart-type (trait léger) de chaque estimateur $y = \mu_{pente}(t_0)$ (-), $y = \mu_{min}(t_0)$ (- -) et $y = \mu_{ener}(t_0)$ (-.). Sont également représentées à titre de référence la valeur idéale $y = x = t_0$ et $y = t_{ener}(t_0)$; il s'agit des diagonales en pointillés clairement visibles sur les FIG. 3.8 [BG] [BD], mais masquées par les estimateurs sur [HG] [HD].

Observations :

- En l'absence de bruit, les biais $\mu_{pente}(t_0) - t_0$ et $\mu_{min}(t_0) - t_0$ sont quasi-nuls et indépendants de la valeur de t_0 (position du début du signal à phase minimale dans la fenêtre), excepté pour les valeurs $t_0 \simeq L$ pour lesquelles le biais $\mu_{pente}(t_0) - t_0$ croît brusquement.
- En présence de bruit, l'hypothèse de «spectre de phase moyenne nulle» n'est vérifiée que pour $t_0 \simeq L/2$, c'est-à-dire là où le SNR est le plus grand et l'influence de la fenêtre minimale. L'intervalle temporel sur lequel le biais de μ_{pente} et μ_{min} est faible

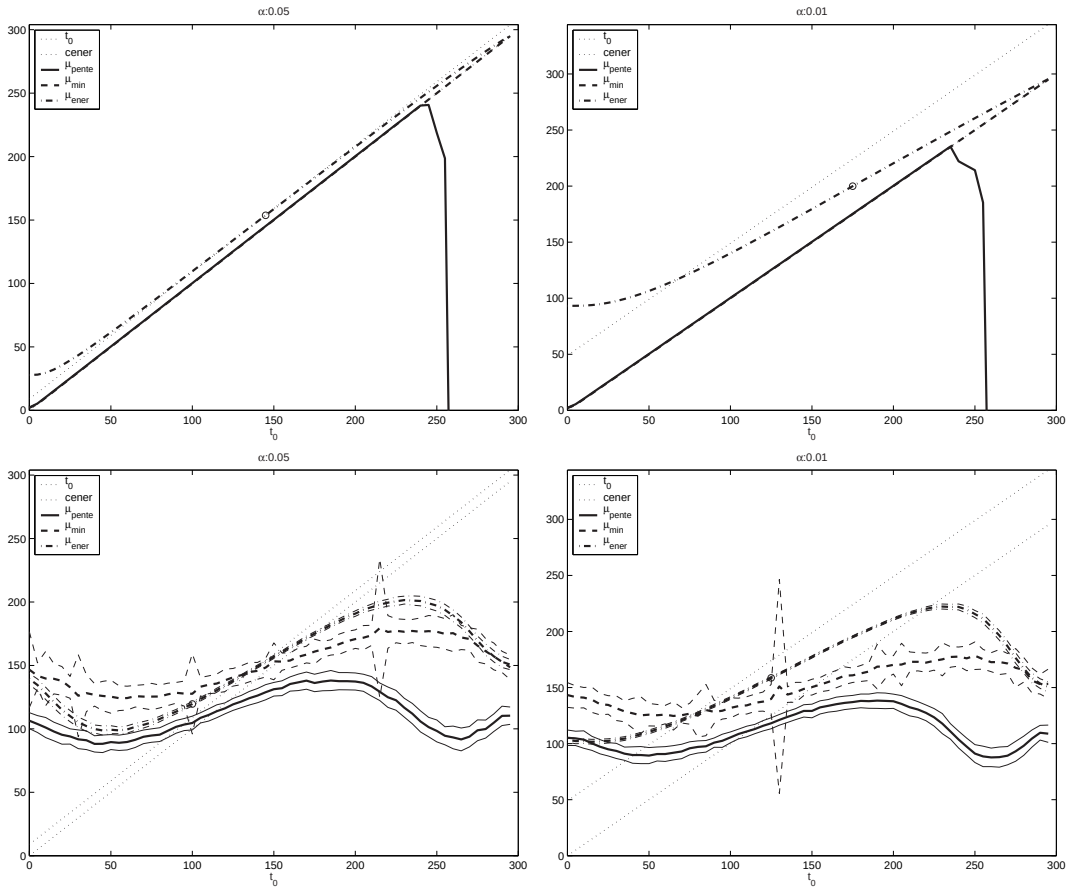


FIG. 3.8 – (ordonnée) Mesure de la moyenne (trait gras) et de la moyenne ± l'écart-type (trait fin) des estimateurs μ_{pente} (—), μ_{min} (- -) et μ_{ener} (- . - .) ; pour différentes positions dans la fenêtre (abscisse) ; pour différents niveaux de bruit : [panneaux du haut] $\sigma_b^2 = 0$, [panneaux du bas] $\sigma_b^2 = 0.1$; pour différents taux d'amortissement de la sinusoïde : [panneaux de gauche], $\alpha = 0.05$ [panneaux de droite] $\alpha = 0.01$; point d'inflexion de μ_{ener} (cercle o sur les courbes)

est petit. Pour les autres valeurs de t_0, l'influence du bruit étant supérieure à celle du signal, le spectre de phase n'est plus de moyenne nulle et les estimateurs μ_{pente} et μ_{min} présente un biais important. La variance des estimateurs μ_{pente} et μ_{min} est également importante.

- En l'absence de bruit, le biais $\mu_{ener}(t_0) - t_0$ dépend du facteur α de la même manière que la différence $t_{ener} - t_0$. Le biais $\mu_{ener}(t_0) - t_0$ est le plus faible pour $t_0 \simeq L$ (ce qui est logique, puisqu'alors le signal observé est plus court). Pour α grand, le biais $\mu_{ener}(t_0) - t_{ener}$ s'annule pour $t_0 \simeq L/2$; c'est là que l'influence de la fenêtre est la plus faible (voir FIG. 3.8 [HG]). Le point d'inflexion de la fonction μ_{ener} (indiqué par un cercle sur la figure) se trouve également à cette position. Pour α petit, le biais $\mu_{ener}(t_0) - t_{ener}$ se décale à gauche de $t_0 = L/2$, et le point d'inflexion à droite de $t_0 = L/2$. Le point d'inflexion correspond à une concentration locale des valeurs de

μ_{ener} qui se situe donc dans l'intervalle $[t_{ener}, t_0]$. La dérivée de μ_{ener} augmente aux extrémités du signal.

- En présence de bruit, (voir FIG. 3.8 [B]), ces propriétés de μ_{ener} restent globalement valables. La variance de μ_{ener} est plus faible que celle de μ_{pente} et μ_{min}.

Conclusion : En l'absence de bruit, le meilleur estimateur de t_0 est μ_{min}. En présence de bruit, μ_{pente} et μ_{min} n'ont un biais faible que sur un intervalle t_0 relativement étroit. A l'inverse le biais de μ_{ener} est plus important mais moins dépendant de t_0 (voir FIG. 3.8 [BG] et [BD]). La variance de l'estimateur μ_{ener} est plus réduite que celle des estimateurs μ_{pente} et μ_{min}. Nous constatons également la présence d'un point d'inflexion de la fonction $\mu_{ener}(t_m)$, point d'inflexion situé dans l'intervalle $[t_{ener}, t_0]$. Ce point d'inflexion indique une concentration temporelle locale des allocations μ_{ener} supérieure au reste de la fenêtre. Ceci est intéressant puisque cette concentration nous fournit dés lors un indice de fiabilité de l'estimation de μ_{ener}. Dans la partie suivante, nous tirons parti de cette dernière propriété.

3.4.2.4 Fonctions de confiance

Nous venons de voir que l'estimateur μ_{ener} de t_{ener} possède un point d'inflexion situé dans l'intervalle $[t_{ener}, t_0]$. Nous proposons deux traitements permettant de tirer parti de cette propriété.

Différentielle temporelle : Dans [PR99b] [PR99a] nous proposons de mesurer la confiance accordée à $\mu(t)$ par calcul de sa différentielle temporelle.

$$\begin{array}{ll} \forall t_m & t_m \rightarrow \mu(t_m) \\ \forall \mu(t_m), \mu(t_{m+1}) & \gamma((\mu(t_m) + \mu(t_{m+1}))/2) = \left| \dfrac{\mu(t_{m+1}) - \mu(t_m)}{t_{m+1} - t_m} \right| \end{array} \tag{3.20}$$

La valeur absolue est utilisée pour prévenir des ré-assignements rétrogrades. $\mu(t)$ prenant ses valeurs dans l'intervalle $[-L/2, L/2]$, la fonction $\gamma(t)$ prend ses valeurs dans l'intervalle $[0, I+L]$ où L est la taille de la fenêtre et $I = t_{m+1} - t_m$ le pas d'avancement de l'analyse. La fonction γ peut être ramenée dans l'intervalle $[0, 1]$

$$\gamma_n = \left(1 - \frac{\gamma}{I+L}\right) \tag{3.21}$$

Observations : L'utilisation d'une différentielle ne fournit qu'une information locale sur la cohérence des assignations temporelles. $\gamma(t)$ ne permet pas de distinguer les singularités faibles (concordance d'assignation sur une région étroite) des singularités fortes (concordance des assignations sur une région large).

Cumul des assignements locaux : Nous proposons une autre méthode reposant sur le cumul des assignements locaux. Pour chaque valeur de $\mu(t_m)$, un cumul des valeurs avoisinantes est effectué. [12] Le cumul est effectué par évaluation aux points $\mu(t_M)$ d'une fonction de Gauss centrée sur $\mu(t_M)$:

$$\boxed{\begin{array}{ll} \forall t_m & t_m \rightarrow \mu(t_m) \\ \forall \mu(t_M) & \gamma(\mu(t_M)) = \displaystyle\sum_{\mu(t_m)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\mu(t_M) - \mu(t_m))^2}{2\sigma^2}\right) \end{array}} \tag{3.22}$$

σ détermine l'horizon temporel considéré autour de chaque point.

Normalisation et valeur de référence : Une fenêtre de type Gauss possède plus de 99% de son énergie concentrée sur la largeur $t = [-3\sigma, 3\sigma]$.

- Dans le cas d'un signal $\mu(t_m)$ dont les valeurs sont uniformément réparties sur la largeur $[-3\sigma, 3\sigma]$, i.e. un signal $s(n)$ sans localisation temporelle [13] , la valeur de γ est égale à $1/I$ (dans lequel I est le pas d'avancement de l'analyse à fenêtre glissante). [14]
- A l'inverse, pour un signal dont l'énergie locale est concentrée en un instant unique μ, chaque fenêtre d'analyse couvrant cet instant (analyse à fenêtre glissante) apporte une contribution de $\frac{1}{\sqrt{2\pi}\sigma}$ à la valeur de γ. Un instant μ étant couvert par L/I fenêtres d'analyse glissante, la valeur totale de γ est égale à $\gamma = \frac{L}{I\sqrt{2\pi}\sigma}$.

Nous définissons la valeur normalisée de γ, notée γ_n comme

$$\boxed{\gamma_n = \gamma \cdot \frac{I\sqrt{2\pi}\sigma}{L}} \tag{3.23}$$

La valeur correspondant à un signal uniformément réparti, c'est-à-dire sans localisation, est égale à

$$\boxed{\gamma_{h,n} = \frac{\sqrt{2\pi}\sigma}{L}} \tag{3.24}$$

Observations : Cette méthode présente, par rapport à la méthode «différentielle temporelle», l'avantage de rendre compte de l'étendue temporelle de la concordance d'assignation et donc de permettre la distinction entre singularités faibles et fortes.

3.4.3 Caractérisation en largeur temporelle

La largeur temporelle est la caractéristique complémentaire à la localisation que nous attribuons aux singularités. La largeur temporelle est théoriquement indépendante de la localisation de la singularité. Cependant, de par l'utilisation d'un algorithme basé sur la concordance des assignations d'énergie (concordance des valeurs $\mu(t_m)$), la largeur temporelle n'est plus indépendante de la localisation, puisque seules les singularités présentant une concordance temporelle importante de μ seront étiquetées comme «singularités», et dés lors leur largeur temporelle sera bornée supérieurement.

La largeur temporelle est calculée comme l'«écart type» σ temporel de l'énergie du signal local :

$$\sigma^2(s, t_m) = \int_{-L}^{L} (t - \mu(s, t_m))^2 |s(t_m, t - t_m)|^2 dt \tag{3.25}$$

Une formulation équivalente dans le domaine fréquentiel s'obtient en sommant la contribution de la dérivée fréquentielle du spectre d'amplitude σ_A^2 [15] et de la variance du retard de groupe σ_ϕ^2 [16] [Coh95]

$$\boxed{\sigma^2(s, t_m) = \underbrace{\int \left(\frac{\partial |S(t_m, \omega)|}{\partial \omega}\right)^2 d\omega}_{\sigma_A^2} + \underbrace{\int (\mu - \tau_g(s, t_m, \omega))^2 |S(t_m, \omega)|^2 d\omega}_{\sigma_\phi^2}} \tag{3.28}$$

Normalisation et valeur de référence : Pour une fenêtre $h(t)$ de taille L, $\sigma(s, t_m, L)$ prend ses valeurs dans l'intervalle $[0\,,\, L/2]$. σ peut donc être normalisé par la demi-taille de fenêtre $L/2$. Dans le cas d'un signal périodique, il est également intéressant d'obtenir une largeur temporelle relativement à la durée de la période fondamentale locale. Dans ce cas le choix d'une longueur de fenêtre égale à un multiple entier de la période fondamentale est intéressant. Nous définissons σ_n comme la version normalisée de σ :

$$\boxed{\sigma_n = \left(1 - \frac{\sigma}{L/2}\right)} \tag{3.29}$$

Une valeur caractéristique de σ_n est celle obtenue dans le cas d'un signal constant (ou dans le cas d'une sinusoïde d'amplitude constante et d'un nombre de périodes suffisant par rapport à la durée L). Dans ce cas σ est déterminé par les caractéristiques de la fenêtre : $\sigma = \sqrt{\frac{\sum_n (nh(n))^2}{\sum_n h^2(n)}}$. Nous notons cette valeur caractéristique σ_h et nous notons sa valeur normalisée $\sigma_{h,n}$:

$$\boxed{\sigma_{h,n} = \left(1 - \frac{\sqrt{\frac{\sum_n (nh(n))^2}{\sum_n h^2(n)}}}{L/2}\right)} \tag{3.30}$$

Pour un signal $s(t)$ donné, une valeur de $\sigma_n(s, t_m, L)$ supérieure à $\sigma_{h,n}$ est signe d'une concentration temporelle importante du signal à cet endroit. σ_n nous renseigne sur la possibilité de décomposer un signal en formes d'onde élémentaires .

3.4.4 Améliorations : soustraction de la contribution du filtre

Même si théoriquement la convolution d'un signal à phase minimale (comme le signal dérivé du débit glottal) par un filtre tous-pôles (comme le filtre buccal) produit une séquence à phase minimale, les pôles du filtre tous-pôles proches du cercle unité (pôles de résonance élevée) produise un étalement temporel du signal. Ceci est gênant dans le cas des méthodes de détection de singularités étudiées puisque

- le centre de gravité temporel t_{ener} d'un signal à phase minimale est d'autant plus éloigné de son début t_0 que le signal est à décroissance lente,
- l'influence de la fenêtre d'analyse sur l'estimation de t_{ener} par μ_{ener} sera d'autant plus grande que le signal est à décroissance lente.

Pour ces raisons il est intéressant de soustraire la contribution du filtre avant détection des singularités. Ceci peut s'obtenir

- soit en déconvoluant le signal $p(t)$ par la réponse impulsionnelle du filtre du système $v(t)$; l'estimation s'effectue alors sur le signal résiduel ;
- soit en soustrayant la contribution du filtre $v(t)$ du retard de groupe du signal $p(t)$ [17] .

Dans les deux cas, le filtre considéré peut être estimé par prédiction linéaire ou, étant donné les propriétés phase minimale du filtre, par filtrage homomorphique.

3.4.4.1 Estimation du filtre du système

◊ ***Filtre de prédiction linéaire***

La méthode de prédiction linéaire [MG76] utilisée est la méthode de Burg [Kay88].

◊ ***Filtrage homomorphique et à phase minimale***

[DYM89], [MMY89] et [YMR91] utilisent les propriétés des signaux à phase minimale afin de déduire du cepstre réel le spectre de phase du filtre du système [OS75]) (voir annexe E).

Le spectre de phase obtenu par filtrage homomorphique peut être utilisé pour l'estimation des résonances et anti-résonances du système. Pour un filtre à phase minimale, les fréquences de résonances sont estimées par les maxima locaux positifs du retard de groupe $\tau_g(\omega)$, tandis que les anti-résonances le sont par les maxima locaux négatifs. L'amplitude absolue de $\tau_g(\omega)$ est proportionnelle à la résonance et donc inversement proportionnelle à la bande passante. Ceci est vrai du moins en l'absence de bruit additif. Diverses méthodes peuvent être utilisées afin de minimiser l'influence de bruit additif.

Observations : Nous avons implémenté deux méthodes d'estimation du filtre homomorphique.

La première méthode repose sur l'observation du signal sur une durée inférieure à une période. Le cepstre réel $c(n)$ obtenu est tronqué sur la partie des quéfrences négatives $c(n) = 0, \forall n > 0$ afin de rendre le cepstre causal (condition de phase minimale). Le spectre de phase (ainsi que le retard de groupe) est obtenu par TF du cesptre tronqué. Nous avons constaté que cette méthode d'estimation est très sensible au positionnement de la fenêtre par rapport au signal. En pratique seule l'estimation obtenue lorsque la fenêtre est proche de l'IFG permet l'estimation des formants. Pour les autres positions, la fonction retard de groupe obtenue est très bruitée.

La deuxième méthode repose sur l'observation du signal sur une durée de l'ordre de 3 périodes fondamentales. Le cepstre réel $c(n)$ obtenu est tronqué sur l'axe négatif $c(n) = 0, \forall n > 0$ et pour les valeurs de quéfrences supérieures à 0.8 périodes fondamentales $c(n) = 0, \forall n >$ T0. La deuxième troncature est réalisée à l'aide d'une fenêtre de Hann. Cette méthode présente une plus grande résistance au positionnement de la fenêtre.

Sur des signaux tests [18] en l'absence de bruit, le filtrage homomorphique permet la détection de résonances très proches difficiles à détecter par une méthode usuelle de type prédiction linéaire (voir Fig. 3.9). Cependant, en présence de bruit [19], il devient extrêmement difficile de distinguer les maxima dus aux résonances du système de ceux dus aux discontinuités du spectre résultant de la présence de bruit (voir Fig. 3.10).

3.4.5 Comparaison des méthodes existantes et proposées

◊ ***Méthodes comparées***

Jusqu'à présent, nous avons présenté dans un même cadre les méthodes de détection de singularités utilisant le retard de groupe. L'objectif était de mettre en évidence les ressemblances entre ces méthodes. Ces méthodes ont cependant été proposées par des auteurs différents. Ci-dessous, nous présentons les méthodes proposées par chaque auteur et les com-

parons expérimentalement.

Méthode GDP : Dans la méthode GDP [SY95], μ est pris comme la pente du spectre de phase déroulé du signal résiduel $e(n)$. La pente est calculée par régression linéaire. Cette méthode ne donne pas lieu à l'estimation d'une largeur temporelle σ.

$$s(n) \rightarrow \boxed{\text{LP-1}} \rightarrow e(n) \rightarrow \boxed{\text{FFT}} \rightarrow \phi(e,\omega) \rightarrow \text{pente}$$

Méthode GDS : Dans la méthode GDS, [PR99b] et [PR99a] μ est pris comme la moyenne du retard de groupe du signal $s(n)$, $\tau_g(s)$, pondéré par l'énergie du signal $|S(\omega)|^2$.

$$1)\ s(n) \rightarrow \boxed{FFT_{th}/FFT_h} \rightarrow \tau_g(s,\omega)$$
$$2)\ \tau_g(s,\omega)\ ,\ |S(\omega)|^2 \rightarrow \mu$$

Méthode GDR : La méthode GDR est identique à la méthode GDS à l'exception du fait que l'estimation s'effectue sur le signal résiduel $e(n)$ (résiduel de prédiction linéaire). [20]

$$1)\ s(n) \rightarrow \boxed{\text{LP-1}} \rightarrow e(n) \rightarrow \boxed{FFT_{th}/FFT_h} \rightarrow \tau_g(e,\omega)$$
$$2)\ \tau_g(e,\omega)\ ,\ |E(\omega)|^2 \rightarrow \mu$$

Méthode GDCCC : Dans la méthode GDCCC [Kaw00], μ est pris comme la moyenne d'un retard de groupe corrigé, pondéré par l'énergie du signal $s(n)$. Le retard de groupe corrigé est obtenu par soustraction du retard de groupe du signal $s(n)$ du retard de groupe du filtre à phase minimale approximé à partir du **C**epstre **C**omplexe : soit $\tau_g(s) - \tau_g(h_{min})$.

$$1)\ s(n) \rightarrow \boxed{FFT_{th}/FFT_h} \rightarrow \tau_g(s,\omega)$$
$$2)\ s(n) \rightarrow \boxed{\text{FFT}} \rightarrow \boxed{\text{Filtrage homomorphique}} \rightarrow \phi(v_{min},\omega) \rightarrow \tau_g(v_{min},\omega)$$
$$3)\ \tau_g(s,\omega) - \tau_g(v_{min},\omega)\ ,\ |S(\omega)|^2 \rightarrow \mu$$

Méthode GDCLP : Nous introduisons cette méthode comme une variante de la méthode GDCCC. L'estimation du filtre est ici obtenue par prédiction linéaire. Le retard de groupe corrigé est obtenu par soustraction du retard de groupe du signal $s(n)$ du retard de groupe du filtre AR obtenu par prédiction linéaire.

$$1)\ s(n) \rightarrow \boxed{FFT_{th}/FFT_h} \rightarrow \tau_g(s,\omega)$$
$$2)\ s(n) \rightarrow \boxed{\text{LP}} \rightarrow v_{LP}(n) \rightarrow \boxed{\text{FFT}} \rightarrow \phi(v_{LP},\omega) \rightarrow \tau_g(v_{LP},\omega)$$
$$3)\ \tau_g(s,\omega) - \tau_g(v_{LP},\omega)\ ,\ |S(\omega)|^2 \rightarrow \mu$$

Les méthodes sont résumées au tableau Tab. 3.1.

3.4.5.1 Paramètres d'analyse

La fenêtre d'analyse utilisée est du type «Hann». Sa taille L est choisie égale à 1.25 $\overline{\text{T0}}$. Le choix $L = 1.25\overline{\text{T0}}$ nous a conduit aux meilleurs résultats expérimentaux sur nos signaux. Ce choix de taille de fenêtre permet de diminuer la détection des maxima secondaires à l'intérieur d'une période, tout en restant proche des conditions d'observation de signaux à phase minimale (un seul IFG présent dans la fenêtre) [21] . Le pas d'avancement de la fenêtre est choisi égal à $L/32$ en notant par L la durée de la fenêtre d'analyse. La taille de σ dans (3.22) est choisie égale au pas d'avancement.

TAB. 3.1 – Méthodes d'estimation des singularités utilisant le retard de groupe

	$\mu = \overline{\tau_g(x)}$	$\mu = \overline{\tau_g(x) \cdot \lvert Y \rvert^2}$
$\tau_g(s)$		GDS [PR99b]
$\tau_g(r)$	GDP [SY95] [YdD98]	GDR
$\tau_g(s) - \tau_g(v_{min})$		GDCCC [Kaw00]
$\tau_g(s) - \tau_g(v_{LP})$		GDCLP

La fréquence minimale considérée du spectre est 600 Hz. Ce choix résulte également d'une constatation expérimentale. Aux figures FIG. 3.11 et FIG. 3.12 page 50, nous montrons les résultats de l'ensemble des fonctions GDS, GDR, GDCLP, GDCCC et GDP en utilisant une fréquence minimale de 0 Hz FIG. 3.11 et de 600 Hz FIG. 3.12. De meilleurs résultats sont obtenus en ne considérant le signal qu'au dessus de 600 Hz. Ceci peut s'expliquer par le fait que le signal en dessous de 600 Hz ne peut être considéré comme un signal à phase minimale [22] .

3.4.5.2 Détection des IFGs

Dans cette expérience, nous nous intéressons à la localisation des maxima des fonctions γ_n correspondant à chaque estimateur GDS, GDR, GDCLP, GDCCC et GDP. Les estimateurs sont comparés sur les mêmes signaux tests que ceux utilisés pour la méthode de la norme de Frobenius :

Numéro de figure	Description	$\overline{f0}$	Laryngographe
FIG. 3.18 page 57	voix d'homme	172 Hz	oui
FIG. 3.20 page 58	voix d'homme	138 Hz	oui
FIG. 3.22 page 59	voix d'homme	121 Hz	oui
FIG. 3.24 page 60	voix de femme	317 Hz	oui
FIG. 3.26 page 61	voix de femme	307 Hz	oui
FIG. 3.28 page 62	voix parlée, signal périodique	111 Hz	non
FIG. 3.30 page 63	vocal fry, signal composé de pulses peu périodiques	40 Hz	non
FIG. 3.32 page 64	sinusoïde, premier formant à f0, forme d'onde symétrique	281 Hz	non
FIG. 3.34 page 65	voix chantée de femme	510 Hz	non
FIG. 3.36 page 66	sons de trompette	350 Hz	non
FIG. 3.38 page 67	sons de violon	260 Hz	non
FIG. 3.40 page 68	sons de piano	260 Hz	non

Les figures de résultats sont placées en dessous de celles utilisant la méthode de la norme de Frobenius de manière à faciliter les comparaisons. Comme pour cette dernière, les résultats sont indiqués après passage des fonctions γ_n dans un filtre structurant $g(n)$ ("Averaged Value Substracted") (voir 3.9), ceci afin de faciliter la lecture. Sa taille est de $T = \overline{T0}$. Les fonctions $g(x)$ sont également indiquées sous forme normalisée ($[0, 1]$).

Chaque figure représente, du haut vers la bas, (1) le signal, (2) $g(\gamma)$ de GDS, (3) $g(\gamma)$ de GDR, (4) $g(\gamma)$ de GDCLP, (5) $g(\gamma)$ de GDCCC et (6) $g(\gamma)$ de GDP. Lorsqu'un signal laryngographique est disponible, celui-ci est superposé en trait léger aux signaux (2) (3) (4) (5).

Observations : L'estimateur GDP produit le plus souvent une fonction d'observation bruitée dans laquelle la localisation de maxima correspondant aux IFGs est rendue difficile par la présence de nombreux autres maxima (voir FIG. 3.20, FIG. 3.24). Son application sur des sons d'instruments ne répondant pas au modèle source/ filtre ne produit pas de bons résultats. Les deux méthodes étudiées de correction du spectre du retard de groupe, GDCLP et GDCCC, ne produisent pas de différence significative par rapport à la méthode sans correction du spectre GDS. La méthode GDR, utilisant le signal résiduel, bien que présentant un nombre important de maxima secondaires, fournit les meilleurs résultats sur des signaux de type voix parlées d'homme. A l'inverse, sur des signaux du type voix parlées de femme (FIG. 3.24, FIG. 3.26, FIG. 3.32 et FIG. 3.34), les résultats ne sont pas bons. La méthode GDS, utilisant directement le signal, bien que présentant un décalage par rapport à

la position exacte de l'IFG, fournit les résultats les plus stables pour l'ensemble des signaux.

3.4.5.3 Etude des fonctions de concentration

Nous illustrons ici les fonctions γ_n et σ_n comme mesures de concentration des formes d'onde en comparaison des fonctions limites $\gamma_{h,n}$ et $\sigma_{h,n}$ correspondant au cas d'un signal sans concentration temporelle d'énergie.

Les signaux suivants sont utilisés :

Numéro de figure	Description	$\overline{\text{f0}}$	Laryngographe
FIG. 3.42	voix parlée, signal périodique	111 Hz	non
FIG. 3.43	sinusoïde, premier formant à f0, forme d'onde symétrique	281 Hz	non
FIG. 3.44	sons de trompette	350 Hz	non

Pour cette expérience, la globalité du spectre est prise en compte, y compris la contribution de la partie basse fréquence ($\leq$ 600 Hz).

Chaque figure représente, du haut vers la bas, (1) le signal, (2) la fonction GDS (3), la fonction GDR (4), la fonction GDCLP (5), la fonction GDCCC et (6) la fonction GDP. Le panneau de gauche représente les fonctions γ_n associées aux estimateurs GDS, GDR, GDCLP, GDCCC et GDP. Dans chacun des sous-panneaux, les lignes verticales représentent les valeurs limites $\gamma_{h,n}$. Le panneau de droite représente les fonctions σ_n associées aux mêmes estimateurs. Dans chacun des sous-panneaux, les lignes verticales représentent les valeurs limites $\sigma_{h,n}$.

Observations : Les différences entre les différents estimateurs sont d'abord illustrées dans le cas d'un signal de synthèse de type CHANT [23] à la figure FIG. 3.41. Une différence de comportement apparaît de manière marquée entre les estimateurs utilisant le signal résiduel, GDR et GDP, et les autres estimateurs. Les maxima de la fonction γ_n de GDS se trouvent en retrait du début des formes d'onde élémentaires . Ce décalage est absent des autres estimateurs du fait de la correction du spectre. La valeur de la fonction γ_n associée aux estimateurs GDR et GDP et de la fonction σ_n associée à GDR est plus importante que celle des fonctions γ_n et σ_n associées aux autres estimateurs. Ceci se comprend en considérant (3.25). GDCLP et GDCCC ne corrigent (3.25) que par soustraction de la contribution du filtre au terme de phase de (3.25), alors que GDR corrige également la contribution du filtre au terme d'amplitude de (3.25). La répercussion de ceci sur les signaux γ_n associés aux estimateurs GDCLP et GDCCC est un effet de fenêtrage plus important sur le signal et une variance locale plus faible de γ_n.

Les figures FIG. 3.42, FIG. 3.43 et FIG. 3.44 illustrent la mesure de concentration γ_n et σ_n sur des signaux réels.

Remarquons que l'estimateur GDP n'a pas de fonction σ_n associée.

3.4.6 Caractérisation dans le plan temps/fréquence

Jusqu'à présent, l'évaluation des fonctions μ (γ) et σ a été effectuée pour l'ensemble des fréquences. Remarquons que dans ce cas, le formalisme fréquentiel que nous avons adopté n'est pas nécessaire, puisque μ et σ peuvent alors être évalués à partir de (D.3) et (3.25).

L'utilisation de formulations fréquentielles (3.16) et (3.28) se justifie par la possibilité offerte d'une caractérisation locale en fréquence du signal. Tant (3.16) que (3.28) peuvent être estimés pour une bande de fréquence limitée du spectre.

Le principal intérêt d'une caractérisation dans le plan temps/ fréquence est de pouvoir obtenir une mesure du caractère «forme d'onde élémentaire » dans le plan temps/ fréquence et donc, dans le cas de l'étude des algorithmes de modification du signal, de permettre de mesurer quelle partie du plan temps/ fréquence peut être modélisée sous la forme de formes d'onde élémentaires . Ceci constitue une approche similaire à celle prise par [d'A89] et [dR89], dans laquelle la modélisation la plus adaptée aux caractéristiques des régions fréquentielles du signal est utilisée. Dans le cas de [d'A89], les modèles utilisés sont le modèle addition de sinusoïde pour la région basse du spectre et Forme d'Onde Formantiques (FOF) pour la région supérieure du spectre.

Nous nous sommes intéressés à deux types de décompositions en bandes de fréquence :
- une décomposition en bandes d'octave
- et une décomposition en bandes formantiques.

3.4.6.1 Décomposition en bandes d'octave

L'axe des fréquences est décomposé en bandes de fréquence de rapport «largeur de bande sur fréquence centrale» constant. Les régions fréquentielles sont caractérisées par les fréquences suivantes : $[Fe/4, Fe/2]$, $[Fe/8, Fe/4]$, $[Fe/16, Fe/8]$, ... dans lequel Fe désigne le taux d'échantillonnage. Ce type de décomposition est illustré à la Fig. 3.13.

3.4.6.2 Décomposition en bandes formantiques

Cette décomposition, utilisée dans [d'A89] et [dR89] s'avère très intéressante mais particulièrement difficile à effectuer du fait de la nécessité de déterminer préalablement les limites (variables dans le temps) de chaque bande formantique. L'estimation des bandes de résonance peut se faire à chaque instant par estimation des fréquences de résonance (détection des fréquences du filtre de prédiction linéaire ou des fréquences des maxima de la fonction retard de groupe du filtre homomorphique), puis par création des trajets temporels de formants en raccordant les estimations faites en des instants discrets.

Malgré son intérêt, nous n'avons pas poussé plus avant ce type de décomposition, étant donné la difficulté d'obtenir des trajets lisses de formants en dehors du cas simple de signaux stationnaires. Ce type de décomposition est illustré à la figure Fig. 3.14.

3.4.6.3 Application de la décomposition en bandes d'octave

Aux Fig. 3.45, Fig. 3.46 et Fig. 3.47, nous illustrons la caractérisation γ_n / σ_n pour une décomposition en 6 bandes d'octave : [0,Fe/64], [Fe/64,Fe/32], [Fe/32,Fe/16], [Fe/16,Fe/8],

[Fe/8,Fe/4] et [Fe/4,Fe/2] dans lequel Fe représente la fréquence d'échantillonnage. Cette caractérisation est appliquée à 4 des signaux préalablement testés :

Numéro de figure	Description	$\overline{f0}$	Laryngographe
FIG. 3.45	voix parlée, signal périodique	111 Hz	non
FIG. 3.46	sinusoïde, premier formant à f0, forme d'onde symétrique	281 Hz	non
FIG. 3.47	son de trompette	350 Hz	non

Soulignons que le fenêtrage étroit appliqué dans le domaine temporel ($L = 1.25$T0) produit un étalement fréquentiel important : largeur du lobe principal à $-6dB_{20} = Bw \simeq 2$f0). De ce fait, les valeurs obtenues dans une bande ne sont pas indépendantes de celles obtenues dans les bandes avoisinantes.

- Sur la FIG. 3.45, nous observons une valeur γ_n qui dépasse périodiquement la valeur limite $\gamma_{h,n}$ dans les bandes 1,2,4,5,6. Il en est de même de la valeur σ_n qui dépasse périodiquement la valeur limite $\sigma_{h,n}$. Ces maxima supérieurs à la valeur limite indiquent qu'à cet endroit, et selon l'observation du signal par une fenêtre de taille 1.25 T0, le signal dans cette bande de fréquence présente une concentration d'énergie supérieure à celle d'une sinusoïde et est donc susceptible d'être représenté adéquatement (c'est-à-dire sans détérioration majeur du fait du fenêtrage) par une forme d'onde élémentaire de largeur T0 centrée en cet endroit.
- Des conclusions équivalentes peuvent être tirées de l'observation de la FIG. 3.47 : les bandes 1,3,4,5,6 peuvent, du fait de la valeur $\sigma_n > \sigma_{h,n}$, être représentée par une forme d'onde élémentaire . Dans la deuxième bande d'octave, le signal présente un double maximum de σ_n par période, signe d'une éventuelle détérioration du signal par le fenêtrage.
- Le signal de la FIG. 3.46 présente une forme d'onde symétrique . Il s'agit de la voyelle «i» prononcée par une voix de femme. La décomposition en bandes d'octave nous montre cependant une structure impulsionnelle pour les trois dernières bandes d'octave. Ces bandes contiennent une énergie faible mais qui peut vraisemblablement être représentée par des formes d'onde élémentaires centrées en les maxima locaux de γ_n.

3.4.7 D'une analyse à bande large à une analyse à bande étroite

Considérons le signal résultant de la somme de trois sinusoïdes harmoniques de ω_0 et de fréquence centrale ω_c :

$$x(t) = A_1 \cos((\omega_c - \omega_0)t + \phi_1) + A_2 \cos(\omega_c t + \phi_2) + A_3 \cos((\omega_c + \omega_0)t + \phi_3) \tag{3.32}$$

Si nous ne considérons que les trois partiels d'amplitude les plus importants, ce signal représente par exemple la réponse d'un système ayant une résonance unique à ω_c excité par une impulsion périodique de pulsation ω_0. Afin de ne pas alourdir les développements, nous considérons égales les amplitudes des trois composantes ($A_1 = A_2 = A_3 = 1$), et nous considérerons ϕ_2 égale à $(\phi_1 + \phi_3)/2$. Nous pouvons réécrire le signal comme

$$x(t) = \cos\left(\omega_0 t + \frac{\phi_3 - \phi_1}{2}\right) \cdot \cos\left(\omega_c t + \frac{\phi_1 + \phi_3}{2}\right) \tag{3.33}$$

Pour peu que $\omega_c > \omega_0$ (correspondant à un formant d'ordre supérieur), le premier terme est généralement désigné par **«signal modulant»** et représente l'enveloppe d'énergie du signal. Ce signal est périodique de période T_0 et peut s'interpréter comme la répétition à une période T_0 d'une forme d'onde élémentaire . Son retard par rapport à $t = 0$ est déterminé par $\omega_0(t - \tau_g)$ dans lequel $\tau_g = \frac{(\phi_3-\phi_1)/2}{\omega_0} = \frac{\phi_3-\phi_1}{(\omega_c+\omega_0)-(\omega_c-\omega_0)}$. Il s'agit du **retard de groupe** qui, dans ce cas, est égal à la différence de phase $\phi_3 - \phi_1$ du modèle sinusoïdal (3.32).

Le deuxième terme opère une translation du signal modulant à la fréquence ω_c et est désigné par **«porteuse»**. Son retard est caractérisé par $\omega_c(t - \tau_\phi)$ dans lequel $\tau_\phi = \frac{(\phi_1+\phi_3)/2}{\omega_c} = \frac{\phi_1+\phi_3}{(\omega_c-\omega_0)+(\omega_c+\omega_0)}$. Il s'agit du **retard de phase** moyen [24] .

Il est évident que (3.32) et (3.33) sont équivalents. La première expression représente le signal par une somme de sinusoïdes, la seconde comme une forme d'onde périodique centrée sur une fréquence de résonance ω_c.

La largeur de bande de (3.33) est déterminée par le terme de modulation d'amplitude [25] et est égale à ω_0. La largeur de bande de chaque composante de (3.32) est théoriquement nulle. La durée sur laquelle le signal est observé détermine donc le modèle. Pour une analyse à bande large, l'observation se fera en terme de (3.33) ; pour une analyse à bande étroite, l'observation se fera en terme de (3.32). Ceci est illustré à la FIG. 3.15, où les spectres à bandes large (-) et étroite (- -) d'un même signal sont superposés. Le spectre de phase à bande étroite coïncide avec celui du spectre à bande large uniquement aux positions des composantes harmoniques. La bande passante théoriquement nulle des composantes de (3.32), traduite en des paliers de phase constante du fait de la convolution par la TF d'une fenêtre de pondération symétrique, rend le retard de groupe localement nul. Le retard de groupe tel que défini précédemment et appliqué à un signal à bande étroite est nul aux positions des composantes d'énergie. Une notion semblable à celle de retard de groupe peut être trouvée par différenciation de la valeur de la phase prise aux différentes harmoniques. Il s'agit donc d'un **retard de phase relatif** entre les différentes harmoniques du signal.

Retard de phase relatif

$$\tau_g(\omega_h, \omega_{h+1}) = -\frac{\phi_{\omega_{h+1}} - \phi_{\omega_h}}{\omega_{h+1} - \omega_h} \tag{3.36}$$

dans lequel ω_h désigne la $h^{\text{ème}}$ harmonique du spectre.

3.4.8 Comparaison de la méthode de détection des singularités utilisant le retard de groupe avec les méthodes d'alignement utilisées en modélisation sinusoïdale

- [Sty98] propose une méthode d'alignement de segments sonores dans le cadre de la concaténation de diphones. Sous sa forme générale, cette méthode consiste à calculer pour chaque segment un **centre de gravité** défini comme

$$\tau = -\frac{\partial\phi}{\partial\omega}|_{\omega=0} \tag{3.37}$$

  Dans [Sty98], on suppose l'égalité du retard de groupe du signal et du signal résiduel.

Cette expression constitue un cas particulier de la méthode du retard de groupe dans lequel le calcul de $\mu(x, t_m, L)$ serait limité à la fréquence nulle.
Sous sa forme approximée, le centre de gravité de [Sty98] est calculé comme

$$\tau = -\phi(\omega_0) \tag{3.38}$$

Il s'agit donc d'un alignement sur la phase du premier partiel (τ tel que $\cos(\omega_0 t+\tau) = 0$)

- Une autre méthode d'alignement de segments sonores dans le cadre de la concaténation de diphones repose sur la **minimisation de la phase d'un filtre**. Dans [Cha88] et [CM89], le décalage τ par rapport à la position de concaténation optimale est estimé dans le domaine fréquentiel par

$$\tau_{\#} = \frac{Fe}{\omega_0}\frac{1}{H}\arg\left[\sum_{h=1}^{H} X_h X_{h+1}^*\right] \tag{3.39}$$

dans lequel $h \in H$ désigne l'ensemble des composantes harmoniques du spectre.
L'expression $\sum_{h=1}^{H} X_h X_{h+1}^*$ qui peut se réécrire $\sum_{h=1}^{H} A_h A_{h+1} e^{j(\omega_{h+1}-\omega_h)}$ constitue également une approximation du retard de groupe. Cependant la pondération est ici différente.

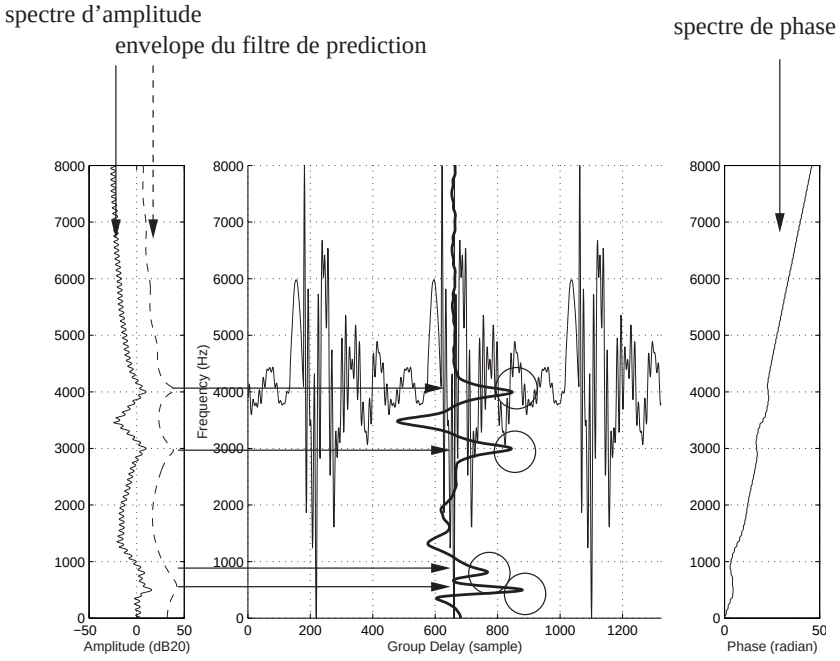


FIG. 3.9 – Détection des formants à partir du retard de groupe du filtre à phase minimale (filtrage homomorphique). [G] Spectre d'amplitude (-) et réponse en amplitude du filtre de prédiction linéaire (- -) [M] Retard de groupe superposé au signal, [D] Spectre de phase. Signal de synthèse : résonance à 500, 800, 3000 et 4000 Hz

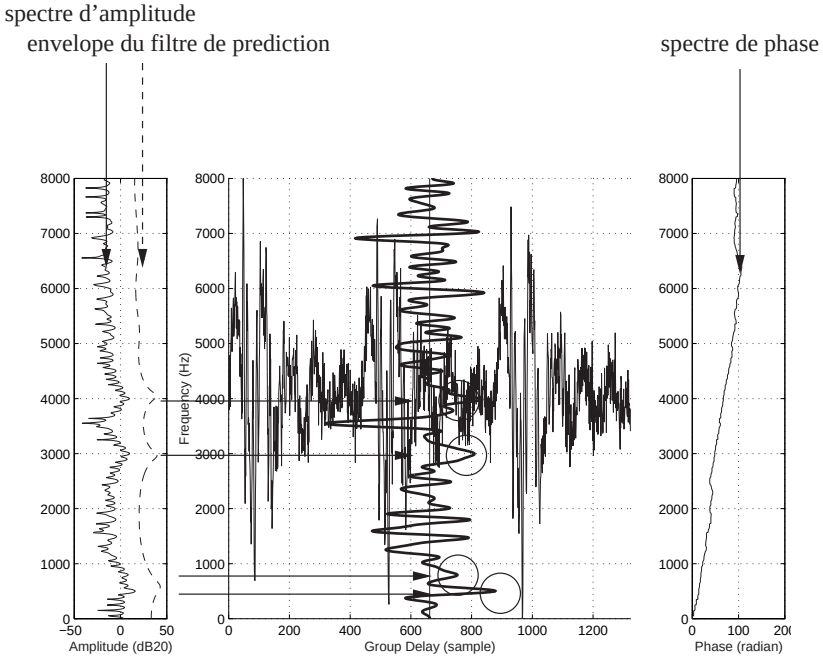


FIG. 3.10 – Détection des formants à partir du retard de groupe du filtre à phase minimale (filtrage homomorphique). Signal de synthèse : résonance à 500, 800, 3000 et 4000 Hz, bruit additif

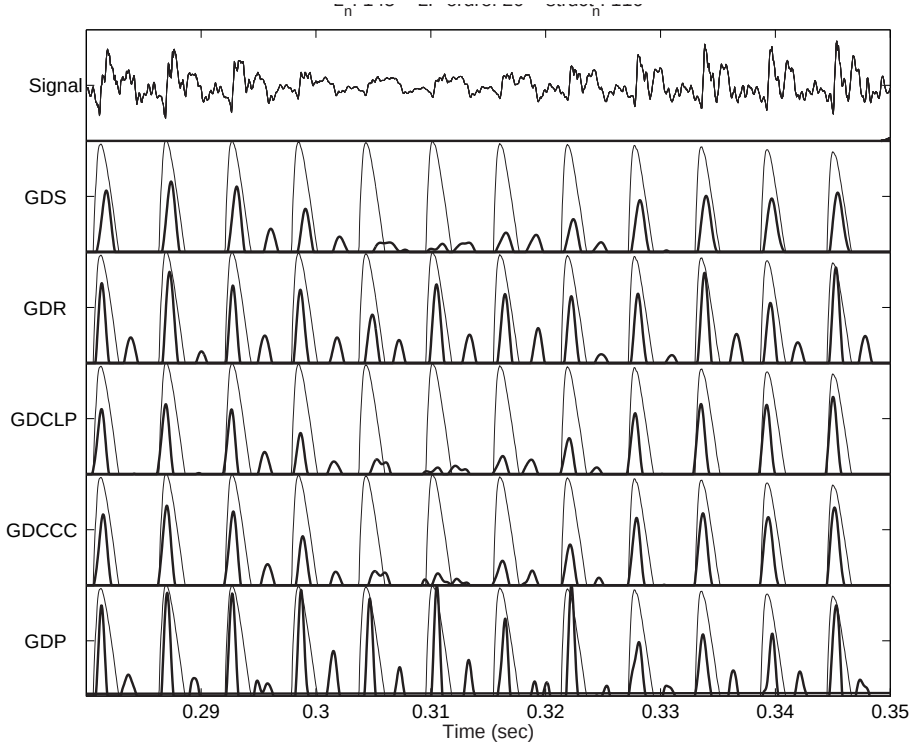


FIG. 3.11 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 0 Hz, Signal= rl001-2

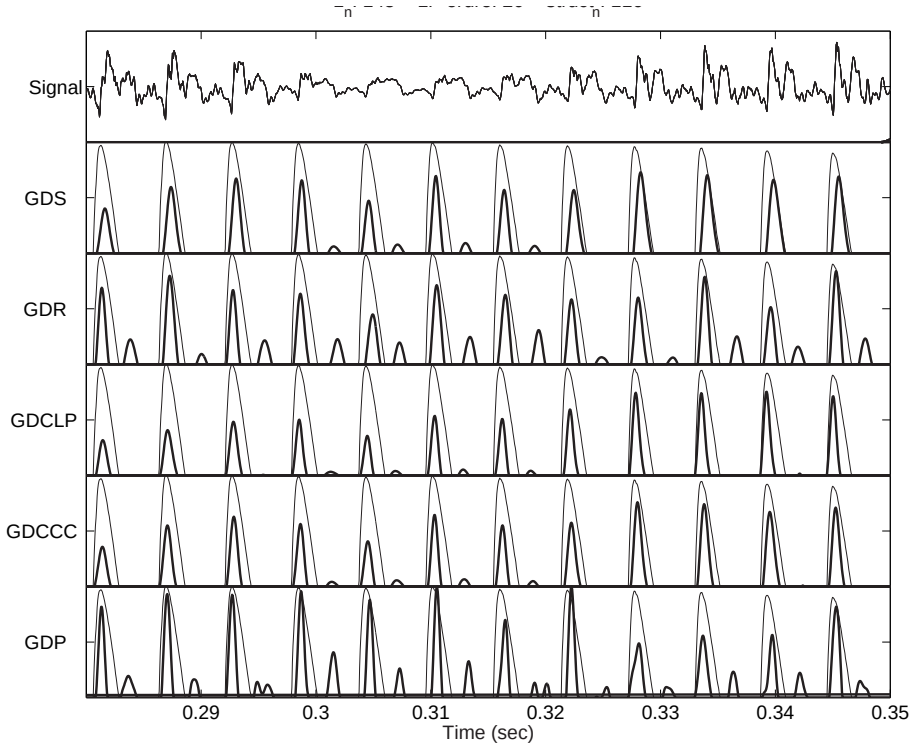


FIG. 3.12 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 600 Hz, Signal= rl001-2

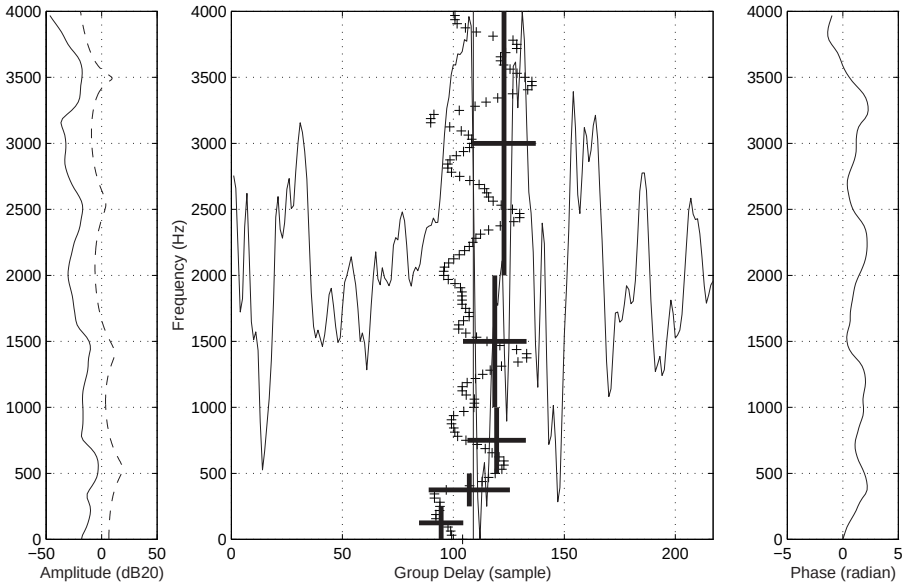


FIG. 3.13 – Décomposition de τ_g en bandes d'octave. Pour chaque bande de fréquence Ω : $\mu(\Omega)$ (traits gras verticaux), $\sigma(\Omega)$ (traits gras horizontaux)

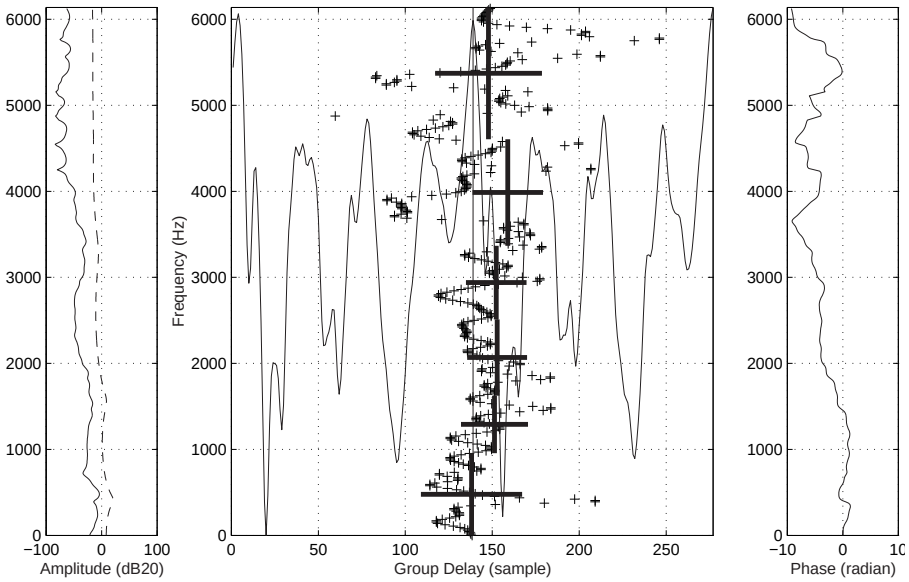


FIG. 3.14 – Décomposition de τ_g bandes formantiques. Pour chaque bande de fréquence Ω : $\mu(\Omega)$ (traits gras verticaux), $\sigma(\Omega)$ (traits gras horizontaux)

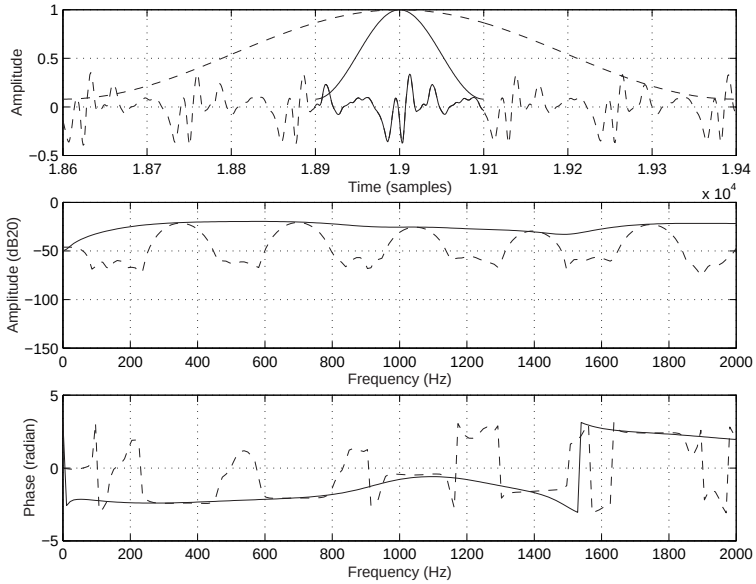


FIG. 3.15 – Spectre d'amplitude et de phase, analyse à bande large (-), analyse à bande étroite (- -)

3.5 Détection des transitoires

Jusqu'à présent, nous avons étudié le problème de la localisation et de la caractérisation de singularités dans un signal. Ces singularités étaient étudiées sous l'angle de la décomposition d'un signal périodique en formes d'onde élémentaires et ont été rapprochées, dans le cas du signal vocal dit «voisé», des «Instants de Fermeture de la Glotte». Dans cette partie, nous nous intéressons à la localisation des singularités du signal qui ne résultent pas d'une excitation périodique, soit qu'ils s'agissent de plosives dans le cas de la voix ou encore d'attaques d'instruments de musique. Nous les désignons sous le terme général de «transitoires».

Même si la modélisation des transitoires est possible [VLM97] [TG00], nous nous intéressons uniquement à leur localisation.

Plusieurs méthodes de détection de transitoires ont été étudiées.

Variation de l'enveloppe spectrale La première méthode étudiée [PR99b] est inspirée de [BB83] et [LG96]. Elle repose sur une mesure de la variation de l'enveloppe spectrale. L'estimation de la position des transitoires est basée sur la variation de la densité spectrale de puissance du signal. Les DSP successives du signal sont calculées. Chaque DSP est considérée comme représentant la loi de probabilité du spectre à un instant donné. Le problème est formulé en utilisant la théorie de l'information. Nous mesurons à chaque instant quelle est la diminution de l'entropie associée à une DSP à un instant t_2 apportée par la connaissance de la DSP à l'instant précédant t_1. Cette mesure est effectuée par calcul de la divergence de Kullback-Leibler [Bas89] .

$$D_{KL} = \int_{\omega=-\pi}^{\pi} K\left[\frac{DSP(t_1,\omega)}{DSP(t_2,\omega)}\right] d\omega \tag{3.40}$$

dans lequel $K = u - \log u - 1$; $DSP(t_1,\omega)$ et $DSP(t_2,\omega)$ représente les densités spectrales de puissance calculées aux instants t_1 et t_2.

Lorsque cette divergence est importante, nous considérons la présence d'une transitoire.

Nous avons observé une grande sensibilité de cette méthode dans les zones bruitées du signal (détection erronée en région bruitée) ainsi que des détections erronées à l'intérieur de segment périodique. La version finalement utilisée est une version modifiée dans laquelle la DSP est lissée dans le domaine cepstral (10 premiers coefficients cepstraux).

Variation d'énergie temporelle La deuxième méthode étudiée [VLM97] [Lev98] repose sur une mesure de la variation d'énergie locale du signal. La puissance du signal estimée à l'aide d'une fenêtre de courte durée est comparée à celle estimée à l'aide d'une fenêtre de longue durée. A la trame m, nous estimons une puissance P1 à l'aide d'une fenêtre de type Hamming de taille égale à 12 msec. Cette estimation P1 est comparée à la puissance P2 obtenue par somme pondérée des puissances estimées aux trames précédentes. Si $P2/P1 > 1$, nous sommes en présence d'une transitoire.

Les performances de cette méthode peuvent être améliorées en ne considérant que la partie du signal au-dessus de 1000 Hz.

Méthode basée sur le retard de groupe Nous définissons les transitoires comme une singularité locale du signal non corrélée avec le reste du signal.

Par cette définition, l'application de l'algorithme de détection des IFGs fondé sur l'utilisation du retard de groupe dans les zones non périodiques (donc non corrélées) permet la localisation des transitoires. Cet algorithme ne permet cependant pas de localiser les

transitoires se trouvant à l'intérieur d'une région périodique. Une deuxième approche consiste à calculer la corrélation de la forme d'onde entourant chaque singularité (telle que détectée au début de ce chapitre) avec les formes d'onde entourant les singularités avoisinantes. Une transitoire est alors définie comme une singularité non corrélée avec les singularités voisines. Malheureusement, la discrimination entre valeurs de corrélation obtenues pour des singularités correspondant à des formes d'onde périodiques et celles correspondants à des formes d'onde non périodiques est faible.

La méthode finalement retenue consiste à appliquer la méthode du retard de groupe calculée sur un horizon plus large. Nous choisissons une durée de 5 T0 pour la fenêtre d'observation choisie.

Les résultats des trois méthodes sont illustrés à la Fig. 3.16. Nous avons obtenus les meilleurs résultats avec l'algorithme du retard de groupe, même si le choix du paramètre «durée de l'observation» de cette méthode s'avère primordial. Un autre avantage de cette méthode tient dans sa précision temporelle.

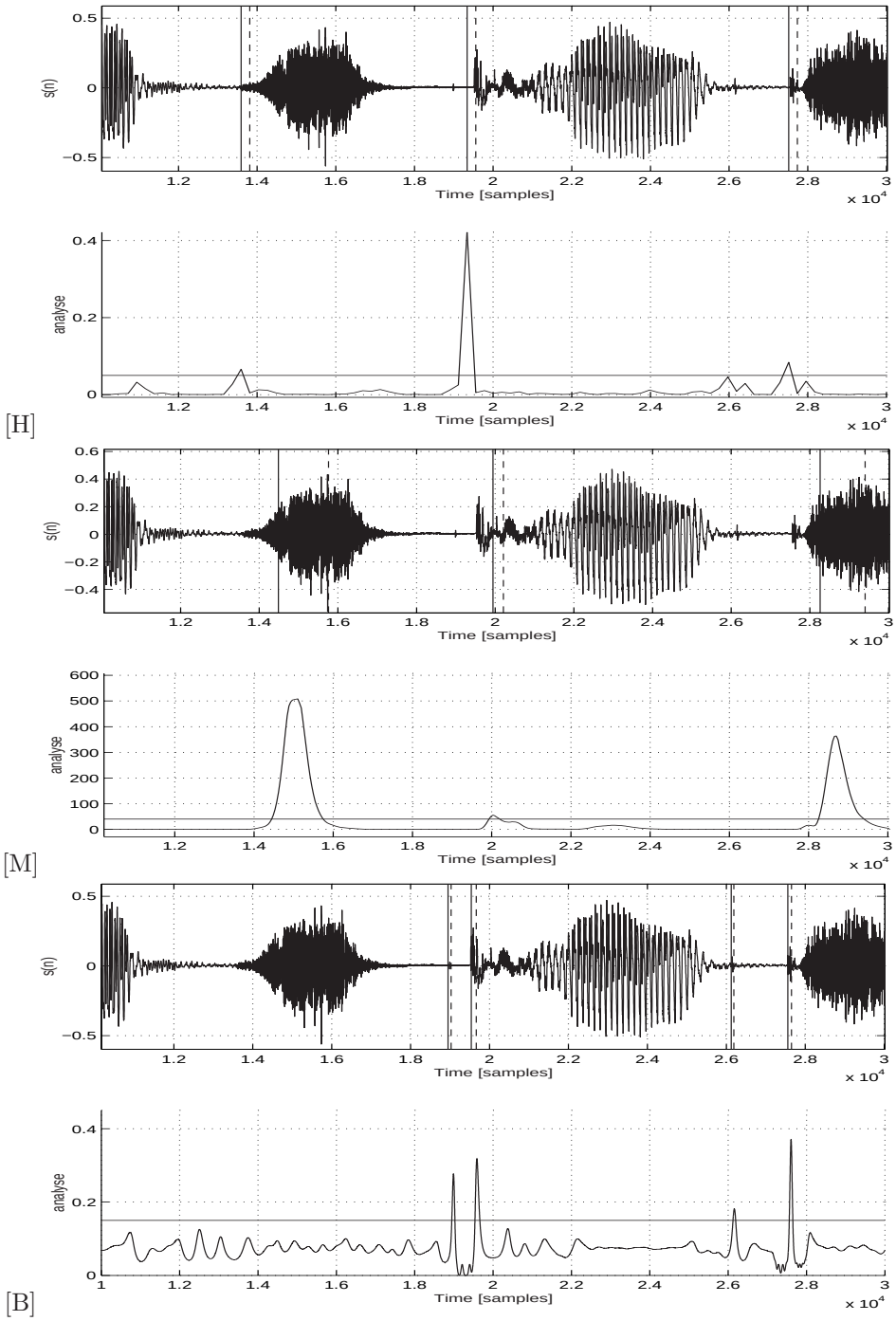


FIG. 3.16 – Détection des transitoires [H] méthode de la variation de l'enveloppe spectrale [M] méthode de la variation d'énergie [B] méthode du retard de groupe

3.6 Conclusion

Dans cette partie, nous avons étudié le problème de la détection des singularités dans un signal sous deux aspects : la détection des singularités périodiques du signal -correspondant aux Instants de Fermeture de la Glotte (IFGs) dans le cas du signal de parole-, la détection des transitoires considérés ici comme des singularités non-périodiques.

Les méthodes de détection d'IFGs fondées sur la rupture du modèle auto-régressifs du signal, en particulier la méthode dite de la «norme de Frobenius», ainsi que celles fondées sur les propriétés phase minimale du signal glottal et sur le retard de groupe, ont été étudiées.

Nous avons montré le bien fondé des méthodes utilisant le retard de groupe à l'aide d'une modélisation simple des formants du conduit buccal. Nous avons proposé un estimateur utilisant le retard de groupe du signal (GDS) ou du signal résiduel (GDR), proposé le cumul des assignements locaux des valeurs de cet estimateur, ainsi que définit ses valeurs limites.

Cette estimateur a ensuite été comparé aux méthodes existantes. Cette comparaison nous permet de conclure au bon comportement et à la robustesse de notre estimateur pour une large classe de signaux.

L'utilisation du retard de groupe permet également la caractérisation de la largeur temporelle des singularités ainsi qu'une caractérisation dans le plan temps/fréquence. Les caractéristiques localisation et largeur temporelle, nous permettrons, dans la suite de cette recherche, le positionnement des marques de découpage du signal PSOLA ainsi qu'une mesure de la détérioration du signal engendrée par le fenêtrage de l'algorithme PSOLA à bande large.

Pour une observation du signal sur un horizon plus large, nous avons montré que notre estimateur du retard de groupe peut également être utilisée pour la localisation des transitoires dans le signal.

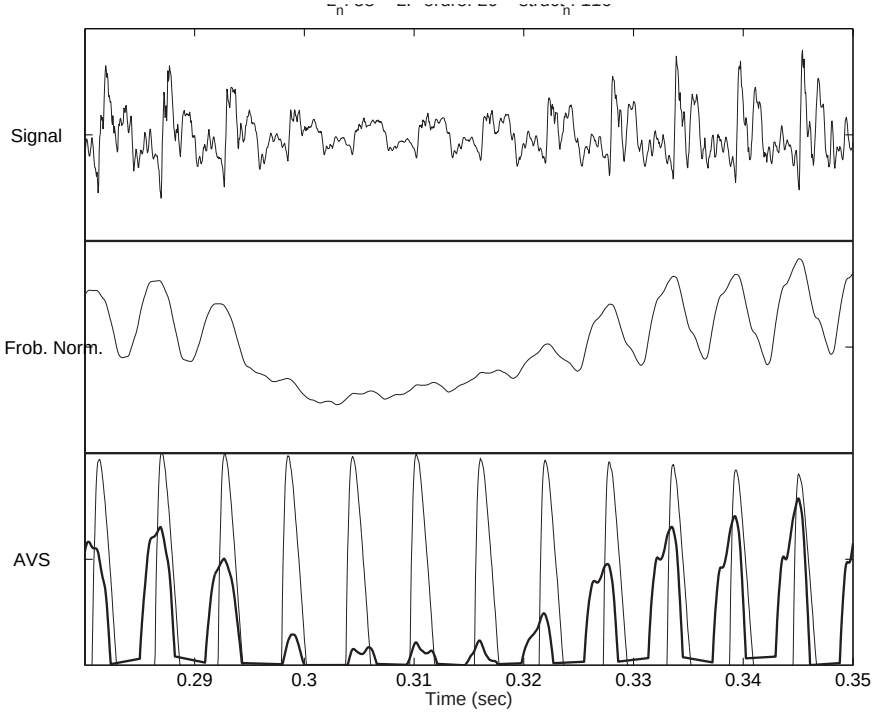


FIG. 3.17 – Méthode de la norme de Frobenius, Signal= rl001-2

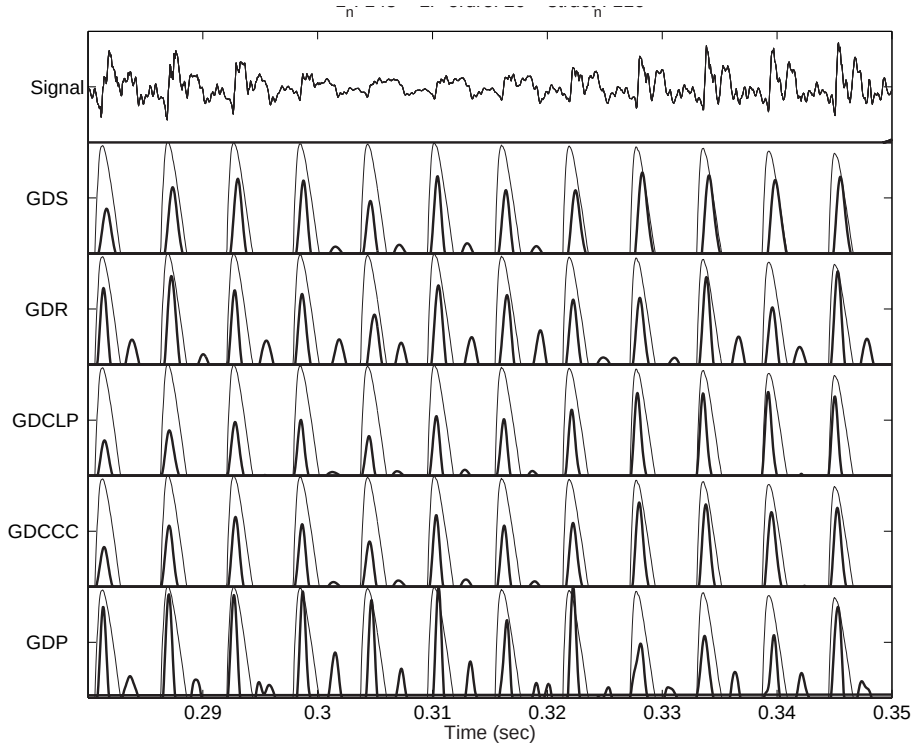


FIG. 3.18 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= rl001-2

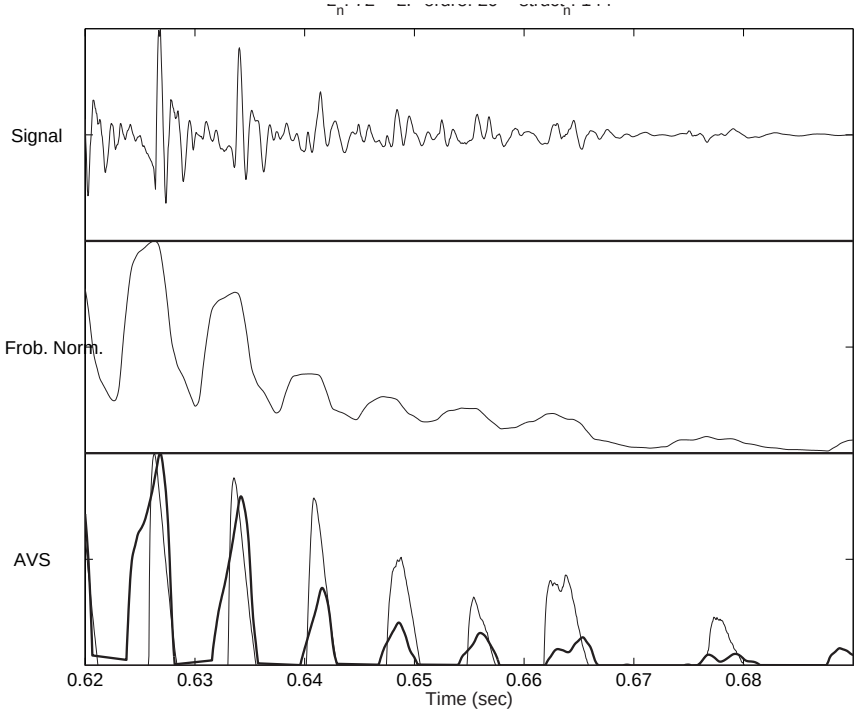


FIG. 3.19 – Méthode de la norme de Frobenius, Signal= rl001-3

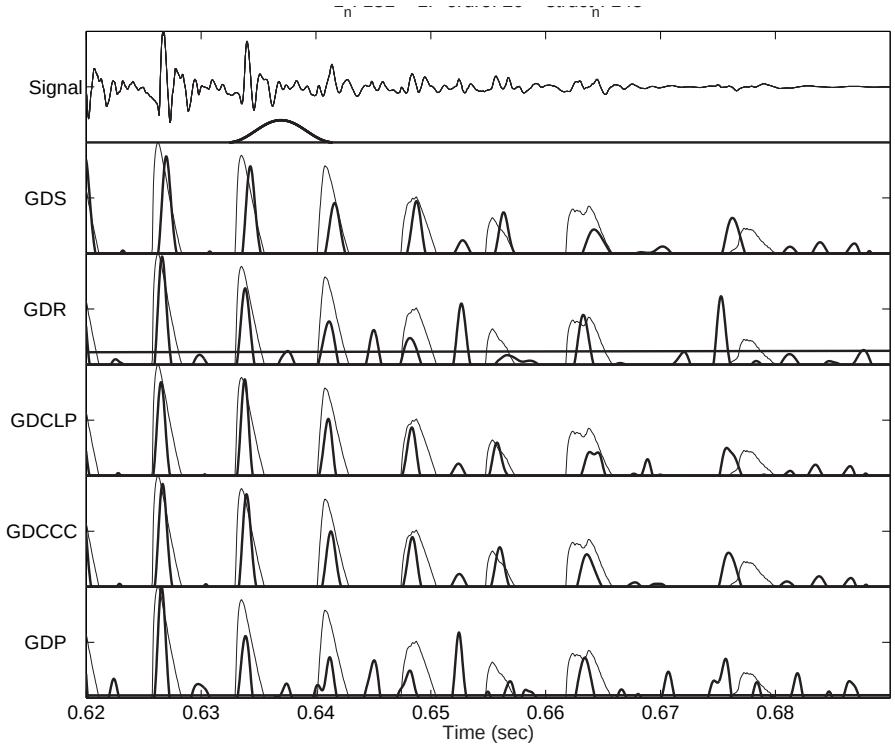


FIG. 3.20 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= rl001-3

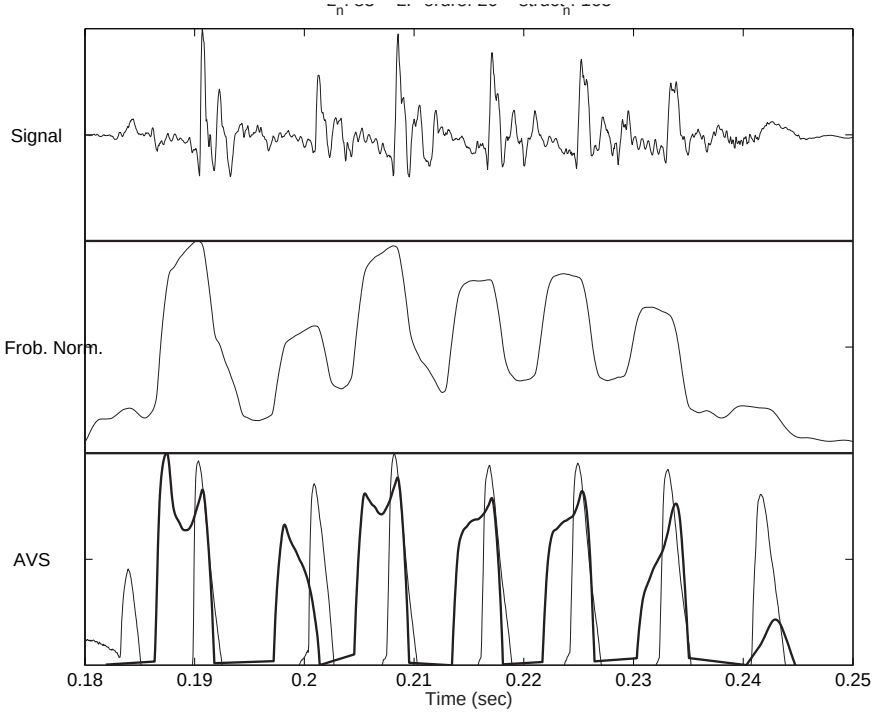


FIG. 3.21 – Méthode de la norme de Frobenius, Signal= rl002-1

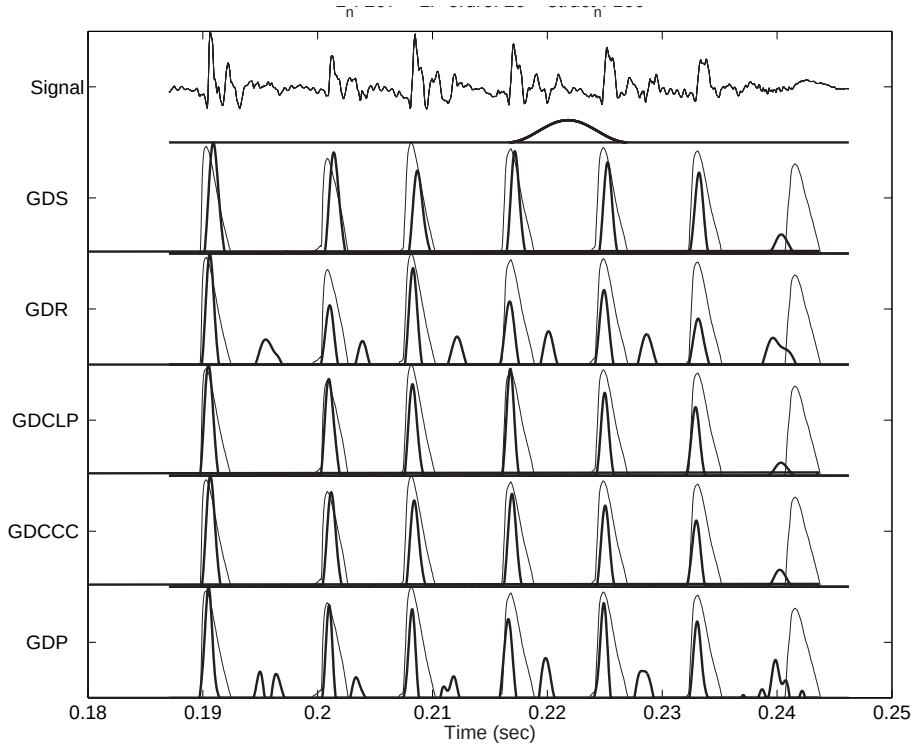


FIG. 3.22 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= rl002-1

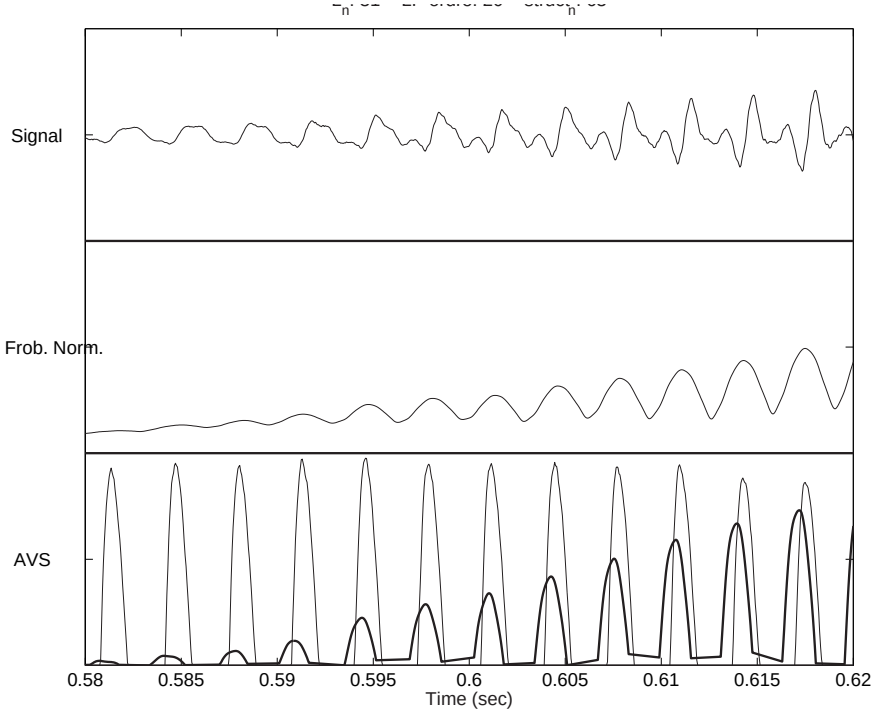


FIG. 3.23 – Méthode de la norme de Frobenius, Signal= sb010-1

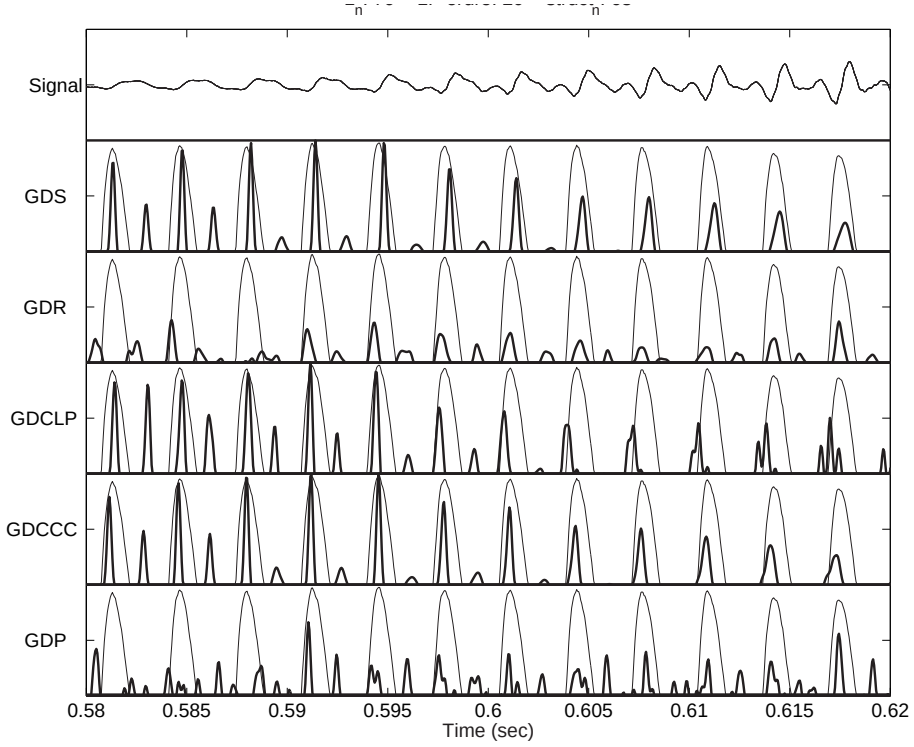


FIG. 3.24 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= sb010-1

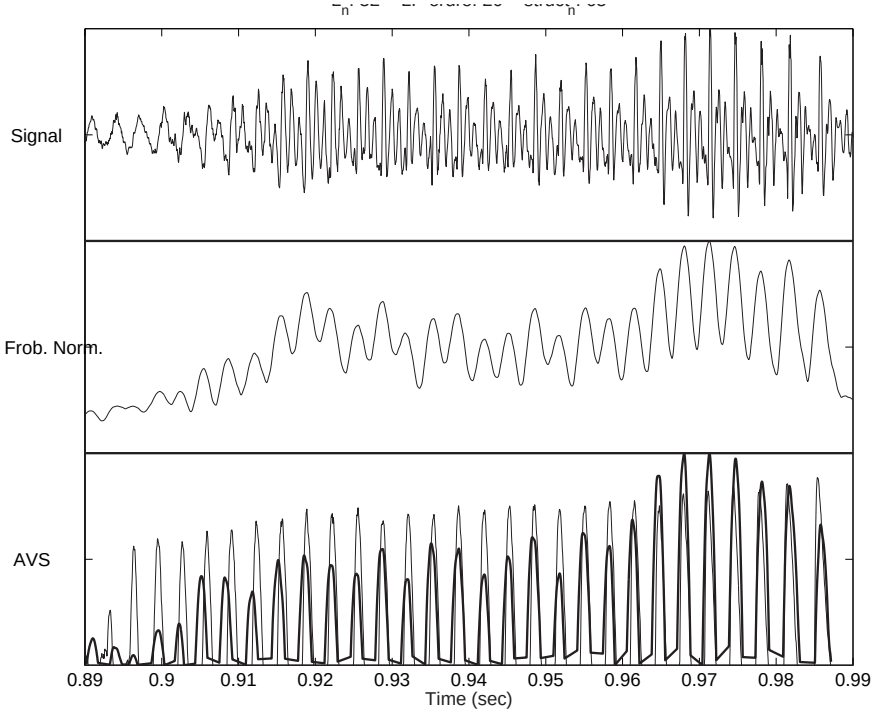


FIG. 3.25 – Méthode de la norme de Frobenius, Signal= sb010-2

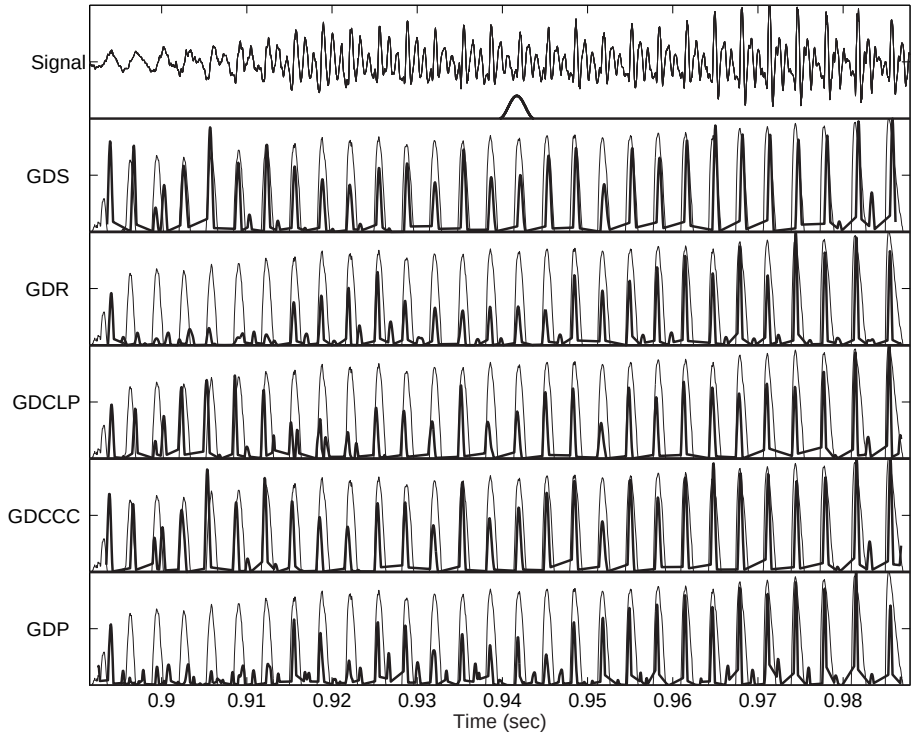


FIG. 3.26 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= sb010-2

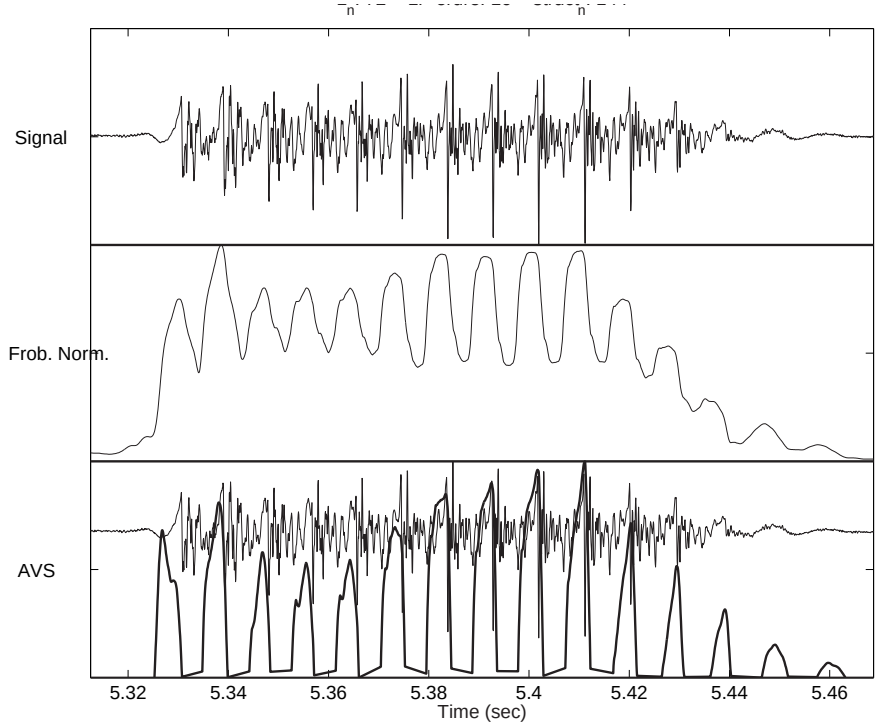


FIG. 3.27 – Méthode de la norme de Frobenius, Signal= speech-85000-87500

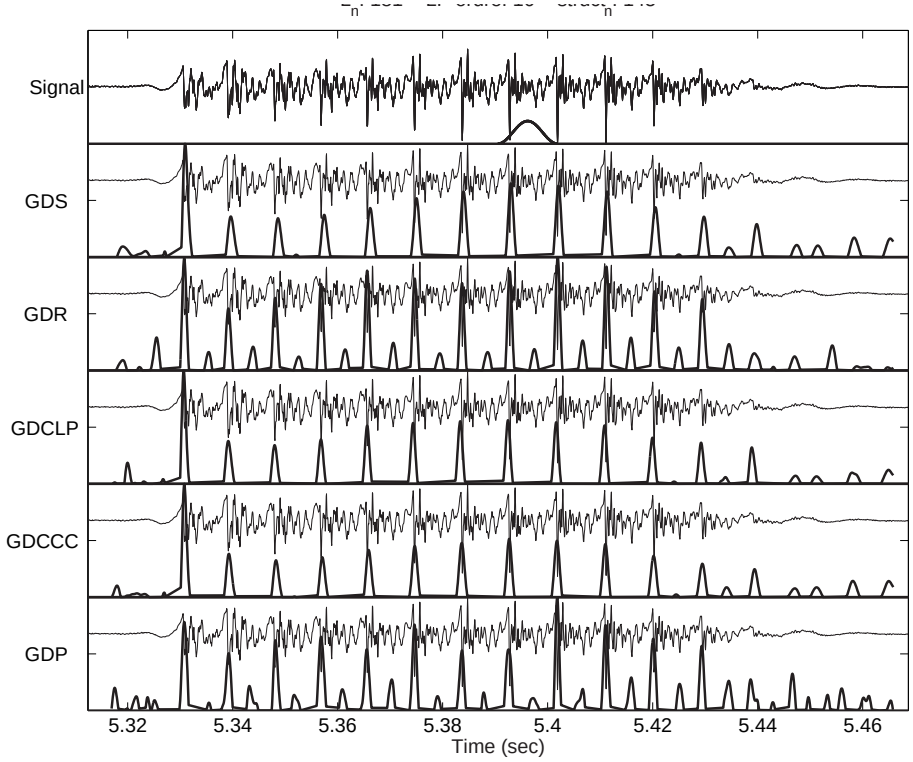


FIG. 3.28 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= speech-85000-87500

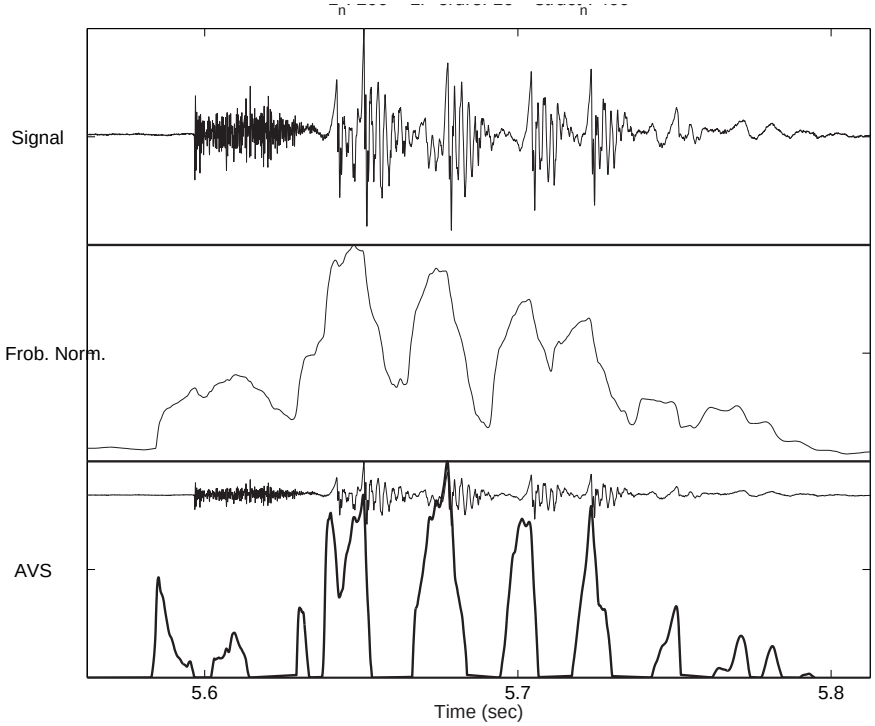


FIG. 3.29 – Méthode de la norme de Frobenius, Signal= speech-89000-93000

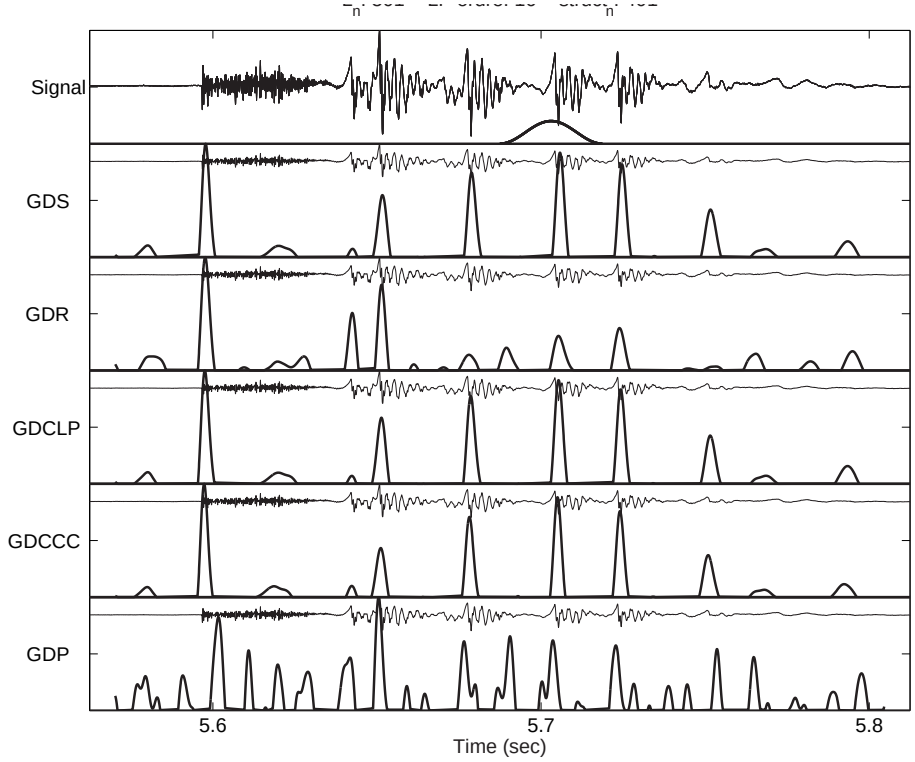


FIG. 3.30 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= speech-89000-93000

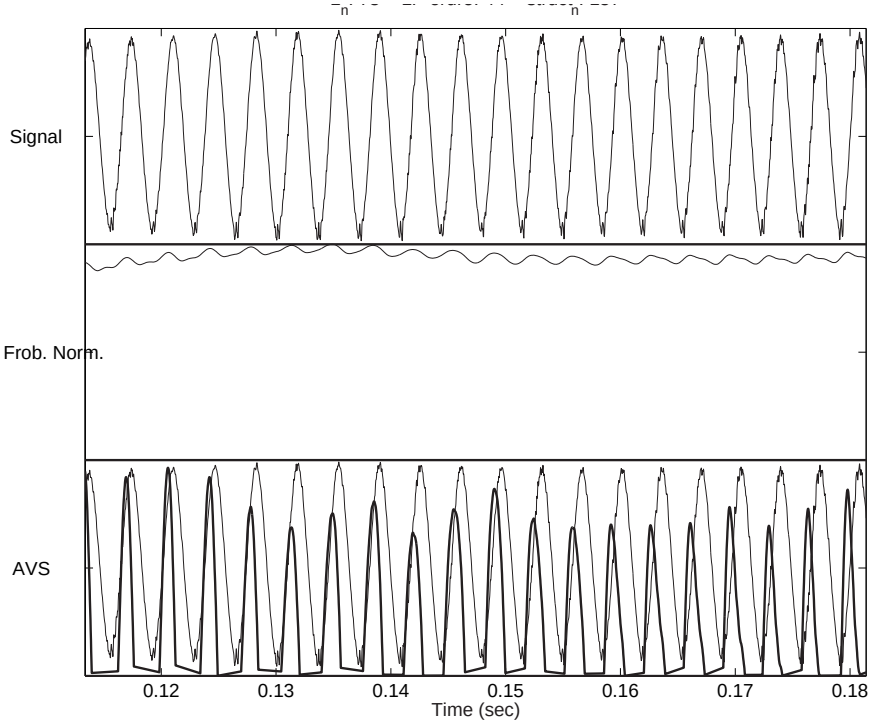


FIG. 3.31 – Méthode de la norme de Frobenius, Signal= vie

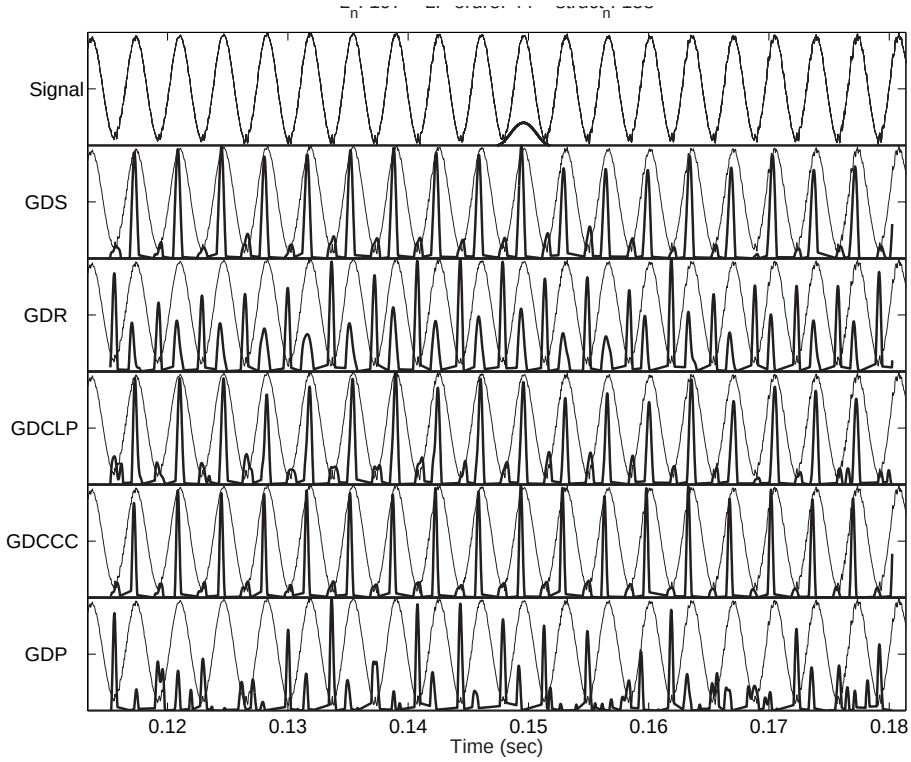


FIG. 3.32 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= vie

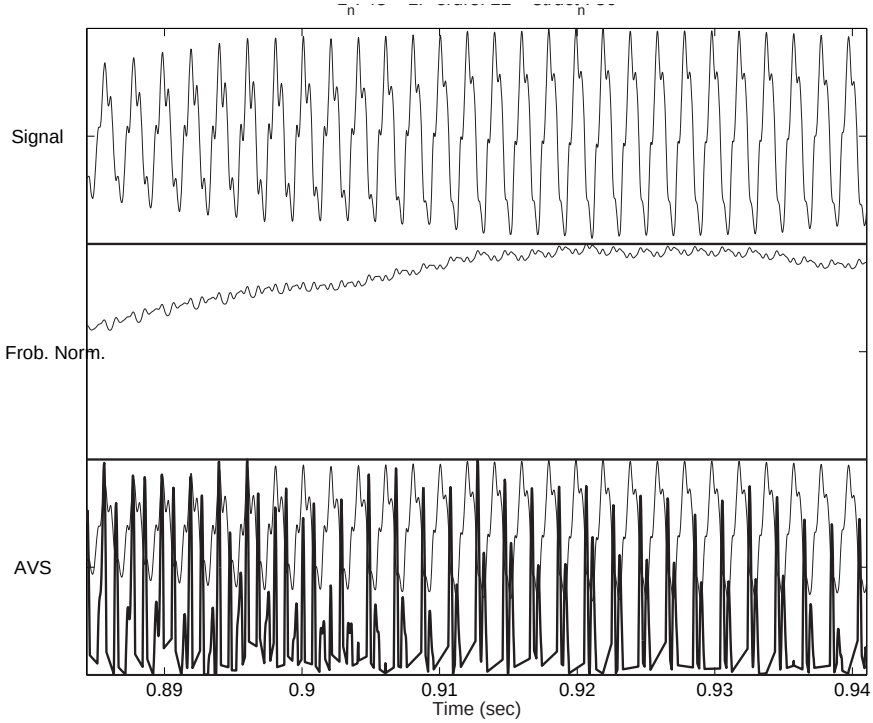


FIG. 3.33 – Méthode de la norme de Frobenius, Signal= venti-aux

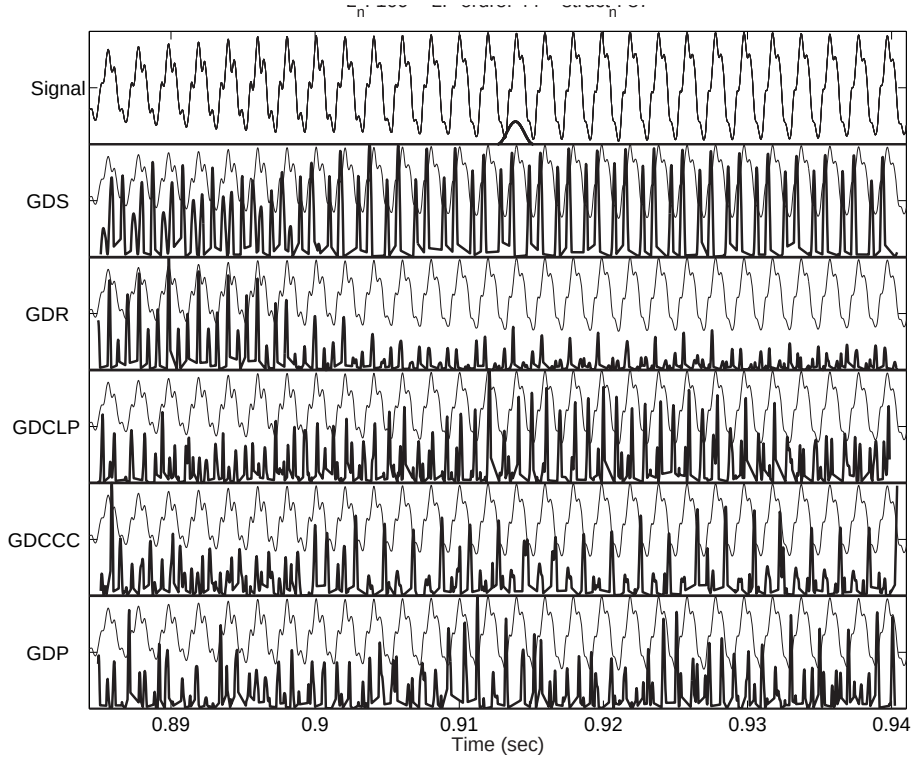


FIG. 3.34 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= venti-aux

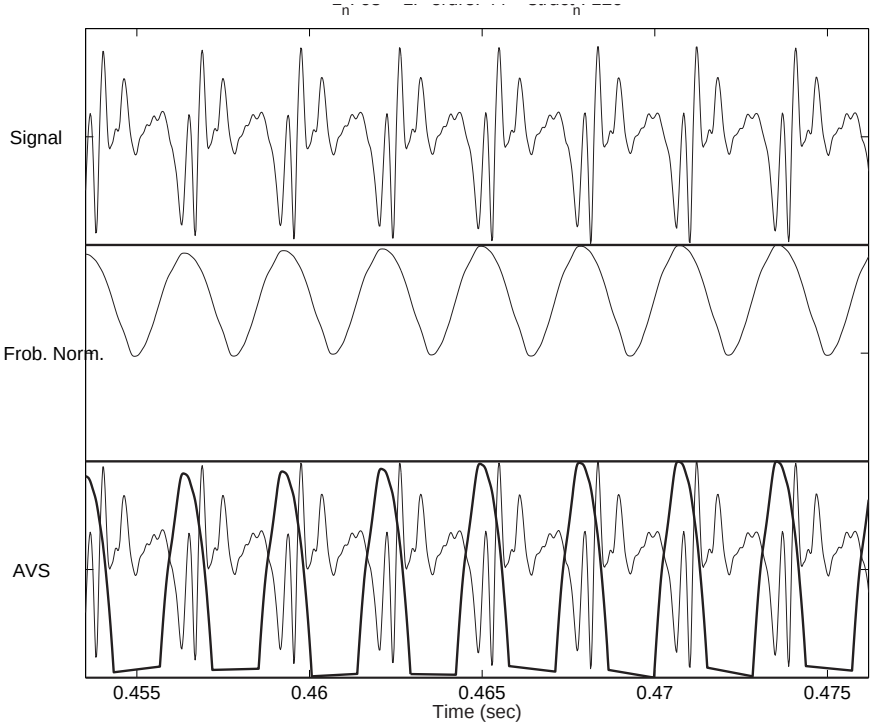


FIG. 3.35 – Méthode de la norme de Frobenius, Signal= trumpet

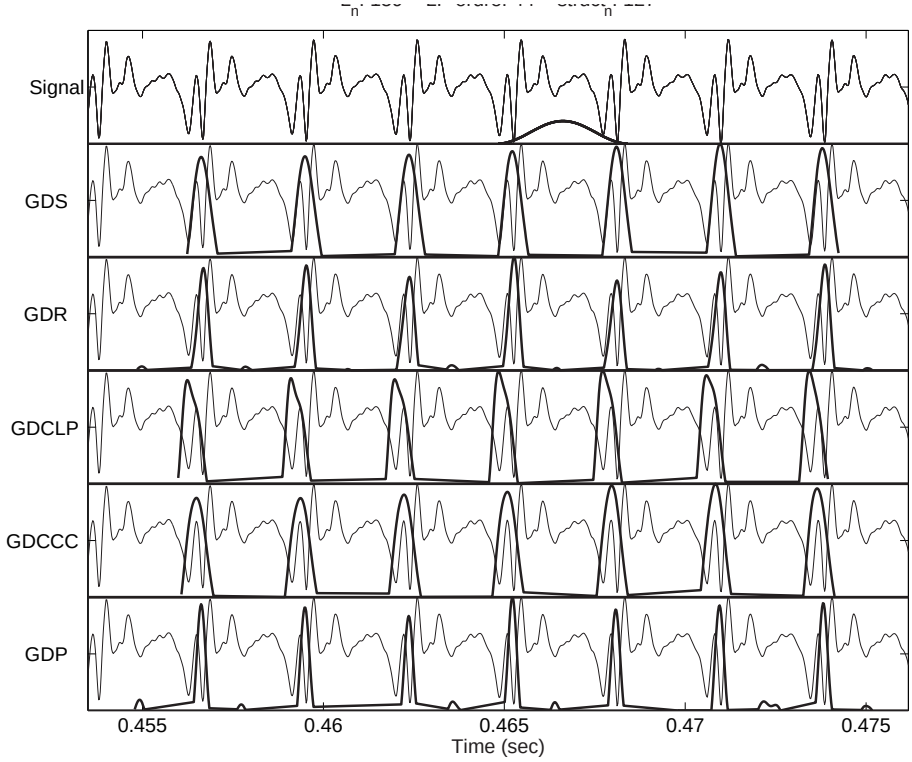


FIG. 3.36 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= trumpet

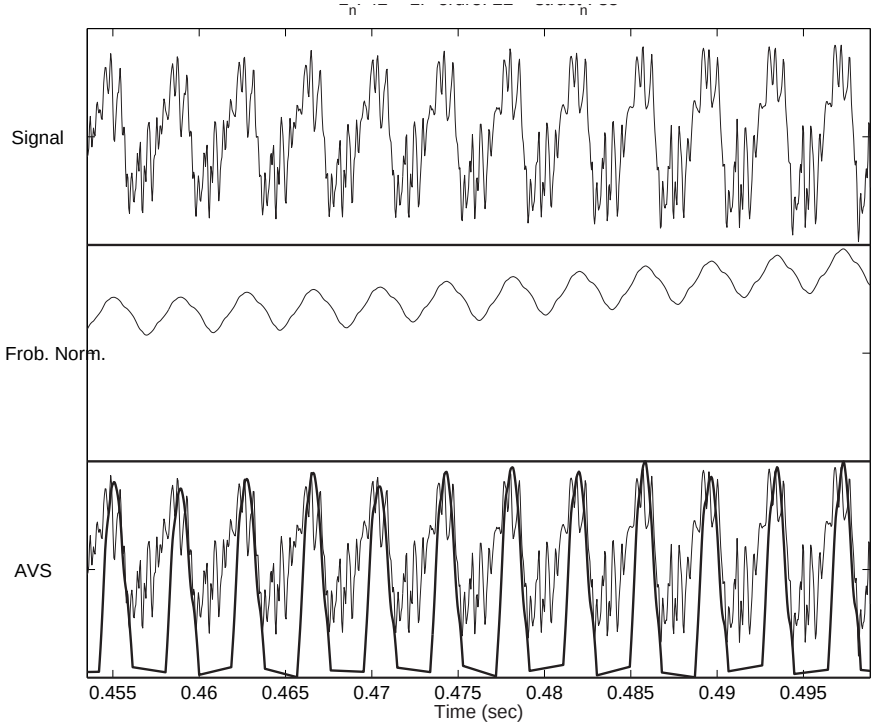


FIG. 3.37 – Méthode de la norme de Frobenius, Signal= violon

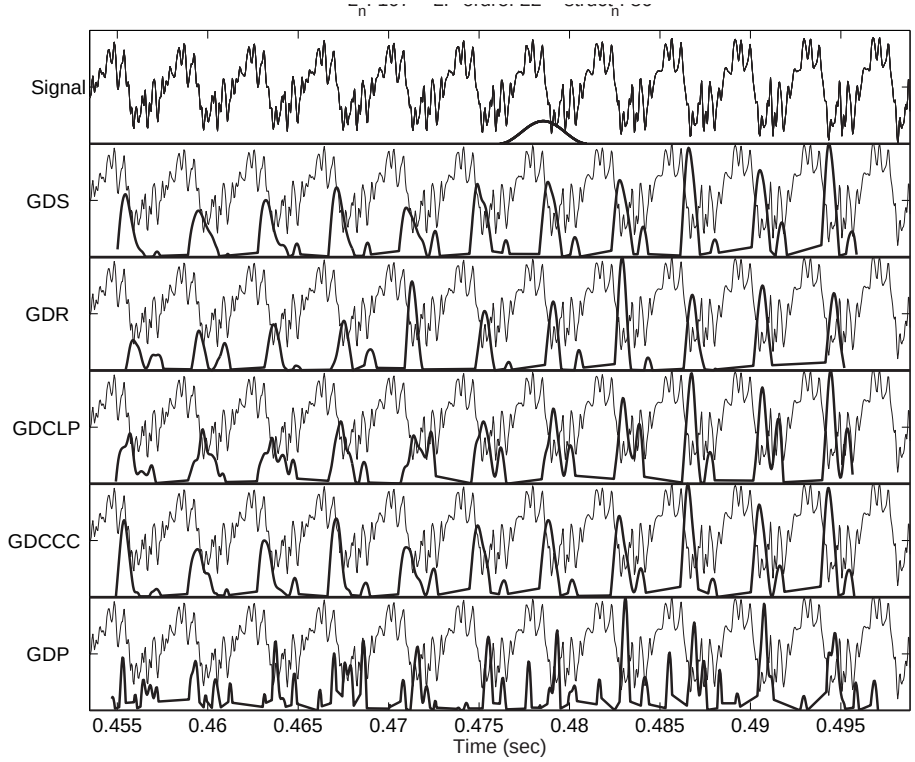


FIG. 3.38 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= violon

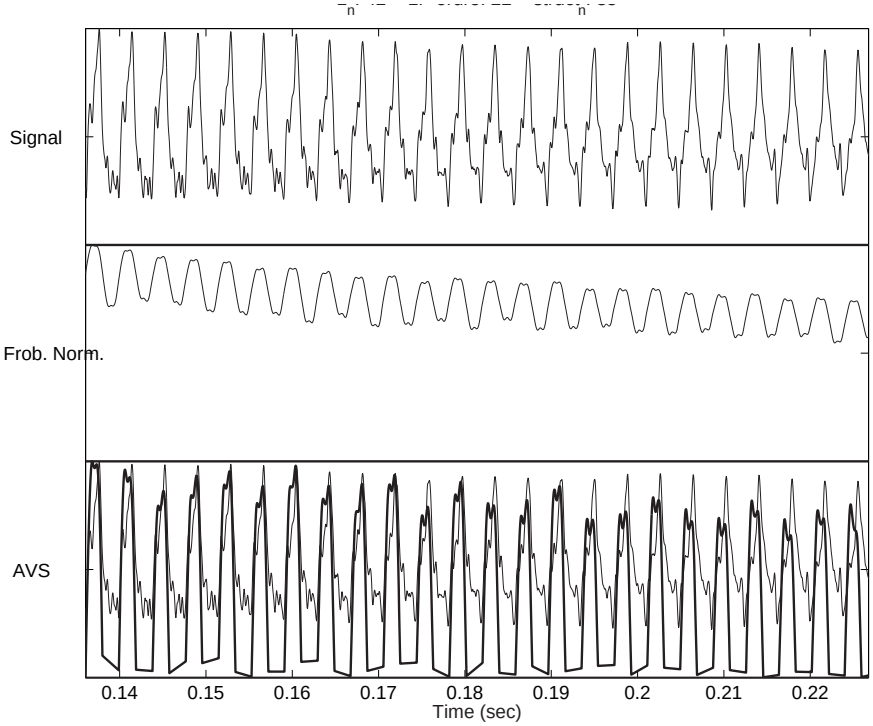


FIG. 3.39 – Méthode de la norme de Frobenius, Signal= piano

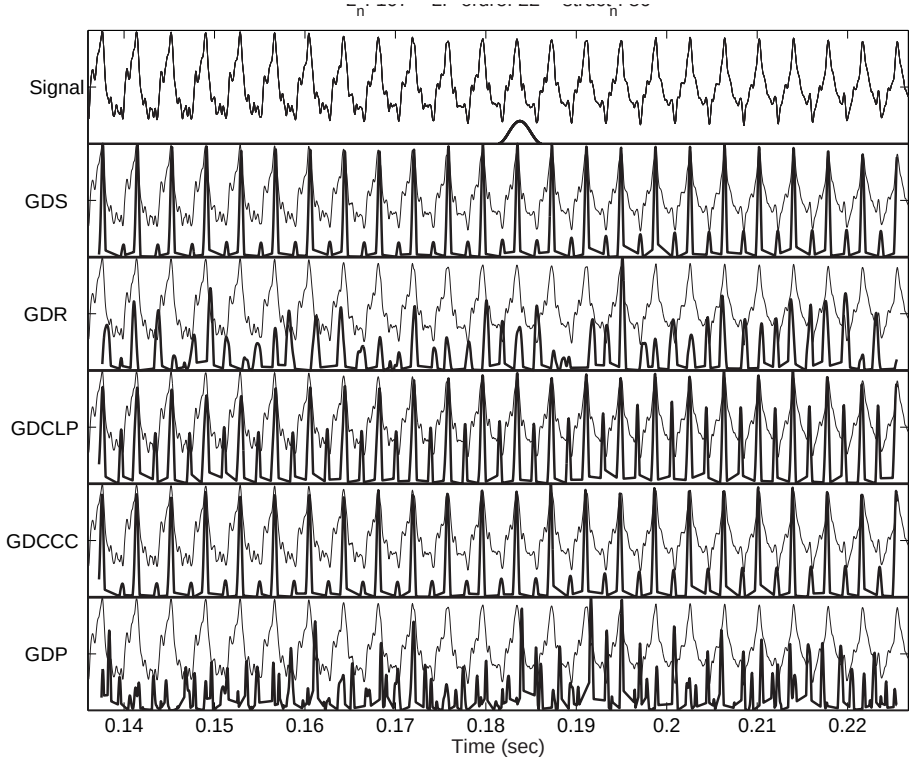


FIG. 3.40 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, Signal= piano

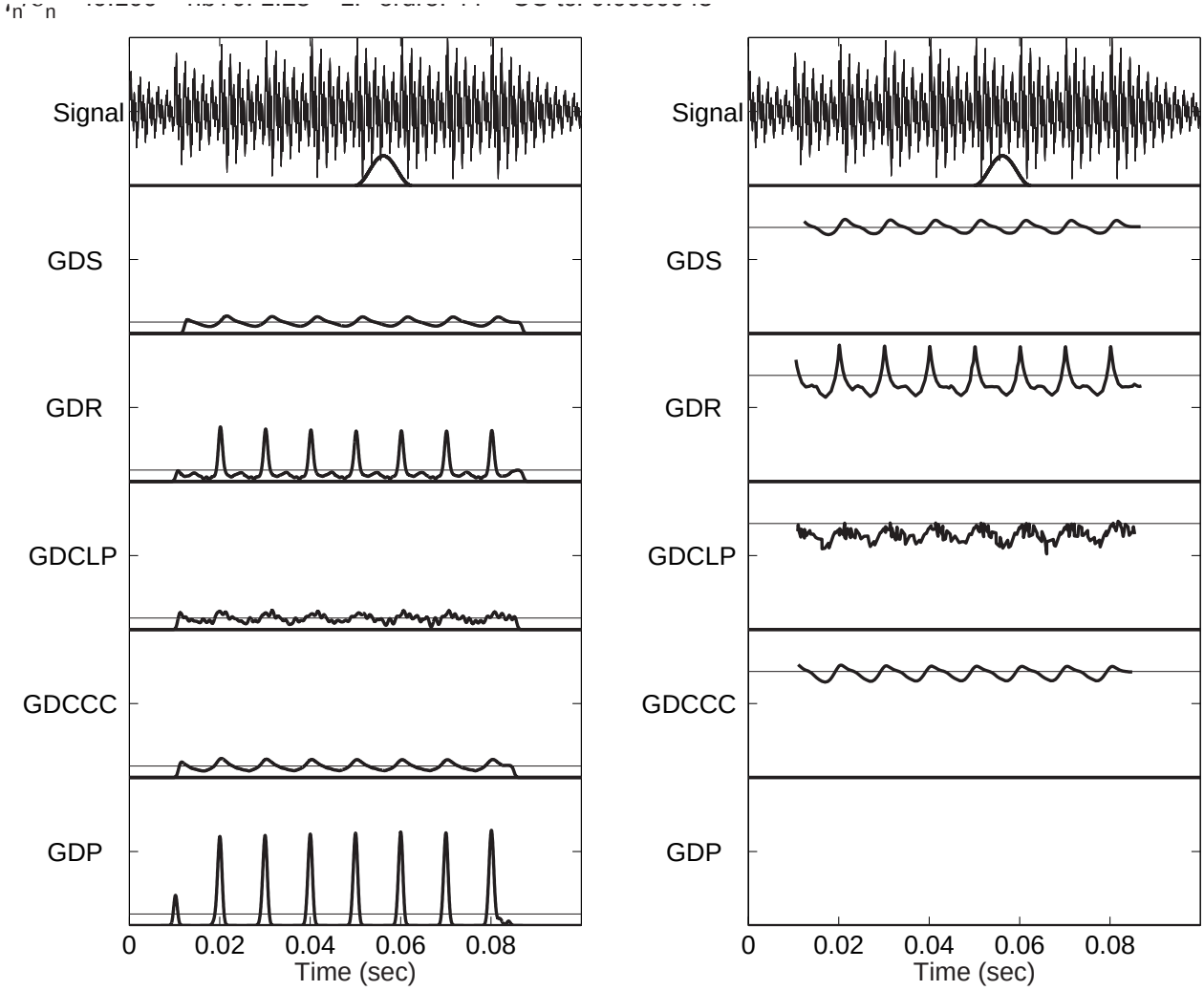


FIG. 3.41 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 0 Hz, [G] γ_n (trait épais), $\gamma_{h,n}$ (traits légers horizontaux) [D] σ_n (trait épais), $\sigma_{h,n}$ (traits légers horizontaux), Signal= formant-chant

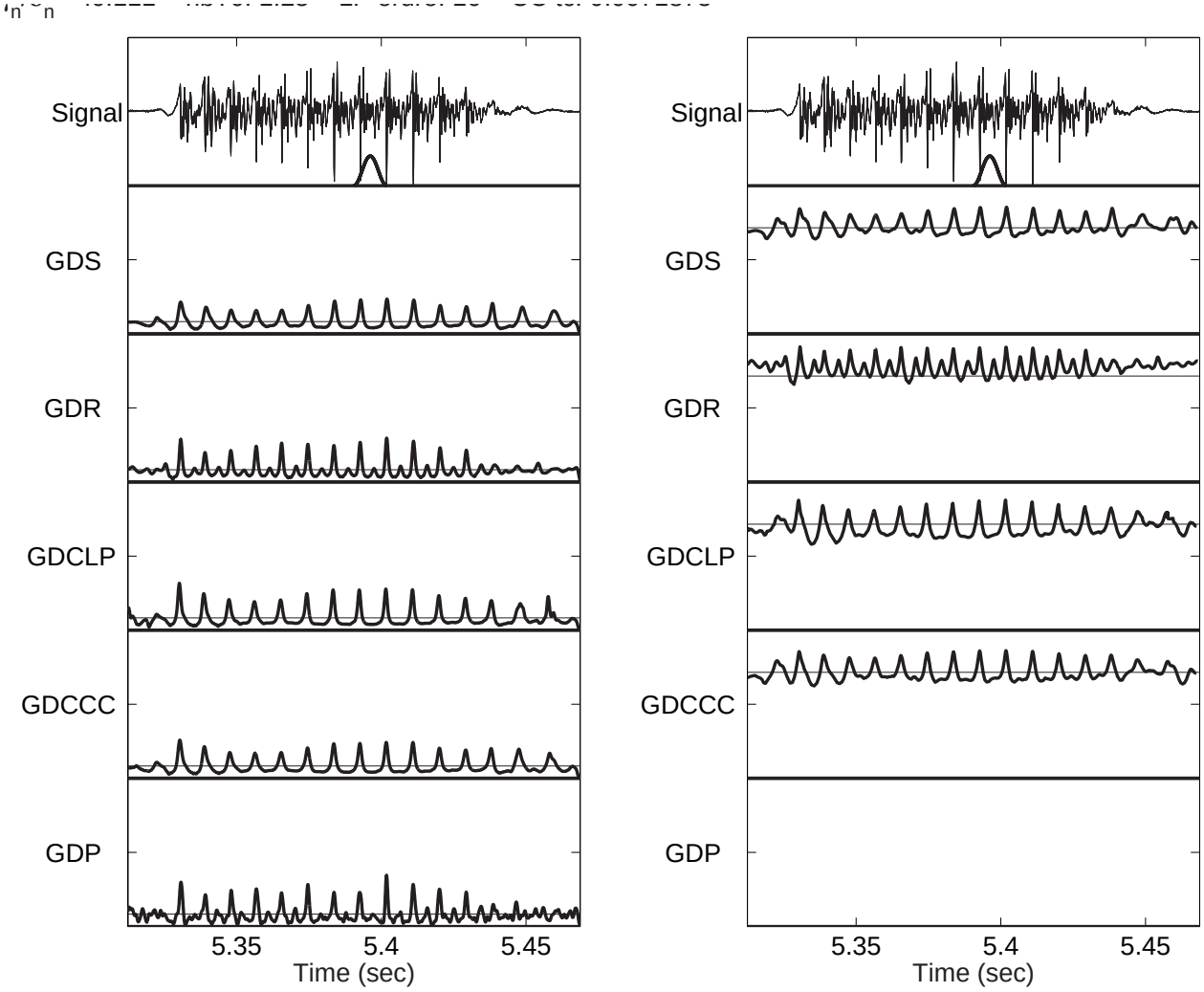


FIG. 3.42 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 0 Hz, [G] γ_n (trait épais), $\gamma_{h,n}$ (traits légers horizontaux) [D] σ_n (trait épais), $\sigma_{h,n}$ (traits légers horizontaux), Signal= speech-85000-87500

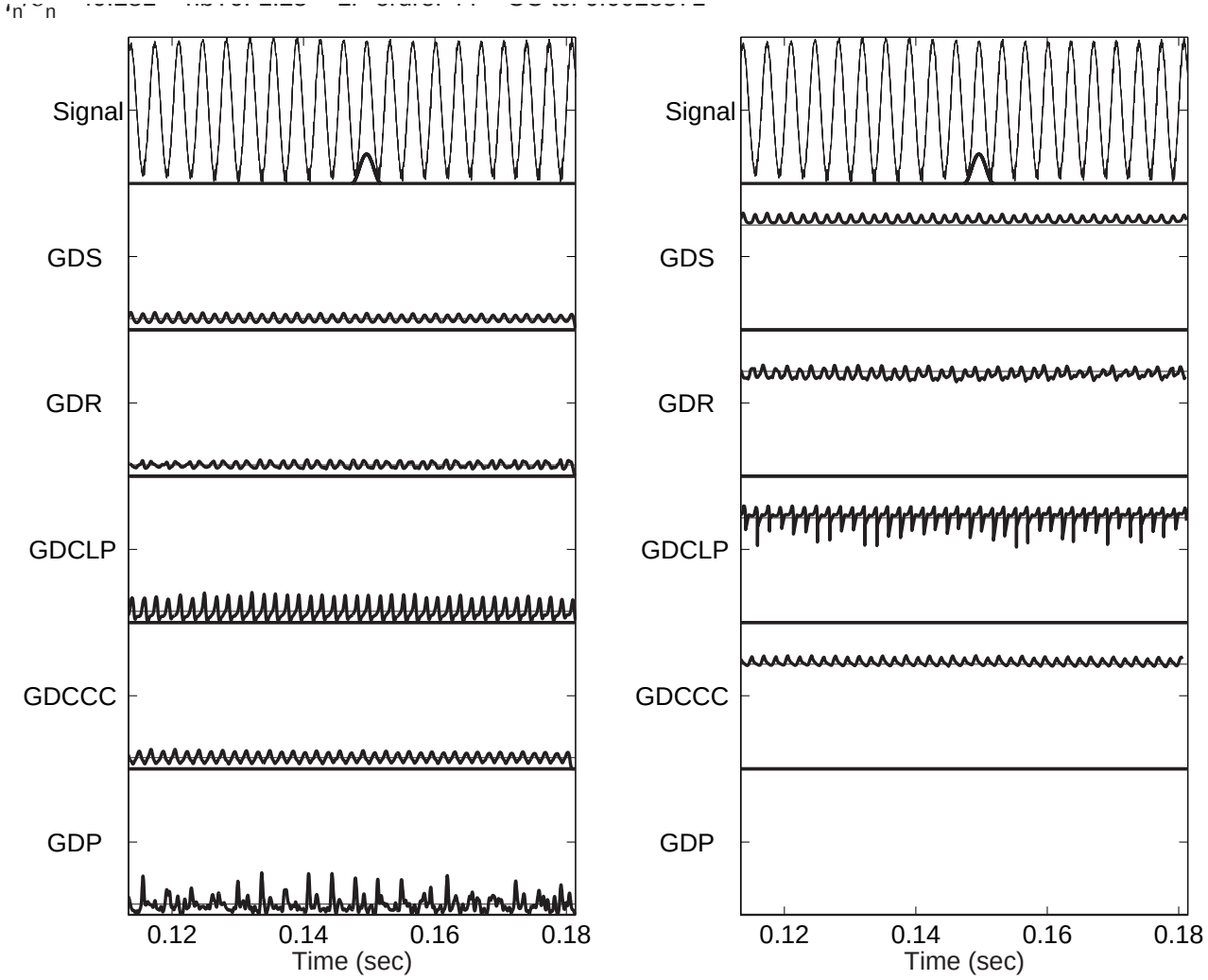


FIG. 3.43 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 0 Hz, [G] γ_n (trait épais), $\gamma_{h,n}$ (traits légers horizontaux) [D] σ_n (trait épais), $\sigma_{h,n}$ (traits légers horizontaux), Signal= vie

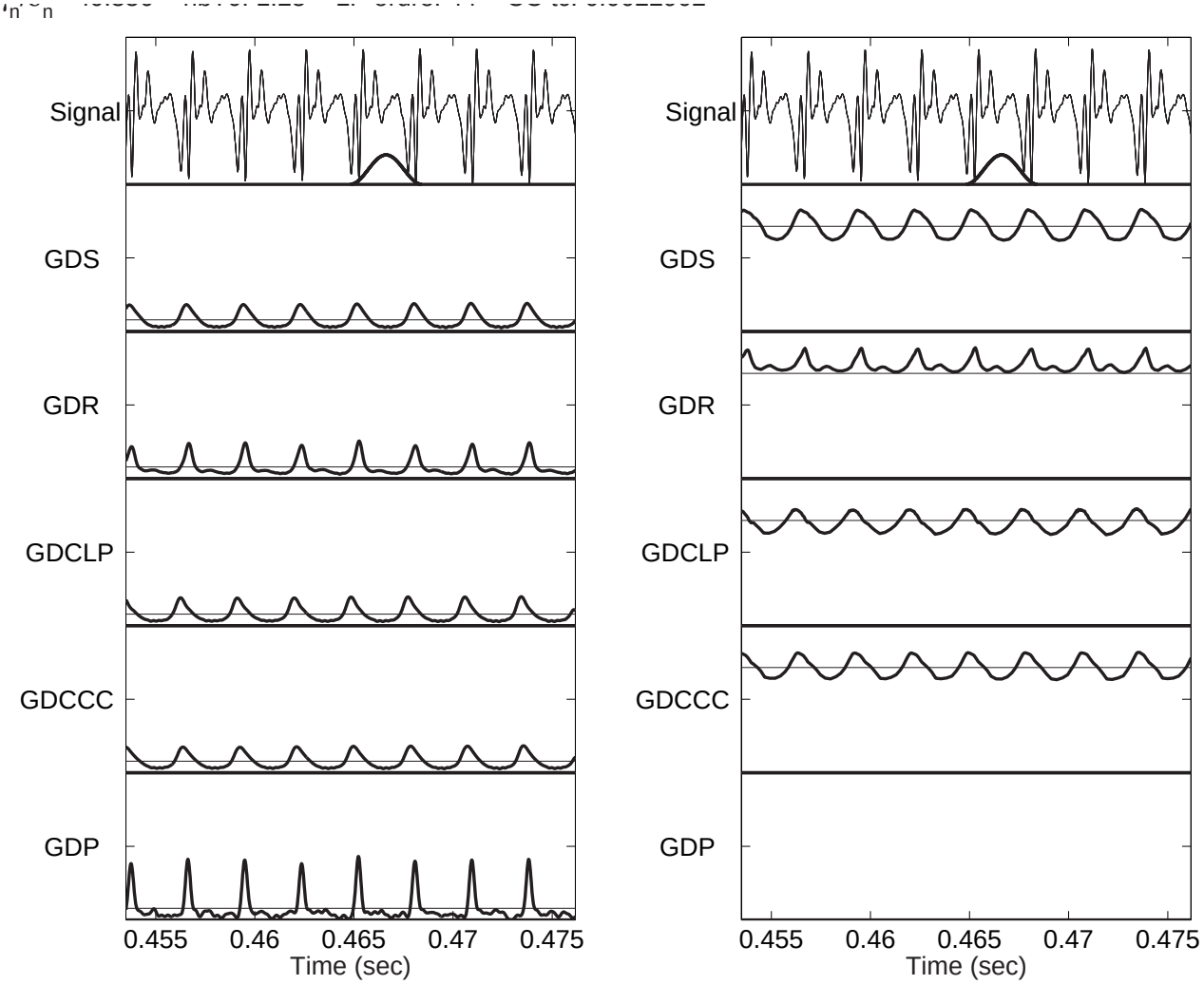


FIG. 3.44 – Méthodes du retard de groupe GDS, GDR, GDCLP, GDCCC, GDP, fréquence minimale de 0 Hz, [G] γ_n (trait épais), $\gamma_{h,n}$ (traits légers horizontaux) [D] σ_n (trait épais), $\sigma_{h,n}$ (traits légers horizontaux), Signal= trumpet

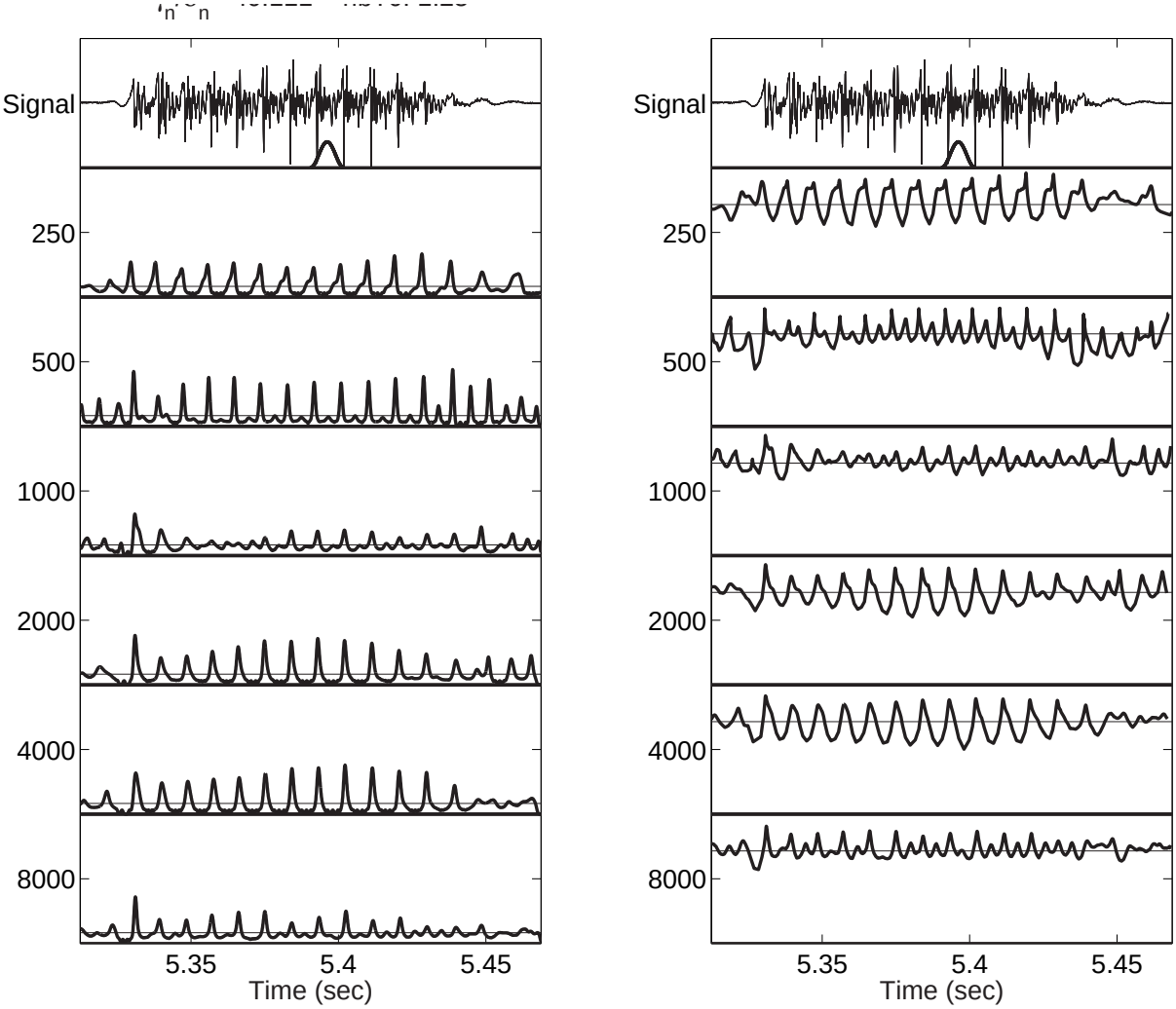


FIG. 3.45 – Méthode du retard de groupe GDS appliquée par bande d'octave : (1) Signal (2) [0,Fe/64] (3) [Fe/64,Fe/32] (4) [Fe/32,Fe/16] (5) [Fe/16,Fe/8] (6) [Fe/8,Fe/4] (7) [Fe/4,Fe/2], Signal : speech-85000-87500

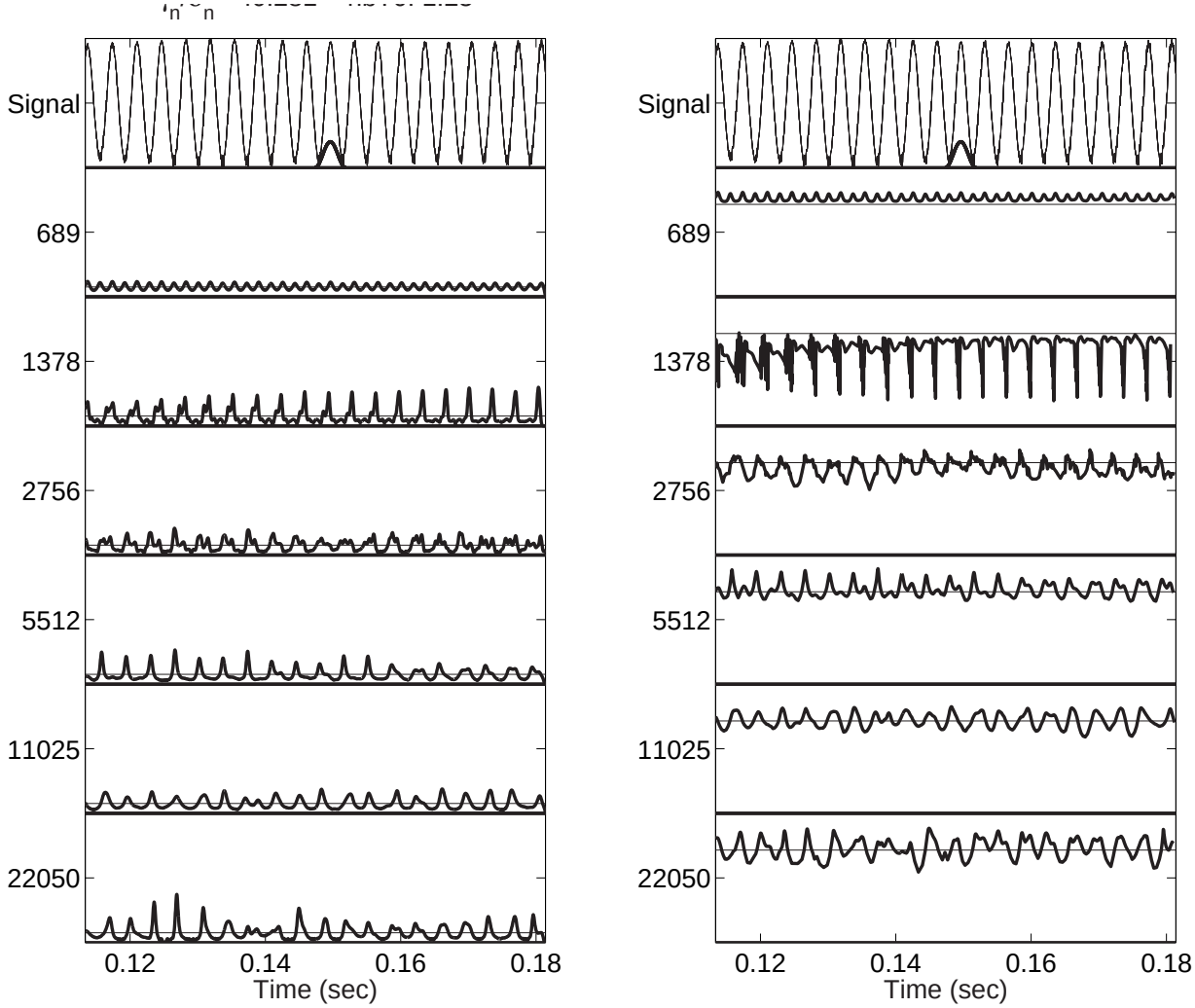


FIG. 3.46 – Méthode du retard de groupe GDS appliquée par bande d'octave : (1) Signal (2) [0,Fe/64] (3) [Fe/64,Fe/32] (4) [Fe/32,Fe/16] (5) [Fe/16,Fe/8] (6) [Fe/8,Fe/4] (7) [Fe/4,Fe/2], Signal : vie

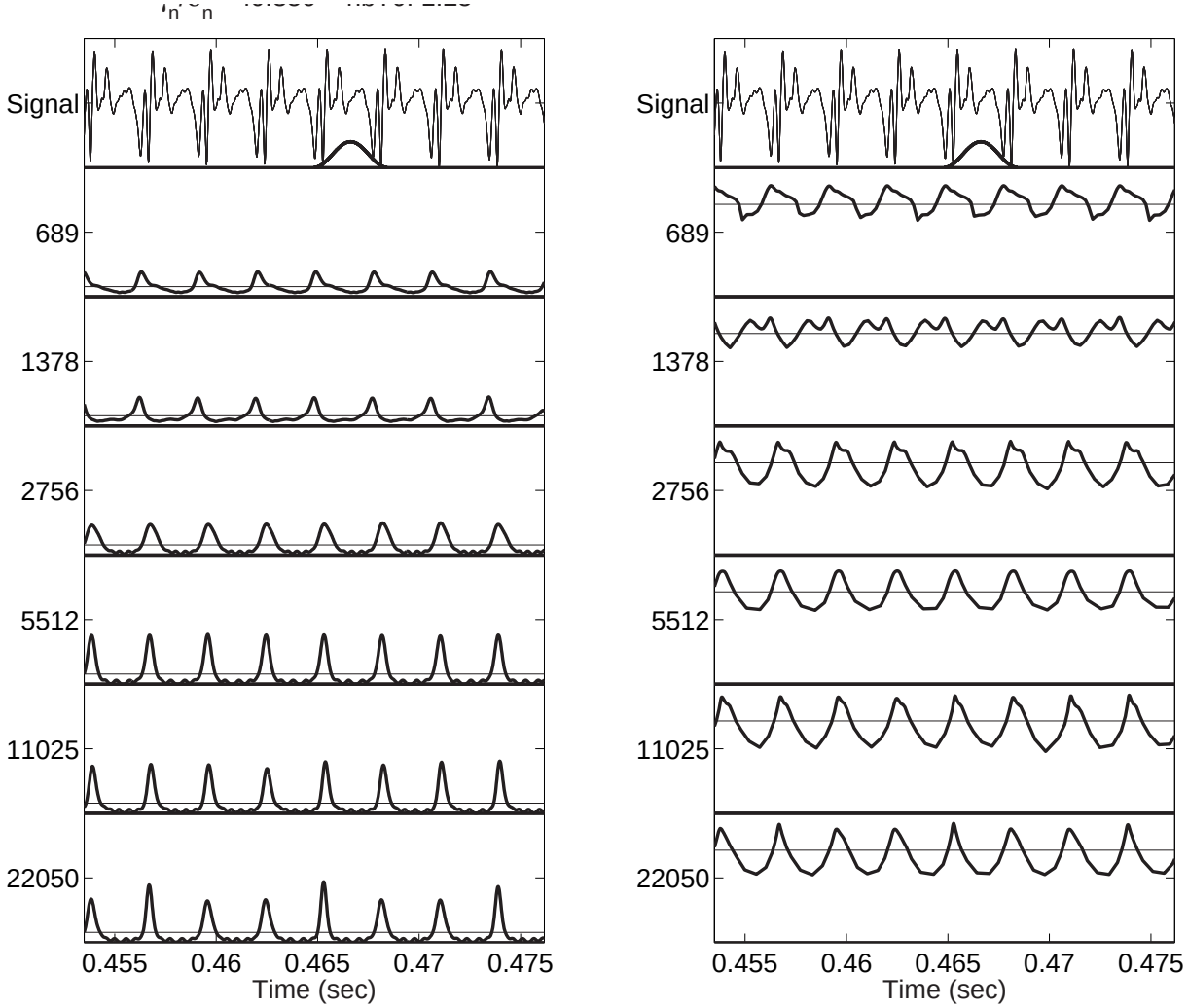


FIG. 3.47 – Méthode du retard de groupe GDS appliquée par bande d'octave : (1) Signal (2) [0,Fe/64] (3) [Fe/64,Fe/32] (4) [Fe/32,Fe/16] (5) [Fe/16,Fe/8] (6) [Fe/8,Fe/4] (7) [Fe/4,Fe/2], Signal : trumpet

Notes de bas de page relatives à la partie 3

1. De manière générale, tout signal périodique pouvant être modélisé sous forme d'un modèle source/filtre et dont la réponse impulsionnelle du filtre est courte par rapport à la période fondamentale (taux d'amortissement élevé) peut faire l'objet d'une décomposition en formes d'onde élémentaires .
2. Une méthode équivalente dans le domaine fréquentiel est proposée par [NMEL95]
3. La norme de Frobenius d'une matrice $\mathbf{S}$ est définie comme

$$|\mathbf{S}|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{p+1} s_{i,j}^2 \tag{3.5}$$

et peut se réécrire comme

$$|\mathbf{S}|_F^2 = Tr(\mathbf{S}^T\mathbf{S}) = \sum_{i=1}^{p+1} \sigma_i^2 \tag{3.6}$$

où Tr désigne la trace et σ_i sont les valeurs propres de la matrice $\mathbf{S}^T\mathbf{S}$.

4. Dans (3.3) la première colonne est considérée comme une combinaison linéaire des autres colonnes. Il est possible de trouver p autres estimateurs en considérant tour à tour chacune des p colonnes comme combinaison linéaire des autres. La famille de solutions ainsi obtenue est appelée famille LLS (Linear Least Squares).
5. l'hypothèse H1 est donc «les deux réalisations sont issues de processus différents»
6. Dans ce cas la durée de l'observation devient trop petite par rapport à l'ordre du modèle auto-régressif.
7. Communication personnelle de Hideki Kawahara.
8. Dans le cas d'un signal **symétrique**, [Kaw00] propose de mesurer l'influence de la fenêtre d'analyse sur l'estimation de μ_{ener} au travers du tracé de la valeur d'assignement μ_{ener} calculée aux instants t_m (centre de la fenêtre d'analyse). Dans le cas d'un signal $x(t)$ de type gaussien symétrique (approximation d'une forme d'onde élémentaire par une cloche) de paramètres (t_x, σ_x) et d'une fenêtre $h(t)$ de type gaussien de paramètres (t_m, σ_h), il est montré que

$$\mu_{ener} = \frac{\sigma_x^2 t_m + \sigma_h^2 t_x}{\sigma_x^2 + \sigma_h^2} \tag{3.17}$$

La valeur de μ_{ener} est donc d'autant moins influencée par la fenêtre que σ_x^2 est petit. Une valeur μ_{ener} égale à t_m et décroissante de part et d'autre de t_m est obtenue quand $\mu_{ener} = t_x$.

9. utilisant l'estimation (D.2) de $\tau_g(x, t_m, L, \omega)$
10. t_{ener} est définit comme $t_{ener} = \frac{\int_{t=-\infty}^{+\infty} t|x(t-t_m)|^2 dt}{\int_{t=-\infty}^{+\infty} |x(t-t_m)|^2 dt}$
11. À titre d'exemple, en l'absence de bruit et dans le cas le plus simple d'une fenêtre parabolique, nous obtenons

$$\mu_{ener} = \frac{\int_t t|s(t_m, t-t_m)|^2 dt}{\int_t |s(t_m, t-t_m)|^2 dt} = \frac{1}{\alpha} \frac{\frac{1}{2}(\alpha^2 L^2 + \frac{3}{2}\alpha L + \frac{3}{4})e^{2\alpha(t_0-L)} - \frac{1}{8}(4\alpha^3 t_0(t_0^2 - L^2) + 2\alpha^2(3t_0^2 - L^2) + 3(1 + 2\alpha t_0))}{\frac{1}{2}(\alpha L + \frac{1}{2})e^{2\alpha(t_0-L)} - \frac{1}{4}(2\alpha^2(t_0^2 - L^2) + (1 + 2\alpha t_0))} \tag{3.19}$$

12. Ce cumul peut être vu comme un filtrage passe-bas des valeurs $\mu(t_m)$. Cependant, du fait de l'obtention de valeurs $\mu(t_m)$ indépendantes du taux d'échantillonnage mais non régulièrement espacées, un filtrage passe-bas traditionnel est exclu.
13. Sans localisation relativement à la largeur L de la fenêtre.
14. Cette approximation de l'intégrale de la fonction de Gauss par une somme est du moins valable pour une densité de valeurs suffisante dans la largeur $[-3\sigma, 3\sigma]$. Pour cette raison un petit pas d'avancement I de l'analyse à fenêtre glissante doit être choisi. Le choix $I = \sigma$ conduit à des résultats satisfaisants.
15. La dérivée du spectre d'amplitude est calculée par différentielle locale

$$\sigma_A^2(s, t_m) = \frac{\sum_{\omega_k \in \Omega} \left(\frac{|S(t_m,\omega_k)| - |S(t_m,\omega_{k-1})|}{2\pi(\nu_k - \nu_{k-1})} \right)^2}{\sum_{\omega_k \in \Omega} |S(t_m, \omega_k)|^2} \tag{3.26}$$

16. La variance du retard de groupe est calculée comme le carré de la déviation standard en fréquence de $\tau_g(\omega)$ pondérée par l'énergie des composantes fréquentielles :

$$\sigma_\phi^2(s, t_m) = \frac{\sum_{\omega_k \in \Omega} \left(\tau_g(s, t_m, \omega_k) - \mu(s, t_m)\right)^2 |S(t_m, \omega_k)|^2}{\sum_{\omega_k} |S(t_m, \omega_k)|^2} \tag{3.27}$$

17. Puisque les estimations de μ_{pente} et μ_{min} dépendent uniquement du retard de groupe du signal et non de son spectre d'amplitude, il n'est pas nécessaire de déconvoluer le signal, mais seulement de soustraire du retard de groupe du signal le retard de groupe du filtre du système.

$$\begin{aligned} s(n) &= e(n) \otimes v(n) \\ S(\omega) &= E(\omega) \cdot V(\omega) \\ \tau_g(S(\omega)) &= \tau_g(E(\omega)) + \tau_g(V(\omega)) \end{aligned} \tag{3.31}$$

18. Signal test composé de 4 résonances aux fréquences 500, 800, 3000 et 4000 Hz
19. Un bruit blanc gaussien d'écart type $\sigma = 0.2$ a été ajouté au signal test
20. Remarquons qu'en considérant le signal résiduel comme un signal entièrement blanchi ($|E(\omega)| = const$) nous pouvons théoriquement négliger la pondération par l'énergie du spectre. Dans ce cas, nous retrouvons la méthode proposée par [YdD98]. En pratique, nous avons constaté de meilleurs résultats lorsque cette pondération est utilisée.
21. [SY95] utilise une taille de fenêtre inférieure à $\overline{\text{T0}}$
22. Communication personnelle du Professeur Hideki Kawahara.
23. Signal composé de trois formants aux fréquences 500, 100 et 3000 Hz et de taux d'atténuation faible
24. Pour rappel le retard de phase est défini comme

$$\tau_\phi(\omega) = -\frac{\phi(\omega)}{\omega} \tag{3.34}$$

$\tau_\phi(\omega)$ est la distance temporelle entre le centre de la fenêtre d'analyse et le maximum du cosinus à la fréquence ω.
25. D'après [Coh95], la largeur de bande est définie par

$$BW^2 = \int \left(\frac{\partial |X(\omega)|}{\partial \omega} \right)^2 dt + \int \left(\frac{\partial \phi(t)}{\partial \omega} - \langle \omega \rangle |X(t)^2| \right) dt \tag{3.35}$$

Chapitre 4

Sinusoïdalité

4.1 Introduction

Dans ce chapitre, nous étudions la représentation d'un signal par une somme de sinusoïde. L'adéquation de la représentation d'une région du plan temps/ fréquence d'un signal par une sinusoïde est généralement désignée par le terme «sinusoïdalité» de cette région. La sinusoïdalité d'une région dépend non seulement des propriétés du signal, mais également des propriétés et des contraintes imposées à la sinusoïde : stationnarité locale (en temps) des paramètres de la sinusoïde, non-stationnarités, modulation d'amplitude (pour la représentation du bruit), ... Chacune de ces contraintes réfère la sinusoïde et donc la «sinusoïdalité» à un «modèle» de sinusoïdes ou «modèle sinusoïdal».

4.1.1 Le modèle sinusoïdal

Sous sa forme la plus générale, le modèle sinusoïdal représente un signal $s(t)$ comme une somme de $H(t)$ sinusoïdes d'amplitudes $A_h(t)$, et de phases $\phi_h(t)$ variables au cours du temps.

$$s(t) = \sum_{h=1}^{H(t)} A_h(t) \cos(\phi_h(t)) \tag{4.1}$$

Chacune des h sinusoïdes représente une région du plan temps/fréquence.

La représentation d'un signal par le modèle (4.1) nécessite l'estimation des paramètres $A_h(t)$ et $\phi_h(t)$ à chaque instant t et pour chaque région h du signal. Afin de rendre cette estimation plus aisée, deux hypothèses sont généralement faites :

Hypothèse 1 : Le signal est supposé évoluer lentement dans le temps et donc les paramètres du modèle correspondant sont supposés varier lentement au cours du temps (signaux passe-bas). Ceci permet de discrétiser leur estimation en des instants disjoints $t_m, t_{m+1}, ...$ [1] . Cependant $\phi_h(t)$ ne peut être considéré comme un paramètre à variation lente, puisque sa dérivée temporelle dépend de la position de la composante h sur l'axe des fréquences. La définition de «variation lente» est donc revue et corrigée par «variation lente relativement à la position sur l'axe des fréquences». $\phi_h(t)$ peut être réécrit comme la contribution de trois termes : un terme dépendant de la position de

h sur l'axe des fréquence, $\omega_h \cdot t$, un terme d'initialisation, $\phi_{0,h}$, et un terme de perturbation, $\phi_{\delta,h}(t)$. Le terme d'initialisation n'intervenant qu'à la première estimation, et les termes de perturbation qu'après l'initialisation nous pouvons combiner ces deux termes en un terme unique $\phi_{0,h}(t_m)$ dans lequel t_m désigne le temps d'observation du signal. Nous pouvons donc écrire $\phi_h(t) = \omega_h \cdot t + \phi_{0,h}$. Les paramètres ω_h et $\phi_{0,h}$ sont maintenant définis comme des paramètres à variation lente.

Cette hypothèse est souvent utilisée dans la littérature pour faire localement en temps une approximation d'ordre 0 (hypothèse de stationnarité locale) des paramètres du modèle, bien que celle-ci ne soit pas justifiée en regard des propriétés du signal. Autour de l'instant t_m, les paramètres $A_h(t)$, $\omega_h(t)$ et $\phi_{0,h}(t)$ sont considérés comme constants. Ils sont dés lors notés $A_{h,m}$, $\omega_{h,m}$ et $\phi_{0,h,m}$.

Hypothèse 2 : Le nombre de sinusoïdes à un instant donné, $H(t)$, est limité.

Le modèle sinusoïdal sous sa forme simplifiée $\hat{s}(t)$ ne constitue plus qu'une approximation du signal $s(t)$:

$$\begin{cases} \hat{s}(t) = \sum_{m} \hat{s}_m(t) & \\ \hat{s}_m(t) = \sum_{h=1}^{H(t_m)} \hat{s}_{m,h}(t) & \\ \hat{s}_{m,h}(t) = A_{h,m} \cos\left((t - t_m)\omega_{h,m} + \phi_{0,h,m}\right) & \text{si } t \in [t_m - L/2, t_m + L/2] \\ \qquad = 0 & \text{sinon} \end{cases} \tag{4.2}$$

4.1.2 Estimation du modèle sinusoïdal

La représentation d'un signal par un modèle sinusoïdal consiste dans un premier temps à déterminer quelles régions du plan temps/ fréquence du signal peuvent être représentées par le modèle sinusoïdal considéré. Si le modèle considéré est (4.2), il s'agit de déterminer quelles régions du plan temps/ fréquence peuvent être représentées par des sinusoïdes d'amplitude et de fréquence constante sur la durée de l'observation.

Dans un second temps, les paramètres du modèle sont estimés pour chacune de ces régions. Si le modèle considéré est (4.2), il s'agit d'estimer les paramètres $A_{h,m}$, $\omega_{h,m}$ et $\phi_{0,h,m}$.

Dans la suite, nous désignons la première étape par le terme «détection» des composantes sinusoïdales, en gardant en tête qu'il ne s'agit pas de rechercher des composantes sinusoïdales dans le signal, mais de rechercher les régions du plan temps/ fréquence pouvant être modélisées par des sinusoïdes. Nous désignons la seconde étape par le terme «estimation» des paramètres du modèle sinusoïdal.

Distinction Détection/ Estimation Il est facile de comprendre que les étapes de «détection» et d'«estimation» ne sont pas indépendantes. En effet, la détection consistant à mesurer dans quelles mesures une région du plan temps/fréquence peut être modélisée par un modèle sinusoïdal, nous devons connaître les paramètres de ce modèle pour mesurer cette adéquation. La détection passe donc également par une étape d'estimation. Dans le cas du modèle (4.2), la détection peut se faire en mesurant l'erreur commise en modélisant une région du plan temps/ fréquence par une sinusoïde d'amplitude, fréquence et phase localement constantes. Le calcul de cette erreur de modélisation nécessite la connaissance préalable de $\omega_{h,m}, A_{h,m}, \phi_{0,h,m}$.

L'estimation des paramètres du modèle sinusoïdal peut se faire pour toutes les régions du plan temps/ fréquence (même pour les régions ne comportant pas de composantes sinusoïdales). Cependant, leur interprétation en tant que paramètres d'un modèle sinusoïdal n'a de sens que pour les régions comportant des composantes sinusoïdales. L'estimation, ou du moins son interprétation, dépend donc d'une détection.

Ceci montre bien la dépendance entre la détection et l'estimation. Dans la suite de ce chapitre, les problèmes de détection et d'estimation sont séparés non pas en fonction des outils utilisés, qui sont généralement les mêmes, mais en fonction des finalités poursuivies. Dans le cas de la détection, les estimateurs sont étudiés en fonction de leur pouvoir discriminant sinusoïde/ non-sinusoïde. Dans le cas de l'estimation, les estimateurs sont étudiés en fonction de leur biais et de leur variance.

⇒ Cette distinction détection/ estimation constitue la base de l'organisation de ce chapitre. Dans la première partie, nous étudions le problème de l'estimation, dans la seconde celui de la détection.

Dualité Temps/ Fréquence Un modèle sinusoïdal de paramètres localement stationnaires, tel que (4.2), n'est en théorie valable que pour un signal dont les propriétés ne changent pas au cours du temps. Néanmoins, il constitue une approximation d'autant plus justifiée que les propriétés du signal évoluent lentement et que la détection/ estimation s'effectue sur une durée courte. Dans ce cas, la stationnarité des paramètres du modèle peut se comprendre comme l'ordre 0 du développement en série limité de la variation des paramètres. La diminution de la durée d'observation se fait cependant au détriment de la résolution fréquentielle (dualité temps/ fréquence) et donc d'un recouvrement potentiel des composantes spectrales. La réduction de résolution fréquentielle peut être contournée en considérant l'ensemble des composantes du spectre simultanément, ce qui permet de tenir compte des recouvrements spectraux.

⇒ Ceci constitue la seconde distinction adoptée dans ce chapitre : prise en compte du recouvrement fréquentiel (résolution globale en fréquence), non-prise en compte (résolution locale en fréquence).

Modèles sinusoïdaux à paramètres non-stationnaires La réduction de la durée de l'observation améliore l'approximation à l'ordre 0 des paramètres du modèle. Une autre possibilité est de prendre en compte les non-stationnarités des paramètres du modèle dans la détection et l'estimation. Nous étudions le passage d'un modèle sinusoïdal d'ordre 0 à un ordre 1 (fréquence et amplitude à variation linéaire sur la durée de l'observation).

⇒ Ceci constitue la troisième distinction adoptée dans ce chapitre : modèle sinusoïdal stationnaire ou non-stationnaire.

L'extension du modèle sinusoïdal pose cependant le problème d'une définition de la mesure de sinusoïdalité. Dans le cas d'un *modèle à paramètres localement stationnaires*, la sinusoïdalité peut être mesurée comme l'erreur ϵ_m commise en modélisant à un instant donné une région du plan temps/ fréquence par UNE sinusoïde (pour un modèle stationnaire il n'existe qu'un modèle sinusoïdal ajusté au spectre par translation en fréquence, en amplitude et en phase). Dans le cas d'un *modèle à paramètres non-stationnaires*, l'erreur ϵ_m est l'erreur commise en modélisant une partie du plan temps/ fréquence par UNE sinusoïde parmi une infinité de sinusoïdes (chacune correspondant à une combinaison de modulations de fréquence et d'amplitude) . Dans ce cas, seule la cohérence de l'estimation des paramètres du modèle

Modèles sinusoïdaux à paramètres constants		
Résolution locale en fréquence (partie 4.2.1.1)		
spectre →	*Maximum local*	→ ω_k
ω_k →	*Interpolation/Régression spectrale*	→ ω_h, A_h
ω_k →	*Fréquence Instantanée*	→ ω_h
ω_h →	*Moindres carrés*	→ A_h, ϕ_h
Résolution globale en fréquence (partie 4.2.1.2)		
$\forall h \in H : \omega_h \rightarrow$	*Moindres carrés*	→ $\forall h \in H : A_h$, ϕ_h
$\forall k \in K : \omega_k \rightarrow$	*Moindres carrés itératifs*	→ $\forall h \in H : \omega_h$, A_h, ϕ_h

Modèles sinusoïdaux à paramètres variables		
Résolution locale en fréquence (partie 4.2.1.2)		
ω_k →	*Mesure de distorsion du spectre*	→ ω_h (+variation), A_h (+variation), phase ϕ_h
Résolution globale en fréquence (partie 4.2.2.4)		
$\forall k \in K : \omega_k \rightarrow$	*Polynôme d'amplitude complexe (Moindres carrés)*	→ $\forall h \in H : \omega_h$ (+variation), A_h (+variation), ϕ_h

TAB. 4.1 – Estimateurs des paramètres d'un modèle sinusoïdal : ω_k fréquence d'un pic, ω_h fréquence d'une sinusoïde, A_h amplitude d'une sinusoïde, ϕ_h phase d'une sinusoïde, h indice des composantes sinusoïdales, H ensemble des composantes sinusoïdales, k point de la TFDCT, K ensemble des points de la TFDCT

au cours du temps nous fournit une information sur la qualité de la modélisation d'une partie du signal par une sinusoïde.

Remarque : *L'approche étudiée dans ce chapitre pour l'étude des estimateurs des modèles sinusoïdaux est celle généralement désignée sous le nom d'approche par trame («frames»). Notons cependant que d'autres approches, différentes de celle-ci, ont récemment été proposées. Ces approches utilisent la globalité du signal afin de maximiser globalement l'adéquation d'un modèle. Ces méthodes tirent bénéfice du fait qu'en traitement de signal de qualité, les analyses en temps-réel ne sont pas toujours requises et donc qu'un traitement causal (causal au sens que l'estimation des paramètres ne dépend pas du futur du signal) n'est pas requis. Parmi ces approches, signalons celles de [AKI95], dans laquelle le signal est suivi en temps à l'aide de filtres passe-bande, les paramètres étant calculés dans chaque bande de fréquence de manière isolée ; l'approche de [Cor99] rajoute à la précédente une analyse du signal à la fois causale et anti-causale.*

4.2 Estimation des paramètres du modèle sinusoïdal

Dans cette partie, nous étudions l'estimation des paramètres des modèles sinusoïdaux. Cette partie est organisée de la manière suivante : les estimateurs sont divisés :

- premièrement, selon le modèle sinusoïdal auquel ils appartiennent : modèle sinusoïdal stationnaire ou non-stationnaire.
- deuxièmement, dans chacune de ces catégories, selon que l'estimation s'effectue en fréquence de manière locale ou globale

L'ensemble des estimateurs étudiés sont regroupés de cette manière à la TAB. 4.1. **Remarque** : *Dans la suite de cette partie, nous appelons :*

- *«finesse fréquentielle» : la distance fréquentielle entre deux «bins» de la TFDCT.*
- *«résolution fréquentielle» : le pouvoir de séparation de deux lignes spectrales pour des conditions d'analyse données.*

Nous ne considérons que des analyses à temps fixé ; aussi les indices t_m ou m sont-ils omis.

4.2.1 Estimateur des modèles sinusoïdaux stationnaires

La représentation d'un signal par le modèle sinusoïdal de paramètres localement stationnaires (4.2) nécessite l'estimation des paramètres $[A_{h,m}, \omega_{h,m}, \phi_{0,h,m}]$ pour chaque région fréquentielle h considérée comme représentable par le modèle (4.2).

Si nous connaissons les fréquences ω_h auxquelles se trouvent les composantes à estimer, le travail est grandement simplifié et se résume à estimer A_h et $\phi_{0,m}$, ce qui peut se faire par minimisation de l'erreur quadratique globale de modélisation ϵ_h. En l'absence d'information relative à ces fréquences, l'estimation commence généralement par une estimation locale en fréquence sur le spectre du signal.

4.2.1.1 Estimation locale en fréquence

◊ ***Détection des pics du spectre***

«Peak Picking» ou détection des maxima locaux du spectre d'amplitude La méthode du «peak picking» [MQ86b] consiste à détecter les maxima locaux du spectre d'amplitude de la TFDCT $|S(k,m)|$. Un pic est défini comme la position d'un point du spectre dont l'amplitude est supérieure à celle de son voisinage : $|S(k)|$ tel que $|S(k,m)| > |S(i,m)|\ \forall i \in [k - I/2, ..., k + I/2]$. La taille de l'intervalle I est choisi en fonction des caractéristiques de la fenêtre de pondération $h(n)$ utilisée, de sa longueur, et d'hypothèses pouvant être faites quand au contenu du signal (comme l'harmonicité). Soit $\omega_{k,h}$ la fréquence d'un pic dont l'amplitude est supérieure à celle de ses voisins. $\omega_{k,h}$ constitue une première approximation de la fréquence de la composante ω_h recherchée. La finesse fréquentielle de cette estimation est inversement proportionnelle à la taille de la fenêtre de pondération utilisée (nous ne considérons pas l'effet du prolongement par zéro pour l'instant) : $\Delta\omega = \frac{Fe}{N=\lfloor Fe \cdot L \rfloor} = \frac{1}{L}$ dans lequel L est la taille de la fenêtre en secondes.

Cette première approximation peut être améliorée grâce à l'utilisation de la technique du prolongement par zéro .

Zéro-padding ou prolongement par zéro Le prolongement par zéro est une technique consistant à compléter le signal temporel par une séquence d'échantillons nuls. L'utilisation première du prolongement par zéro est l'obtention d'un signal d'une longueur égale à $N = 2^x$ (x entier) permettant la décomposition dyadique propre aux algorithmes de calcul rapide de la Transformée de Fourier, «Fast Fourier Transform». Nous définissons un facteur de prolongement par zéro comme $ZP = N/2^x$ (ZP entier) dans lequel x est le plus petit entier permettant $2^x \geq L \cdot Fe$. Le facteur $ZP = 1$ correspond donc à la complétion implicite de la FFT.

Les facteurs $ZP > 1$ sont souvent utilisés en estimation afin d'améliorer la lisibilité du spectre pour un coût de calcul très faible. Il est important de noter que le prolongement par zéro *ne rajoute pas d'information au spectre* ; il n'augmente pas la résolution spectrale (pourvoir de séparation de deux lignes spectrales), mais seulement la finesse spectrale (distance fréquentielle entre deux «bins» de la TFDCT). Son utilisation permet une lecture plus commode du spectre et en particulier un déroulement du spectre de phase plus aisé. Ceci se comprend aisément en considérant le prolongement par zéro comme une interpolation (filtrage à bande limitée) des points du spectre [2] . Nous passons d'une finesse spectrale $\Delta\omega = \frac{Fe}{L_\#}$ (sans prolongement par zéro) à $\Delta\omega = \frac{Fe}{N}$ (avec prolongement par zéro).

Limitation du prolongement par zéro : le prolongement par zéro ne constitue qu'une amélioration faible pour l'estimation des fréquences. En effet, pour obtenir une précision de l'ordre de 1Hz pour un signal à 44100Hz, et une fenêtre de durée 10ms, il faudrait utiliser un facteur $ZP = 86$.

Le prolongement par zéro fournit une interpolation quasi-parfaite des points du spectre, mais évaluée seulement en un nombre discret de fréquences. Dans le paragraphe suivant, nous nous intéressons à un autre type d'interpolation : l'interpolation permettant la création d'un modèle continu de la forme du spectre. L'obtention d'un modèle continu permet l'obtention d'une fréquence continue indépendante de la finesse spectrale (mais pas de la résolution spectrale).

◊ *Estimateurs morphologiques : création d'un modèle continu de la forme du spectre*

L'interpolation [SS90] et la régression [MD92] locale du spectre sont deux techniques souvent utilisées pour affiner l'estimation des paramètres $[A_h, \omega_h, et\phi_h]$.

Le voisinage de $\omega_{k,h}$ sert à l'estimation d'un polynôme dont l'ordre et la forme sont choisis de manière à s'approcher de la forme de la TF d'une sinusoïde pondérée par une fenêtre d'analyse. Le polynôme permet de déterminer une nouvelle estimation de ω_h (position du maximum du polynôme) et de A_h (maximum du polynôme). ϕ_h peut être estimé de manière équivalente sur le spectre de phase voisins de $\omega_{k,h}$.

Interpolation : L'interpolation repose sur l'estimation d'un polynôme d'ordre p passant par $p + 1$ points fixés. La résolution étant linéaire, l'estimation du polynôme implique un coût de calcul faible. Cependant, un compromis est à trouver entre robustesse au bruit (impliquant un nombre élevé de points) et régularité du polynôme (impliquant

un ordre faible).

Régression : La régression repose sur l'estimation d'un polynôme d'ordre p passant près d'un nombre de points $p' > p$. Comme $p' > p$, la résolution est non-linéaire et se résout par minimisation d'une erreur quadratique. La technique de régression se caractérise par un coût d'estimation plus élevé mais une plus grande robustesse au bruit. Dans la suite, nous distinguerons deux types de méthodes de régression :

- celles augmentant le nombre d'observations p' ;
- celles diminuant le nombre d'inconnues p ; ceci est possible en faisant certaines hypothèses sur le signal, comme la symétrie du spectre autour de la position de la fréquence ω_h

[MD92] propose l'utilisation d'un polynôme d'ordre 2 en raison de la similarité de forme entre un polynôme d'ordre 2 et le spectre de puissance autour de la position de la composante sinusoïdale. L'estimation peut s'effectuer de deux manières :

Interpolation du spectre de puissance sur trois points,

$$\begin{cases} \omega_h = \omega_k + \dfrac{\Delta}{2} \dfrac{P_{k-1} - P_{k+1}}{P_{k-1} - 2P_k + P_{k+1}} \\ P_h = P_k - \dfrac{1}{8} \dfrac{(P_{k-1} - P_{k+1})^2}{P_{k-1} - 2P_k + P_{k+1}} \end{cases} \tag{4.3}$$

dans lequel

- Δ représente la finesse spectrale $\Delta = \omega_k - \omega_{k-1}$.
- $P_k = A_k^2 = |S_k|^2$ représente la puissance

Régression du spectre de puissance sur trois points. La régression est obtenue en réduisant le nombre d'inconnues. La forme du spectre de puissance autour de la position de la composante sinusoïdale est considérée comme symétrique et peut donc être approximée par un polynôme dont seuls les termes pairs sont non-nuls. [MD92] propose l'utilisation d'une parabole $(\omega - \omega_h)^2 = -4s(p - p_h)$ dont le paramètre s (demi-distance entre la directrice de la parabole et son foyer) est déterminé à priori (s dépend du type et de la longueur de la fenêtre de pondération utilisée, ainsi que de la puissance P_h).

$$\begin{cases} \omega_h = \omega_k + \dfrac{s}{\Delta}(P_{k+1} - P_{k-1}) \\ P_h = \dfrac{\Delta^2}{6s} + \dfrac{\omega_h^2}{4s} + \dfrac{\sum P_i}{3} \end{cases} \tag{4.4}$$

Cependant, le paramètre s, intervenant dans l'estimation de la puissance, dépend lui-même de la puissance. Pour cette raison, [MD92] propose 1) la normalisation préalable du spectre local en utilisant un autre estimateur de P_h tel que l'interpolation, 2) le calcul de s pour une puissance unitaire et finalement 3) le calcul de ω_h et P_h par (4.4).

Proposition : Nous montrons que l'utilisation d'un polynôme d'ordre 2 est justifiée d'un point de vue théorique dans le cas de l'utilisation du **spectre de log-amplitude** et pour le cas particulier d'une **fenêtre de pondération gaussienne**. Considérons le signal $x(t) = A_h e^{j(\omega_h t + \phi_0)}$, fenêtré par $h(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(t/\sigma)^2}$. Le spectre de log-amplitude s'exprime alors

$$\log(|S(\omega)|) = \log(A_h) - \frac{\sigma^2}{2}(\omega - \omega_h)^2 \tag{4.5}$$

dans lequel σ est l´écart-type de la fenêtre de Gauss. (4.5) constitue très exactement l'expression d'une parabole.

Dans le cas de l'utilisation jointe du spectre de log-amplitude et de la fenêtre de Gauss, la solution (4.4) reste valable, mais bénéficie des avantages suivants :

- L'utilisation du spectre de log-amplitude rend la forme de la fenêtre indépendante de l'amplitude de la composante cherchée ;
- L'utilisation d'une fenêtre de Gauss permet une définition exacte du paramètre s de la parabole : $s = \frac{1}{2(2\pi\sigma)^2}$.

Dans le cas de l'utilisation d'une fenêtre cosinusoïdale ou d'une échelle non-logarithmique d'amplitude, s (σ) doit être déterminé de manière empirique pour chaque type de fenêtre d'analyse (voir annexe J).

L'utilisation du spectre de log-amplitude et d'une fenêtre de pondération de Gauss constitue la base de la méthode que nous proposerons plus loin **«Mesure de la distorsion du spectre complexe»**. Dans ce modèle, la forme du lobe n'est pas imposée puisqu'elle est fonction des variations du signal. Ce n'est que dans le cas où le signal peut être considéré sans variations que nous pouvons imposer cette forme et obtenir une expression similaire à celle de [MD92] :

$$\begin{cases} \omega_h = \omega_k + \dfrac{s}{\Delta}(\log(A_{k+1}) - \log(A_{k-1}))) \\ \log(A_h) = \dfrac{\Delta^2}{6s} + \dfrac{\omega_h^2}{4s} + \dfrac{\sum \log(A_i)}{3} \end{cases} \tag{4.6}$$

◊ ***Fréquence instantanée***

Les estimateurs étudiés jusqu'à présent reposaient uniquement sur l'observation du spectre d'amplitude. L'estimateur «fréquence instantanée» repose théoriquement, quant à lui, uniquement sur l'observation de la phase.

Dans le cas d'un signal *mono-composante*, la fréquence instantanée est définie comme la dérivée de la phase par rapport au temps : $\omega_\phi = \frac{\partial\phi(t)}{\partial t}$. Elle mesure «la fréquence moyenne autour du temps t» et peut être calculée par dérivée de l'argument $\phi(n)$ du signal analytique correspondant au signal réel $x(n)$: $x(n) + jx_H(n) = A(t)e^{j\phi(n)}$, dans lequel H désigne la Transformée de Hilbert.

Dans le cas de signaux *multi-composantes*, nous distinguons autant de fréquences instantanées qu'il y a de composantes simultanément présentes dans le signal. Afin de les estimer, [AKI95] et [Cor99] effectue un filtrage du signal en bandes de fréquences afin de permettre la détermination de la fréquence instantanée dans chaque bande de fréquence par dérivée de l'argument du signal analytique correspondant. Dans le cadre de la synthèse sinusoïdale, le calcul de la fréquence instantanée se fait par différentiation de la phase d'une composante h estimée sur deux trames successives : $\omega_\phi(\omega_h) = \frac{\omega_h(t_{m+1}) - \omega_h(t_m)}{t_{m+1} - t_m}$. Elle implique l'appariement des composantes d'une trame à l'autre. Il s'agit également de la technique utilisée dans le vocodeur de phase (dans ce cas, phase de fréquences fixes donc sans nécessité d'appariement d'une trame à l'autre).

L'estimation de la fréquence instantanée par rapport de deux TFs est proposée par de nombreux auteurs ([Cha88], [PR99b], [Fit99]) et peut être directement rapprochée du réassignement fréquentiel [AF95] [3] . La fréquence instantanée est dans ce cas estimée par le

rapport de la TFDCT du signal pondéré par la dérivée temporelle de la fenêtre $dh \triangleq \frac{\partial h(n)}{\partial n}$ [4] sur la TFDCT du signal pondéré par la fenêtre $h \triangleq h(n)$.

$$\omega_r(\omega_k) = \omega_k - \Im\left\{\frac{STFT_{dh}(\omega_k)STFT_h^*(\omega_k)}{|STFT_h(\omega_k)|^2}\right\} \tag{4.9}$$

4.2.1.2 Estimation globale en fréquence

◊ ***Intérêt d'une estimation globale en fréquence***

La résolution spectrale, définie comme le pouvoir de résoudre (ou encore de pouvoir estimer les paramètres de) deux sinusoïdes de fréquences proches, dépend de la durée sur laquelle l'observation est effectuée ainsi que du type de la fenêtre de pondération utilisée. Deux composantes aux fréquences f_1 et f_2 ne pourront être estimées correctement que si leur représentation fréquentielle ne se recouvrent «pas trop». Une condition parfois utilisée est le non-recouvrement à -6dB_{20} de leur lobe principal [5] . Dans le cas où le signal est considéré comme harmonique, ceci correspond à la condition $\text{Bw} = \frac{\text{Cw}}{L} < \text{f0}$, c'est-à-dire revient à choisir une taille de fenêtre d'analyse de durée $L > \text{Cw} \cdot T_0$. À titre d'exemple, pour une fenêtre de type Blackman ($\text{Cw} = 2.35$), nous obtenons $L > 2.35\text{T0}$.

Ceci, du moins, est vrai en ne considérant pas l'influence des deux composantes au travers de leur phase respective. En effet, pour de petites tailles de fenêtre, la relation de phase des composantes ω_k détermine également la résolution atteignable. Ceci est illustré à la Fig. 4.1 pour trois relations de phase critiques :

- [G] phases identiques ($\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2)$), engendrant l'addition des composantes aux fréquences intermédiaires $(\omega_1 + \omega_2)/2$, $(\omega_2 + \omega_3)/2$ et donc la plus mauvaise résolution,
- [M] phases en quadrature ($\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2) + \frac{\pi}{2} + p2\pi$), engendrant l'influence minimale des composantes,
- [D] opposition de phase ($\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2) + (2p+1)\pi$), engendrant la soustraction des composantes aux fréquences intermédiaires et donc le creusement maximum.

En l'absence d'information sur la relation de phase des composantes voisines, une taille de fenêtre assurant le non-recouvrement à -12dB_{20} est préférable, [6] nécessitant donc l'utilisation de fenêtres de durée plus importante.

L'utilisation de fenêtres de durée courte, afin de diminuer l'erreur commise en approximant les paramètres du signal par l'ordre 0, n'est possible qu'en prenant en compte les recouvrements dus à la basse résolution dans l'estimation. Ces méthodes de résolution globale sont étudiées dans cette partie.

◊ ***Estimation par minimisation d'un critère d'erreur de modélisation***

L'estimation des paramètres A_h, ω_h et $\phi_{0,h}$ s'effectue pour toutes les composantes $h \in H$ simultanément de manière à minimiser un critère d'erreur de modélisation entre signal réel $s(t)$ et signal modélisé.

Remarque : *L'ensemble des critères proposés peuvent également être calculés de manière locale en*

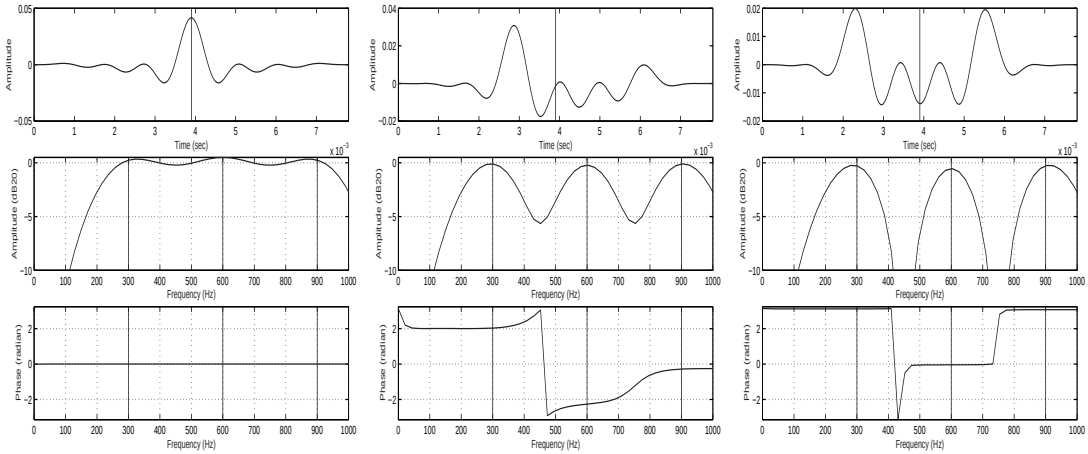


FIG. 4.1 – Influence de la relation de phase des composantes sur la résolution spectrale : signal et spectre d'amplitude/ phase pour un signal composé de 3 sinusoïdes à 300, 600, 900 Hz d'amplitude 1, fenêtre de Blackman, $L_{sec} = 2.35T0$: [G] $\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2)$, [M] $\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2) + \frac{\pi}{2}$, [D] $\phi(\omega_1) = \phi(\omega_3) = \phi(\omega_2) + \pi$

fréquence. La résolution locale étant un cas particulier de la résolution globale, nous présentons seulement la résolution globale.

◇ *A. Choix d'un critère d'erreur*

Le modèle de signal stationnaire (4.2) est utilisé dans [GL88], alors qu'un modèle stationnaire prenant en compte le bruit est proposé par [MA94] [7] .

Dans les deux cas, différents critères d'erreur sont envisageables :

Minimisation de l'erreur entre spectres d'amplitude :

$$\epsilon = \int \left(|S(\omega)|^n - |\hat{S}(\omega)|^n \right) d\omega \tag{4.13}$$

n est généralement choisi égal à 2 et représente donc une minimisation d'énergie. Ce critère ne permet que l'estimation des amplitudes.

Minimisation de l'erreur entre spectres complexes :

$$\epsilon = \int |S(\omega) - \hat{S}(\omega)|^2 d\omega \tag{4.14}$$

Ce critère permet l'estimation des amplitudes et des phases.

La comparaison des modèles [GL88] et [MA94] ainsi que des deux critères d'erreurs (4.13) et (4.14) est faite dans [Dut94]. L'auteur conclut en une mauvaise estimation de l'amplitude par les critères basés sur une minimisation de l'énergie (4.13). L'auteur conclut également en une meilleure estimation de l'amplitude par le critère de [MA94] (du moins quand le signal présente du bruit). Cependant il conclut en une différence très faible d'un point de vue

perceptif entre la resynthèse utilisant l'estimation de [MA94] et (4.14). Pour cette raison et en vue de sa plus grande simplicité, nous choisissons le critère (4.14).

◇ *B. Estimation des amplitudes et des phases*

La TF du modèle de signal (4.2) s'exprime

$$\hat{S}(\omega_k) = \sum_{h=1}^{H} \frac{A_h}{2} \left(e^{j\phi_{0,h}} H(\omega_k - \omega_h) + e^{-j\phi_{0,h}} H(\omega_k + \omega_h) \right) \tag{4.15}$$

et peut se réécrire

$$\hat{S}(\omega_k) = \sum_{h=1}^{2H} B_h(\omega_k)\theta_h \tag{4.16}$$

dans lequel nous notons

$$\begin{cases} B_h(\omega_k) = H(\omega_k - \omega_h) + H(\omega_k + \omega_h) \\ B_{h+H}(\omega_k) = j\left[H(\omega_k - \omega_h) - H(\omega_k + \omega_h)\right] \\ \theta_h = \dfrac{A_h}{2}\cos(\phi_{0,h}) \\ \theta_{h+H} = \dfrac{A_h}{2}\sin(\phi_{0,h}) \end{cases} \tag{4.17}$$

Soit sous forme matricielle

$$\underline{\hat{S}} = \underline{\underline{B}}\,\underline{\theta} \tag{4.18}$$

La minimisation de ϵ par rapport à A_h et ϕ_h se fait par minimisation des dérivées partielles $\frac{\partial ||\underline{S} - \underline{\hat{S}}||^2}{\partial \underline{\theta}}(\underline{\hat{\theta}}) = 0$ et conduit à l'optimum pour

$$\boxed{\underline{\hat{\theta}} = (\underline{\underline{B}}^{\mathcal{H}}\underline{\underline{B}})^{-1}\underline{\underline{B}}^{\mathcal{H}}\underline{S}} \tag{4.19}$$

Les amplitudes et phases sont finalement données par

$$\begin{cases} A_h = 2\sqrt{\theta_h^2 + \theta_{l+L}^2} \\ \phi_{0,h} = \text{atan}\left(\dfrac{\theta_{l+L}}{\theta_h}\right) \end{cases} \tag{4.20}$$

◇ *C. Estimation des fréquences*

La résolution du système pour les fréquences est plus compliquée, puisque $\hat{S}(\omega)$ n'est pas linéaire dans les fréquences ω_h. La solution proposée dans [DT96] et [DH97] consiste à effectuer un développement en série limitée de la TF $H(\omega)$ de la fenêtre de pondération autour d'une première approximation de la fréquence de la sinusoïde. Soit Ω_h cette première

approximation, ω_h la «vraie» fréquence de la sinusoïde et Δ_h le décalage entre les deux : $\omega_h = \Omega_h + \Delta_h$. En utilisant le développement en série limitée, nous pouvons réécrire $H(\omega_k - \omega_h)$ comme

$$H(\omega_k - \omega_h) = H(\omega_k - \Omega_h) - H'(\omega_k - \Omega_h)\Delta_h + o(\Delta_h^2) \tag{4.21}$$

Soit $\hat{S}(\omega)$ le modèle obtenu à partir de la première approximation des fréquences Ω_h. Soit $\tilde{S}(\omega)$ l'approximation cherchée de $S(\omega)$ obtenue à partir des fréquences inconnues ω_h. $\tilde{S}(\omega)$ s'exprime

$$\tilde{S}(\omega_k) = \hat{S}(\omega_k) - \left(\sum_h \frac{A_h}{2} (e^{j\phi_{0,h}} H'(\omega_k - \Omega_h) + e^{-j\phi_{0,h}} H'(\omega_k + \Omega_h)) \right) \cdot \Delta_h \tag{4.22}$$

Soit sous forme matricielle en notant $\underline{\Theta}$ le vecteur $[\Delta_1, ...\Delta_H]$.

$$\underline{\tilde{S}} = \underline{\hat{S}} + \underline{\underline{C}}\,\underline{\Theta} \tag{4.23}$$

L'erreur à minimiser est

$$\epsilon = ||\underline{S} - \underline{\tilde{S}}||^2 = ||\underline{S} - (\underline{\hat{S}} + \underline{\underline{C}}\,\underline{\Theta})||^2 \tag{4.24}$$

La minimisation de ϵ par rapport à Δ_h conduit à l'optimum

$$\boxed{\underline{\hat{\Theta}} = (\underline{\underline{C}}^{\mathcal{H}}\underline{\underline{C}})^{-1}\underline{\underline{C}}^{\mathcal{H}}(\underline{S} - \underline{\hat{S}})} \tag{4.25}$$

Ceci nous donne les Δ_h, donc la nouvelle estimation des fréquences : $\omega_h = \Omega_h + \Delta_h$.

L'estimation de ω_h se fait par itération en remplaçant, à chaque itération, Ω_h par la valeur obtenue à l'itération précédente $\Omega_h + \Delta_h$. La convergence est assez rapide pour autant que la fréquence initiale se trouve sur le lobe principal de la composante [Hel97] [8] .

4.2.2 Modèles sinusoïdaux non-stationnaires

4.2.2.1 Introduction

Ci-après, nous étudions l'effet et la prise en compte des non-stationnarités du signal dans le modèle sinusoïdal. Dans le modèle sinusoïdal «classique», les paramètres sont supposés localement constants (hypothèse de stationnarité locale). Cette hypothèse, bien que constituant une approximation raisonnable pour les parties tenues des sons musicaux, n'est que rarement vérifiée pour le reste (attaques, transitions, vibrato, trémolo, variations de formants, ...).

Nous commençons par illustrer l'effet sur le spectre des non-stationnarités de type variation linéaire de fréquence et d'amplitude. Ensuite, afin de permettre une estimation des paramètres dans des conditions de signaux non-stationnaires, deux approches sont étudiées :

- Nous proposons un modèle sinusoïdal «étendu» (variation des paramètres à l'ordre 1). L'estimation des paramètres de ce modèle s'effectue localement en fréquence et repose sur une «mesure de distorsion du spectre complexe» ; elle appartient donc à la classe des estimateurs morphologiques.
- Une autre approche est présentée, permettant la prise en compte des non-stationnarités du signal dans une résolution globale. Malgré les avantages procurés par cette approche, tel la prise en compte des recouvrements spectraux, elle ne fait pas référence explicitement à un modèle sinusoïdal à variation lente. Ceci à pour conséquence une non-régularité des paramètres (oscillation des polynômes).

4.2.2.2 Illustration de l'effet des non-stationnarités

Nous considérons l'effet d'une modulation de fréquence et d'amplitude linéaire et symétrique par rapport au centre de l'observation. Le signal utilisé est une sinusoïde de 300 Hz d'amplitude unitaire tantôt modulée en fréquence et tantôt en amplitude (fréquence d'échantillonnage = 44100 Hz, longueur du signal = 0.36 sec, type de fenêtre = Blackman).

À titre de référence, les spectres d'amplitude et de phase en l'absence de modulations sont indiqués à la FIG. 4.2.

Modulation de fréquence : Une modulation de fréquence (FIG. 4.4) résulte en un élargissement symétrique du lobe central de la composante. L'amplitude du lobe diminue de manière à compenser son élargissement. La courbure du spectre de phase dépend du signe de la modulation de fréquence.

Modulation d'amplitude : Une modulation d'amplitude (FIG. 4.5) ne modifie pas le spectre d'amplitude lorsque la fréquence reste constante. Le spectre de phase prend la forme d'un polynôme d'ordre 3 dont le point d'inflexion se situe à la fréquence du pic et dont le signe de la courbure dépend du signe de la modulation. [1].

Modulation de fréquence et d'amplitude : La présence des deux modulations simultanément (FIG. 4.3) résulte en un élargissement et un décalage fréquentiel du lobe central de la composante. Le sens du décalage dépend à la fois du signe de la modu-

[1] Le signe de la courbure est en rapport direct avec le **retard de groupe** : pente de phase négative pour un retard de groupe positif (concentration de l'énergie à droite du centre de la fenêtre ou modulation positive de l'amplitude), pente de phase positive pour un retard de groupe négatif (concentration de l'énergie à gauche du centre de la fenêtre ou modulation négative de l'amplitude)

lation de fréquence et de celle de l'amplitude. Un phénomène similaire apparaît sur le spectre de phase.

4.2.2.3 Estimation locale en fréquence : mesure de la distorsion du spectre complexe

♢ ***Introduction***

Les observations faites dans la partie 4.2.2.2 quant aux effets d'une modulation de fréquence et d'amplitude sur le spectre d'amplitude et de phase ont été également observées dans [Mas96]. L'auteur propose l'utilisation d'un réseau de neurones, entraîné sur des signaux tests et les spectres complexes correspondants, afin de déterminer les paramètres de modulation ayant engendré la déformation observée du spectre.

L'étude analytique des déformations du spectre complexe dans le cadre de la synthèse sinusoïdale a été étudiée par [MA86] [GBM+96] [GBM+96] et [Gri99] (remarquons cependant qu'il s'agit, dans le cas d'une fenêtre de Gauss, de l'exemple de chirp de fréquence rencontré régulièrement dans la littérature). L'étude est faite dans le cadre d'une modulation linéaire de fréquence. Chaque composante sinusoïdale du spectre est modélisée par un terme d'amplitude constante A_h, de fréquence constante ω_h, de variation linéaire de fréquence Δ_h/π et d'une phase $\phi_{0,h}$:

$$\hat{x}(t) = Ae^{j(\phi_{0,h} + \omega_h t + \Delta_h t^2)} \tag{4.30}$$

La fenêtre de pondération est choisie de manière à permettre une résolution analytique simple. Le choix se porte sur une fenêtre de type Gauss de moyenne $t = 0$ et de variance σ^2 [2] :

$$h(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}} \tag{4.31}$$

La Transformée de Fourier du modèle sinusoïdal fenêtré $\hat{s}(t) = \hat{x}(t) \cdot h(t)$ s'écrit en notation module/ argument :

$$\hat{S}(\omega) = A\alpha(\omega)e^{j(\beta(\omega) + \phi_{0,h})} \tag{4.32}$$

dans laquelle

$$\begin{cases} \alpha(\omega) = & D_h^{-\frac{1}{4}} e^{-\frac{\sigma^2}{2}\frac{(\omega-\omega_h)^2}{D_h}} \\ \beta(\omega) = & -\Delta_h\sigma^4\frac{(\omega-\omega_h)^2}{D_h} + \frac{\text{atan}(2\Delta_h\sigma^2)}{2} \end{cases} \tag{4.33}$$

où D_h est défini comme $D_h \triangleq 1 + 4\Delta_h^2\sigma^4$

Les expressions des spectres de log-amplitude ($\log(A)+\log(\alpha(\omega))$) et de phase ($\beta(\omega)+\phi_{0,h}$) sont des polynômes d'ordre 2 en ω. Il s'agit de paraboles dont l'abscisse du foyer est en ω_h

[2] Signalons, comme autre avantage, que la fenêtre de Gauss est la fenêtre réalisant le meilleur compromis localisation temporelle/ localisation fréquentielle

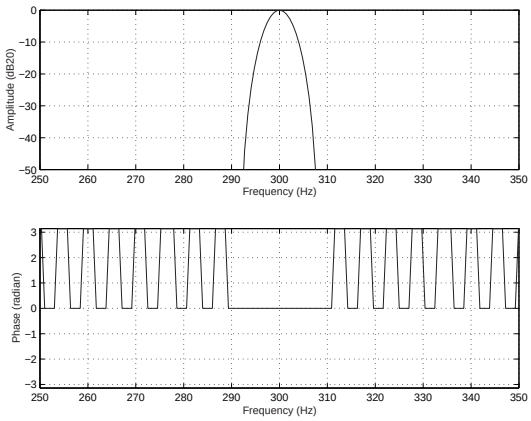


FIG. 4.2 – Spectre d'amplitude et de phase d'un signal de fréquence et d'amplitude constantes, fenêtre=blackman, durée de l'observation=360 msec

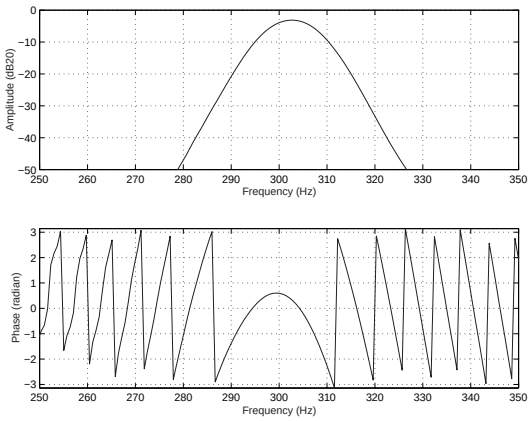
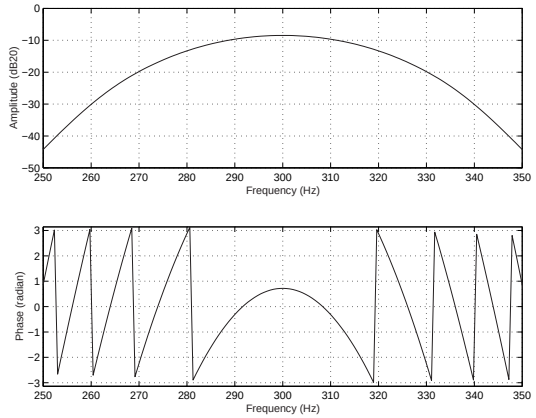


FIG. 4.3 – Spectre d'amplitude et de phase d'un signal de fréquence croissante (100 Hz/sec) et d'amplitude croissante (8 lin/sec)



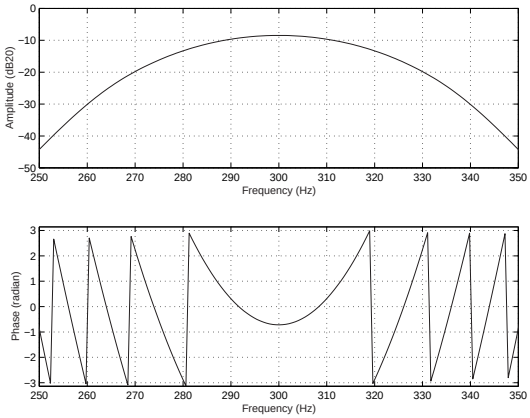
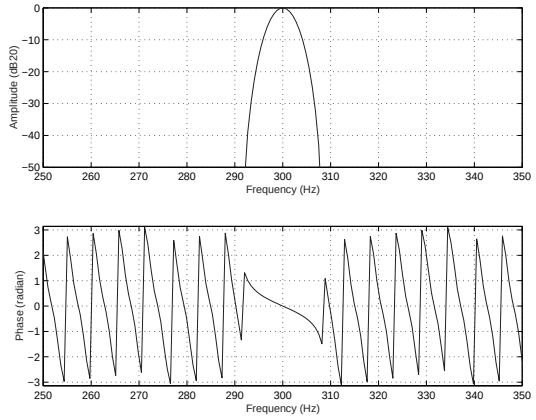


FIG. 4.4 – Spectre d'amplitude et de phase d'un signal d'amplitude constante et de fréquence [G] croissante (300 Hz/sec) [D] décroissante (-300 Hz/sec)



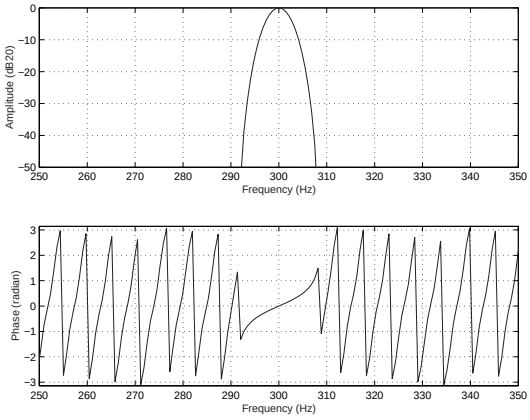


FIG. 4.5 – Spectre d'amplitude et de phase d'un signal de fréquence constante et d'amplitude [G] croissante (4 lin/sec) [D] décroissante (-4 lin/sec)

et dont le facteur de forme dépend de la modulation de fréquence et de la taille effective de la fenêtre σ.

Pour un spectre donné $S(\omega)$, nous pouvons déterminer deux polynômes d'ordre 2 approximant le mieux possible (au sens d'un critère de minimisation d'erreur ou simplement par interpolation) la région du spectre de log-amplitude et de phase autour de chaque pic ω_k. Les paramètres Δ_h et ω_h se déterminent alors à partir des coefficients de ce polynôme : soit par résolution du système découlant de l'expression du spectre de log-amplitude (4.33), soit à partir de celui découlant de l'expression du spectre de phase (4.33), soit encore à partir des deux.

Le modèle que nous proposons dans la suite est une extension du modèle proposé par [MA86] dans le cas d'une modulation simultanée de fréquence et d'amplitude. Ce modèle reprenant comme cas particulier celui de [MA86] (lorsque la modulation d'amplitude est nulle), nous ne détaillons pas les solutions de [MA86].

◊ *Modèle sinusoïdal proposé*

Soit $\hat{x}(t)$ le modèle sinusoïdal de fréquence linéaire $(\omega_h + 2\Delta_h t)$ et d'amplitude linéaire $(a_{0,h} + a_{1,h}t)$:

$$\boxed{\hat{x}(t) = (a_{0,h} + a_{1,h}t)e^{j(\phi_{0,h} + \omega_h t + \Delta_h t^2)}} \tag{4.34}$$

La Transformée de Fourier du modèle sinusoïdal fenêtré $\hat{s}(t) = \hat{x}(t) \cdot h(t)$ s'écrit en notation module/ argument :

$$\hat{S}(\omega) = \alpha(\omega)\alpha'(\omega)e^{j(\beta(\omega) + \beta'(\omega) + \phi_{0,h})} \tag{4.35}$$

dans lequel $\alpha(\omega)$ et $\beta(\omega)$ sont donnés par (4.33) et $\alpha'(\omega)$ et $\beta'(\omega)$ par

$$\begin{cases} \alpha'(\omega) = & \frac{1}{D_h}\sqrt{(a_{0,h}D_h + a_{1,h}\Delta_h\sigma^4 2(\omega - \omega_h))^2 + (a_{1,h}\sigma^2(\omega - \omega_h))^2} \\ \beta'(\omega) = & \text{atan}\left(\frac{-a_{1,h}\sigma^2(\omega - \omega_h)}{a_{0,h}D_h + a_{1,h}\Delta_h\sigma^4 2(\omega - \omega_h)}\right) \end{cases} \tag{4.36}$$

Comparativement au modèle de [MA86], $\alpha'(\omega)$ et $\beta'(\omega)$ ne sont plus des polynômes d'ordre 2. De ce fait, la représentation du spectre de log-amplitude $(\log(\alpha(\omega)) + \log(\alpha'(\omega))$) et du spectre de phase $(\beta(\omega) + \beta'(\omega) + \phi_{0,h})$ ne sont également plus des polynômes d'ordre 2. La résolution du système pour ω_h, $a_{1,h}/a_{0,h}$ et Δ_h n'est donc pas immédiate. Nous proposons le développement en série limitée de $\alpha'(\omega)$ et $\beta'(\omega)$.

Choix de l'ordre du développement en série : Au vu de la FIG. 4.5, lors d'une modulation d'amplitude, le spectre de phase prend la forme d'un polynôme d'ordre 3 sur le lobe principal. Un développement d'ordre 3 en phase et d'ordre 2 en log-amplitude semble donc indiqué. Se pose cependant un problème de sur-détermination, puisque le système possède dans ce cas 7 équations pour 5 inconnues. Les développements à l'ordre 2 en log-amplitude et en phase conduisent à un système de 6 équations à 5 inconnues :

Soit les développements à l'ordre 2

- Développements en série limitée du terme de log-amplitude

$$log(\alpha(\omega)\alpha'(\omega)) \simeq \boxed{\left(-\frac{\sigma^2}{2D_h} + X\right)x^2} + \frac{a_{1,h}}{a_{0,h}}\frac{2\Delta_h\sigma^4}{D_h}x + \log(a_{0,h}) - \frac{1}{4}\log(D_h) + O(x^3) \tag{4.37}$$

dans lequel x est défini comme $x \triangleq (\omega - \omega_h)$ et X comme $X \triangleq \frac{a_{1,h}^2}{a_{0,h}^2}\frac{\sigma^4}{2D_h^2}(2 - D_h)$

- Développements en série limitée du terme de phase

$$\beta(\omega) + \beta'(\omega) \simeq \boxed{\left(-\frac{\Delta_h\sigma^4}{D_h} + Y\right)x^2} - \frac{a_{1,h}}{a_{0,h}}\frac{\sigma^2}{D_h}x + \phi_{0,h} + \frac{1}{2}\text{atan}(2\Delta_h\sigma^2) + O(x^3) \tag{4.38}$$

dans lequel x est défini comme $x \triangleq (\omega - \omega_h)$ et Y comme $Y \triangleq 2\frac{a_{1,h}^2}{a_{0,h}^2}\frac{\sigma^6\Delta_h}{D_h^2}$

En utilisant les expressions (4.37) et (4.38) et en approximant la forme du spectre de log-amplitude du signal observé $S(\omega)$ près d'un pic par $P_{\log|S|}(\omega) = a_{\log|S|}\omega^2 + b_{\log|S|}\omega + c_{\log|S|}$ et la forme du spectre de phase par $P_{\phi(S)}(\omega) = a_{\phi(S)}\omega^2 + b_{\phi(S)}\omega + c_{\phi(S)}$, nous obtenons deux systèmes de 3 équations pour 5 inconnues. Si nous supposons ω_h connu, chacun des deux systèmes conduit à une estimation indépendante de $a_{1,h}/a_{0,h}$ et de Δ_h à partir des deux premières équations de chaque système (la troisième équation servant respectivement à l'estimation de $a_{0,h}$ et de $\phi_{0,h}$).

$$\begin{cases} a_{\log|S|} = & -\dfrac{\sigma^2}{2D_h} + \boxed{X} \\ b_{\log|S|} = & 2\omega_h\dfrac{\sigma^2}{2D_h} + \dfrac{a_{1,h}}{a_{0,h}}\dfrac{2\Delta_h\sigma^4}{D_h}\boxed{-2\omega_h X} \\ c_{\log|S|} = & -\omega_h^2\dfrac{\sigma^2}{2D_h} + \log(a_{0,h}) - \dfrac{1}{4}\log(D_h) - \omega_h\dfrac{a_{1,h}}{a_{0,h}}\dfrac{2\Delta_h\sigma^4}{D_h} + \boxed{\omega_h^2 X} \end{cases} \tag{4.39}$$

$$\begin{cases} a_{\phi(S)} = & -\dfrac{\Delta_h\sigma^4}{D_h}\boxed{+Y} \\ b_{\phi(S)} = & 2\omega_h\dfrac{\Delta_h\sigma^4}{D_h} - \dfrac{a_{1,h}\sigma^2}{a_{0,h}D_h}\boxed{-2\omega_h Y} \\ c_{\phi(S)} = & -\omega_h^2\dfrac{\Delta_h\sigma^4}{D_h} + \phi_{0,h} + \dfrac{1}{2}\text{atan}(2\Delta_h\sigma^2) + \omega_h\dfrac{a_{1,h}}{a_{0,h}}\dfrac{\sigma^2}{D_h}\boxed{+\omega_h^2 Y} \end{cases} \tag{4.40}$$

L'estimation simultanée de ω_h, $a_{1,h}/a_{0,h}$ et Δ_h nécessite le couplage des deux premières équations de chaque système. Pour résoudre ce système de 4 équations à 3 inconnues nous avons imposé $\sigma^2 = -2\frac{G}{a_{\log|S|}}$, ce qui correspond à l'annulation d'un terme de modulation d'amplitude symétrique par rapport à $t = 0$ (voir annexe K). Cependant le système correspondant au développement à l'ordre 2 s'avère difficilement soluble ; aussi proposons-nous la solution pour le développement à l'ordre 1 (l'ordre 1 correspond aux développements (4.37) et (4.38) ainsi qu'aux systèmes (4.39) et (4.40) sans les parties encadrées).

Finalement la solution est donnée ci-dessous.

Mesure de distorsion du spectre complexe

- Modèle de signal $\hat{x}(t) = (a_{0,h} + a_{1,h}t)e^{j(\phi_{0,h} + \omega_h t + \Delta_h t^2)}$
- Fenêtre de pondération de type Gauss d'écart-type σ
- Forme du spectre du signal fenêtré approximée par :
  spectre de log-amplitude $P_{\log|S|}(\omega) = a_{\log|S|}\omega^2 + b_{\log|S|}\omega + c_{\log|S|}$,
  spectre de phase $P_{\phi(S)}(\omega) = a_{\phi(S)}\omega^2 + b_{\phi(S)}\omega + c_{\phi(S)}$,

$$\begin{cases} \dfrac{a_{1,h}}{a_{0,h}} = \dfrac{1}{2}\dfrac{b_{\phi(S)}a_{\log|S|} - a_{\phi(S)}b_{\log|S|}}{G} \\ \omega_h = -\dfrac{1}{2}\dfrac{b_{\log|S|}a_{\log|S|} + a_{\phi(S)}b_{\phi(S)}}{G} \\ \Delta_h = -\dfrac{1}{4}\dfrac{a_{\phi(S)}}{G} \end{cases} \tag{4.41}$$

dans lequel $G = a^2_{\log|S|} + a^2_{\phi(S)}$.

$$\begin{cases} \log(a_{0,h}) = P_{\log|S(\omega_{\max})|} + \dfrac{1}{4}\log(D_h) - \dfrac{a^2_{1,h}}{a^2_{0,h}}\dfrac{2\Delta_h^2\sigma^6}{D_h} \\ \phi_{0,h} = P_{\phi(S(\omega_{\max}))} - \dfrac{1}{2}\mathrm{atan}(2\Delta_h\sigma^2) + \dfrac{a^2_{1,h}}{a^2_{0,h}}\dfrac{2\Delta_h\sigma^4}{D_h}(1 + 2\Delta_h^2\sigma^4) \end{cases} \tag{4.42}$$

dans lequel $D_h = 1 + 4\Delta_h^2\sigma^4$

◊ *Étude des biais*

À partir des résultats établis au paragraphe précédent, nous montrons le biais engendré par des modulations d'amplitude et de fréquence sur l'estimation des paramètres du modèle, si ceux étaient estimés sans prise en compte des non-stationnarités.

Remarque : Les biais proposés sont ceux qui résulteraient d'une modulation sous forme de variation linéaire sur la durée de l'observation, c'est-à-dire une variation symétrique par rapport au milieu de la fenêtre.

Les résultats sont résumés à la Fig. 4.6.

Fréquence de la sinusoïde : La fréquence de la sinusoïde est généralement calculée comme la position du maximum du polynôme d'ordre 2 approximant la forme du pic du spectre de log-amplitude. Dans ce cas

$$\omega_{\max} = -\frac{b_{\log|S|}}{2a_{\log|S|}} = \omega_h \boxed{+\frac{a_{1,h}}{a_{0,h}}2\Delta_h\sigma^2} \tag{4.43}$$

Le biais de l'estimateur $\omega_{\max}$ par rapport à ω_h n'apparaît qu'en présence d'une modulation simultanée de fréquence et d'amplitude. Dans ce cas, il est proportionnel à la grandeur de la modulation d'amplitude et de fréquence et à la taille effective de la

fenêtre σ, c'est-à-dire à la variation totale sur la durée de l'observation. Le signe du biais dépend du signe des modulations d'amplitude et de fréquence.

Amplitude de la sinusoïde : L'amplitude de la sinusoïde est généralement estimée comme l'amplitude au maximum du polynôme d'ordre 2 approximant la forme du pic du spectre de log-amplitude. Dans ce cas

$$P_{\log(|S(\omega)|)}(\omega_{\max}) = \log(a_{0,h}) \boxed{-\frac{1}{4}\log(D_h) + \frac{a_{1,h}^2}{a_{0,h}^2}\frac{2\Delta_h^2\sigma^6}{D_h}} \tag{4.44}$$

Le biais de l'estimateur $P_{\log(|S(\omega)|)}(\omega_{\max})$ par rapport à $\log(a_{0,h})$ n'apparaît que dans le cas d'une modulation de fréquence (sous estimation de l'amplitude, puisque $D_h \geq 1$)) et peut être accentué/diminué par une modulation simultanée d'amplitude (surestimation de l'amplitude). Dans les deux cas, il dépend de la taille effective de la fenêtre.

Valeur du spectre de log-amplitude en ω_0 :

$$P_{\log(|S(\omega)|)}(\omega_h) = \log(a_{0,h}) - \frac{1}{4}\log(D_h) \tag{4.45}$$

Le biais de l'estimateur $P_{\log(|S(\omega)|)}(\omega_h)$ par rapport à $\log(a_{0,h})$ ne dépend plus que de la modulation de fréquence (sous estimation de l'amplitude).

Phase de la sinusoïde :

$$P_{\phi(S(\omega))}(\omega_{\max}) = \phi_{0,h} + \frac{1}{2}\text{atan}(2\Delta_h\sigma^2) \boxed{-\frac{a_{1,h}^2}{a_{0,h}^2}\frac{2\Delta_h\sigma^4}{D_h}(1+2\Delta_h^2\sigma^4)} \tag{4.46}$$

Le biais de l'estimateur $P_{\phi(S(\omega))}(\omega_{\max})$ par rapport à $\phi_{0,h}$ n'apparaît que dans le cas d'une modulation de fréquence, et peut être accentué/diminué par une modulation simultanée d'amplitude. Dans les deux cas, il dépend de la taille effective de la fenêtre.

Valeur du spectre de phase en ω_h :

$$P_{\phi(S(\omega))}(\omega_h) = \phi_{0,h} + \frac{1}{2}\text{atan}(2\Delta_h\sigma^2) \tag{4.47}$$

Le biais de l'estimateur $P_{\phi(S(\omega))}(\omega_h)$ par rapport à $\phi_{0,h}$ ne dépend plus que de la modulation de fréquence.

4.2.2.4 Estimation globale en fréquence : polynôme d'amplitude complexe

Afin de prendre en compte les non-stationnarités du signal dans le modèle sinusoïdal, [Lar89] propose de remplacer la constante A_h, représentant l'amplitude constante sur la durée de l'observation, par un polynôme complexe d'ordre Q. Le polynôme étant complexe, il permet la représentation à la fois des modulations d'amplitude et de fréquence sur la durée de l'observation.

Le signal $x(t)$ est approximé par une somme de H sinusoïdes de fréquence ω_h $(1 \leq h \leq H)$, chacune modulée par un polynôme d'ordre Q en n (n = numéro d'échantillon) de coefficients complexes $c_{h,q}$ $(0 \leq q \leq Q)$: $c_{h,k} = a_{h,k} + jb_{h,k}$.

$$\boxed{\hat{x}(n) = \sum_{h=1}^{H}\left(c_{h,0} + c_{h,1}n + \ldots + c_{h,Q}n^Q\right)e^{j\omega_h\frac{n}{Fe}}} \tag{4.48}$$

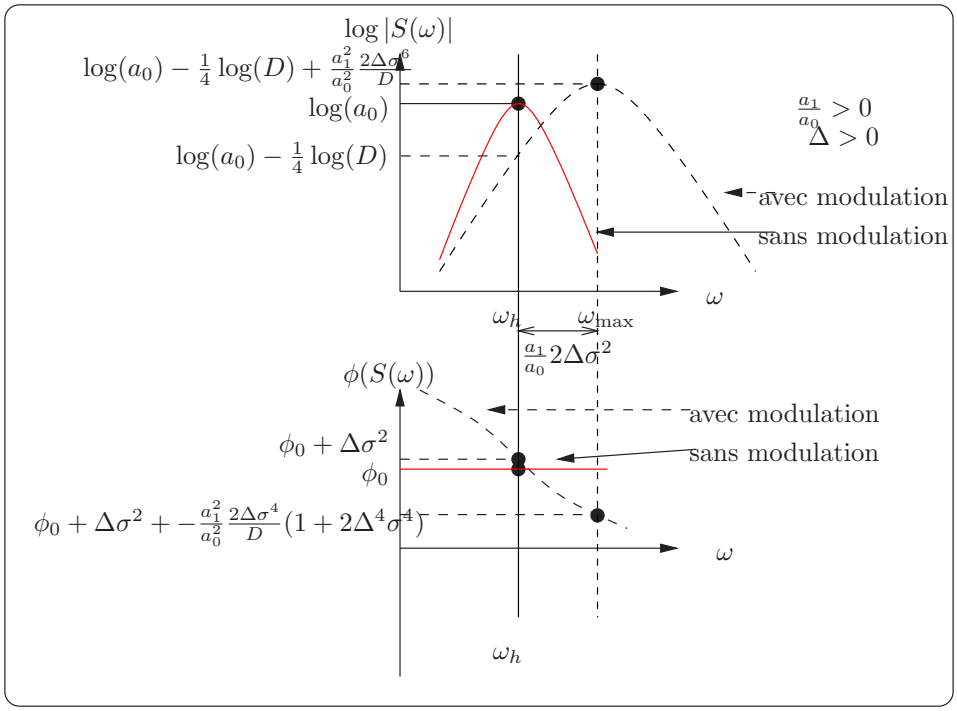


FIG. 4.6 – Biais engendré par les modulation de fréquence et d'amplitude selon (4.41)

(4.48) se réécrit en termes de modulation d'amplitude et de phase :

$$\hat{x}(n) = \sum_{l=1}^{L} \left(A_h(n) e^{j\phi_h(n)} \right) e^{j\omega_h \frac{n}{Fe}} \tag{4.49}$$

Nous montrons que pour l'ordre Q = 1, le polynôme complexe ne permet pas la représentation d'une variation de fréquence indépendante de celle d'amplitude. En effet, pour Q = 1, la fréquence instantanée $\frac{\partial \phi(n)}{\partial n} = \frac{\omega}{Fe} + \frac{b_1 a_0 - a_1 b_0}{A^2(n)}$ varie de manière inversement proportionnelle au carré de l'amplitude instantanée $A(n) = \sqrt{(a_0 + a_1 n)^2 + (b_0 + b_1 n)^2}$.

Le passage à l'ordre Q = 2 permet une variation indépendante de la fréquence et de l'amplitude. Pour Q = 2, nous obtenons les paramètres suivants (l'indice h est omis afin de ne pas alourdir la notation) :

Amplitude instantanée :

$$A(n) = \sqrt{(a_0 + a_1 n + a_2 n^2)^2 + (b_0 + b_1 n + b_2 n^2)^2} \tag{4.50}$$

Dérivée de l'amplitude instantanée :

$$\frac{\partial A(n)}{\partial n} = \frac{(a_0 a_1 + b_0 b_1) + (a_1^2 + 2a_0 a_2 + b_1^2 + 2b_0 b_2)n + 3(a_1 a_2 + b_1 b_2)n^2 + 2(a_2^2 + b_2^2)n^3}{A(n)} \tag{4.51}$$

Phase instantanée :

$$\phi(n) = \omega \frac{n}{Fe} + \text{atan}\left(\frac{b_0 + b_1 n + b_2 n}{a_0 + a_1 n + a_2 n} \right) \tag{4.52}$$

Phase initiale :

$$\phi(0) = \text{atan}\left(\frac{b_0}{a_0} \right) \tag{4.53}$$

Fréquence instantanée

$$\frac{\partial \phi(n)}{\partial n} = \frac{\omega}{Fe} + \frac{(b_1 a_0 - a_1 b_0) + 2(b_2 a_0 - a_2 b_0)n + (b_2 a_1 - b_1 a_2)n^2}{A^2(n)} \tag{4.54}$$

Variation de la fréquence instantanée

$$\begin{aligned} \frac{\partial^2 \phi(n)}{\partial n^2} = & \frac{[2(b_2 a_0 - a_2 b_0) + 2n(b_2 a_1 - b_1 a_2)]}{A^2(n)} \\ & - 2\frac{[(b_1 a_0 - a_1 b_0) + 2(b_2 a_0 - a_2 b_0)n + (b_2 a_1 - b_1 a_2)n^2]\frac{\partial A(n)}{\partial n}}{A^3(n)} \end{aligned} \tag{4.55}$$

Observations Le nombre de degrés de liberté du modèle est grand. Ceci est à la fois un avantage puisqu'une grande variété de signaux peut être représentée, et un inconvénient puisque cette liberté se traduit par la modélisation de bruit dans les composantes sinusoïdales. Le modèle ne se réfère pas explicitement à une variation d'ordre délimité de fréquence et

d'amplitude. Même si une moyenne et une variation linéaire d'amplitude et de fréquence peuvent s'en déduire, cette non-spécification se traduit le plus souvent par une oscillation des paramètres sur la durée de l'observation (voir FIG. 4.7).

Le système se résout de manière globale pour l'ensemble des composantes permettant la prise en compte des recouvrements spectraux. La résolution est effectuée par inversion de matrice, ce qui pose le problème du conditionnement de la matrice (conditionnement mauvais pour des composantes proches en fréquence), ainsi que du coût de calcul nécessité (inversion d'une matrice de taille $(H * Q * 2, L)$ où H est le nombre de composantes et L la durée de l'observation).

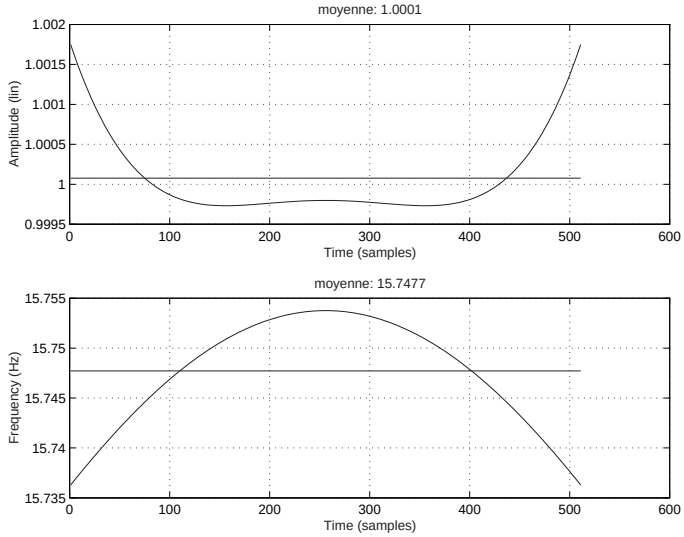


FIG. 4.7 – Estimation de l'amplitude et de la fréquence (+ moyenne temporelle) par la méthode du polynôme d'amplitude complexe. Signal : sinusoïde d'amplitude constante au cours du temps égale à 1, et de fréquence constante au cours du temps égale à 15.75 Hz, fréquence d'initialisation : $f_h = 16$ Hz

4.2.3 Comparaison des méthodes d'estimation de fréquence et d'amplitude

Le comportement des différents estimateurs de fréquence et d'amplitude, présentés dans cette partie, a été étudié sur un ensemble de signaux tests représentant différentes conditions d'analyse : échelle du spectre, résolution spectrale, finesse spectrale, présence de bruit, signaux non-stationnaires.

Les principaux résultats de cette expérience sont résumés ici. Les résultats complets sont accessibles au lecteur intéressé sur simple demande.

Les estimateurs étudiés sont :

i	interpolation du spectre sur $[\omega_{k-1}\omega_k\omega_{k+1}]$ (voir équation (4.3))
r	régression du spectre sur $[\omega_{k-2}\omega_k\omega_{k+2}]$
R	régression du spectre sur $[\omega_{k-1}\omega_k\omega_{k+1}]$ en imposant la forme du spectre autour de ω_h (voir équation (4.4) et (4.6))
d	mesure de distorsion du spectre complexe (voir équation (4.41))
p	fréquence instantanée (voir équation (4.9))
h	moindres carrés itératifs (voir équation (4.25))
l	moindres carrés polynôme d'amplitude complexe (voir équation (4.48))

Le **premier objectif** de cette comparaison est de déterminer quels sont les **paramètres** d'analyse optimaux pour les estimateurs morphologiques **i** , **r** , **R** et **d** : spectre en échelle de puissance ou logarithmique, utilisation ou pas de prolongement par zéro , fenêtre de type Blackman ou Gauss.

Le **second objectif** de cette comparaison est de déterminer quels sont les **meilleurs estimateurs** (parmi ceux considérés) pour chacun des types suivants de signaux : signaux stationnaires, signaux non-stationnaires (modulation de fréquence et d'amplitude), signaux à composantes multiples dans le cas d'une analyse à bande large (recouvrements spectraux).

La comparaison est effectuée sur des signaux tests.

Dans chaque cas, nous étudions l'influence du rapport signal à bruit (SNR) ainsi que l'influence du décalage (δ) de la fréquence de la composante de la sinusoïde par rapport aux fréquences discrètes de la TFDCT.

La comparaison est effectuée sous forme de calcul de biais, de variance et d'erreur quadratique moyenne (mse) des estimateurs de fréquence et d'amplitude sur un ensemble de 1000 réalisations.

Remarque à propos de l'estimation de l'amplitude pour p : *L'estimateur* **p** *est un estimateur uniquement de fréquence. Étant donné ses bonnes propriétés (mse faible constatées a posteriori), nous l'utilisons afin de tester un estimateur d'amplitude. Cette estimateur d'amplitude est l'estimateur de minimisation de l'erreur quadratique de modélisation* **locale** *en fréquence :*

$$A_h = \frac{\sum_{\omega_k \in W_h} S(\omega_k) H(\omega_h - \omega_k)}{\sum_{\omega_k \in W_h} |H(\omega_k)|^2}$$

Nous renvoyons le lecteur à l'annexe L pour de plus amples détails sur le protocole de cette expérience.

4.2.3.1 Détermination des paramètres d'analyse optimaux pour les estimateurs morphologiques

◊ ***Expe1 : Paramètres d'analyse***

Cette expérience nous permet de conclure que :

- L'utilisation d'une échelle logarithmique ainsi que d'un facteur de prolongement par zéro supérieur à 1 permettent de diminuer significativement le biais de l'ensemble des estimateurs morphologiques.
- L'utilisation d'une fenêtre de type Gauss de facteur de troncature faible ($L = 12\sigma$) permet de diminuer l'ensemble des biais par rapport à l'utilisation d'une fenêtre de Blackman. En présence de bruit, l'utilisation d'une fenêtre de type Gauss n'apporte que peu de bénéfice. A l'inverse l'utilisation d'une fenêtre de type Gauss de facteur de troncature élevé ($L = 6\sigma$) augmente le biais des estimateurs par rapport à l'utilisation d'une fenêtre de Blackman.
- Parmi les estimateurs morphologiques considérés, l'estimateur **R** possède le biais minimal en l'absence de bruit. En présence de bruit, les estimateurs **i** et **d** possèdent la valeur mse minimale. L'estimateur de variance minimale est l'estimateur **r** .

Ceci est illustré aux figures Fig. 4.8, Fig. 4.9, Fig. 4.10. Chacune de ces figures correspond à un rapport signal à bruit (SNR) différent :

- Fig. 4.8 : SNR = ∞ dB
- Fig. 4.9 : SNR = 1.5 dB
- Fig. 4.10 : SNR = -1.5 dB

Sur chaque figure, l'abscisse représente les différents estimateurs morphologiques considérés : **i** , **r** , **R** et **d** . Les figures de gauche correspondent à l'estimation de la fréquence, celles de droite à l'estimation de l'amplitude. En l'absence de bruit, seules les valeurs de biais sont indiquées. En présence de bruit, sur chaque figure est indiquée du haut vers le bas : le biais, l'écart-type et la racine carrée de la valeur mse. Les valeurs indiquées de biais, d'écart-type et de mse sont les valeurs moyennes pour l'ensemble des décalages δ. Les valeurs sont indiquées en échelle logarithmique.

La signification des traits sur chacune des figures Fig. 4.8, Fig. 4.9 et Fig. 4.10 est la suivante :

- Estimation sur le spectre de puissance (+ –),
- Estimation sur le spectre de log-amplitude (+ - -),
- Estimation sur le spectre de log-amplitude et utilisation d'un facteur de prolongement par zéro =2 (+ -.),
- Estimation sur le spectre de log-amplitude, utilisation d'un facteur de prolongement par zéro =2 et d'une fenêtre de type Gauss-$L = 12\sigma$ (+ . . .),
- Estimation sur le spectre de log-amplitude, utilisation d'un facteur de prolongement par zéro =2 et d'une fenêtre de type Gauss-$L = 6\sigma$ (x –).

◊ ***Expe2 : Influence du décalage δ sur l'estimation***

Le décalage δ entre la fréquence de la sinusoïde et la fréquence discrète la plus proche de la TFDCT influence la qualité des estimateurs morphologiques. Ceci a été montré par [MD92]. Contrairement à [MD92], nous constatons que tous les estimateurs morphologiques

i , **r** , **R** et **d** sont influencés de manière équivalente par δ (variations relatives [9] identiques).

Ceci est illustré aux FIG. 4.11 (en l'absence de bruit, SNR=∞) et FIG. 4.12 (en présence de bruit, SNR=1.5) pour une estimation sur le spectre de puissance et une fenêtre de type Blackman. L'abscisse de chacune de ces figures représente le paramètre δ, l'ordonnée représente la valeur absolue du biais de l'estimateur (à gauche : estimateur de fréquence, à droite : estimateur d'amplitude).

En l'absence de bruit, le biais de l'ensemble des estimateurs de fréquence est virtuellement nul en $\delta = 0$. Pour les estimateurs d'interpolation **i** et **d** , il est également nul en $\delta = \pm 1/2$. Ceci se comprend en constatant qu'en ces positions les estimateurs morphologiques sont indépendants du paramètre de forme du lobe et donc de son approximation par une parabole. Pour les estimateurs d'interpolation **r** et **R** , le biais est maximum autour de $\delta = \pm 1/2$ (position où le paramètre de forme du lobe a le plus de poids dans l'expression mathématique). Pour les mêmes raisons, le biais des estimateurs d'amplitude est minimum en $\delta = 0$ et maximum en $\delta = \pm 1/2$.

En présence de bruit, les valeurs mse de fréquence et d'amplitude sont minimum en $\delta = 0$ et maximum en $\delta = \pm 1/2$. Dans ce cas, le maximum de la valeur mse de fréquence s'explique non plus du fait de la mauvaise approximation de la forme du lobe par une parabole, mais du fait qu'il s'agit du décalage pour lequel la forme du lobe est le plus influencé par le bruit.

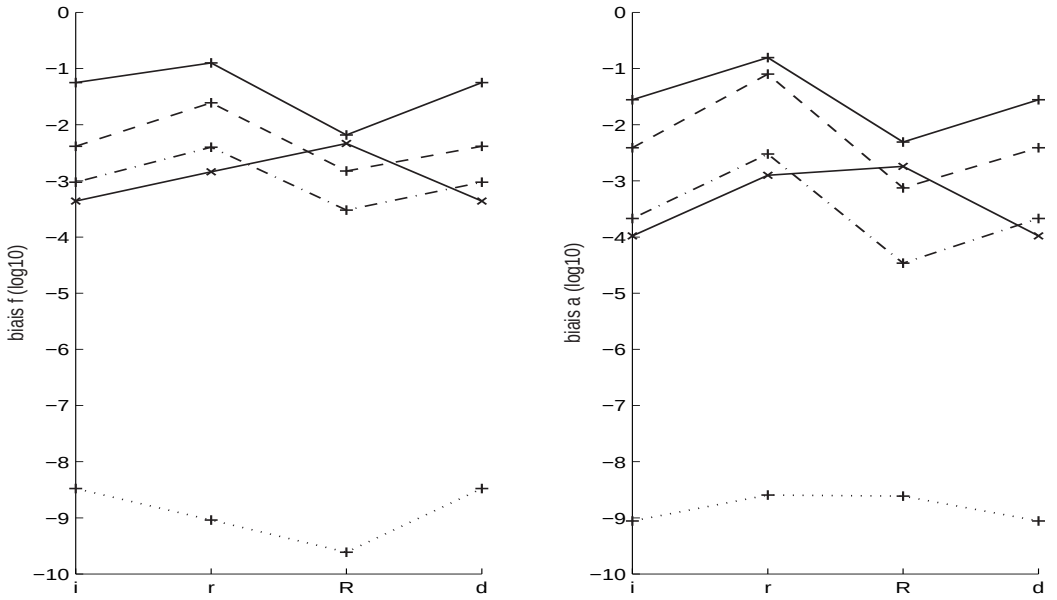


FIG. 4.8 – Expe1 : Moyenne-sur-δ des valeurs absolues des biais des estimateurs morphologiques **i** , **r** , **R** et **d** pour différents paramètres d'analyse (voir texte) en l'absence de bruit (SNR=∞))

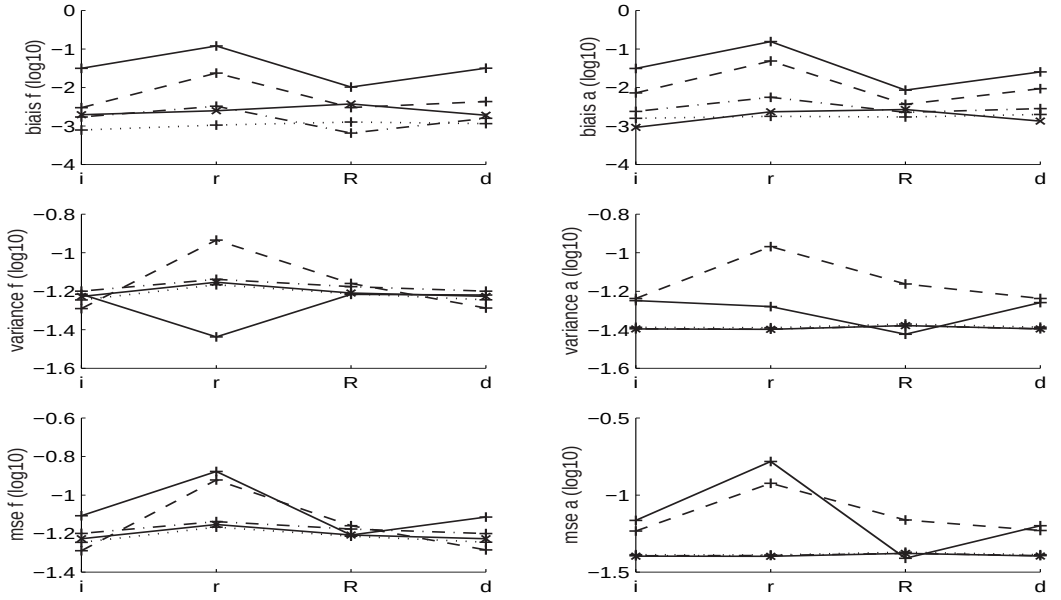


FIG. 4.9 – Expe1 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs morphologiques **i** , **r** , **R** et **d** pour différents paramètres d'analyse (voir texte) en présence de bruit (SNR=1.5, 1000 réalisations)

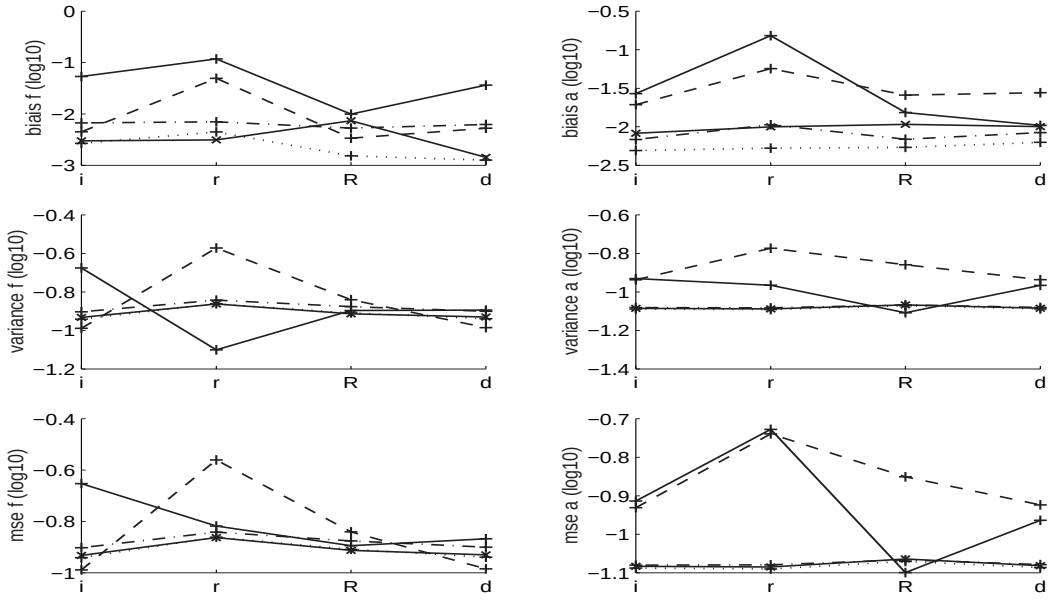


FIG. 4.10 – Expe1 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs morphologiques **i** , **r** , **R** et **d** pour différents paramètres d'analyse (voir texte) en présence de bruit (SNR=-1.5, 1000 réalisations)

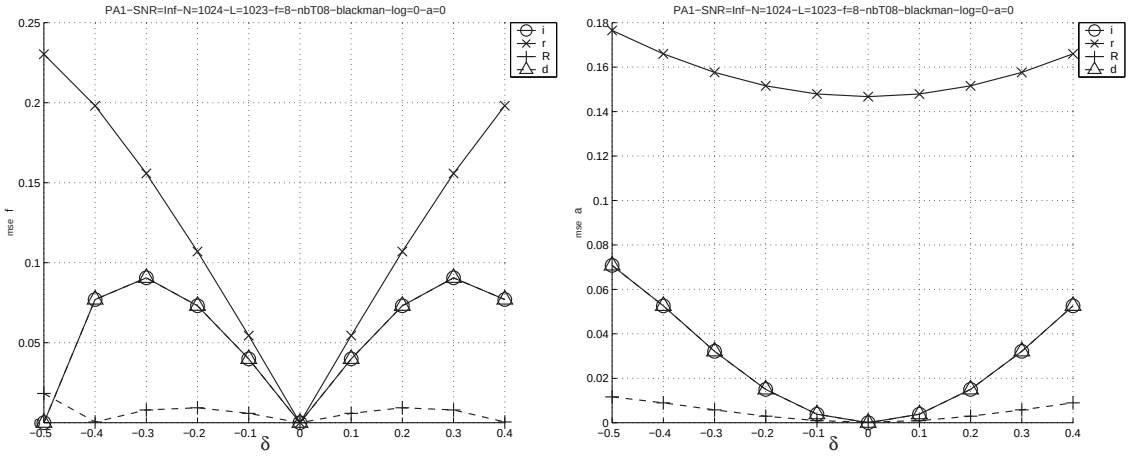


FIG. 4.11 – Expe2 : Variation des biais (ordonnées) des estimateurs morphologiques de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** et **d** en fonction du facteur de décalage δ (abscisse) pour un spectre en échelle de puissance, une fenêtre de type Blackman et en l'absence de bruit (SNR=∞)

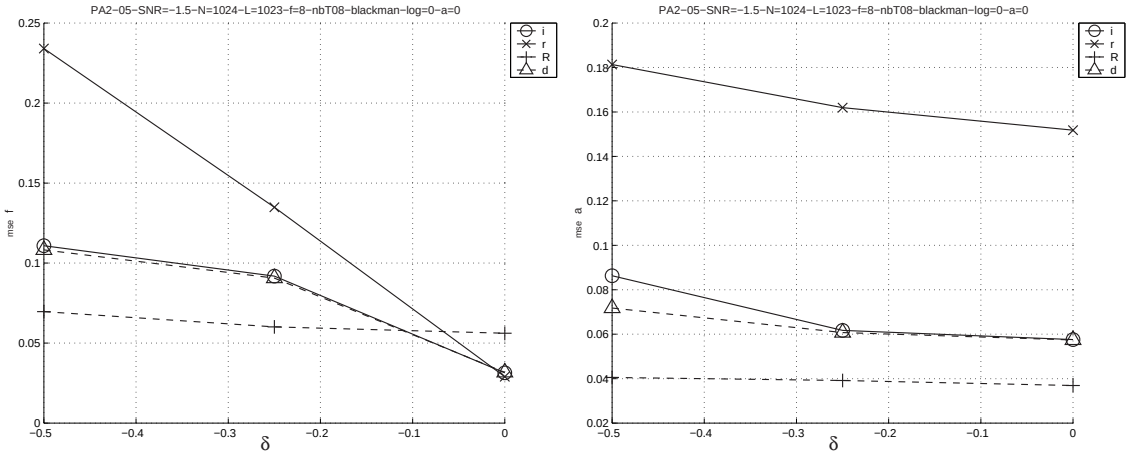


FIG. 4.12 – Expe2 : Variation des valeurs mse (ordonnées) des estimateurs morphologiques de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** et **d** en fonction du facteur de décalage δ (abscisse) pour un spectre en échelle de puissance, une fenêtre de type Blackman et en présence de bruit (SNR=1.5)

4.2.3.2 Comparaison des estimateurs pour différents types de signaux

Nous comparons l'ensemble des estimateurs (morphologiques ou non) pour différents types de signaux : signaux stationnaires, non-stationnaires (modulations de fréquence et d'amplitude) et signaux à composantes multiples dans le cas de basses résolutions (recouvrements spectraux). Les paramètres d'analyse utilisés pour les estimateurs morphologiques sont ceux résultant de l'étude précédente (spectre en échelle logarithmique, facteur de prolongement par zéro =2, fenêtre de type Blackman). Les paramètres d'analyse pour les estimateurs non morphologiques sont ceux discutés dans l'annexe L (fenêtre rectangulaire pour **h** et **l** , fenêtre de Hamming pour **p**).

♢ ***Expe3 : Signal stationnaire***

Dans le cas d'un signal stationnaire et d'une résolution fréquentielle normale (observation de 8 périodes fondamentales), les meilleurs estimateurs de fréquence sont **h** et **p** . Les meilleurs estimateurs d'amplitude sont **R** , **h** et **p** . En présence de bruit, les meilleurs estimateurs sont les estimateurs de moindres carrés, **h** et **l** , ainsi que la fréquence instantanée **p** et l'estimateur d'amplitude par moindre carré local en fréquence **p** .

Ceci est illustré à la Fig. 4.13. Les différents traits représentent les différents rapports signal à bruit :

- SNR=∞ (+ –)
- SNR=1.5 (+ - -)
- SNR=-1.5 (+ -.)

L'étude du comportement des estimateurs de fréquence et d'amplitude en fonction de δ nous montre que les estimateurs **h** et **p** sont virtuellement indépendants de δ. Ceci est illustré aux Fig. 4.14 (en l'absence de bruit, SNR=∞) et Fig. 4.15 (en présence de bruit, SNR=1.5).

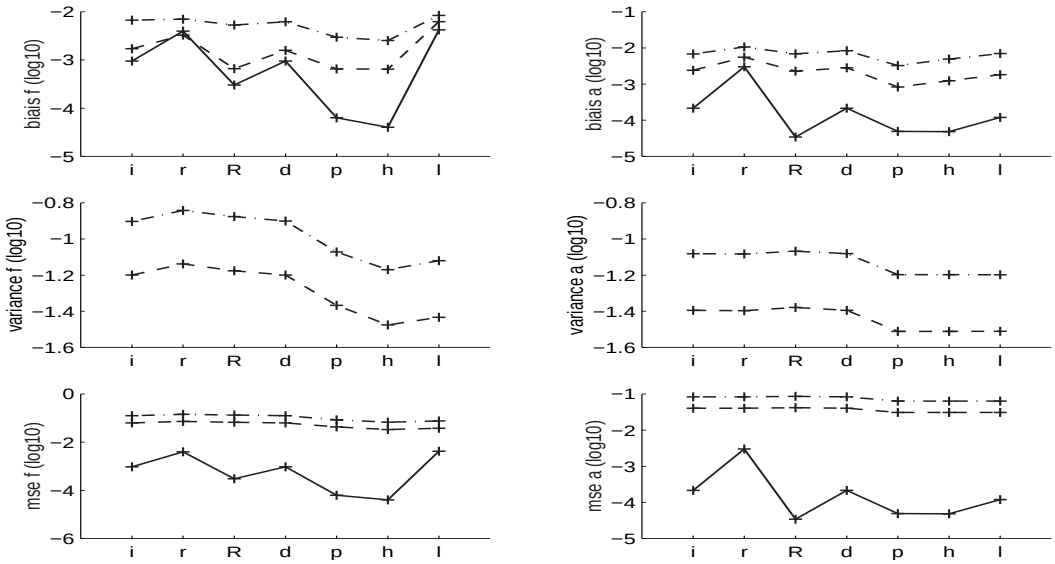


FIG. 4.13 – Expe3 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** , **d** , **p** , **h** , **l** pour trois niveaux de bruit différents (SNR=∞, 1.5 et -1.5) (voir texte)

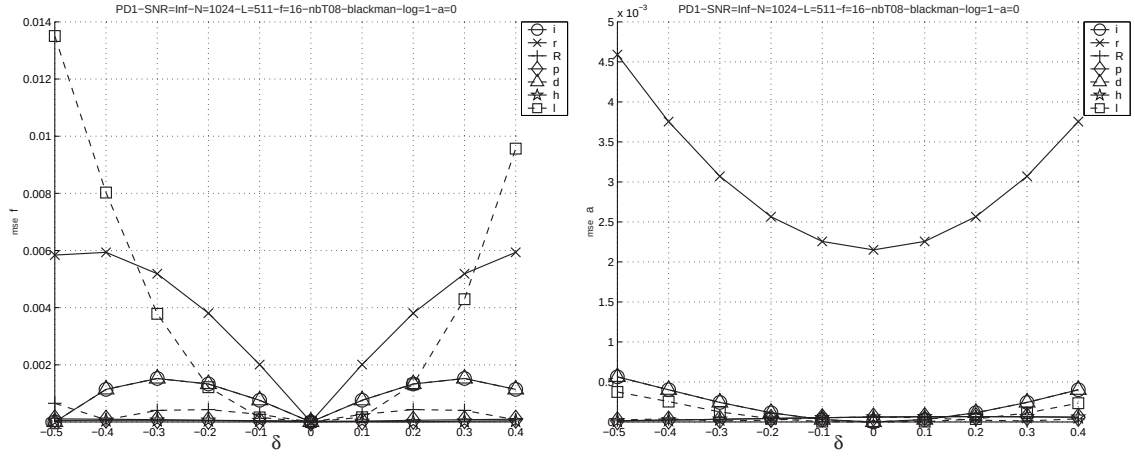


FIG. 4.14 – Expe3 : Variation des biais (ordonnées) des estimateurs de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** et **d** , **p** , **h** , **l** en fonction du facteur de décalage δ (abscisse) en l'absence de bruit (SNR=∞)

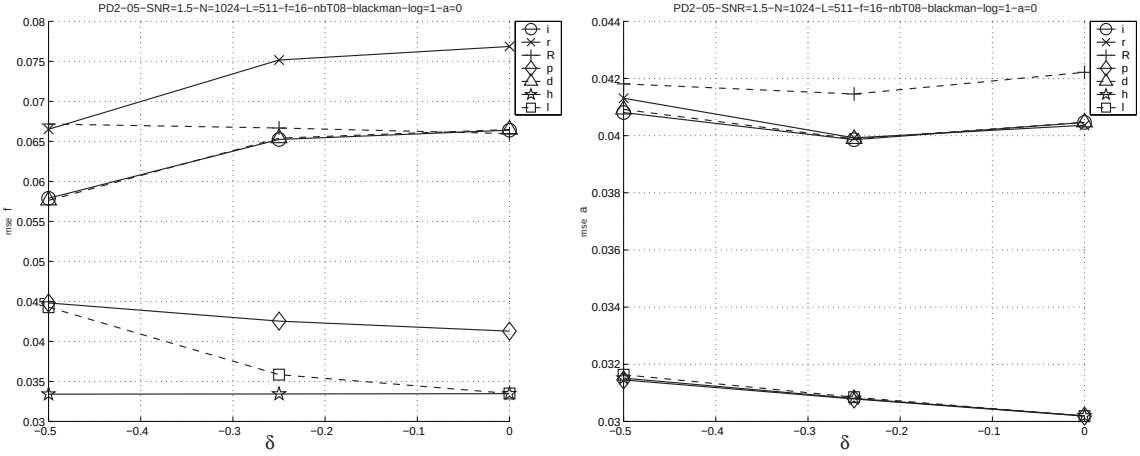


FIG. 4.15 – Expe3 : Variation des valeurs mse (ordonnées) des estimateurs morphologiques de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** et **d** , **p** , **h** , **l** en fonction du facteur de décalage δ (abscisse) en présence de bruit (SNR=1.5)

◊ ***Expe4 : Signal non-stationnaire***

Dans le cas d'un signal présentant simultanément des modulations de fréquence et d'amplitude, le meilleur estimateur tant de fréquence que d'amplitude est **d** . En présence de bruit, l'estimateur **l** , de par sa variance faible, constitue également un estimateur intéressant. Des deux estimateurs permettant l'estimation des paramètres de modulation (estimateurs du taux de variation linéaire de fréquence et d'amplitude), **d** et **l** , le meilleur estimateur est **d** (précision relative de l'estimation du taux de modulation de fréquence : 0.85 pour **d** , 0.65 pour **l**; de modulation d'amplitude : 0.99 pour **d** , 0.35 pour **l**).

L'utilisation d'une fenêtre de Gauss de paramètre $L = 12\sigma$ au lieu d'une fenêtre de Blackman permet de diminuer encore davantage le biais de l'estimateur **d** de fréquence, d'amplitude et de paramètres de modulations. Une fenêtre de type Gauss de paramètre $L = 6\sigma$ permet également de diminuer ces biais mais dans une moindre mesure. Le passage d'une fenêtre de Blackman à une fenêtre de Gauss est donc intéressant dans le cas de signaux à modulation. Le gain est cependant faible en présence de bruit.

Ceci est illustré aux FIG. 4.16, FIG. 4.17 et FIG. 4.18. Chacune des trois figures FIG. 4.16, FIG. 4.17 et FIG. 4.18 correspond à un rapport signal à bruit différent :

- FIG. 4.16 : SNR=∞
- FIG. 4.17 : SNR=1.5
- FIG. 4.18 : SNR=-1.5

La signification des traits sur chacune des figures FIG. 4.16, FIG. 4.17 et FIG. 4.18 est la suivante :

- fenêtre de type blackman (+ –)
- fenêtre de type gauss-12σ (+ - -)
- fenêtre de type gauss-6σ (+ - .)

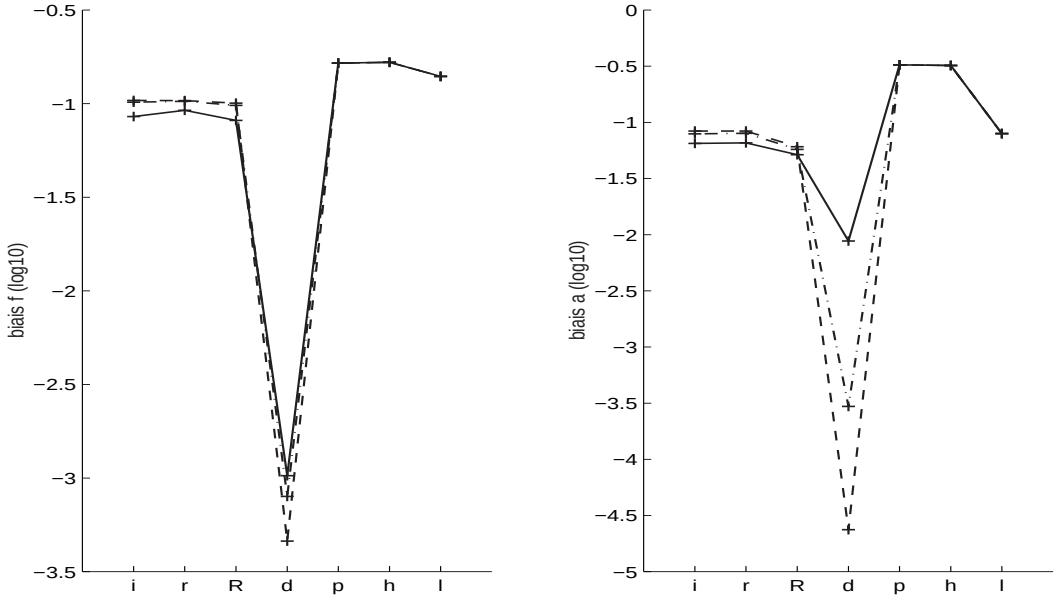


FIG. 4.16 – Expe4 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs **i** , **r** , **R** , **d** , **p** , **h** et **l** pour trois types de fenêtres différents (voir texte) en l'absence de bruit (SNR=∞)

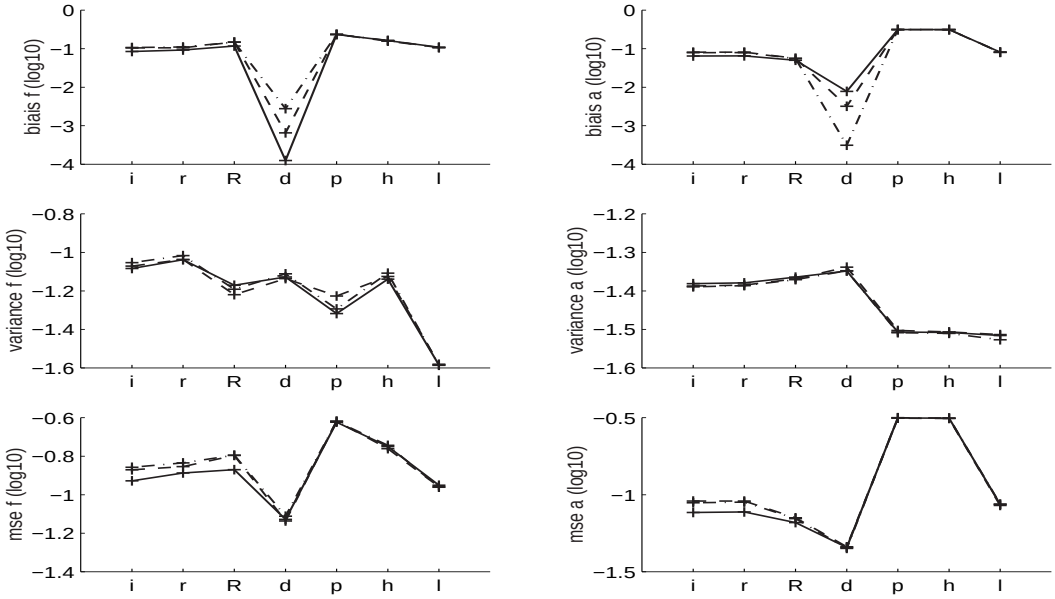


FIG. 4.17 – Expe4 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs **i** , **r** , **R** , **d** , **p** , **h** et **l** pour trois types de fenêtres différents (voir texte) en présence de bruit (SNR=1.5)

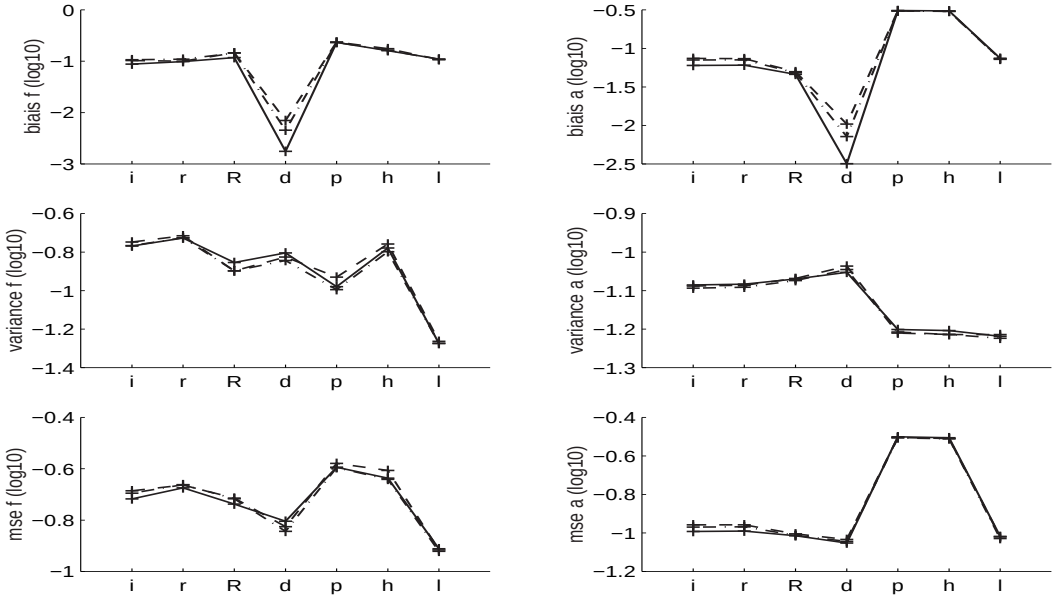


FIG. 4.18 – Expe4 : Moyenne-sur-δ des valeurs absolues des biais, variances et mse des estimateurs **i** , **r** , **R** , **d** , **p** , **h** et **l** pour trois types de fenêtres différents (voir texte) en présence de bruit (SNR=-1.5)

◊ *Expe5 : Signal à composantes multiples, analyse à bande large*

Dans la dernière expérience, nous nous intéressons à l'estimation de fréquence et d'amplitude dans le cas de signaux à composantes multiples et lorsque la résolution spectrale est faible (observation de 3 périodes fondamentales, donc recouvrements spectraux).

Dans le cas de résolutions spectrales faibles, les relations de phase entre composantes adjacentes jouent un rôle déterminant sur l'estimation tant de fréquence que d'amplitude. Ceci est illustré à la figure FIG. 4.19. Le signal est composé de trois composantes aux fréquences ω_1, ω_2 et ω_3 en rapports harmoniques. L'amplitude des trois composantes est prise par simplicité égale à 1. La phase de la deuxième composante ϕ_2 est gardée égale à 0. L'abscisse de la FIG. 4.19 représente les différentes relations de phase possibles entre les phases ϕ_1 et ϕ_3 et la phase ϕ_2 également représentées dans le tableau ci-dessous.

$\phi(\omega_1)$	0	$\frac{\pi}{2}$	π	0	$\frac{\pi}{2}$	π	0	$\frac{\pi}{2}$	π
$\phi(\omega_2)$	0	0	0	0	0	0	0	0	0
$\phi(\omega_3)$	0	0	0	$\frac{\pi}{2}$	$\frac{\pi}{2}$	$\frac{\pi}{2}$	π	π	π

L'ordonnée représente le biais introduit sur l'estimation de la fréquence de ω_2 (figure de gauche) et de l'amplitude a_2 (figure de droite).

Ceci montre que les estimateurs locaux en fréquence (estimateurs morphologiques et estimateur des moindres carrés local en fréquence), à l'exception de **p** , sont influencés par les relations de phase des composantes.

Les estimateurs de résolution globale en fréquence, **h** et **l** , ne sont que faiblement influencés par les composantes voisines et sont dans ce cas les meilleurs estimateurs avec l'estimateur de fréquence local en fréquence **p** .

A la figure FIG. 4.20, nous montrons les valeurs mse des estimateurs dans un des cas critiques de relation de phase ($\phi_2 - \phi_1 = \pi, \phi_3 - \phi_1 = 0$) et en présence de bruit.

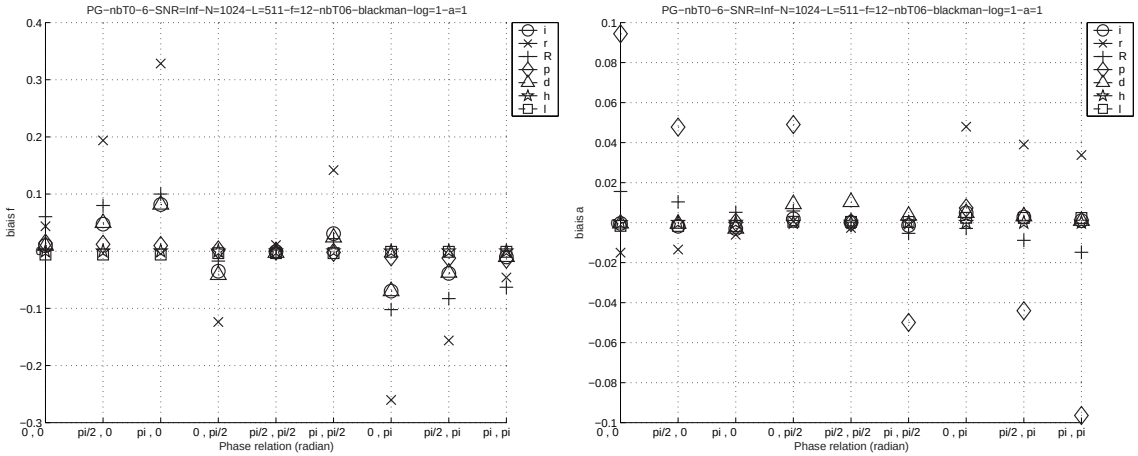


FIG. 4.19 – Expe5 : Influence des relations de phase entre composante étudiée et composantes voisines de la composante étudiée (abscisse) sur la valeur du biais (ordonnée) des estimateurs de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** , **d** , **p** , **h** et **l** en l'absence de bruit (SNR=∞)

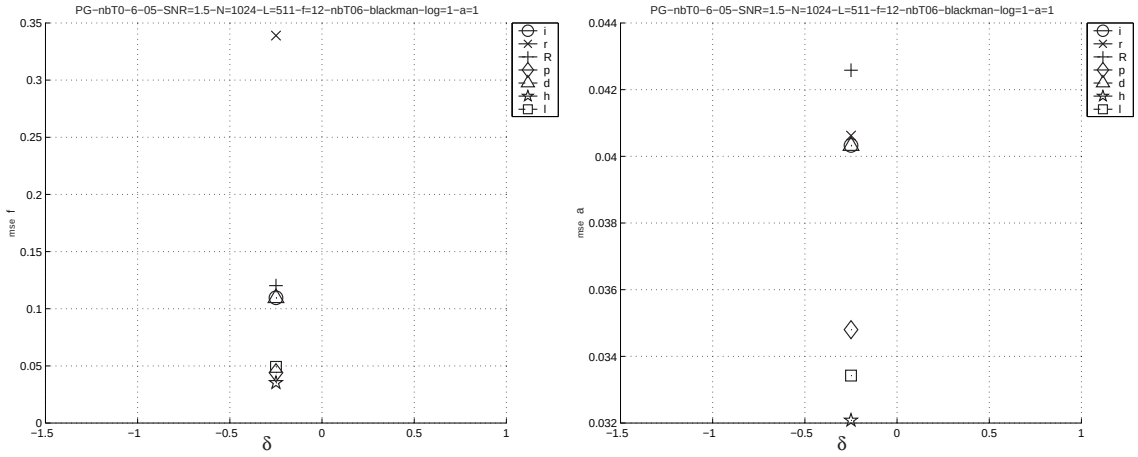


FIG. 4.20 – Expe5 : Valeur mse des estimateurs de fréquence (panneau de gauche) et d'amplitude (panneau de droite) **i** , **r** , **R** , **d** , **p** , **h** et **l** en présence de bruit (SNR=1.5) dans le cas d'une relation de phase entre composantes $(\phi 1, \phi 2, \phi 3) = (\pi, 0, 0)$

4.2.4 Conclusion

Dans cette partie, nous avons étudié les estimateurs les plus communément utilisés en modélisation du signal musical. Ces estimateurs ont été divisés en plusieurs classes, selon que l'estimation s'effectue de manière locale ou globale en fréquence, en utilisant la forme du spectre - estimateurs morphologiques - ou en minimisant un critère d'erreur de modélisation - estimateur des moindres carrés , que l'estimation permette la prise en compte de variations du signal ou non.

Nous avons proposé un modèle sinusoïdal fondé sur la mesure de la distorsion du spectre, et dont les paramètres d'amplitude et de fréquence varient de manière linéaire sur la durée de l'observation.

Une comparaison des estimateurs sur une base de signaux tests et dans divers conditions d'analyse nous permet de conclure sur les points suivants :

Parmi les quatres estimateurs morphologiques considérés : **r** , **R** et **i** / **d** , l'utilisation des estimateurs **i** / **d** semble s'imposer. Ceci pour plusieurs raisons :

1. le besoin de déterminer la forme de la parabole (paramètre **s** de la parabole) pour l'estimateur **R** . Cette détermination n'est pas aisée et doit s'effectuer pour chaque rapport L (longueur de fenêtre) sur N (nombre de points) de la TFDCT.
2. l'imposition de la forme de la parabole dans **R** provoque un biais important en présence de modulation. Du fait de la non-imposition de la forme du lobe, l'estimateur **i** est moins influencé par la présence de modulations, alors que l'estimateur **d** corrige l'estimation de fréquence et d'amplitude en fonction de son estimation des modulations.
3. l'action conjointe du passage en échelle logarithmique et du prolongement par zéro font que le biais de **d** est d'un ordre de grandeur acceptable (supérieur à la valeur de **R** mais inférieur à la valeur de **R** en échelle puissance et en l'absence de prolongement par zéro).

Ceci est donc en contradiction avec les conclusions de [MD92].

Le choix doit maintenant s'effectuer entre

- un estimateur morphologique, permettant la prise en compte de modulations de fréquence et d'amplitude, et permettant de ce fait des observations sur des durées plus longues (évitant de ce fait le problème de recouvrement spectraux)
- un estimateur du type moindre carré de résolution globale en fréquence, permettant une estimation sur des fenêtres de courtes durées et permettant de ce fait d'améliorer l'approximation à l'ordre 0 des paramètres du signal.

4.3 Détection de composantes sinusoïdales

4.3.1 Introduction

Dans la partie précédente, nous nous sommes intéressés à l'estimation des paramètres de modèles sinusoïdaux (stationnaires ou non) supposant la localisation des régions du plan temps/fréquence représentables par des sinusoïdes, connue. Dans cette partie, nous nous intéressons à cette localisation que nous appelons, de manière quelque peu abusive [10] , détection de sinusoïdes.

Bien qu'il existe de nombreuses méthodes permettant cette localisation, nous nous intéressons seulement à deux classes de méthodes de détection de sinusoïdes : fondée sur l'erreur de modélisation ou de représentativité d'un signal par un modèle, et fondée sur l'erreur de spécification d'un modèle étant donné un ensemble d'observations issues d'un signal.

Erreur de modélisation ou de représentativité La première classe de méthodes est fondée sur le calcul de l'erreur commise en modélisant une région du plan temps/fréquence par un modèle sinusoïdal particulier. Le terme **erreur de modélisation** généralement utilisé prête à confusion, puisque cette erreur dépend non seulement de la qualité de l'estimation des paramètres du modèle sinusoïdal considéré (**erreur d'estimation ou de spécification**) et de la propension du modèle sinusoïdal considéré à représenter la région du plan temps/fréquence considérée (**erreur de représentativité ou caractère réducteur du modèle**). Ainsi, une sinusoïde noyée dans du bruit, quelle que soit la qualité de l'estimation de ses paramètres (erreur d'estimation nulle) aura une erreur de modélisation importante du fait de sa faible représentativité de l'énergie du signal (caractère réducteur faible). L'erreur de représentation dépend non seulement de la spécification du modèle (à la fois du choix du modèle et de l'estimation de ses paramètres), mais également de la largeur de la région T/F considérée [11] .

À chaque modèle sinusoïdal, nous pouvons associer une erreur de modélisation. Pour cela, nous calculons le pourcentage de l'énergie d'une région du plan T/F non expliquée par le modèle sinusoïdal considéré. Cependant, parmi l'ensemble des modèles, celui de fréquence et d'amplitude constante donne lieu à une interprétation particulièrement intéressante. En effet, nous montrons plus loin que pour ce modèle, l'erreur de modélisation est équivalente à la corrélation du spectre complexe par la réponse fréquentielle de la fenêtre d'analyse. Cette corrélation complexe, ne nécessitant pas l'estimation des paramètres du modèle, constitue donc un estimateur de sinusoïdalité [12] (un estimateur du caractère représentable par une sinusoïde) et nous fournit donc un indice pouvant être utilisé pour la détection.

Erreur de spécification La deuxième classe de méthodes reflète d'avantage le caractère «détection» que «représentation». Ces méthodes permettent, dans le cas théorique d'une estimation parfaite, la détection de sinusoïdes dans du bruit. Le modèle sinusoïdal est un modèle fréquentiel mais également un modèle d'évolution temporelle de composantes fréquentielles. En ce sens, la validité d'un modèle sinusoïdal doit se vérifier non pas seulement pour un spectre donné (validité interne ou validité pour l'ensemble des observations ayant permis la création du modèle), mais également dans l'évolution temporelle de ses paramètres (validité externe ou extérieure aux observations ayant permis la création du modèle). Les méthodes de cette classe reposent généralement sur

des critères de régularité du modèle qui découlerait des estimations si celles-ci étaient considérées comme appartenant à un modèle sinusoïdal. L'estimation des paramètres d'un modèle sinusoïdal est effectuée à chaque instant pour un ensemble de régions candidates (points du spectre supposés appartenir à une composante sinusoïdale). Ces estimations sont ensuite utilisées pour la création de trajets sinusoïdaux. La régularité de ces trajets est calculée et permet de déterminer dans quelle mesure les estimations ayant permis la création de ce trajet correspondent effectivement aux paramètres d'un modèle sinusoïdal. Ces méthodes ne donnent aucune information concernant la représentativité (le caractère réducteur) d'une région du plan T/F par un modèle sinusoïdal.

Ce type de méthode ne nécessite pas de définition de largeur fréquentielle W_h de la région considérée ni du nombre H de composantes considérées.

4.3.2 Erreur de modélisation ou de représentativité

L'erreur de modélisation est une mesure très utilisée dans le domaine de la parole afin d'estimer (dans le temps et en fréquence) les régions du signal dites «voisées» par opposition aux régions dites «non-voisées» (Voiced/ UnVoiced) [GL88]. De manière simplifiée, une région est dite voisée quand le signal qu'elle renferme résulte de la vibration des cordes vocales. Une région est dite non-voisée quand elle ne résulte pas de l'action des cordes vocales mais d'autres mécanismes de production du son (plosive, bruit de constriction du conduit buccal, ...). La vibration des cordes vocales étant quasi-périodique, pour un conduit buccal à variation lente le signal résultant peut être représenté par une somme de sinusoïdes. L'erreur de modélisation mesure, dans une succession de bandes de fréquence, l'erreur commise en modélisant le signal par une somme de sinusoïdes. Pour chaque bande de fréquence une prise de décision voisé/non-voisé peut être effectuée en fonction de la valeur de l'erreur de modélisation.

4.3.2.1 Définition

Autour d'un temps t_m et pour une région W_h, l'**erreur de modélisation** est définie comme l'erreur commise en modélisant la région W_h par un ensemble de sinusoïdes. Soit $S(\omega_k)$ le spectre d'un signal, $\hat{S}(\omega_k)$ le spectre discrétisé d'un modèle sinusoïdal correspondant à S, et $\hat{S}_h(\omega_k)$ celui de la $h^{\text{ème}}$ composante de ce modèle. L'erreur de modélisation *locale* en fréquence autour du temps t_m et dans la région fréquentielle W_h est définie par

$$\epsilon_h = \frac{\sum_{\omega_k \in W_h} \left| S(\omega_k) - \sum_{h \text{ tel que } \omega_h \in W_h} \hat{S}_h(\omega_k) \right|^2}{\sum_{\omega_k \in W_h} |S(\omega_k)|^2} \tag{4.56}$$

Remarque : *Puisque la somme est limitée, cette expression suppose l'orthogonalité des composantes du modèle (recouvrements spectraux négligeables). L'expression complète doit comporter également la contribution de l'axe négatif des fréquences.*

Dans le cas d'un signal périodique, une erreur de modélisation peut être calculée pour chacune des harmoniques du signal. Pour une fréquence fondamentale ω_0, nous pouvons définir les bandes de fréquence comme $W_h = [(h-0.5)\omega_0, (h+0.5)\omega_0]$. L'erreur de modélisation *globale* en fréquence est obtenue en choisissant W_h égale à l'axe des fréquences.

Nous reviendrons plus loin sur la détermination de W_h qui revête un caractère important, mais nous voudrions maintenant expliquer la raison de l'utilisation d'une erreur de modélisation pour la détection de sinusoïde.

Dans le cas particulier où le modèle sinusoïdal considéré est un modèle de fréquence et d'amplitude constantes ($\hat{S}_h(\omega_k) = A_h H(\omega_h - \omega_k)$), l'erreur de modélisation donne lieu à une interprétation en termes de ressemblance du signal à une sinusoïde.

4.3.2.2 Corrélation complexe

Telle que définie jusqu'à présent, l'erreur de modélisation peut se comprendre comme l'erreur commise en projetant le spectre du signal sur un ensemble de fonctions constitué de sinusoïdes pondérées. Nous avons considéré cette base comme étant a priori connue. Dans le cas du signal vocal, il s'agissait de sinusoïdes d'amplitude et de fréquence constantes centrées sur les harmoniques de la fréquence fondamentale du signal. Dans le cas général d'un signal non nécessairement périodique, cette base est a priori inconnue. La détection de sinusoïde consiste à déterminer cette base.

Si nous considérons les fonctions de cet ensemble comme orthogonales (pas de recouvrements spectraux), la détermination de ces fonctions peut s'effectuer de manière locale en fréquence. Reste qu'il existe une infinité de base à explorer pour une région fréquentielle donnée [13] .

Si de plus nous considérons les fonctions de base correspondant à un modèle sinusoïdal de paramètres constants, la détermination de ces fonctions se résume à la détermination de la fréquence centrale des fonctions. Dans le cas d'un signal pondéré par une fonction $h(t)$, les fonctions de base sont définies comme l'ensemble des facteurs de transposition en fréquence possibles de $h(n)$. Nous cherchons donc les facteurs de transposition permettant de minimiser l'erreur commise en projetant le signal sur cette base. Ceci peut s'effectuer par calcul du maximum de la **corrélation complexe** entre la région W_h considérée du spectre et la réponse fréquentielle de la fenêtre de pondération pour l'ensemble des facteurs de transposition ω tel que $\omega \in W_h$.

Nous définissons la fonction de corrélation (complexe) [Rod97] entre la réponse fréquentielle de la fenêtre d'observation $H(\omega)$ centrée en ω_h et la TFDCT du signal $S(\omega_k)$ comme

$$\Gamma(\omega_h) =< S, H^* > \tag{4.57}$$

En considérant la fenêtre de pondération symétrique par rapport à l'instant d'analyse et donc sa TF réelle, son calcul sur l'intervalle W_h s'exprime

$$\boxed{\Gamma(\omega_h, W_h) = \sum_{\omega_k \in W_h} H(\omega_h - \omega_k) \cdot S(\omega_k)} \tag{4.58}$$

dans lequel W_h désigne la bande de fréquence considérée. [14] .

Nous définissons les normes L^2 de H et de S sur W_h comme

$$\|H\|^2_{\omega_h, W_h} = \sum_{\omega_k \in W_h} |H(\omega_h - \omega_k)|^2 \qquad \|S\|^2_{\omega_h, W_h} = \sum_{\omega_k \in W_h} |S(\omega_k)|^2 \tag{4.59}$$

Le coefficient de sinusoïdalité est défini comme

$$\boxed{|c(\omega_h, W_h)| = \frac{|\Gamma(\omega_h, W_h)|}{\|H\|_{\omega_h, W_h} \cdot \|S\|_{\omega_h, W_h}}} \tag{4.60}$$

Remarque : *W_h est défini comme un intervalle en fréquence continue. L'application de (4.60) pour un spectre discret nécessite la quantification de la valeur de W_h en un nombre discret de fréquences (bins*

du spectre discret). Il va de soi que l'utilisation d'un facteur de prolongement par zéro supérieur à 1 permet de diminuer l'erreur de quantification commise. La même remarque est à formuler pour $\|H\|^2_{\omega_h,W_h}$ *et* $\|S\|^2_{\omega_h,W_h}$.

(4.60) peut être réécrit

$$|c(\omega_h, W_h)| = \sum_{\omega_k \in W_h} H^0(\omega_h - \omega_k) \cdot S^0(\omega_k) \tag{4.61}$$

dans lequel H^0 et S^0 sont les versions normées de H et S.

La relation existant entre la corrélation complexe et l'erreur de modélisation s'exprime (voir annexe M) :

$$\boxed{\epsilon_h = 1 - |c_h|^2} \tag{4.62}$$

Une interprétation naïve des valeurs de ϵ et c nous conduirait aux conclusions suivantes :
- $\epsilon(\omega_h) = 0$, $|c(\omega_h)| = 1$ indique la présence d'une sinusoïde pure de fréquence et d'amplitude constantes sur la largeur L de la fenêtre,
- $\epsilon(\omega_h) > 0$, $|c(\omega_h)| < 1$ indique la présence de bruit, d'autres composantes sinusoïdales au voisinage de ω_h ou d'une (ou plusieurs) composante(s) non-stationnaire(s) au voisinage de ω_h sur L.

Cependant, cette interprétation ne prend pas en compte le rôle déterminant de W_h et de L dans la signification de ces valeurs.

4.3.2.3 Influence de la longueur d'observation temporelle L et fréquentielle W_h

Comme dans le cas de l'estimation sinusoïdale, le choix de la durée de l'observation temporelle L et fréquentielle W_h est capital pour la détection sinusoïdale. Mais à l'inverse du choix effectué en estimation sinusoïdale (choix de W_h sur base du non-recouvrement des lobes principaux des composantes sinusoïdales), aucun choix n'est évident en détection sinusoïdale. La seule contrainte imposée est que le choix de W_h soit celui permettant la plus grande discrimination entre composantes sinusoïdales (ou régions pouvant être représentées par des sinusoïdes) et composantes non-sinusoïdales.

Dans l'annexe N, nous donnons un développement simplifié (mais rendant malgré tout compte du comportement observé) de :
- ϵ_S défini comme l'espérance de l'erreur de modélisation commise en modélisant une sinusoïde à partir de l'observation d'une sinusoïde en présence de bruit,
- ϵ_{NS} défini comme l'espérance de l'erreur de modélisation commise en modélisant une sinusoïde à partir de l'observation de bruit.

Ces espérances sont calculées en fonction de la durée L de l'observation temporelle, de la largeur W_h de l'observation fréquentielle, ainsi que du rapport signal à bruit SNR. La différence entre les deux erreurs ϵ_S et ϵ_{NS} nous donne la discrimination que l'on peut espérer obtenir entre composantes sinusoïdales et non-sinusoïdales.

Soient L la durée de l'observation temporelle (en échantillons) et W_h la largeur de l'observation fréquentielle (en Hz) ne renfermant au maximum qu'UNE composante sinusoïdale, σ_b^2 la variance du bruit, N la taille de la FFT. Nous définissons un *rapport signal à bruit SNR* comme $SNR = \frac{A^2/2}{\sigma^2}$ dans lequel A est l'amplitude de la sinusoïde et σ^2 la variance du bruit.

Nous définissons également une *bande passante équivalente* B_ω telle que $B_w = Fe \cdot ENBW$ où ENBW est l'Equivalent Noise Bandwidth défini par Harris dans [Har78] (voir annexe N).

La discrimination sinusoïde/bruit (différence entre ϵ_S, espérance de l'erreur dans une région comportant une composante sinusoïdale, et ϵ_{NS}), espérance de l'erreur dans une région ne comportant pas de composante sinusoïdale), s'écrit (voir annexe N)

$$d = \epsilon_S - \epsilon_{NS} = \underbrace{\left(1 - \frac{SNR}{SNR + 2ENBW}\right)}_{\epsilon_S} - \underbrace{\left(1 - \frac{ENBW}{\frac{W_h}{Fe}}\right)}_{\epsilon_{NS}} \tag{4.63}$$

Cette formule est valable dans le cas ou $W_h < B_w$. Dans le cas contraire ($W_h > B_w$), nous avons

$$d = \epsilon_S - \epsilon_{NS} = \underbrace{\left(1 - \frac{SNR}{SNR + 2\frac{W_\omega}{Fe}}\right)}_{\epsilon_S} - \underbrace{\left(1 - \frac{ENBW}{\frac{W_h}{Fe}}\right)}_{\epsilon_{NS}} \tag{4.64}$$

Remarquons d'abord que la discrimination ne dépend pas de la finesse fréquentielle N.

◊ *Variation de ϵ_S en fonction des paramètres*

L'erreur de modélisation ϵ_S dépend du rapport signal à bruit SNR. Plus le SNR est élevé, plus ϵ_S tend vers 0 et est indépendant des autres paramètres. Pour un SNR de 0, ϵ_S tend vers 1. Pour des valeurs plus faibles de SNR, ϵ_S dépend des autres paramètres.

Pour une observation fréquentielle W_h réduite ($W_h < B_w$), l'erreur de modélisation ϵ_S dépend également de ENBW. ENBW dépend lui-même de la durée L de l'observation temporelle. Puisque ENBW diminue quand L augmente, l'erreur ϵ_S diminue donc quand la durée de l'observation est plus grande.

Pour une observation fréquentielle W_h plus large ($W_h > B_w$), ϵ_S dépend de W_h, et la valeur de ϵ_S augmente quand la largeur fréquentielle de l'observation W_h augmente.

◊ *Variation de ϵ_{NS} en fonction des paramètres*

La valeur de ϵ_{NS} est indépendante du SNR. ϵ_S ne dépend que de la durée L de l'observation temporelle (au travers de ENBW) et de la largeur W_h de l'observation fréquentielle . Au plus la durée de l'observation temporelle L est grande, au plus l'erreur ϵ_{NS} tend vers 1. Pour un L donné l'erreur ϵ_{NS} croît lorsque W_h croît.

◊ *Conclusion : variation de la discrimination en fonction des paramètres*

La discrimination dépend du SNR ainsi que de la durée de l'observation. Une augmentation de W_h ($W_h > B_w$) provoque simultanément une augmentation de l'erreur ϵ_S et de l'erreur ϵ_{NS}. Une augmentation de L diminue ϵ_S et augmente ϵ_{NS}, ce qui augmente la discrimination.

◊ ***Illustrations***

Nous illustrons l'augmentation de la discrimination lorsque la largeur de l'observation fréquentielle W_h croît aux FIG. 4.21, FIG. 4.22 et FIG. 4.23. Le signal est un signal test composé de deux sinusoïdes en rapport harmonique (fréquences normalisées de 0.0625 et 0.125) noyé dans un bruit blanc gaussien additif ($SNR = 0dB_{20}$). La fenêtre utilisée est de Blackman, la durée de l'observation correspond à 8 périodes fondamentales. La simulation est effectuée pour 100 réalisations de bruit. Les figures du haut représentent le logarithme du spectre de puissance moyenné pour l'ensemble des réalisations, celles du bas représentent le coefficient de corrélation $|c(\omega)|$ (et non pas l'erreur de modélisation) moyenné pour l'ensemble des réalisations.

- Le choix de W_h pour la FIG. 4.21 correspond à la largeur à $-12dB_{20}$ du lobe de la fenêtre : (**Critère 1**) $W_h = \{\omega_k$ tel que $20 \log 10(H(\omega_h - \omega_k)) > -12\}$.
- Le choix de W_h pour la FIG. 4.22 correspond à (**Critère 2**) $W_h = \omega_0$ (critère identique à celui utilisé pour le coefficient de voisement en parole).
- Le choix de W_h pour la FIG. 4.23 correspond à la valeur la plus importante possible sans entrer dans le lobe de la composante voisine à plus de $-12dB_{20}$ (**Critère 3**).

Les largeurs correspondant à W_h sont indiquées par des traits horizontaux sur les spectres de log-amplitude L'incidence d'une modification de W_h se manifeste tant sur les régions comportant des composantes sinusoïdales (diminution de la valeur de $c(\omega)$ de 0.99 pour FIG. 4.21 à 0.85 pour FIG. 4.23) que sur les régions ne comportant pas de composantes sinusoïdales (diminution de 0.72 pour FIG. 4.21 à 0.38 pour FIG. 4.23). La discrimination sinusoïdes/non-sinusoïdes est augmentée de 0.27 à 0.47. Dans le cas présent, la distance entre les composantes était supposée connue. Ceci a permis d'augmenter W_h au maximum possible (Critère 3) sans interférer avec une composante sinusoïdale voisine. Dans le cas général cette distance est cependant inconnue.

Nous illustrons l'augmentation de la discrimination lorsque la durée de l'observation temporelle L croît aux FIG. 4.24 et FIG. 4.25. Pour la FIG. 4.24 l'observation est de 4 périodes fondamentales. Le $c(\omega)$ moyen est de 0.59 en région non-sinusoïdale et de 0.95 en région sinusoïdale. Pour la FIG. 4.25 l'observation est de 8 périodes. Le $c(\omega)$ moyen est de 0.45 en région non-sinusoïdale et de 0.9 en région sinusoïdale. La discrimination passe donc de 0.35 à 0.45 pour un doublement de durée d'observation et dans les conditions d'un SNR=0 (le $c(\omega)$ moyen en région non-sinusoïdale est indépendant du niveau de bruit ; par contre le $c(\omega)$ moyen en région sinusoïdale dépend du niveau de bruit).

4.3.2.4 Conclusion

L'erreur de modélisation est un critère issu du domaine de la parole (voicing/unvoicing coefficient). Son équivalence à une corrélation complexe (dans le cas d'un modèle sinusoïdal stationnaire) permet d'envisager son utilisation pour la détection des sinusoïdes. Cependant, même dans le cas d'un modèle stationnaire, l'erreur de modélisation reste mal définie du fait de sa dépendance envers la taille de la région T/F considérée. Dans le cas d'un signal harmonique, cette taille peut être déterminée sur base de la fréquence fondamentale et du respect de l'hypothèse de stationnarité locale du signal. Dans le cas général d'un signal non-nécessairement harmonique, cette taille doit permettre la discrimination maximale entre les régions représentables par une sinusoïde et celles qui ne le sont pas. L'allongement fréquentiel de cette région ne permet d'augmenter cette discrimination que dans le cas de rapports signal à bruit élevés. De plus, cet allongement est limité du fait de l'hypothèse d'orthogonalité des

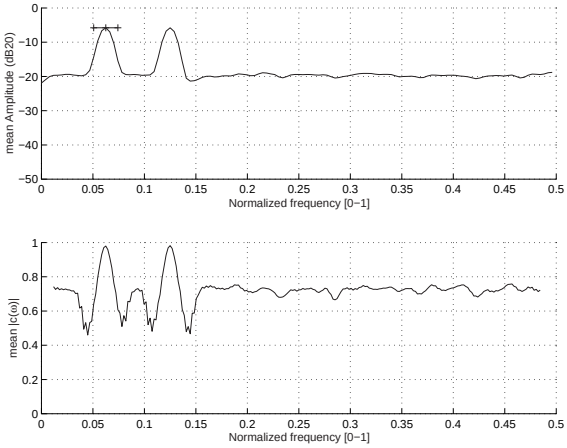


FIG. 4.21 – Spectre d'amplitude et coefficient $|c(\omega)|$ pour deux sinusoïdes aux fréquences normalisées 0.0625 et 0.125 et un bruit blanc gaussien additif : SNR = 0dB10, L=8 périodes, W_h par critère 1.

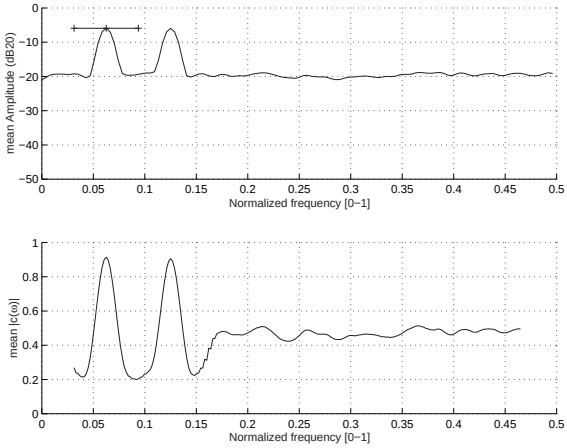


FIG. 4.22 – Spectre d'amplitude et coefficient $|c(\omega)|$ pour deux sinusoïdes aux fréquences normalisées 0.0625 et 0.125 et un bruit blanc gaussien additif : SNR = 0dB10, L=8 périodes, W_h par critère 2.

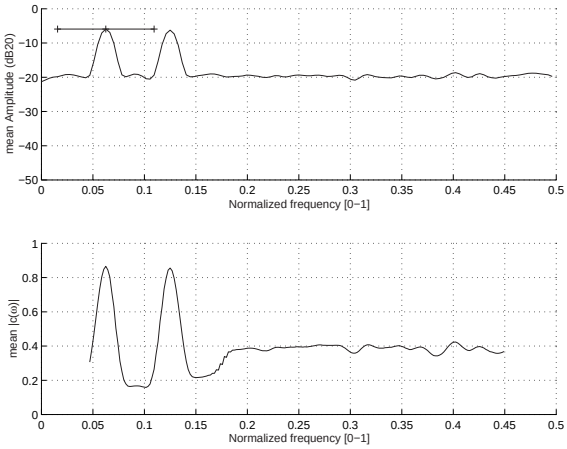


FIG. 4.23 – Spectre d'amplitude et coefficient $|c(\omega)|$ pour deux sinusoïdes aux fréquences normalisées 0.0625 et 0.125 et un bruit blanc gaussien additif : SNR = 0dB10, L=8 périodes, W_h par critère 3.

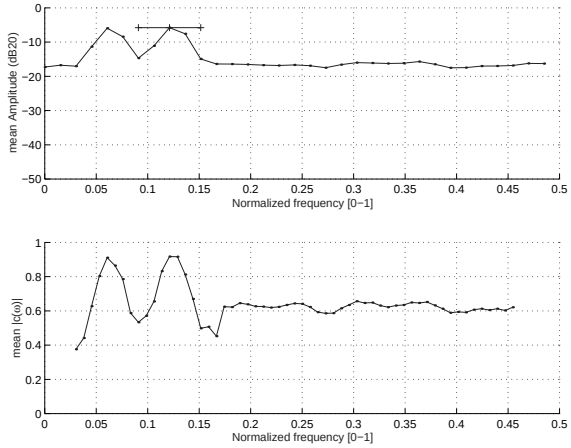


FIG. 4.24 – Spectre d'amplitude et coefficient $|c(\omega)|$ pour deux sinusoïdes aux fréquences normalisées 0.0625 et 0.125 et un bruit blanc gaussien additif : SNR = 0dB10, L=4 périodes, W_h par critère 2.

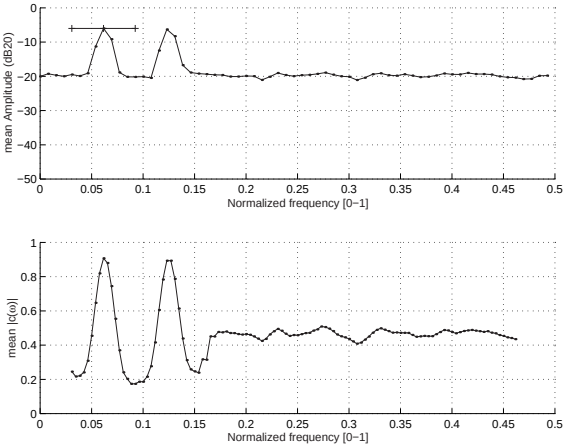


FIG. 4.25 – Spectre d'amplitude et coefficient $|c(\omega)|$ pour deux sinusoïdes aux fréquences normalisées 0.0625 et 0.125 et un bruit blanc gaussien additif : SNR = 0dB10, L= 8 périodes, W_h par critère 2.

composantes et de la non-connaissance de la séparabilité des composantes. L'allongement temporel de cette région permet d'augmenter la discrimination dans tous les cas, mais il est contraint par l'hypothèse de stationnarité locale.

L'erreur de modélisation ou erreur de représentativité ne suffit pas à elle seule à détecter parmi l'ensemble des composantes du spectre celles représentables par des sinusoïdes. Son utilisation viendra en complément d'une autre méthode de détection.

4.3.3 Erreur de spécification

A l'inverse de l'erreur de modélisation qui mesure à quel point un signal est bien représenté compte tenu d'un modèle, l'erreur de spécification mesure à quel point le choix d'un modèle et de ses paramètres est bien spécifié compte tenu d'un signal.

Pour bien comprendre le terme «erreur de spécification», prenons l'exemple d'un ensemble de données (des observations) réparties dans le plan x,y. Nous cherchons la relation $f(x)$ reliant x et y. Si nous spécifions cette relation comme linéaire, nous pouvons estimer les paramètres du modèle linéaire pour cet ensemble d'observations. Pour un ensemble d'observation plus large, nous pouvons mesurer l'erreur commise en modélisant y à partir du modèle linéaire précédent et estimer à quel point le modèle était bien spécifié. Dans le cas de la détection de sinusoïdes, nous considérons un ensemble d'observations comme les estimations d'amplitude, de fréquence et de phase (de dérivée d'amplitude et de fréquence pour un modèle sinusoïdal d'ordre 1) obtenues à l'aide d'un estimateur donné et à un instant donné. Nous désirons comparer la validité du modèle issu de cet ensemble d'observation quand ce modèle est appliqué à un autre ensemble d'observation.

Le choix d'un autre ensemble d'observations peut correspondre aux observations faites au même instant à l'aide d'un autre estimateur indépendant du premier. Nous pouvons ainsi comparer les observations faites à l'aide de l'estimateur «position des maxima d'amplitude du spectre» et celles faites à partir de l'estimateur «fréquence instantanée». Le modèle sinusoïdal issu du premier ensemble d'observation est comparé à celui issu du second ensemble d'observation.

Les deux ensembles d'observations peuvent correspondre aux observations faites par un même estimateur cette fois, mais en deux instants distincts t_m et t_{m+1}. Le modèle sinusoïdal correspondant aux observations à l'instant t_m est comparer à celui correspondant aux observations faites à l'instant t_{m+1}. Cette approche rend d'avantage compte de la définition exacte du modèle sinusoïdal donnée dans la partie 4.1 : «paramètres à variation lente». En effet, l'erreur de modélisation telle que vu précédemment ne prend en compte que la conséquence de cette définition : «paramètres localement constants».

4.3.3.1 Création de trajets de sinusoïdes

Les algorithmes de création de trajets de sinusoïdes («partial tracking») peuvent être considérés comme faisant partie de cette classe de mesure d'erreurs de spécification par comparaison d'observations en des instants disjoints. En effet, le but de ces algorithmes est de déterminer, à partir d'une succession d'ensemble d'estimations, les sous-ensembles tels que la spécification du modèle sinusoïdal à partir des sous-ensembles résultants soit correcte. Le «correct» répond au critère de «paramètre à variation lente» par celui de «régularité temporelle».

Soit $\mathbb{H}_m$ l'ensemble des composantes fréquentielles estimées en un instant t_m. Ces composantes correspondent à l'estimation des paramètres d'un modèle sinusoïdal à l'instant t_m pour chaque région du spectre. Ces estimations ont été effectuées sans connaissance à priori du contenu des régions. Ces estimations ne correspondent pas nécessairement à des régions représentables par des sinusoïdes. Soit $h_m \in \mathbb{H}_m$ une composante fréquentielle estimée à l'instant t_m, caractérisée par les paramètres A_h, ω_h, ϕ_h (également par les paramètres $\Delta\omega_h, \Delta A_h$ dans le cas d'un modèle sinusoïdal d'ordre 1). Il s'agit maintenant de déterminer le sous-

ensemble de composantes de $\mathbb{H}_m$ tel que les paramètres de ce sous-ensemble permettent la création d'un modèle sinusoïdal dont l'évolution temporelle répond à un certain critère de régularité. Les critères de régularité le plus souvent rencontrés sont : continuité fréquentielle ou/et d'amplitude, continuité de dérivée de fréquence ou/et d'amplitude, continuité de phase.

Si nous définissons dans un premier temps un «état» comme l'ensemble des paramètres associés à une estimation h_m, un trajet de sinusoïde est alors caractérisé par une succession d'états telle que la transition entre ces états soit la plus lisse possible ou telle que la probabilité de transition d'un état à un autre soit maximale.

Plusieurs points différencient les algorithmes de création de trajets :

La définition d'un état : Dans [MQ86b] et [SS90] (méthode du cône en fréquence), un état est défini comme l'ensemble des paramètres associés à une estimation h_m en un instant t_m.

Dans [Gar92], un état est défini comme l'ensemble des paramètres associés à un couple d'estimation (h_m, h'_{m+1}) en deux instants successifs t_m et t_{m+1}.

La définition des probabilités de transition entre états : Dans [MQ86b] et [SS90], la probabilité de transition (le score) est établie sur base d'un critère de continuité temporelle de fréquence

$$Ptrans(h'_{m+1}|h_m) = \exp\left(-\frac{|\omega_{h',m+1} - \omega_{h,m}|}{\sigma^2_\omega}\right) \tag{4.65}$$

Dans [Gar92], la probabilité de transition définit la probabilité du passage d'un couple $[h_m, h'_{m+1}]$ à un couple $[h'_{m+1}, h'_{m+2}]$. La probabilité de transition est définie sur base d'une continuité temporelle des dérivées de fréquence et d'amplitude

$$\begin{aligned} &Ptrans([h'_{m+1}, h_{m+2}"]|[h_m, h'_{m+1}]) = \\ &\exp\left(-\frac{(\Delta\omega_{h',m+1} - \Delta\omega_{h,m})^2}{\sigma^2_{\Delta\omega}} - \frac{(\Delta A_{h',m+1} - \Delta A_{h,m})^2}{\sigma^2_{\Delta a}}\right) \end{aligned} \tag{4.66}$$

Les dérivées $\Delta\omega_{h',m+1}$ et $\Delta a_{h',m+1}$ sont obtenues par différenciation des valeurs en t_{m+1} et t_m : $\Delta\omega_{h',m+1} = (\omega_{h,m+1} - \omega_{h,m})/T$ et $\Delta A_{h',m+1} = (A_{h,m+1} - A_{h,m})/T$.

Dans ces définitions de probabilité de transition, σ_ω, $\sigma_{\Delta\omega}$ et $\sigma_{\Delta a}$ sont définis comme des paramètres du modèle et spécifient en fréquence et en dérivées la régularité du trajet souhaitée.

Le type d'optimisation : Dans [MQ86b] et [SS90], l'optimisation est effectuée de manière locale en temps et en fréquence. Chaque trajet est considéré de manière isolée des autres et chaque transition entre états est indépendante de la précédente.

Dans [Gar92], l'optimisation est effectuée de manière globale en temps et en fréquence. L'ensemble des trajets et de leur évolution dans le temps est pris en compte dans l'optimisation. Pour cela, le problème est reformulé sous forme d'un modèle de chaîne de Markov cachée.

4.3.3.2 Méthode proposée

◊ ***Introduction***

La méthode que nous proposons ne prétend pas être optimale de tous points de vue, en particulier elle ne fait pas usage d'une optimisation globale en temps et fréquence. Cependant elle introduit de nouveaux éléments importants accentuant les contraintes propres à un modèle sinusoïdal. Sa reformulation sous forme de modèle de Markov caché afin de permettre une optimisation globale en améliorerait sûrement les performances. Les éléments nouveaux introduits sont :

- introduction de probabilités d'observation des états,
- définition de probabilités de transition sur base de comparaison et continuité d'estimateur.

De la même manière que dans [Gar92], nous définissons dans notre méthode un état comme un couple d'estimation $[h_m, h_{m+1}]$. Nous introduisons une **probabilité d'observation** d'un état, i.e. la probabilité d'observer le couple $[h_m, h_{m+1}]$ à un instant donné. Cette probabilité d'observation repose sur un critère de continuité de fréquence/amplitude et de dérivée de fréquence/amplitude [15] . Cette définition est autorisée par l'utilisation de notre modèle sinusoïdal d'ordre 1 puisque l'ajout d'un terme de pente au modèle permet de regarder la courbure de pente au sein d'un même état [16] . La prise en compte simultanée des valeurs de fréquence/amplitude ainsi que de leur dérivée est effectuée par calcul des courbures maximales des polynômes en fréquence et amplitude d'ordre 3 satisfaisant chacun aux 4 conditions limites respectives ($\omega_{h,m}$, $\omega_{h',m+1}$, $\Delta\omega_{h,m}$, $\Delta\omega_{h',m+1}$ pour le polynôme de fréquence ; $A_{h,m}$, $A_{h',m+1}$, $\Delta A_{h,m}$, $\Delta A_{h',m+1}$ pour le polynôme d'amplitude).

Pour un état donné $[h_m, h'_{m+1}]$, nous calculons une **probabilité de transition** vers un état $[h'_{m+1}, h_{m+2}"]$ sur base de la mesure de la distance euclidienne entre un modèle de fréquence passant par les observations $\omega_{h,m}, \omega_{h',m+1}, \omega_{h",m+2}$ et un modèle de fréquence dérivé de l'évolution de l'observation des phases $\phi_{h,m}, \phi_{h',m+1}, \phi_{h",m+2}$. Le choix de cette probabilité de transition est effectué de manière à la rendre indépendante de la probabilité d'observation.

◊ *Probabilité d'observation : méthode de courbure polynomiale*

La probabilité que nous proposons dans [PR99a] [PR99b] repose simultanément sur une continuité de valeur et de dérivée des fréquences et des amplitudes.

Soit t_m et t_{m+1} deux temps d'analyse successifs, soit $\mathbb{H}$ (respectivement $\mathbb{H}'$) l'ensemble des pics détectés à la trame t_m (respectivement t_{m+1}). Soit $h \in \mathbb{H}$ (respectivement $h' \in \mathbb{H}'$) un pic particulier caractérisé par ses paramètres $\omega_{h,m}, \Delta\omega_{h,m}, a_{h,m}, \Delta a_{h,m}$ (respectivement $\omega_{h',m+1}, \Delta\omega_{h',m+1}, a_{h',m+1}, \Delta a_{h',m+1}$). Dans la suite, nous omettrons les indices de référence au temps (m et $m+1$) sachant que h réfère implicitement à l'instant t_m et h' à l'instant t_{m+1}.

Pour chaque état $[h, h']$, nous calculons la courbure d'un polynôme d'ordre 3 ($P_\omega(t) = at^3 + bt^2 + ct + d$) reliant les pics h et h' en respectant les conditions aux limites de fréquence suivantes (voir FIG. 4.26) :

$$\begin{cases} P_\omega(t_m) = \omega_{h,m} \\ P'_\omega(t_m) = \Delta\omega_{h,m} \\ P_\omega(t_{m+1}) = \omega_{h',m+1} \\ P'_\omega(t_{m+1}) = \Delta\omega_{h',m+1} \end{cases} \tag{4.67}$$

Le même calcul est effectué pour un polynôme $P_a(t)$ en amplitude, de conditions aux limites

$a_{h,m}$, $\Delta a_{h,m}$, $a_{h',m+1}$ et $\Delta a_{h',m+1}$.

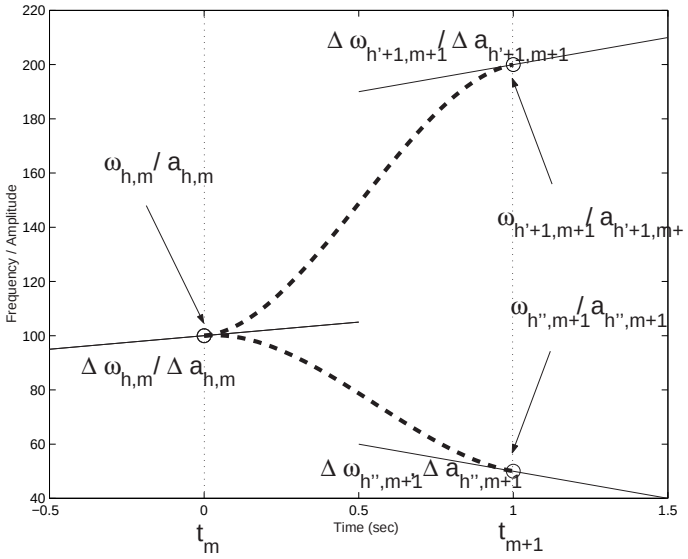


FIG. 4.26 – Calcul de la probabilité d'observation pour deux états $[h_m, h'_{m+1}]$ et $[h_m, h_{m+1}"]$ par courbure polynomiale, polynôme d'ordre 3 de conditions aux limites $\omega_{h,m}$, $\Delta\omega_{h,m}$, $\omega_{h',m+1}$

Une courbure faible garantit une proximité en fréquence (resp. amplitude), une proximité en pente de fréquence (resp. amplitude) et un trajet le plus lisse possible.

Nous nous intéressons au maximum de la valeur absolue de la courbure du polynôme sur l'intervalle (t_m, t_{m+1}) atteinte aux extrémités de l'intervalle. Les valeurs de la courbure de P_ω en t_m et t_{m+1} sont :

$$\begin{aligned} P_\omega"(t_m) &= 2\frac{(t_m - t_{m+1})(2\Delta\omega_h + \Delta\omega_{h'}) + 3(\omega_{h'} - \omega_h)}{(t_m - t_{m+1})^2} \\ P_\omega"(t_{m+1}) &= -2\frac{(t_m - t_{m+1})(\Delta\omega_h + 2\Delta\omega_{h'}) + 3(\omega_{h'} - \omega_h)}{(t_m - t_{m+1})^2} \end{aligned} \tag{4.68}$$

Ce maximum est noté $c_\omega(h, h') : c_\omega(h, h') = \max_t |P_\omega"(t)|$ (resp. $c_a(h, h')$ pour le polynôme en amplitude).

Afin d'obtenir une répartition équivalente des scores sur l'axe des fréquences et des amplitudes, nous normalisons $c_\omega(h, h')$ et $c_a(h, h')$ par leur position respective :

$$c_\omega(h, h') = \frac{c_\omega(h, h')}{0.5(w_h + w_{h'})} \qquad c_a(h, h') = \frac{c_a(h, h')}{0.5(a_h + a_{h'})} \tag{4.69}$$

La probabilité associée à un couple (h, h') est alors définie par

$$Pobs([h_m, h'_{m+1}]) = \exp\left(-\frac{c_\omega^2(h, h')}{\sigma_\omega^2} - \frac{c_a^2(h, h')}{\sigma_a^2}\right) \tag{4.70}$$

dans lequel σ_ω^2 et σ_a^2 sont des paramètres du modèle et déterminent la régularité souhaitée en fréquence et en amplitude.

◊ *Probabilité de transition*

Dans [PR98], nous proposons une mesure de sinusoïdalité (une mesure de «bonne spécification» du modèle) prenant en compte simultanément la comparaison d'estimateur et la continuité temporelle des estimations. Cette mesure repose sur la comparaison d'un modèle d'évolution de fréquence obtenu à partir des observations des fréquences, et d'un modèle d'évolution de fréquence obtenu à partir de l'observation des phases. Cette mesure est intéressante dans notre cas, puisqu'elle utilise une observation - la phase - non utilisée pour le calcul des probabilités d'observation. [17]

De manière à permettre la prise en compte de composantes sinusoïdales de fréquences variables, un modèle linéaire de fréquence, i.e. un modèle quadratique d'évolution de la phase, est utilisé pour calculer la mesure de sinusoïdalité. De ce fait trois observations de phase sont nécessaires, ce qui correspond aux trois phases correspondant à un couple d'états.

Les différentes étapes de cette méthode sont explicitées :

1. Soit $[h_{m-1}, h_m]$ et $[h_m, h_{m+1}]$ les deux états considérés correspondant à des observations aux temps t_{m-1}, t_m et t_{m+1}. Chaque état h_m est caractérisé par une fréquence $\omega_{h,m}$ et une phase $\phi_{h,m}$. Notons D la distance temporelle (supposée constante) entre deux observations successives : $t_{m+1} - t_m = t_m - t_{m-1} = D$
2. Le polynôme de phase d'ordre deux en t, $\phi(t)$, passant par les estimations $\phi_{h,m-1}, \phi_{h,m}, \phi_{h,m+1}$ est déterminé. Les phases ont été préalablement déroulées de manière à garantir la distance fréquentielle minimale : $(\phi_{h,m}^{uw} - \phi_{h,m-1})/D \simeq (\omega_{h,m-1} + \omega_{h,m})/2$ et $(\phi_{h,m+1}^{uw} - \phi_{h,m}^{uw})/D \simeq (\omega_{h,m} + \omega_{h,m+1})/2$ (uw désigne «unwrapped» ou déroulé).
3. La dérivée première du polynôme de phase est utilisée pour calculer une estimation de la fréquence instantanée obtenue à partir de la dérivée de la phase $\omega_\phi(t) = \frac{\partial\phi(t)}{\partial t}$ aux temps t_{m-1}, t_m, t_{m+1} :

$$\begin{cases} \omega_\phi(t_{m-1}) = \dfrac{1}{2D}(4\phi_{h,m}^{uw} - 3\phi_{h,m-1} - \phi_{h,m+1}^{uw}) \\ \omega_\phi(t_m) = \dfrac{1}{2D}(\phi_{h,m+1}^{uw} - \phi_{h,m-1}) \\ \omega_\phi(t_{m+1}) = \dfrac{1}{2D}(\phi_{h,m-1} - 4\phi_{h,m}^{uw} + 3\phi_{h,m+1}^{uw}) \end{cases} \tag{4.71}$$

[18]

4. La distance euclidienne entre l'estimation des fréquences ω et la fréquence instantanée ω_ϕ, obtenue à partir de la dérivée de la phase, est calculée (voir Fig. 4.27)

$$de = \sqrt{\sum_{i=-1}^{1} (\omega_{h,m+i} - \omega_\phi(t_{m+i}))^2} \tag{4.73}$$

La distance euclidienne de donne une mesure de sinusoïdalité (une mesure de la «bonne spécification» du modèle), étant données les estimations de fréquence et de phase et selon un modèle de variation linéaire de la fréquence.

La probabilité de transition entre les états $[h_{m-1}, h_m]$ et $[h_m, h_{m+1}]$ est alors définie par

$$Ptrans([h_m, h_{m+1}]|[h_{m-1}, h_m]) = \exp\left(-\frac{de^2([h_{m-1}, h_m, h_{m+1}])}{\sigma_{de}^2}\right) \tag{4.74}$$

dans lequel σ_{de}^2 est un paramètre du modèle.

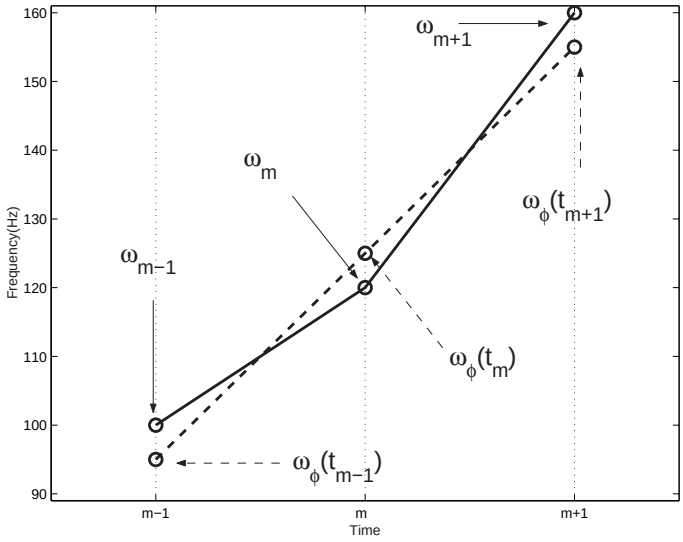


FIG. 4.27 – Calcul de la probabilité de transition $[h_{m-1}, h_m]$ / $[h_m, h_{m+1}]$ par distance euclidienne entre modèle d'évolution temporelle de fréquence $\omega_{m-1}, \omega_m, \omega_{m+1}$ et modèle d'évolution de fréquence obtenu par dérivée de la phase $\omega_\phi(t_{m-1}), \omega_\phi(t_m), \omega_\phi(t_{m+1})$

◊ ***Méthodes complètes***

A titre indicatif, nous donnons les étapes d'estimation nécessaires qui précèdent la méthode de création de trajets sinusoïdaux.

Méthode globale d'analyse sinusoïdale d'un son

- Détection des maxima locaux du spectre en chaque instant t_m
- Autour de chaque maximum, estimation des paramètres du modèle sinusoïdal d'ordre un (estimateur de mesure de distorsion du spectre complexe)
- Algorithme de création de trajets

Algorithme de création de trajets

Pour chaque couple d'instants successifs t_m, t_{m+1}, nous créeons une matrice de probabilités d'observation des états $[h_m, h'_{m+1}]$, $\forall h \in \mathbb{H}_m$ et $\forall h' \in \mathbb{H}'_{m+1}$

Initialisation d'un trajet l : Choix d'une composante h_m d'initialisation du trajet l. Les composantes d'initialisation sont choisies parmi l'ensemble des composantes n'ayant

pas encore servi pour l'initialisation. Une composante déjà utilisée dans un trajet peut à l'inverse servir d'initialisation à un nouveau trajet.

Recherche des états $[h_{m-1}, h'_m]$ et $[h'_m, h_{m+1}"]$ tels que le produit
$Ptotal = Pobs([h_{m-1}, h'_m]) \cdot Ptrans([h'_m, h_{m+1}"]|[h_{m-1}, h'_m]) \cdot Pobs([h'_m, h_{m+1}"])$
soit maximal (voir FIG. 4.28). Nous associons la probabilité $Ptotal$ au trajet l à l'instant t_m.

En pratique, pour accélérer l'algorithme, seuls sont considérés les trois états de part et d'autre de h_m de probabilité d'observation maximale.

Poursuite du trajet l à partir de l'état $[h_m, h_{m+1}]$: Nous recherchons $h_{m+2}"'$ tel que $Ptotal = Pobs([h'_m, h_{m+1}"]) \cdot Ptrans([h_{m+1}", h_{m+2}"']|[h'_m, h_{m+1}"]) \cdot Pobs([h_{m+1}", h_{m+2}"'])$
soit maximal (voir FIG. 4.29). Nous associons la probabilité $Ptotal$ au trajet l à l'instant t_{m+1}.

Terminaison du trajet l : Si la probabilité totale d'une transition entre états est inférieure à un seuil, le trajet est arrêté.

Résolution des conflits : Un conflit a lieu lorsqu'une même composante h_m appartient à deux trajets différents. Dans le cas où toutes les composantes seraient considérées comme sinusoïdales, les conflits ne devraient pas être résolus, puisque les conflits vont à l'encontre de signaux du type croisement de trajets. Néanmoins, nous ne pouvons garantir à 100% que les composantes inclues dans un trajet sont sinusoïdales. Dans le cas où ces composantes ne seraient pas sinusoïdales, elles peuvent donner lieu à un trajet se recoupant avec un trajet réellement sinusoïdal. Il importe donc de régler ces conflits.

La résolution de conflit pour une composante h_m appartenant à deux trajets l et l' est effectuée sur base de la probabilité locale de leur trajet l et l' en t_m.

Paramètres du modèle : Les paramètres du modèle sont σ_ω, σ_a et σ_{ed}, ainsi que la probabilité minimale d'existence d'un trajet.

Remarque importante vis-à-vis du choix d'un produit de probabilités locales plutôt que cumulées : Dans une résolution du type algorithme de Viterbi, à chaque instant t_m et pour chaque état I_m nous recherchons l'état I_{m-1} à l'instant précédent t_{m-1} ayant la plus grande probabilité $Pcumul(I_{m-1}) \cdot Ptrans(I_m|I_{m-1})$ dans lequel $Pcumul$ désigne la probabilité «cumulée» de I_{m-1}. La probabilité cumulée associée à I_m est alors prise égale à $Pcumul(I_m) = Pcumul(I'_{m-1}) \cdot Ptrans(I_m|I'_{m-1}) \cdot Pobs(I_m)$. La terminaison de l'algorithme consiste à rechercher le I_m de probabilité cumulée maximale et à remonter dans le sens inverse du temps l'ensemble des transitions de manière à reconstituer le trajet.

Par comparaison, dans notre algorithme la probabilité cumulée est remplacée par la probabilité locale d'observation : $Pcumul(I_m) = Pobs(I_m)$.

Ceci est justifié par le fait que la probabilité cumulée est mal définie dans le cas d'un signal où les différentes composantes fréquentielles n'apparaissent pas simultanément. De ce fait, les trajets sont de longueurs différentes et la probabilité cumulée d'un trajet dépend de sa longueur. Les probabilités cumulées de deux trajets de longueurs différentes ne sont donc pas comparables. [19] Dans l'algorithme de Viterbi, toutes les chaînes d'états sont supposées de longueur égale. Dans l'utilisation de l'algorithme de Viterbi faite par [Gar92], des trajets de «pics non-existents» ainsi que des probabilités de transitions entre pics non-existents sont définis de manière à ce que toutes les chaînes d'états soient de longueur égale. Même si cette solution devra être explorée, la solution retenue pour l'instant pour notre méthode est

de ne pas cumuler les probabilités. De ce point de vue, l'optimisation n'est pas globale en temps, comme elle n'est pas globale en fréquence du fait de la non-considération simultanée de l'ensemble des trajets.

Du fait du remplacement de la probabilité cumulée par la probabilité locale d'observation ($Pcumul(I_m) = Pobs(I_m)$), la sélection finale de l'algorithme de Viterbi (sélection sur le critère de probabilité cumulée maximale en fin de trajet) est, dans notre algorithme, effectuée à chaque instant.

◊ *Application de la méthode proposée*

Aux Fig. 4.30 etFig. 4.31 sont illustrés les résultats obtenus avec la méthode de suivi de trajets sinusoïdaux proposée. Sur les deux panneaux du haut de la Fig. 4.30, nous avons de plus représenté les fréquences des pics estimés (panneau de gauche) ainsi que les fréquences et pente de fréquence des pics estimés (panneau de droite). Le bénéfice de l'estimation des pentes de fréquence apparaît ici clairement par l'apparition visuelle des trajets sinusoïdaux.

De manière générale, les résultats obtenus à l'aide de l'algorithme sont bons, même si le paramètrage de l'algorithme (choix de σ_ω, σ_a, σ_{de}, seuil de réjection d'un trajet en fonction de sa probabilité, durée minimale d'un trajet, nombre de pics de probabilité faible tolérés dans un trajet) s'avère crucial et difficile.

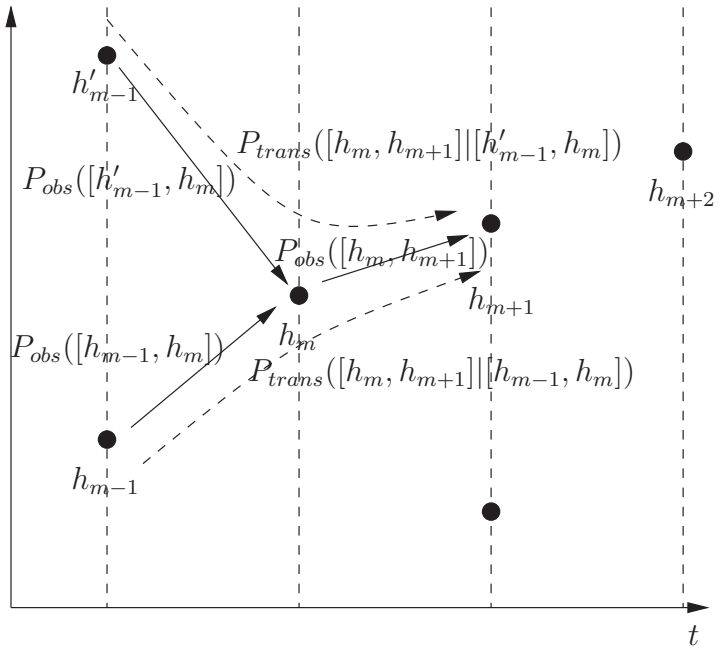


FIG. 4.28 – Algorithme de création de trajet : initialisation : recherche du couple d'états $[h_{m-1}, h_m]$ / $[h_m, h_{m+1}]$ tel que la probabilité totale : $Pobs([h_{m-1}, h_m]) \cdot Ptrans([h_m, h_{m+1}]|[h_{m-1}, h_m]) \cdot Pobs([h_m, h_{m+1}])$ soit maximale

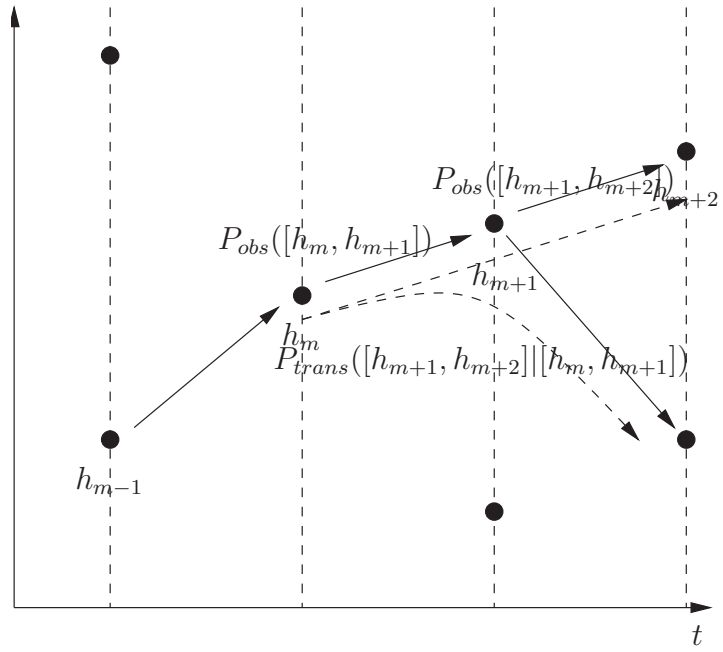


FIG. 4.29 – Algorithme de création de trajet : propagation : recherche de l'état $[h_{m+1}, h_{m+2}]$ tel que la probabilité totale : $Pobs([h_m, h_{m+1}]) \cdot Ptrans([h_{m+1}, h_{m+2}]|[h_m, h_{m+1}]) \cdot Pobs([h_{m+1}, h_{m+2}])$ soit maximale

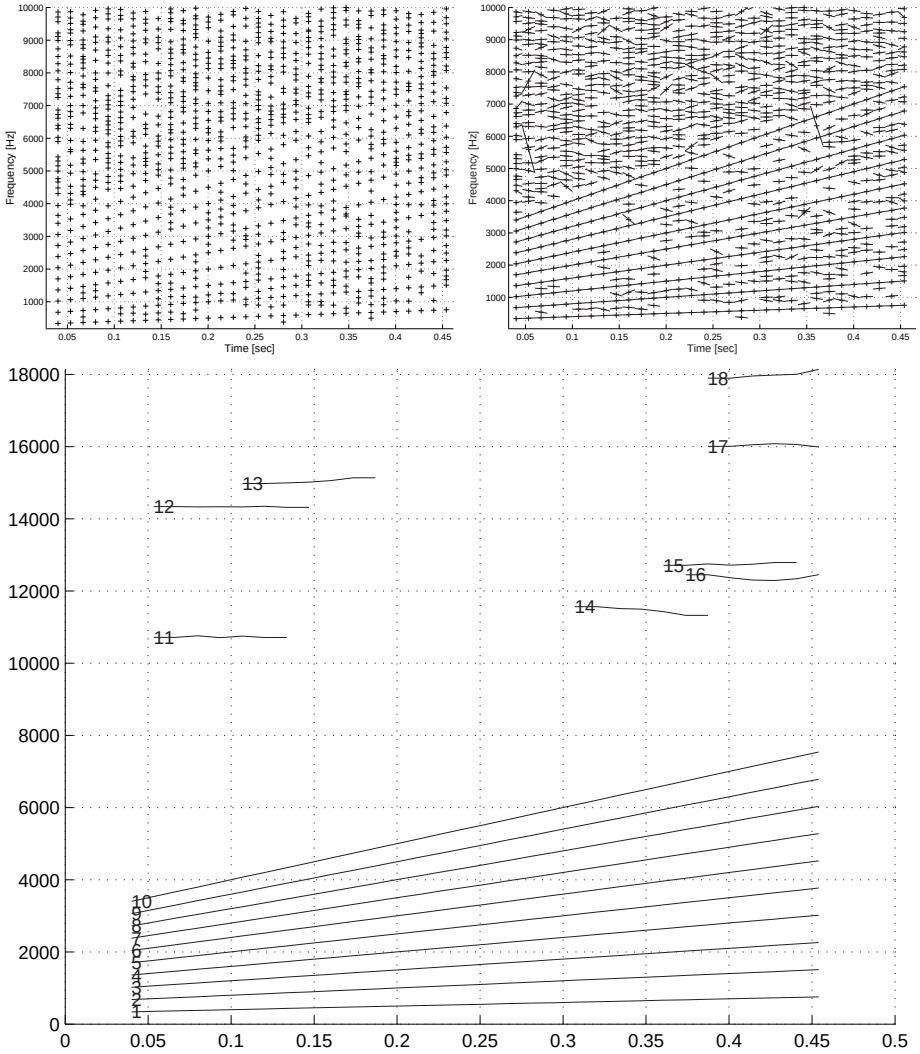


FIG. 4.30 – [HG] Représentation des fréquences estimées [HD] Représentation des fréquences et pentes de fréquences estimées [B] Algorithme proposé pour le suivi de trajets sinusoïdaux proposé. Signal= chirps de fréquence + bruit blanc additif

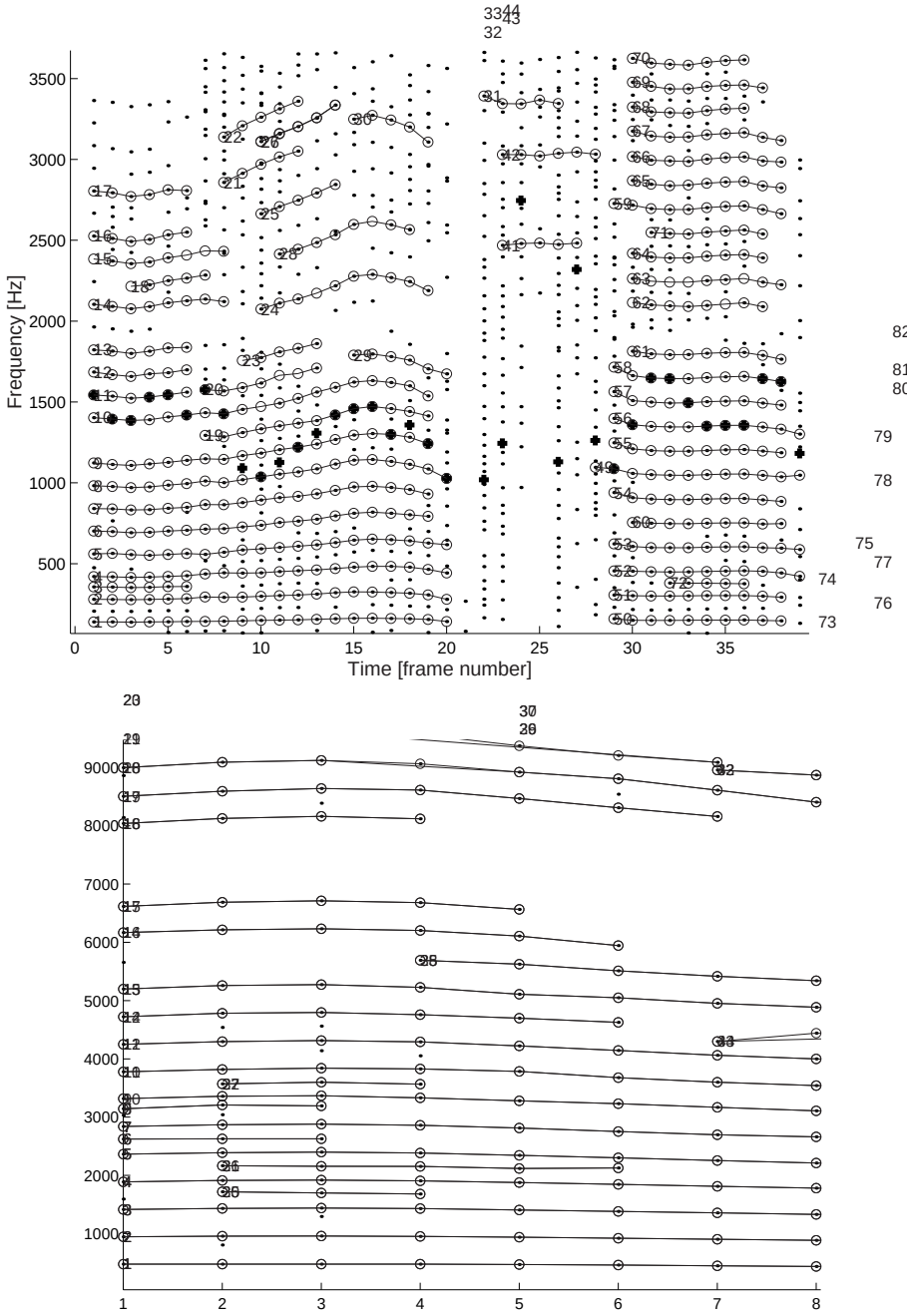


FIG. 4.31 – Algorithme proposé pour le suivi de trajets sinusoïdaux proposé : application sur un signal de voix parlée et de voix chantée [H] Signal= speech, [B] Signal= tibetextract

4.3.4 Conclusion

Dans cette partie, nous nous sommes intéressés à la localisation des régions du plan temps/fréquence représentables par des sinusoïdes. Ce problème a été abordé sous deux approches différentes.

La première approche repose sur le calcul de l'erreur commise en modélisant une région du plan temps/fréquence par un modèle sinusoïdal. Dans le cas où ce modèle serait de paramètres localement constants (hypothèse de stationnarité locale), nous avons montré l'équivalence entre cette erreur et la corrélation complexe entre le spectre local du signal et la réponse fréquentielle de la fenêtre d'observation utilisée. Nous avons montré que la valeur de l'erreur de modélisation dépend non seulement du contenu du signal (présence ou absence d'une sinusoïde dans la région considérée, rapport signal à bruit, variations du signal), mais également de la définition d'une largeur temporelle et fréquentielle d'observation. L'amélioration de la discrimination sinusoïde/bruit peut s'obtenir par élargissement de la région d'observation. Cependant ceci se fait au détriment de l'hypothèse d'orthogonalité (séparation des composantes a priori inconnue) des composantes du signal et de l'hypothèse de stationnarité locale. La taille de la région d'observation est donc bornée supérieurement par prudence. Pour un rapport signal à bruit donné, nous ne pouvons dés lors garantir une discrimination suffisante entre composantes sinusoïdales et non-sinusoïdales. De ce fait, l'erreur de modélisation doit être utilisée en complément d'une autre méthode permettant d'effectuer le reste de la discrimination. Une détection de bruit (i.e. du bruit considéré de manière erronée comme une sinusoïde) étant moins contraignante pour la suite de l'algorithme qu'une non-détection de sinusoïde (i.e une sinusoïde considérée de manière erronée comme du bruit), une valeur de seuillage élevée de l'erreur de modélisation est utilisée. Les composantes de valeurs inférieures à ce seuil sont transmises à un algorithme en aval.

La deuxième approche repose sur le calcul d'une erreur de spécification du modèle. Cette erreur mesure à quel point la spécification d'un modèle sinusoïdal correspondant à un ensemble d'observations (estimations à un instant donné) peut être étendu à un autre ensemble d'observations (estimations à l'instant suivant). L'erreur de spécification permet de rendre compte de manière plus exacte de la définition du modèle sinusoïdal, «paramètre à variations temporelles lentes», non prise en compte par l'hypothèse de «stationnarité locale» de l'erreur de modélisation. L'erreur de spécification d'un modèle est présentée sous l'angle du problème de création de trajets de sinusoïdes à partir d'observations et devant répondre à certains critères de régularité temporelle.

Une nouvelle méthode de création de trajets sinusoïdaux utilisant une double contrainte de régularité est proposée. Cette méthode repose sur la définition d'états comme l'association de deux composantes estimées en deux instants distincts. Une probabilité d'observation d'un état est définie sur base de la courbure maximale du polynôme joignant les composantes de l'état en respectant les conditions de limites dérivées. Une probabilité de transition entre deux états est calculée par comparaison de l'évolution temporelle de fréquences observées et des fréquences instantanées calculées à partir des observations de phase. La probabilité totale d'un trio de composantes en trois instants successifs est définie sur le produit de la probabilité d'observation de chacun de ces états et de la probabilité de transition entre ces états. Ceci correspond à une définition de sinusoïdalité sur base d'un trio de composantes.

Notes de bas de page relatives à la partie 4

1. Lors de la resynthèse du signal, la variation temporelle des paramètres du modèle entre les instants d'estimation est obtenue par interpolation.

2. **Explication :** Soit $x(n)$ un signal analysé autour du temps m à l'aide d'une fenêtre de pondération $h(n)$ de longueur L. Le signal fenêtré $s(n)$ s'exprime par $s(n) = x(n)h(n-m)$ et n'est défini que sur une longueur L. Notons ce signal $s_L(n)$. Le prolongement par zéro consiste à ajouter au signal une séquence d'échantillons de valeur nulle. Soit $z_{N-L}(n)$ une séquence nulle de longueur $N-L$. Appelons $sz_N(n)$ la séquence résultant de la concaténation de $s_L(n)$ et de $z_{N-L}(n)$.

Interprétation fréquentielle possible : la TFDCT de $sz_N(n)$, étant le fenêtrage de la séquence $s_N(n)$ par une fenêtre rectangulaire de longueur L, est équivalente à la convolution de la TFDCT de $s_N(n)$ par la TFDCT d'une fenêtre rectangulaire de longueur L (TFDCT = sinus cardinal = filtre passe-bas idéal, premier zéro en $f = \frac{Fe}{L}$) (voir annexe F). Ce filtrage, étant à bande limitée, produit une interpolation parfaite des points du spectre.

3. Le **ré-assignement fréquentiel** est une méthode permettant d'améliorer la lisibilité des représentations temps/ fréquence. Dans une analyse de type Fourier à court terme «classique», chaque valeur de la TFDCT est assignée à la position fréquentielle correspondant aux fréquences discrètes de la TFDCT ($\omega_k = 2\pi k/N$). Le ré-assignement fréquentiel consiste à assigner la valeur d'énergie initialement assignée à ω_k aux centres de gravité locaux du spectre ω_r :

$$\omega_r(x;t,\omega_k) = \Re\left\{\frac{\int_\xi \xi X(\xi)H^*(\xi-\omega_k)e^{j\xi t}d\xi}{\int_\xi X(\xi)H^*(\xi-\omega_k)e^{j\xi t}d\xi}\right\} \tag{4.7}$$

(**??**) est égale à la fréquence instantanée (voir annexe G)

$$\omega_r(x;t,\omega_k) = \frac{\partial}{\partial t}\phi(x,t,\omega_k) \tag{4.8}$$

4. Nous ne considérons que les fenêtres dérivables

5. La notation $\mathrm{dB}_{20} = 20 \cdot \log_{10}$ exprime un rapport de puissance en échelle logarithmique. Le non recouvrement à $-6\mathrm{dB}_{20}$ signifie donc le non-recouvrement des composantes voisines à plus de la moitié de leur puissance

6. À $-12\mathrm{dB}_{20}$, même dans le pire des cas (celui pour lequel la contribution des phases est additive à la fréquence intermédiaire), pour deux composantes d'amplitude unitaire, l'amplitude à la fréquence intermédiaire est égale à -1/4 + -1/4 = -1/2

7. [MA94] propose la décomposition du signal sur une base de fonction exponentielle complexe (pour la partie déterministe), et une base de fonction à bande étroite (NBBFs) pour la partie bruit. Le modèle proposé s'exprime

$$\hat{s}(t) = \sum_h a_h \cos(h\omega_0 t + \phi_{0,k}) + \sum_h b_h \epsilon_h(t) \cos(h\omega_s t + \varphi_h) \tag{4.10}$$

dans lequel $h\omega_s$ désigne la fréquence centrale de la $h^{\text{ème}}$ fonction à bande étroite, φ_h sa phase aléatoire, $\epsilon_h(t) = u_h(t) * h_L(t)$ une enveloppe passe-bas aléatoire, $u_h(t)$ un bruit blanc, et $h_L(t)$ la réponse impulsionnelle du filtre à bande étroite d'une fonction à bande étroite.

L'estimation des paramètres du modèle s'effectue par minimisation successive de deux critères d'erreur :

1. minimisation de

$$\epsilon_1 = \int |S(\omega) - \hat{S}_p(\omega)|^2 d\omega \tag{4.11}$$

La minimisation conduit à l'estimation des phases et amplitudes des sinusoïdes. Cependant, seule l'estimation des phases des sinusoïdes sont gardées. Leur amplitude est estimée simultanément avec l'amplitude de la partie aléatoire.

2. minimisation de

$$\epsilon_2 = \int (|S(\omega)| - |S_p(\omega)| - E\{|S_r(\omega)|\})^2 d\omega \tag{4.12}$$

8. **Cas particulier : résolution locale en fréquence par hypothèse d'orthogonalité des fonctions de base**

(4.19) peut s'écrire sous forme d'un système d'équations de Yule-Walker [Dut93] :

$$(\underline{\underline{B}}^{\mathcal{H}}\underline{\underline{B}})\hat{\theta} = \underline{\underline{B}}^{\mathcal{H}}\underline{S}. \tag{4.26}$$

Dans ce cas le terme (i, j) de la matrice $\underline{\underline{B}}^{\mathcal{H}}\underline{\underline{B}}$ s'écrit $\int_\omega H(\omega - \omega_{hi})H^*(\omega - \omega_{hj})d\omega$. C'est le produit scalaire des fonctions de base de la décomposition : la TF de la fenêtre d'analyse $H(\omega)$ centrée sur les fréquences ω_{hi} des différentes composantes cherchées.

Le terme i de $\underline{\underline{B}}^{\mathcal{H}}\underline{S}$ s'écrit $\int_\omega S(\omega)H(\omega - \omega_{hi})d\omega$. C'est le produit scalaire du signal sur les bases de la décomposition, c'est-à-dire la TF de la fenêtre d'analyse $H(\omega)$ centrée sur les fréquences ω_{hi} des différentes composantes cherchées.

Pour une résolution fréquentielle suffisante, les fonctions de base $H(\omega - \omega_{hi})$ sont orthogonales. Dans ce cas, la matrice $\underline{\underline{B}}^{\mathcal{H}}\underline{\underline{B}}$ est une matrice diagonale dont les éléments valent $\int H^2(\omega - \omega_{hi})d\omega$. (**??**) se réécrit alors

$$\int H^2(\omega - \omega_{hi})d\omega \cdot \theta_{hi} = \int S(\omega)H(\omega - \omega_{hi})d\omega \tag{4.27}$$

ce qui donne

$$\hat{\theta}_{hi} = \frac{\int S(\omega)H(\omega - \omega_{hi})d\omega}{\int H^2(\omega - \omega_{hi})d\omega} \tag{4.28}$$

De même (4.25) se réécrit $(\underline{\underline{C}}^{\mathcal{H}}\underline{\underline{C}})\hat{\underline{\Theta}} = \underline{\underline{C}}^{\mathcal{H}}(\underline{S}\hat{\underline{S}})$. En supposant les fonctions de base $H'(\omega - \omega_{hi})$ indépendantes, (4.25) se réécrit

$$\hat{\Theta}_{hi} = \frac{A_{hi}}{2} e^{j\phi_{0,hi}} \frac{\int (S(\omega) - \hat{S}(\omega))H'(\omega - \omega_{hi})d\omega}{\int H'^2(\omega - \omega_{hi})d\omega} \tag{4.29}$$

9. Variation relative = variation du paramètre ramené à sa valeur moyenne sur l'ensemble des δ

10. Abusive puisque les signaux musicaux ne sont généralement pas engendrés par la superposition de sinusoïdes et que nous ne cherchons donc pas des sinusoïdes mais des régions représentables par des sinusoïdes

11. Ceci se comprend en constatant que, pour une observation temporelle de durée fixée et un spectre de N points, N/2 erreurs de modélisation peuvent être définies en chaque point du spectre discret. L'*erreur de modélisation 1* est l'erreur commise en modélisant une largeur fréquentielle égale à 1 bin par une composante. Puisque chacun des points d'un spectre discret peut être expliqué par une sinusoïde de fréquence, amplitude et phase déterminées, cette erreur est toujours nulle. L'*erreur de modélisation N/2* est l'erreur commise en modélisant tout le spectre (largeur N/2) par une seule composante. L'erreur de modélisation locale en fréquence dépend donc de la définition de la largeur fréquentielle considérée. De la même manière, l'erreur de modélisation globale en fréquence (considérant simultanément l'ensemble des composantes et l'ensemble des fréquences) dépend du nombre H de composantes considérées dans le modèle.
12. un estimateur de sinusoïdalité et non pas un estimateur des paramètres d'un modèle sinusoïdal
13. Cette infinité de fonctions de l'ensemble correspond aux paramètres non fournis par la projection, tels que les paramètres de fréquence centrale et de modulation de fréquence et d'amplitude.
14. La similitude entre (4.58) et la transformée de Fourier d'un signal pondérée par une fonction fenêtre au carré : $(x(n)\cdot h(n))\cdot h(n) = s(n)\cdot h(n) \rightleftharpoons S(\omega)\otimes H(\omega)$ n'est qu'apparente, puisque dans (4.58) la convolution n'est calculée que localement sur l'intervalle W_h
15. De la même manière, nous pourrions introduire une probabilité d'observation dans la méthode de [MQ86b] et [SS90]. Pour un état h_m, la probabilité d'observer cet état peut être choisie égale à la corrélation complexe $|c(\omega_{h,m})|$.
16. Dans le cas de l'utilisation d'un modèle sinusoïdal d'ordre 0 [Gar92], la continuité de dérivée de fréquence/amplitude nécessite la considération de deux états. C'est la raison pour laquelle, dans [Gar92], la continuité de pente est utilisée comme probabilité de transition entre états.
17. Si pour h_{m+1} fixé, $[h_m, h'_{m+1}]$ et $[h'_{m+1}, h_{m+2}"]$ sont les états de probabilité d'observation maximale, alors toutes probabilités de transition basée sur la continuité de fréquence/amplitude donnera également $Ptrans([h_{m+1}, h_{m+2}]|[h_m, h_{m+1}])$ comme les états de probabilité de transition maximale. Un critère sur la continuité des fréquences et amplitudes est donc redondant puisque déjà utilisé par les probabilités d'observation.
18. Expression dans le cas d'un pas d'avancement non constant : Soit $D = t_m - t_{m-1}$ et $E = t_{m+1} - t_m$

$$\begin{cases} \omega_\phi(t_{m-1}) = \dfrac{(\phi^{uw}_{h,m} - \phi^{uw}_{h,m+1})D^2 + (\phi^{uw}_{h,m} - \phi_{h,m-1})2DE + (\phi^{uw}_{h,m} - \phi_{h,m-1})E^2}{DE(D+E)} \\ \omega_\phi(t_m) = -\dfrac{(\phi^{uw}_{h,m} - \phi^{uw}_{h,m+1})D^2 + (\phi_{h,m-1} - \phi^{uw}_{h,m})E^2}{DE(D+E)} \\ \omega_\phi(t_{m+1}) = -\dfrac{(\phi^{uw}_{h,m} - \phi^{uw}_{h,m+1})D^2 + (\phi^{uw}_{h,m} - \phi^{uw}_{h,m+1})2DE + (\phi^{uw}_{h,m} - \phi_{h,m-1})E^2}{DE(D+E)} \end{cases} \tag{4.72}$$

19. Imaginons la naissance d'une composante à l'instant t=10 ; cette composante atteint un état I à l'instant t=11. Sa probabilité cumulée est très grande puisqu'elle n'a pas eu le temps de décroître sur le trajet t=10 / t=11. Supposons qu'un autre trajet commençant en t=1 atteigne également I à l'instant 11. Ce second trajet aura une probabilité cumulée plus faible que celle du premier trajet du fait de sa longueur différente.

Chapitre 5

Caractérisation des signaux périodiques/harmoniques

5.1 Introduction

La caractérisation des signaux en période fondamentale et harmonicité est certainement l'un des sujets de traitement de signal audio les plus étudiés à l'heure actuelle. Ceci provient du fait que le signal vocal (parole et voix chantée), durant la période d'activité des cordes vocales, peut être considéré comme un signal harmonique. De plus, de nombreux signaux (non-mixtes et monophoniques) d'instruments de musique répondent à un modèle harmonique. L'hypothèse d'harmonicité permet de simplifier grandement les modèles de signaux, tant lors de l'analyse que de la synthèse. De ce fait, de nombreux modèles basés sur l'hypothèse d'harmonicité sont proposés dans la littérature.

Aux deux types de modélisations étudiés dans le cadre de cette recherche - décomposition en formes d'onde élémentaires temporelles et décomposition fréquentielle (famille de sinusoïdes) - correspondent les notions de période fondamentale (répétition d'une même forme d'onde quasi-similaire au cours du temps) et de fréquence fondamentale (structuration des composantes spectrales telle qu'une fréquence explique la majorité des composantes du spectre).

Durant cette recherche, nous n'avons pas développé de nouvelle méthode de caractérisation des signaux en période fondamentale et harmonicité. Aussi, nous contentons-nous de présenter succinctement les bases des deux méthodes d'estimation de la fréquence fondamentale que nous avons utilisées. Nous expliquons ensuite les hypothèses qui peuvent être faites afin de simplifier l'estimation d'un modèle sinusoïdal appelé dans ce cas modèle sinusoïdal harmonique. Nous détaillons ensuite l'estimation des coefficients dits de «voisement» et d'inharmonicité.

5.2 Estimation de la période/fréquence fondamentale

Nous avons utilisé deux méthodes d'estimation de la période fondamentale d'un signal :
- Une méthode temporelle : la méthode dite de l'auto-corrélation du signal local. Cette

méthode est utilisée pour les signaux de paroles de locuteurs masculins (fréquence fondamentale basse).
- Une méthode fréquentielle : la méthode dite du maximum de vraisemblance. Cette méthode est utilisée pour les signaux de paroles de locutrices féminines, de voix chantée et pour les sons d'instruments de musique.

5.2.1 Méthode de l'auto-corrélation

Cette méthode repose sur le calcul de la fonction d'auto-corrélation du signal local. La **fonction d'auto-corrélation** s'exprime

$$\hat{r}(k) = \frac{1}{N-k} \sum_{n=0}^{N-k-1} x_n x_{n+k} \tag{5.1}$$

Cette expression, bien que non normalisée par rapport à l'énergie du signal local ($\hat{r}(k)$ n'est pas nécessairement égale à 1 pour un signal d'auto-corrélation maximale) permet une implémentation à très faible coût de calcul par FFT/IFFT. [1]

La **fonction d'auto-corrélation normalisée** en énergie s'exprime par

$$\bar{r}(k) = \frac{\sum_{n=0}^{N-k-1} x_n x_{n+k}}{\sqrt{\sum_{n=0}^{N-k-1} x_n^2} \sqrt{\sum_{n=0}^{N-k-1} x_{n+k}^2}} \tag{5.3}$$

La fonction d'auto-corrélation normalisée prend ses valeurs dans l'intervalle $[0, 1]$, indépendamment de l'énergie du signal. De la sorte, elle peut être utilisée pour le calcul d'un coefficient de périodicité. $\bar{r}(k)$ nécessite cependant un coût de calcul plus élevé que $\hat{r}(k)$, du fait que son estimation ne peut se faire par FFT/IFFT.

Dans ces deux formules, N détermine l'horizon temporel considéré. Pour un signal périodique de période T0, la fonction d'auto-corrélation possède des pics aux multiples de T0. Ce sont ces pics que nous recherchons. Pour un pic en $k = k_0$, la fréquence fondamentale estimée est égale à f0 $= \frac{Fe}{k_0}$.

La résolution en fréquence de la méthode d'auto-corrélation dépend de la fréquence d'échantillonnage Fe. La résolution en fréquence varie de manière non-linéaire en fonction de la position du maximum de la fonction d'auto-corrélation. Soit k_0 le maximum de $\hat{r}(k)$; la résolution fréquentielle obtenue sur f0 à partir de $k0$ est égale à Δf0 $= \frac{Fe}{k0-1} - \frac{Fe}{k0} = \frac{Fe}{(k0-1)k0}$. Cette limitation peut cependant être contournée par interpolation des points de la fonction $\hat{r}(k)$ autour de k_0. Dans notre implémentation, nous effectuons une interpolation quadratique autour du maximum de la fonction d'auto-corrélation.

Avantage : Cette méthode repose sur la recherche du décalage temporel engendrant la similarité maximale entre deux formes d'onde . L'utilisation de durées d'observation extrêmement courtes est possible (2.5 périodes fondamentales). Une connaissance approximative préalable de la période fondamentale est donc requise.

Inconvénient : Par son utilisation directe de la forme d'onde du signal, cette méthode est très sensible aux variations spectrales du signal (transitions de formants), ainsi qu'à la présence de bruit additif dans le signal (fricatives voisées, ...). Diverses techniques ont été proposées afin de réduire cette sensibilité (traitement sur le résiduel de prédiction linéaire, techniques dites de «clippage» du signal). La méthode que nous avons utilisée

consiste à ne considérer que la partie basse fréquence du signal (¡ 1000 Hz) [2] . Ceci permet de diminuer simultanément la sensibilité aux variations spectrales et au bruit.

5.2.2 Méthode du «maximum de vraisemblance»

Dans la méthode du «maximum de vraisemblance» [Dov94] , nous cherchons la fréquence fondamentale f0 telle que la probabilité d'observer les composantes fréquentielles estimées étant donné f0 soit maximale. Pour un candidat f0, nous définissons un ensemble de bandes de fréquence W_h centrées sur les multiples du f0 candidat et de largeur égale à f0. Dans chaque bande de fréquence nous calculons la probabilité d'observer les composantes fréquentielles étant donné f0. Pour une bande W_h centrée en hf0, cette probabilité est la somme de deux probabilités :

1. la probabilité d'observer les composantes fréquentielles dans le cas où l'harmonique hf0 est présente dans W_h,
2. la probabilité d'observer les composantes fréquentielles dans le cas où l'harmonique hf0 est absente dans W_h.

Les distributions de chaque bande de fréquence sont supposées indépendantes et la probabilité totale d'observer le spectre étant donné un f0 candidat est donc donnée par le produit des probabilités des bandes de fréquence. La maximisation est obtenue de manière itérative en parcourant l'ensemble des f0 candidats.

Avantage : Du fait du traitement dans le domaine fréquentiel et de la prise en compte d'harmoniques manquantes, cette méthode est robuste aux transitions de formants ainsi qu'à l'absence d'énergie du signal à la fréquence de f0.

Inconvénient : La méthode repose sur une détection préalable des composantes fréquentielles et nécessite de ce fait une taille d'observation temporelle supérieure à celle utilisée par la méthode d'auto-corrélation. La conséquence en est un certain lissage temporel des estimations.

5.3 Modèle sinusoïdal harmonique

Au chapitre précédent, nous avons présenté le modèle sinusoïdal sous sa forme générale dans le cas d'un signal non nécessairement harmonique. L'estimation des paramètres ainsi que les méthodes de détection des composantes sinusoïdales ne reposaient pas sur une hypothèse d'harmonicité. Ainsi le critère de détection et de création de trajets sinusoïdaux par erreur de spécification reposait sur un critère de régularité temporelle des paramètres estimés. Dans le cas d'un signal harmonique, l'analyse sinusoïdal peut être grandement simplifiée [MQ86b] [GL88] [Sty96] [GS97] [Oud98].

Dans le cas d'un signal harmonique, la détection des composantes peut être aidée du fait de la connaissance de l'existence d'une relation harmonique entre les fréquences des composantes recherchées. La localisation fréquentielle des composantes recherchées est connue (à un facteur d'imprécision près dépendant du taux d'inharmonicité du signal), facilitant le problème de l'estimation des paramètres du modèle.

La détection, quant à elle, bénéficie de la connaissance du nombre maximal de composantes présentes dans le spectre à un instant donné, ainsi que de la connaissance de la largeur W_h des bandes fréquentielles devant être représentées par une sinusoïde (voir partie 4.3).

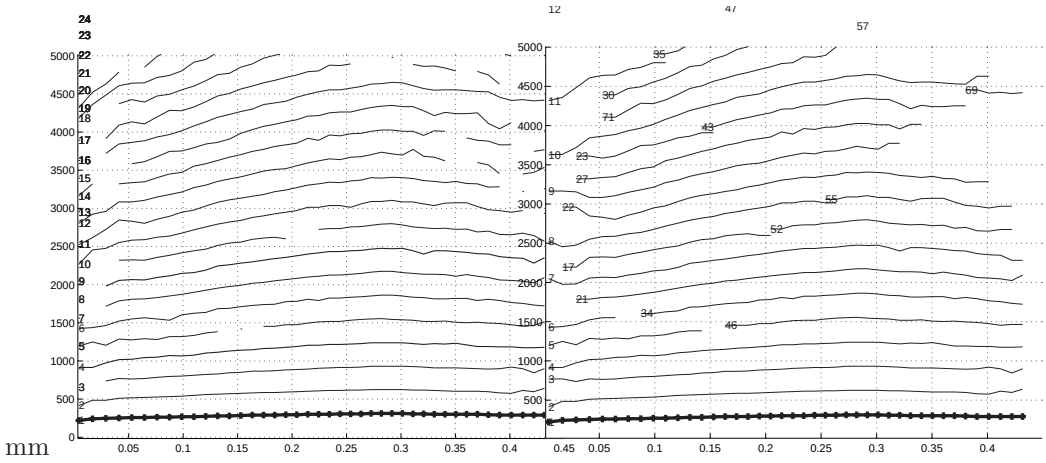


FIG. 5.1 – Estimation des trajets sinusoïdaux : utilisation de l'hypothèse d'harmonicité. Signal = «vie»

FIG. 5.2 – Estimation des trajets sinusoïdaux : utilisation du critère de régularité temporelle. Signal = «vie»

Finalement, dans un cas simple, nous ne cherchons qu'une composante par bande. De plus, les indices i des trajets sinusoïdaux sont rangés par ordre croissant de fréquence ; il n'y a donc jamais de croisement de trajets en modèle sinusoïdal harmonique.

Dans le cas d'un signal harmonique, l'index i est attribué à la $i-1$ harmonique de la période fondamentale. La $i-1$ harmonique est considérée comme la sinusoïde dont la fréquence ω_h est la plus proche de $i\omega_0$. Pour cela, nous recherchons la sinusoïdale présente dans la région fréquentielle $W_h = [i\omega_0 - \alpha\omega_0, i\omega_0 + \alpha\omega_0]$. Si plusieurs composantes sont présentes dans W_h, celle d'amplitude la plus importante est choisie. Si aucune composante n'est présente, l'indice i n'est pas assigné. Cette méthode est appelée «méthode du crible harmonique» . Elle consiste à tamiser l'ensemble des composantes estimées au moyen d'un peigne constitué des multiples de ω_0. Le seul paramètre à déterminer est la largeur de la grille du tamis $\alpha \in [0, 0.5]$. Une valeur faible de α accentue l'harmonicité du signal.

L'algorithme que nous avons utilisé dans le cas de signaux harmonique est assez semblable à l'algorithme de création de trajet proposé en partie 4.3. Les différences essentielles résident dans le fait que la localisation des trajets est a priori connue, et que le numéro d'un trajet est déterminé par le rapport de sa fréquence à la fréquence fondamentale ; ainsi un numéro de trajet est uniformément attribué à une composante du signal du début à la fin du signal. Le calcul de la probabilité d'observation (courbure polynomiale) et de la probabilité de transition (distance euclidienne) est utilisé pour mesurer la régularité du trajet au cours du temps et non pour déterminer l'appariement des pics. L'amplitude d'un trajet dont la probabilité cumulée est faible localement sera ramenée à 0.

5.4 Caractérisation en voisement/inharmonicité

Nous nous intéressons à deux types de caractérisation. La première correspond à ce qui est appelé en traitement de la parole **«coefficient de voisement»**. Ce coefficient est une mesure de l'activité des cordes vocales. Comme la vibration des cordes vocales produit (dans

la majorité des cas) un signal quasi-périodique, ce coefficient est parfois appelé coefficient de périodicité ou d'harmonicité. Ceci peut prêter à confusion, puisque le coefficient de voisement est avant tout une mesure, à chaque instant et dans une succession de bandes de fréquence, de l'énergie expliquée par la périodicité du signal et non de l'aspect périodique même. En ce sens, le coefficient de voisement est un cas particulier (dans le cas d'un signal harmonique) de l'erreur de modélisation étudiée au chapitre précédent. Un résumé des méthodes d'estimation du coefficient de voisement est proposé dans [Rd96].

A l'inverse, le terme périodicité/harmonicité ne s'oppose pas seulement à celui de bruit mais également à celui d'**inharmonicité** (composantes lentement variable dans le temps, donc représentables par des sinusoïdes mais sans relation harmonique entre elles).

Pour éviter cette ambiguïté, nous parlerons uniquement dans la suite :
- de «coefficient de voisement» pour la mesure de l'erreur de modélisation du signal par un modèle harmonique,
- de «coefficient d'inharmonicité» pour la mesure du caractère harmonique ou non du signal

Nous prenons les définitions suivantes :

coefficient de voisement local : mesure de l'énergie expliquée par la période fondamentale à un temps donné et pour une bande de fréquence donnée. Cette mesure peut s'effectuer par calcul de l'**erreur de modélisation** pour une succession de bandes de fréquence centrées sur les multiples de la fréquence fondamentale et de largeur égale à la fréquence fondamentale ([GL88]).

$$\epsilon(\omega_h) = \frac{\sum_k \text{ tel que } {}_{\omega_k \in W_h} |S(\omega_k) - \hat{S}(\omega_k)|^2}{\sum_k \text{ tel que } {}_{\omega_k \in W_h} |S(\omega_k)|^2} \tag{5.4}$$

dans lequel
- $S(\omega_k)$ est la transformée de Fourier du signal fenêtré $s(t) = x(t)h(t - t_m)$
- $\hat{S}(\omega_k)$ est la transformée de Fourier du modèle harmonique
- W_h est défini comme $0.5(\omega_{h-1} + \omega_h) \leq k < 0.5(\omega_h + \omega_{h+1})$

Le coefficient de voisement est défini comme

$$vois(\omega_h) = 1 - \epsilon(\omega_h) \tag{5.5}$$

Dans le cas d'un modèle harmonique, la valeur prise par ω_h correspond à $h\omega_0$. Cependant une tolérance à l'inharmonicité du signal est souvent introduite. Dans ce cas ω_h s'exprime $h\omega_0(1 + \Delta)$. Selon la valeur de Δ, le coefficient de voisement tend à se rapprocher de l'erreur de modélisation (au sens défini au chapitre précédent) ou d'un coefficient d'inharmonicité. Ceci est illustré à la FIG. 5.3

Utilisation : Le coefficient de voisement local permet de distinguer, dans le cas de l'utilisation d'un modèle harmonique (sinusoïdal harmonique ou PSOLA), les régions fréquentielles du spectre dont l'énergie peut être majoritairement expliquée par le modèle. Le reste de l'énergie du spectre sera représenté par un autre modèle non-harmonique (soit par du bruit, soit un modèle sinusoïdal inharmonique).

coefficient de voisement global : mesure de l'énergie expliquée par la période fondamentale à un temps donné pour l'ensemble des fréquences. Il est obtenu par le calcul de $1 - \epsilon$ pour K_h égal à l'ensemble du spectre.

Utilisation : Le coefficient de voisement global permet la segmentation du signal en segments temporels voisé/non-voisé. Cette segmentation est nécessaire pour le marquage PSOLA (voir partie suivante) ; la contrainte de périodicité du marquage PSOLA n'est en effet appliquée que dans les segments voisés.

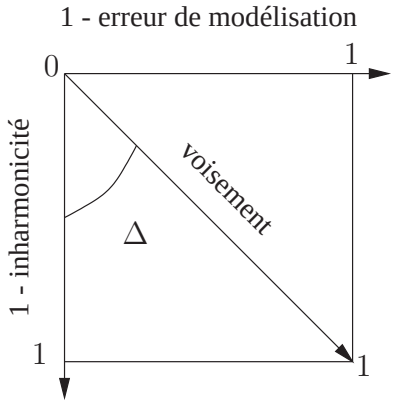


FIG. 5.3 – Erreur de modélisation, coefficient de voisement et coefficient d'inharmonicité.

coefficient d'inharmonicité local : mesure de la déviation des composantes fréquentielles par rapport à un modèle de signal purement harmonique étant donné une fréquence fondamentale.

$$inharmo(\omega_h) = \frac{|\omega_h - h\omega_0|}{\omega_0/2} \tag{5.6}$$

dans lequel ω_h est la fréquence la plus proche du $h^{\text{ème}}$ harmonique de la fréquence fondamentale : $\omega_h = h(1\pm\Delta)\omega_0$; Δ détermine la tolérance du modèle à l'inharmonicité ; ω_0 représente la fréquence fondamentale (non nécessairement égale à ω_1).

Puisque l'inharmonicité maximale (étant donné une fréquence fondamentale) est égale à la demi-fréquence fondamentale, nous normalisons celle-ci par $\omega_0/2$.

coefficient d'inharmonicité global : moyenne des coefficients d'inharmonicité locaux pondérés par l'énergie des composantes.

$$inharmo = \frac{1}{\omega_0/2} \frac{\sum_{h=1}^{H} |S(\omega_h)|^2 \cdot (\omega_h - h\omega_0)^2}{\sum_{h=1}^{H} |S(\omega_h)|^2} \tag{5.7}$$

Le coefficient d'inharmonicité permet de distinguer dans le cas d'un signal représentable par des sinusoïdes, la propension de ces sinusoïdes à être en rapport harmonique.

Utilisation : Le coefficient d'inharmonicité détermine dans quelle mesure un modèle harmonique (sinusoïdal harmonique ou PSOLA) peut être utilisé.

Calcul d'une «fréquence de coupure» voisé/non-voisé Le coefficient de voisement local tel que défini plus haut s'avère difficilement applicable sans un traitement aval. Pour une composante h donnée, les variations temporelles de $\epsilon(\omega_h)$ rendent son utilisation directe difficile. Aussi avons-nous préféré le passage par une fréquence dite «fréquence de coupure» voisé/non-voisé. Nous définissons la fréquence de coupure comme la fréquence à partir de laquelle le modèle harmonique n'explique que faiblement l'énergie du spectre du signal. Cette fréquence divise le spectre en une région basse fréquence dont l'énergie est considérée

comme principalement expliquée par le modèle harmonique et un région haute fréquence dont l'énergie est peu expliquée par le modèle harmonique.

La méthode utilisée pour la détermination de la fréquence de coupure est une méthode du type «risque minimum» [WP94], également utilisée dans [Oud98]. Pour chaque région K_h autour d'une harmonique, nous calculons l'erreur de modélisation d'énergie du signal par le modèle harmonique. Pour chaque région, la valeur $\epsilon(h)$ est comparée à un seuil. La comparaison permet la prise de décision région voisée/non-voisée. Dans un second temps, nous cherchons à déterminer la fréquence de coupure telle que la somme

- nzv : du nombre de bande h considérée comme voisée à gauche de la fréquence
- nznv : du nombre de bande h considérée comme non-voisée à droite de la fréquence

soit maximale.

Illustrations : Le coefficient de voisement global ainsi que la segmentation en régions voisées et non-voisées sont illustrés à la figure Fig. 5.4 pour un signal de voix parlée. La figure du haut illustre le signal et la segmentation résultante de l'application d'un seuil sur la fonction de voisement globale. La figure du bas illustre le coefficient de voisement global.

L'estimation de la fréquence de coupure voisé/non-voisé sur le même signal est représentée à la figure Fig. 5.5 en superposition à son spectrogramme.

Le coefficient de voisement local ainsi que l'inharmonicité locale sont illustrés à la Fig. 5.6 pour un son de piano. Remarquons l'inharmonicité croissante des composantes à mesure que leur fréquence augmente.

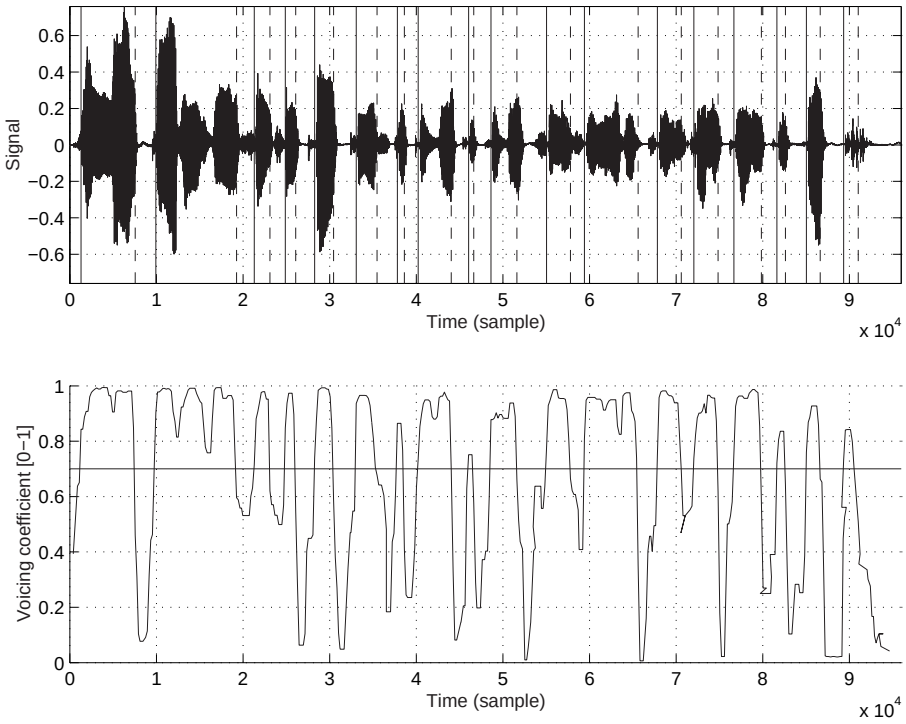


FIG. 5.4 – Calcul du coefficient de voisement global. [H] signal et marques de début (-) et de fin (- -) de régions voisées. [B] fonction temporelle de voisement global. Signal= speech

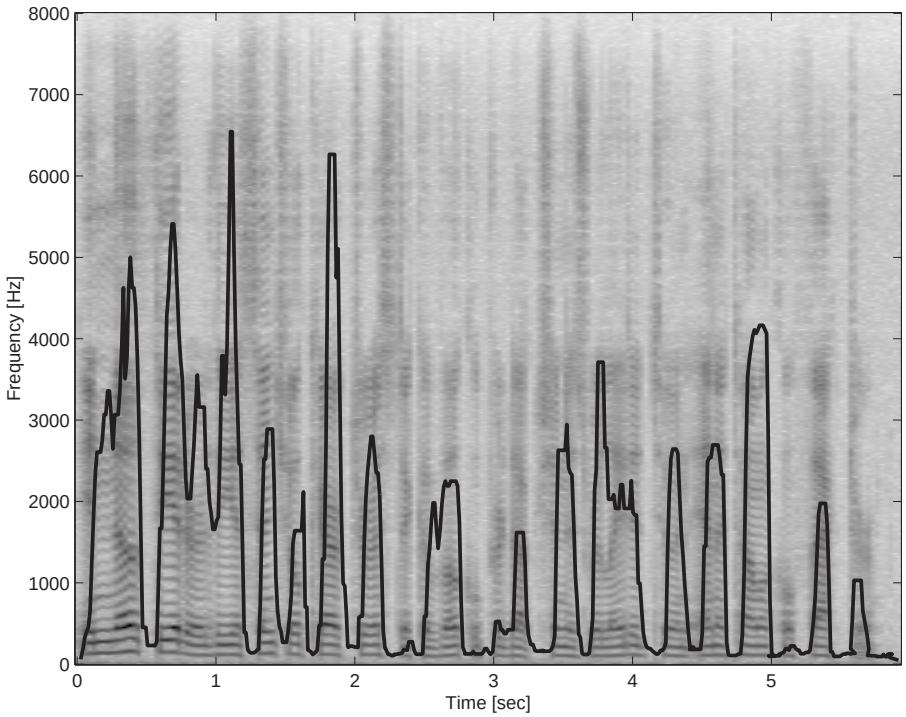


FIG. 5.5 – Calcul de la «fréquence de coupure» voisé/non-voisé en fonction du coefficient de voisement local. Signal= speech

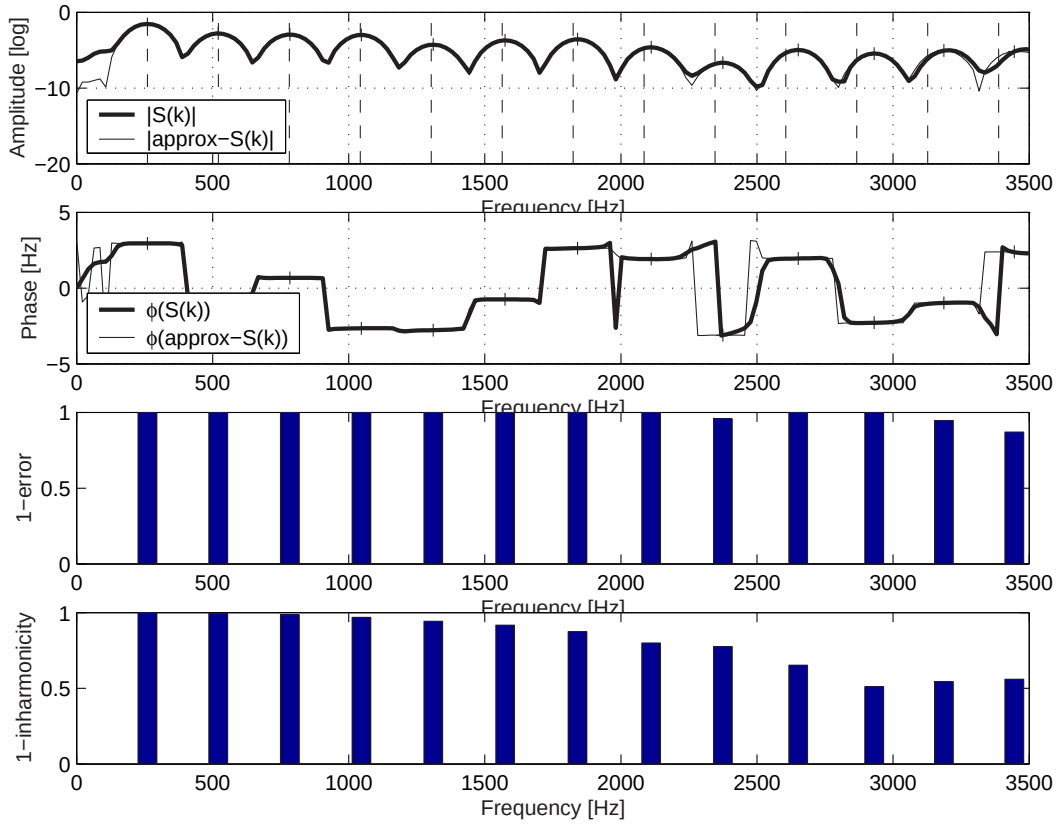


FIG. 5.6 – Du haut vers le bas : spectre d'amplitude et spectre de phase du signal (trait gras) et du modèle (trait fin), coefficient de voisement, (1-coefficient d'inharmonicite). Signal= piano

Notes de bas de page relatives à la partie 5

1. Calcul de la fonction d'auto-corrélation par FFT/IFFT :

$$\begin{aligned} \tilde{r}(k) &= \frac{1}{N-k} FFT^{-1}\left(|FFT(x(n))|^2\right) \\ \hat{r}(k) &= \tilde{r} \qquad \forall k \in [0, L] \end{aligned} \tag{5.2}$$

dans lequel L est la taille de l'observation temporelle et $x(n)$ le signal.

Remarque : *Afin d'éviter le repliement des bords de la fonction d'auto-corrélation dû au traitement dans le domaine fréquentiel, il est nécessaire d'appliquer un facteur de prolongement par zéro . On choisit un facteur de prolongement par zéro d'ordre 2.*

2. Ceci est implémenté de manière économique par utilisation de l'algorithme FFT/IFFT de calcul de la fonction d'auto-corrélation et une mise à zéro des coefficients de la partie supérieur du spectre.

Chapitre 6

Marquage de singularités périodiques

6.1 Introduction

Dans la partie 3 de ce rapport, nous avons étudié la détection des singularités du signal. Ces singularités seront utilisées dans la suite pour l'algorithme de modification du signal PSOLA ; ceci afin de découper le signal en formes d'onde élémentaires . Ce découpage s'effectue par multiplication du signal par une succession de fenêtres de pondération $h(t)$ de tailles proportionnelles à la période fondamentale locale du signal et centrées en des marqueurs m_i placés de manière synchrone à la période fondamentale locale du signal. Dans le cas particulier de l'algorithme PSOLA-bande-large que nous utiliserons, le facteur de proportionnalité est égal à deux. De ce fait, le positionnement des marqueurs m_i doit être proche des maxima locaux de la forme d'onde , de manière à diminuer la détérioration engendrée par le fenêtrage.

Dans [MV95], l'éloignement maximal par rapport au maximum d'énergie de la forme d'onde est estimé à 10 % de la période fondamentale. A l'inverse, la contrainte de périodicité est considérée comme impérative. Dans [KK97], cet éloignement est estimé, sur des bases expérimentales, à 25% de la période fondamentale locale.

Dans le cas d'un signal de parole, les singularités trouvées dans la partie 3 sont proches temporellement des instants de fermeture de la glotte (IFG) et donc proches des maxima locaux du signal. Cependant, nous ne pouvons garantir que la distance entre deux singularités est rigoureusement égale à la période fondamentale locale. Pour d'autres types de signaux, comme ceux de voix chantées ou ceux d'instruments de musique, la localisation des singularités peut être erronée ou encore ne plus être périodique.

Pour toutes ces raisons, nous étudions dans ce chapitre les algorithmes permettant de satisfaire au mieux à deux contraintes :

- la contrainte d'une inter-distance entre marqueurs égale à la période fondamentale locale,
- la contrainte de la localisation des marqueurs proches des maxima locaux de l'énergie.

Les résolutions doivent s'effectuer sur l'ensemble d'un intervalle, puisque le déplacement d'un marqueur en vue de satisfaire un critère d'énergie engendre une modification des périodes gauches et droites du marqueur.

Marquage pour PSOLA ou pour LP-PSOLA Comme nous le verrons dans la partie modification du signal de ce rapport, l'algorithme PSOLA peut être appliqué directement sur le signal (méthode PSOLA) ou sur le signal résiduel (méthode LP-PSOLA). Il est donc logique de faire correspondre le marquage au signal qui sera utilisé par l'algorithme de modification. Pour une synthèse de type PSOLA, le marquage s'effectuera par un estimateur observant le signal : GDS. Pour une synthèse de type LP-PSOLA, il se fera par un estimateur observant le signal résiduel : norme de Frobenius ou GDR.

Lorsqu'une modélisation auto-régressive est applicable, le marquage sur le signal résiduel présente plusieurs avantages.

Premièrement l'estimation des IFGs est facilitée (voir partie 3).

Deuxièmement, les contraintes périodicité/énergie sont plus faciles à satisfaire pour un marquage sur le signal résiduel. En effet, si nous notons $e(t) = \sum_i \delta(t - i\text{T0})$ la succession des IFGs dans le cas d'un signal de parole de période fondamentale T0 constante, les maxima locaux des fonctions d'observation opérant sur le signal résiduel seront proches des maxima de $e(t)$ et seront donc équidistants de T0. En notant $v_\tau(t)$ la réponse impulsionnelle à l'instant t du filtre (variable au cours du temps) du conduit vocal excité à l'instant τ, le signal résultant du filtrage du signal résiduel peut s'écrire $s(t) = \sum_i \delta(t-i\text{T0}) * v_{iT}(t)$. Le décalage des maxima locaux de $s(t)$ par rapport aux maxima de $e(t)$ dépendent maintenant de la variation au cours du temps du filtre $v(t)$ et ne sont donc plus nécessairement équidistants de T0.

Notons D la différence entre la distance entre deux maxima locaux et la période fondamentale T0. Dans le cas d'une évolution lente du filtre $v(t)$, D évolue lentement et cette évolution peut être suivie par un décalage progressif du positionnement des marqueurs m_i. Dans le cas d'une évolution rapide du filtre, la variations de D ne peuvent plus être suivies. Dans ce cas, une autre stratégie est plus adaptée, celle du marquage de manière indépendante du segment précédant la transition et de celui succédant la transition [1]

Le décalage de D est illustré à la figure Fig. 6.1 pour un signal de voix parlée féminine. La partie du haut illustre le marquage sur le signal $s(t)$ et illustre la modification de D due à une transition de formant. La partie du bas illustre le marquage sur le signal résiduel et le fait que D reste quasi-constant lors de la transition. La figure Fig. 6.2 illustre le marquage sur la fonction d'énergie du signal (partie haute) et du signal résiduel (partie basse) pour un signal de voix parlée masculine.

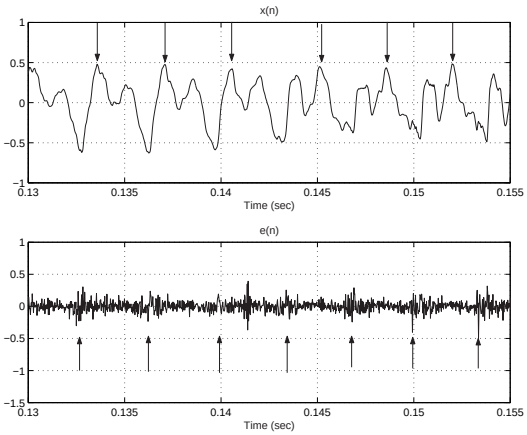


FIG. 6.1 – Contraintes périodicité/énergie pour le positionnement des marques. La modification de la forme d'onde sur le signal [H] empêche le respect des contraintes périodicité/énergie sur le signal. [B] Ces contraintes peuvent être respectées sur le signal résiduel. Signal= voix de femme «abstractionmlf», transition du «l» au «a» de la phrase «l'abstraction de **la** vie te donne raison»

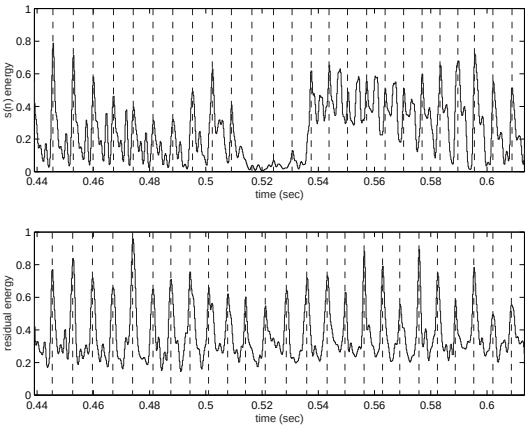


FIG. 6.2 – [H] Marquage sur l'énergie du signal, [B] marquage sur l'énergie du résiduel. Signal= locuteur masculin «Kara»

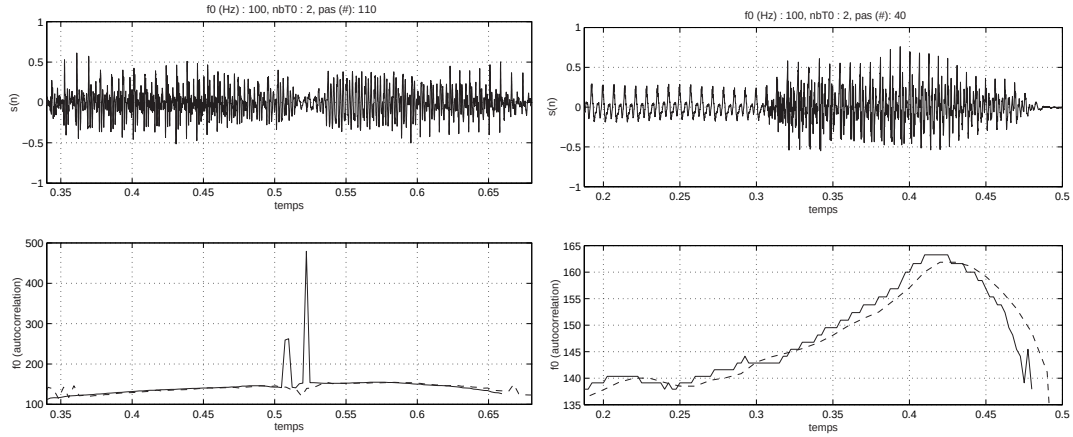


FIG. 6.3 – [H] Signal, [B] fréquence fondamentale f0 estimée par méthode d'auto-corrélation du signal (trait plein), par méthode fréquentielle (pointillés) pour [G] Signal= Kara, [D] Signal= speech

6.2 Détection des maxima locaux de la fonction d'énergie

Nous cherchons le maximum de l'énergie local à chaque période sous la contrainte d'une distance entre deux maxima successifs proche de la période fondamentale locale.

6.2.1 Méthode propagative

Définissons un vecteur d'instants périodiques $\Theta = [\theta_0, \theta_1, ..., \theta_i, ...]$. La valeur des θ_i est déterminée de manière récursive. Pour cela, nous définissons autour de l'instant θ_i, un intervalle $I_i = \left[\theta_i - \frac{\mathrm{T0}_{i-1}}{\alpha}, \theta_i + \frac{\mathrm{T0}_i}{\alpha}\right]$, dans lequel $\mathrm{T0}_i$ est la période fondamentale locale et α détermine la longueur de l'intervalle. Le maximum de l'énergie dans l'intervalle I_i est noté t_i. θ_{i+1} est déterminé par $\theta_{i+1} = t_i + \mathrm{T0}_i$ (voir FIG. 6.4).

Commentaires : *Avantage :* Le marquage est capable de s'adapter à la modification de la forme d'onde au cours du temps. *Désavantage :* Pour des valeurs α petites, l'algorithme peut donc dévier progressivement d'un marquage périodique. Une erreur de marquage à un instant donné se propage sur la suite du segment.

De par son adaptabilité au contenu du signal, cet algorithme semble particulièrement adapté au marquage PSOLA sur le signal.

6.2.2 Méthode vectorielle

Définissons un vecteur d'instants périodiques $\Theta = [\theta_0, \theta_1, ..., \theta_i, ...]$ de telle manière que $\theta_{i+1} - \theta_i = \mathrm{T0}_i$ (voir FIG. 6.5). Autour de chaque instant θ_i, nous définissons un intervalle $I_i = \left[\theta_i - \frac{\mathrm{T0}_{i-1}}{\alpha}, \theta_i + \frac{\mathrm{T0}_i}{\alpha}\right]$, dans lequel α détermine la longueur de l'intervalle. Le maximum de l'énergie dans chaque intervalle I_i est noté t_i.

Commentaires : *Avantage :* Le centre des intervalles I_i reste rigoureusement espacé de T0. Une erreur de marquage ne peut donc pas se propager sur la suite du segment. *Désavantage :* Le marquage ne s'adaptant pas à la modification de la forme d'onde , si la forme d'onde change, le passage du marquage d'un maximum local au marquage d'un maximum local secondaire est possible.

Cet algorithme est particulièrement adapté au marquage LP-PSOLA sur le résiduel.

6.2.3 Choix de la taille de l'intervalle I

I_i est défini comme l'intervalle $\left[\theta_i - \frac{\mathrm{T0}_{i-1}}{\alpha}, \theta_i + \frac{\mathrm{T0}_i}{\alpha}\right]$, où θ_i désigne le centre de l'intervalle et α définit la taille de l'intervalle. α doit être borné inférieurement à 2, puisque $\alpha < 2$ conduirait à utiliser une portion de signal identique d'un I à l'autre. Pour des valeurs très élevées de α, la différence entre les résultats obtenus en utilisant la méthode propagative et la méthode vectorielle sont très minimes. Un compromis est à trouver entre le choix d'un grand

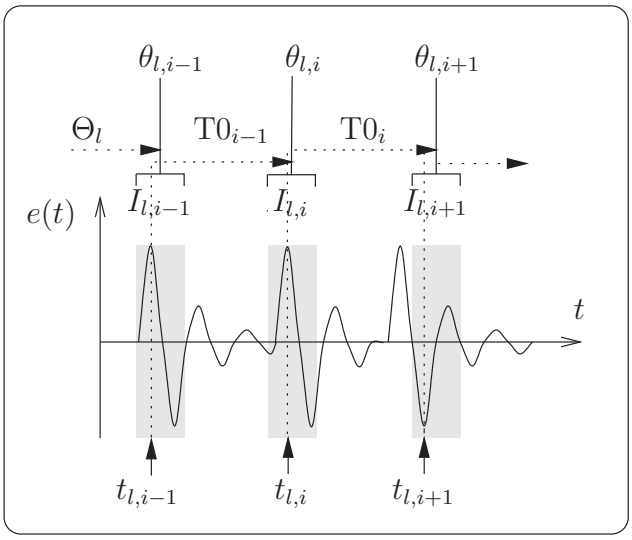


FIG. 6.4 – Détection des maxima locaux de la fonction d'énergie : méthode propagative

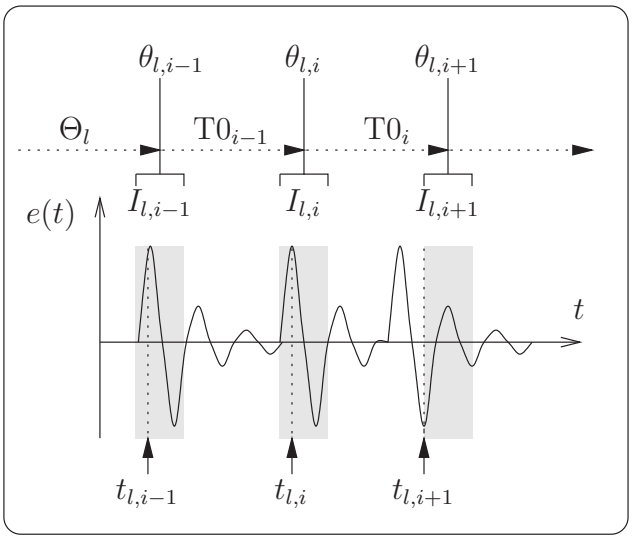


FIG. 6.5 – Détection des maxima locaux de la fonction d'énergie : méthode vectorielle

I (exemple : voir FIG. 6.6 [G]), qui permettra de suivre au mieux l'évolution des maxima de la forme d'onde , - quitte à perdre la synchronie à la période fondamentale -, et le choix d'un petit I (exemple : voir FIG. 6.6 [D]), qui renforcera la périodicité mais ne garantira plus la sélection du maximum local à chaque période. Une valeur de $\alpha = 4$ semble un bon compromis amenant des résultats satisfaisant dans la majorité des cas.

6.2.4 Vecteur de balayage

Cet algorithme vient en complément de l'algorithme vectoriel. En effet, dans le cas où I est choisi petit, l'algorithme vectoriel dépend fortement de l'initialisation du vecteur, et une mauvaise initialisation biaisera la détection sur le reste du segment. Afin de parer à cela, l'algorithme est initialisé (choix de la valeur θ_0) sur plusieurs maxima locaux. Les résultats qui correspondent à l'initialisation produisant un marquage dont la somme de l'énergie du signal à la position des marques est maximale (voir FIG. 6.7) sont finalement retenus. Cet algorithme peut être rapproché des algorithmes de minimisation par recherche aléatoire [WP94].

Soit t_0 l'initialisation du vecteur Θ. Notons $t_{l,0}$ les différents temps d'initialisation du vecteur Θ, Θ_l les vecteurs correspondants, $I_{l,i}$ les intervalles correspondants et $t_{l,i}$ les maxima locaux aux $I_{l,i}$. Pour des intervalles de taille I et un signal de période fondamentale $\mathrm{T0}_0$ en t_0, le nombre d'initialisations différentes possibles est égale à :

$$L = \frac{\mathrm{T0}_0}{I} = \alpha \tag{6.1}$$

Pour chaque vecteur $\Theta_l = [\theta_{l,0}, \theta_{l,1}, ..., \theta_{l,i}, ...]$, nous calculons la somme des valeurs de l'énergie aux temps $t_{l,i}$: $\sigma_l = \sum_i e(t_{l,i})$. Finalement les marques d'énergie sont déterminées comme celles correspondant au vecteur Θ_l dont le σ_l est le plus grand : $\tau_i = t_{l',i}$ où $l' = \arg\max_{l \in L} \sigma_l$.

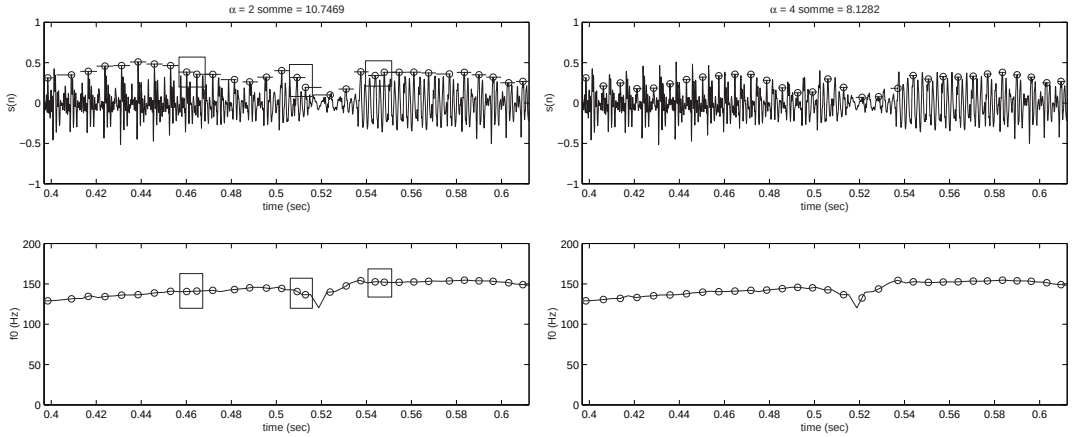


FIG. 6.6 – Choix de la taille de l'intervalle I : [G] $\alpha = 2$ (Signal= Kara), [D] $\alpha = 4$ (Signal= Kara)

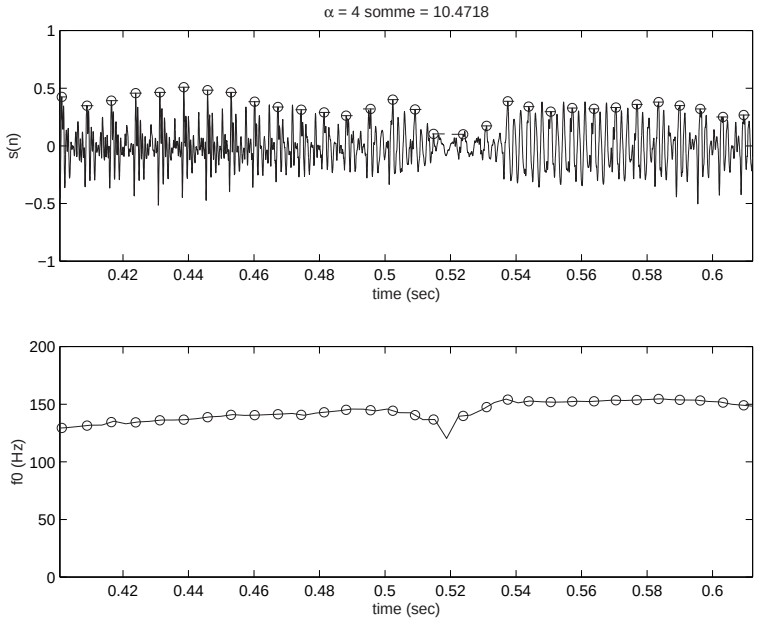


FIG. 6.7 – Vecteur de balayage : $\alpha = 2$ (Signal= Kara)

6.3 Satisfaction des deux contraintes

Les deux algorithmes présentés dans ce paragraphe visent à satisfaire les deux contraintes suivantes :

contrainte de périodicité : la distance entre deux marques successives est égale à la période fondamentale locale $m_{i+1} - m_i = T0_i$

contrainte d'énergie : l'éloignement des marques par rapport aux maxima locaux d'énergie est minimal $m_i = \tau_i$

Le premier algorithme proposé procède de manière itérative en satisfaisant les contraintes en alternance, jusqu'à la stabilisation de la position des marques.

Le deuxième algorithme procède par minimisation d'une erreur, prenant simultanément en compte la contrainte de périodicité et celle d'énergie.

La deuxième méthode proposée garantit un positionnement optimal en fonction des contraintes données. Cependant, en raison de la nécessité d'une inversion de matrice pour l'obtention de la solution, cette méthode peut s'avérer lourde pour des segments comportant un nombre important de marques.

6.3.1 Algorithme itératif

Dans [Bas95], il est proposé un algorithme itératif pour la résolution des contraintes énergie/périodicité. Cet algorithme, dont nous nous sommes inspirés, ne converge cependant que vers la période fondamentale et donc n'effectue pas de réelle optimisation. Nous l'avons corrigé et proposons dans la suite la version corrigée.

6.3.1.1 Notations

Nous notons (voir FIG. 6.8) :

- t_i^j la position de la $i^{\text{ème}}$ marque à la $j^{\text{ème}}$ itération,
- $c_i^j = \frac{t_i^j + t_{i+1}^j}{2}$ la position milieu entre la marque i et la marque $i+1$ à la $j^{\text{ème}}$ itération,
- $\hat{T0}_i^j$ la période fondamentale déterminée par la distance entre t_i^j et t_{i+1}^j à la $j^{\text{ème}}$ itération,
- $T0_i^j$ la période fondamentale locale vraie au temps $(t_i^j + t_{i+1}^j)/2$

6.3.1.2 Initialisation

Soit t_i^0 la position initiale de la $i^{\text{ème}}$ marque. Cette position est choisie égale à la position du maximum local de la fonction d'énergie : $t_i^0 = \tau_i$.

6.3.1.3 Itération

La modification des valeurs est effectuée selon la fonction non-linéaire suivante, empêchant des variations brusques d'une itération à la suivante :

$$x^{j+1} = x^j \cdot \left[\frac{y}{x^j}\right]^\theta \qquad \theta \in [0,1] \tag{6.2}$$

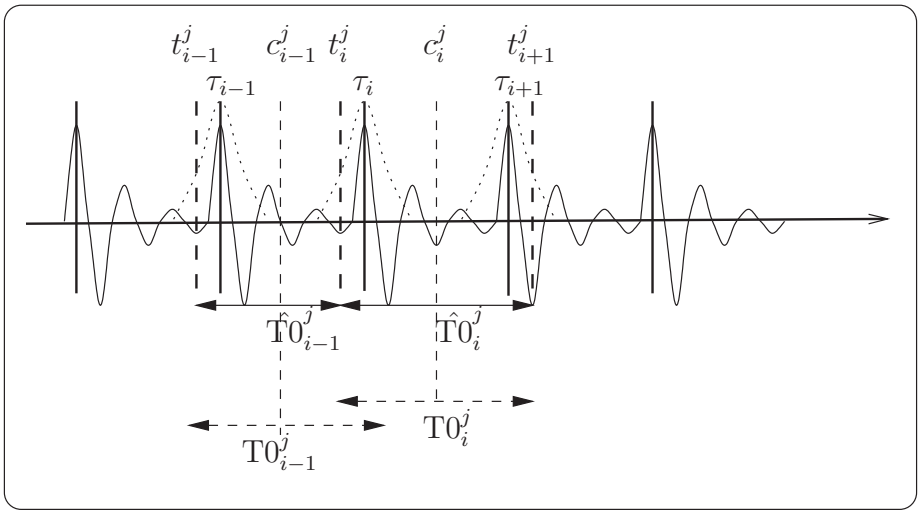


FIG. 6.8 – Algorithme itératif

À chaque itération, nous déplaçons les marques de manière à satisfaire à la

contrainte de périodicité en respectant la contrainte d'énergie : les marques t_i^j sont déplacées de manière à faire tendre $\hat{\mathrm{T0}}_{i-1}^j$ et $\hat{\mathrm{T0}}_i^j$ vers $\mathrm{T0}_{i-1}^j$ et $\mathrm{T0}_i^j$, mais en se gardant d'éloigner t_i^j trop loin de τ_i.

contrainte d'énergie en respectant la contrainte de périodicité : les marques t_i^j sont déplacées de manière à faire tendre t_i^j vers τ_i, mais en se gardant d'éloigner trop $\hat{\mathrm{T0}}_{i-1}^j$ et $\hat{\mathrm{T0}}_i^j$ de $\mathrm{T0}_{i-1}^j$ et $\mathrm{T0}_i^j$.

6.3.1.4 Contrainte de périodicité en respectant la contrainte d'énergie

Chaque déplacement de marque t_i^j influençant la période de gauche $\hat{\mathrm{T0}}_{i-1}^j$ et de droite $\hat{\mathrm{T0}}_i^j$, les deux contributions doivent être prises en compte simultanément.

La prise en compte simultanée de la contribution des modifications apportées à la période de gauche et à la période de droite sur la $i^{\text{ème}}$ marque à la $j^{\text{ème}}$ itération nous fournit la position de la marque à la $(j+1)^{\text{ème}}$ itération :

$$\boxed{t_i^{j+1} = t_i^j + \frac{\hat{\mathrm{T0}}_{i-1}^j}{2}\left[\left(\frac{\mathrm{T0}_{i-1}^j}{\hat{\mathrm{T0}}_{i-1}^j}\right)^{\gamma} - 1\right] - \frac{\hat{\mathrm{T0}}_i^j}{2}\left[\left(\frac{\mathrm{T0}_i^j}{\hat{\mathrm{T0}}_i^j}\right)^{\gamma} - 1\right]} \tag{6.3}$$

Respect de la contrainte d'énergie : γ permet de minimiser l'effet du déplacement (dû à la satisfaction de la contrainte de périodicité) de la marque sur la contrainte d'énergie. γ est fonction du déplacement de la $i^{\text{ème}}$ marque par rapport à τ_i, ceci proportionnellement à

la période locale :

$$\gamma = f\left(\frac{|s_i^{j+1} - \tau_i|}{\frac{1}{2}(\mathrm{T0}_{i-1} + \mathrm{T0}_i)}\right) \tag{6.4}$$

La valeur de s_i^{j+1} choisie pour le calcul de γ est celle qui correspondrait à la satisfaction complète de la contrainte de périodicité, i.e.

$$s_i^{j+1} = t_i^j + \frac{\mathrm{T0}_{i-1}^j}{2} - \frac{\mathrm{T0}_i^j}{2} \tag{6.5}$$

γ mesure donc la dégradation de la contrainte d'énergie si la modification de la position des marques de manière à satisfaire aux périodes était appliquée à 100 %. γ pondère le déplacement de la marque t_i^j de manière à limiter cette dégradation.

La fonction $f(x)$, définie plus loin, est à décroissance infinie. De ce fait, γ ne peut faire que ralentir la convergence des marques vers une position totalement synchrone à la période fondamentale, d'où la nécessité de la deuxième étape de l'itération.

6.3.1.5 Contrainte d'énergie en respectant la contrainte de périodicité

La deuxième partie de l'itération tend à satisfaire la contrainte d'énergie. La marque t_i^j est déplacée de manière à tendre vers sa position d'énergie maximale τ_i selon

$$\boxed{t_i^{j+1} = t_i^j \left(\frac{\tau^i}{t_i^j}\right)^{\delta}} \tag{6.6}$$

Respect de la contrainte de périodicité : δ permet de minimiser l'effet du déplacement (dû à la satisfaction de la contrainte d'énergie) de la marque sur la contrainte de périodicité. δ est fonction de la modification apportée à $\hat{\mathrm{T0}}_{i-1}^j$ et à $\hat{\mathrm{T0}}_i^j$ par rapport à $\mathrm{T0}_{i-1}^j$ et à $\mathrm{T0}_i^j$, ceci proportionnellement à la période locale.

$$2\delta = f\left(\frac{|(s_i^{j+1} - t_{i-1}^j) - \mathrm{T0}_{i-1}|}{\mathrm{T0}_{i-1}}\right) + f\left(\frac{|(t_{i+1}^j - s_i^{j+1}) - \mathrm{T0}_i|}{\mathrm{T0}_i}\right) \tag{6.7}$$

La valeur de s_i^{j+1} choisie pour le calcul de δ est celle qui correspondrait à la satisfaction complète de la contrainte d'énergie, i.e.

$$s_i^{j+1} = \tau_i \tag{6.8}$$

δ mesure donc la dégradation de la contrainte de périodicité si la modification de la position de la marque de manière à satisfaire la contrainte d'énergie était appliquée à 100 % ; δ pondère le déplacement de la marque t_i^j de manière à limiter cette dégradation.

6.3.1.6 Choix des fonctions de contraintes énergie / périodicité

- Pour γ, la fonction $f(x)$ est une mesure du déplacement de la marque. Elle est choisie égale au produit de deux fonctions :

$$f(x) = \alpha(x) \cdot \beta(x) \tag{6.9}$$

  - $\beta(x)$ est une fonction tenant compte de l'éloignement de la marque par rapport à τ_i, et
  - $\alpha(x)$ est une fonction prenant en compte la position de la marque dans le segment. Cette dernière permet une pondération différente selon la position de la marque dans le segment, et permet par exemple de fixer les marques proches des bords du segment afin de faciliter la convergence de l'algorithme.
- Pour δ, la fonction $f(x)$ est une mesure de la modification de la période fondamentale, elle est choisie égale à la fonction $\beta(x)$

$$f(x) = \beta(x) \tag{6.10}$$

◊ ***Fonction*** $\beta(x)$

$\beta(x)$ est choisi de manière à ne pas bloquer complètement le mouvement d'une marque. L'usage de fenêtres de pondération classiques est donc exclu. De même, les fonctions de type Gauss sont exclues pour raison de décroissance trop rapide.

La fonction retenue est la suivante (voir Fig. 6.9) :

$$\beta(x) = \frac{1}{1 + b0\left[\frac{x - x_0}{\hat{\text{T0}}}\right]^2} \tag{6.11}$$

dans laquelle $x = t_i^j$ et $x_0 = \tau_i$

◊ ***Fonction*** $\alpha(x)$

Pour $\alpha(x)$, l'utilisation de fenêtres de pondération classiques (Hamming, Hanning) n'est pas appropriée, puisque celles-ci possèdent des points d'inflexion, ce qui est dangereux dans un algorithme de convergence. La fonction utilisée est un arc de sinus (voir Fig. 6.10) :

$$\alpha(x) = \alpha_0 sin\left(\frac{\pi x}{L}\right) \tag{6.12}$$

dans laquelle $x = t_i^j$ et L est la longueur du segment de signal analysé

6.3.1.7 Combinaison des deux contraintes

◊ ***Contraintes en alternance***

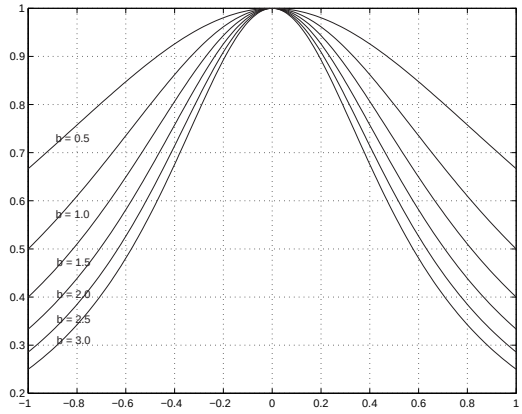


FIG. 6.9 – Contrainte $\beta(x)$ pour différentes valeurs de $b = b0$, valeurs normalisées : $x_0 = 0$ et $T = 1$

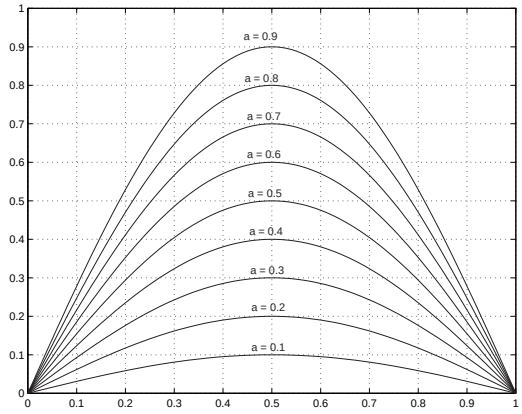


FIG. 6.10 – Contrainte $\alpha(x)$ pour différentes valeurs de $a = \alpha_0$, valeurs normalisées : $L = 1$

Cet algorithme ne converge pas (oscillation infinie). Le déplacement provoqué par une contrainte est rattrapé par l'autre.

◊ ***Contraintes simultanées : formulation 1***

$$t_i^{j+1} = \underbrace{t_i^j \left(\frac{\tau_i}{t_i^j}\right)^{\delta}}_{\text{contrainte énergie}} + \underbrace{\frac{\hat{\mathrm{T0}}_{i-1}^{j}}{2}\left[\left(\frac{\mathrm{T0}_{i-1}^{j}}{\hat{\mathrm{T0}}_{i-1}^{j}}\right)^{\gamma} - 1\right] - \frac{\hat{\mathrm{T0}}_{i}^{j}}{2}\left[\left(\frac{\mathrm{T0}_{i}^{j}}{\hat{\mathrm{T0}}_{i}^{j}}\right)^{\gamma} - 1\right]}_{\text{contrainte périodicité}} \tag{6.13}$$

L'algorithme est arrêté lorsque le carré de la somme des déplacements d'une itération à l'itération suivante passe en dessous de la résolution numérique du calculateur :

$$\sum_i \left(t_i^{j+1} - t_i^j\right)^2 < 2.2204e - 16 \tag{6.14}$$

Les erreurs suivantes sont calculées :

- erreur sur la périodicité :

$$\boxed{\epsilon_p = \sum_i \left(\hat{\mathrm{T0}}_i^{j+1} - \mathrm{T0}_i\right)^2} \tag{6.15}$$

- erreur sur l'énergie :

$$\boxed{\epsilon_e = \sum_i \left(t_i^{j+1} - \tau_i\right)^2} \tag{6.16}$$

Nous désignons par $b0p$ la valeur de $b0$ utilisée pour δ (dégradation de la périodicité) et $b0e$ celle utilisée pour γ (dégradation de l'énergie).

Nous n'avons pas testé la convergence de l'algorithme d'un point de vue théorique. A l'inverse nous avons testé la convergence sur un signal test dont les conditions initiales (positions des maxima d'énergie τ_i, période fondamentale voulue $\mathrm{T0}_i$) rendent difficile la satisfaction des contraintes. Ce signal est illustré aux figures FIG. 6.11, FIG. 6.12 et FIG. 6.13.

Dans le tableau suivant, nous indiquons, pour différents choix des paramètres $b0p$ et $b0e$, le nombre d'itérations nécessaires pour atteindre le critère d'arrêt (6.14). Nous indiquons également les erreurs ϵ_p et ϵ_r obtenues lors de l'arrêt.

$b0p = b0e = b0$	nombre d'itérations nécessaires	ϵ_p	ϵ_e
¡ 6	oscillation infinie		
6	331	0.0028	0.0028
10	57	0.0027	0.0029
13, 14	minimum 39	0.0026	0.0029
20	51	0.0026	0.0030
50	107	0.0025	0.0030
100	191	0.0025	0.0031
...			

Observations : Cet algorithme converge pour les valeurs de $b0 > 10$. Pour une valeur plus faible de $b0$, les fenêtres de contraintes énergie/périodicité étant trop larges, l'algorithme ne parvient pas à se stabiliser et les valeurs oscillent indéfiniment. Des valeurs plus grandes de $b0$ facilitent la convergence de l'algorithme, mais le choix d'un $b0$ élevé (écart autorisé faible à chaque itération) ralentit la convergence. Pour le choix $b0p = b0e = b0$, les deux erreurs ϵ_p et ϵ_e sont minimisées de manière (quasi) équivalente.

◊ ***Contraintes simultanées : formulation 2***

$$\begin{aligned} s_i^{j+1} &= t_i^j + \underbrace{\frac{\hat{\mathrm{T0}}_{i-1}^j}{2}\left[\left(\frac{\mathrm{T0}_{i-1}^j}{\hat{\mathrm{T0}}_{i-1}^j}\right)^{\gamma} - 1\right] - \frac{\hat{\mathrm{T0}}_i^j}{2}\left[\left(\frac{\mathrm{T0}_i^j}{\hat{\mathrm{T0}}_i^j}\right)^{\gamma} - 1\right]}_{\text{contrainte périodicité}} \\ t_i^{j+1} &= \underbrace{s_i^{j+1}\left(\frac{\tau_i}{s_i^{j+1}}\right)^{\delta}}_{\text{contrainte énergie}} \end{aligned} \tag{6.17}$$

Pour le même signal que précédemment (FIG. 6.11, FIG. 6.12 et FIG. 6.13), nous avons observé les nombres d'itérations nécessaires suivants pour atteindre le critère d'arrêt (6.14) :

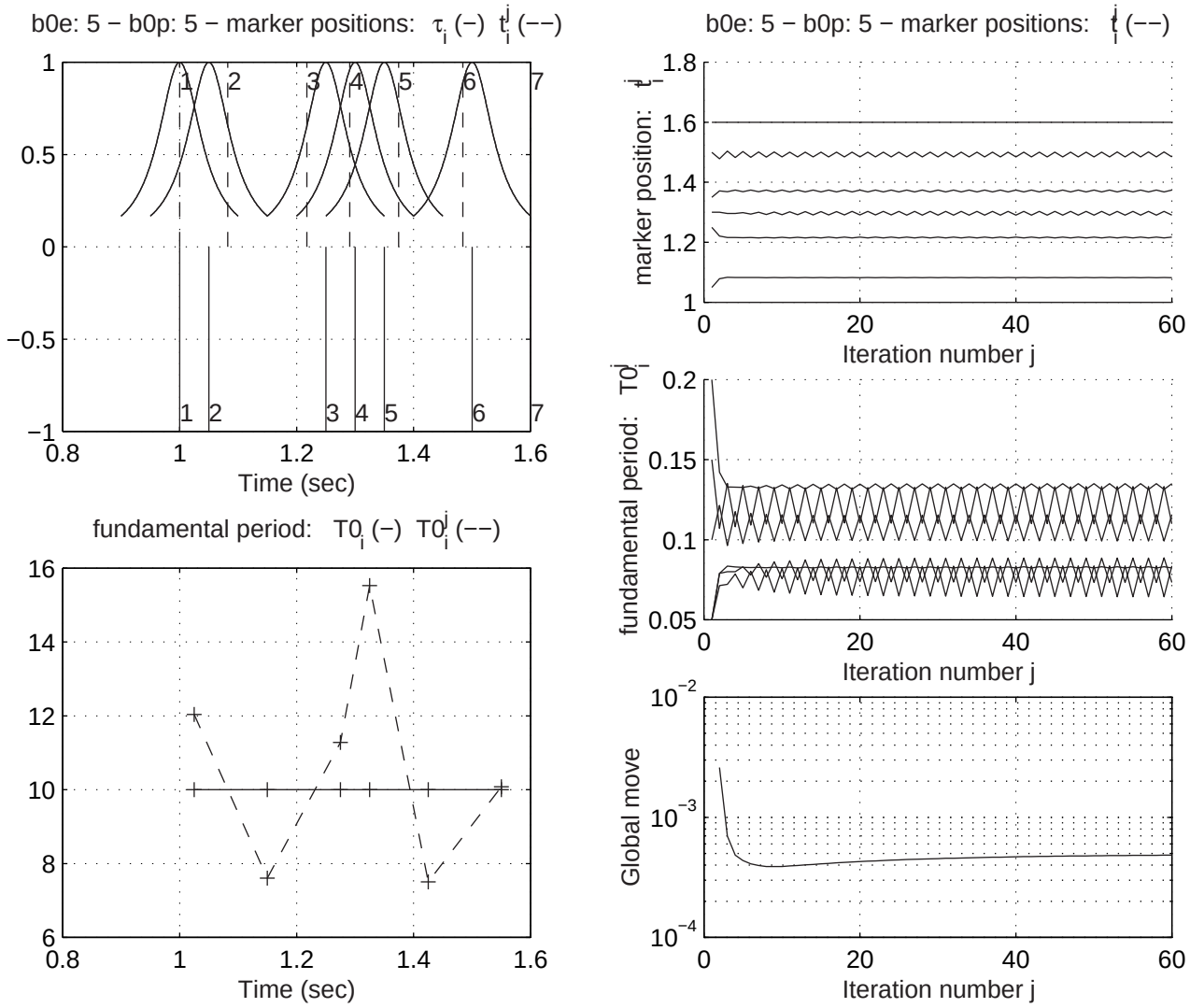


FIG. 6.11 – Algorithme itératif, formulation 1, étude de la convergence pour $b0 = 5$
[HG] position initiale des marques τ_i (trait plein), position des marques m_i après N itérations (trait pointillé)
[BG] période fondamentale voulue $T0_i$ (trait plein), période fondamentale $\hat{T0}_i$ obtenue après N itérations (trait pointillé)
[HD] évolution de la position des marqueurs m_i en fonction du nombre d'itérations
[MD] évolution des périodes fondamentales $\hat{T0}_i$ en fonction du nombre d'itérations
[BD] déplacement global des marqueurs m_i d'une itération à la suivante selon (6.14)

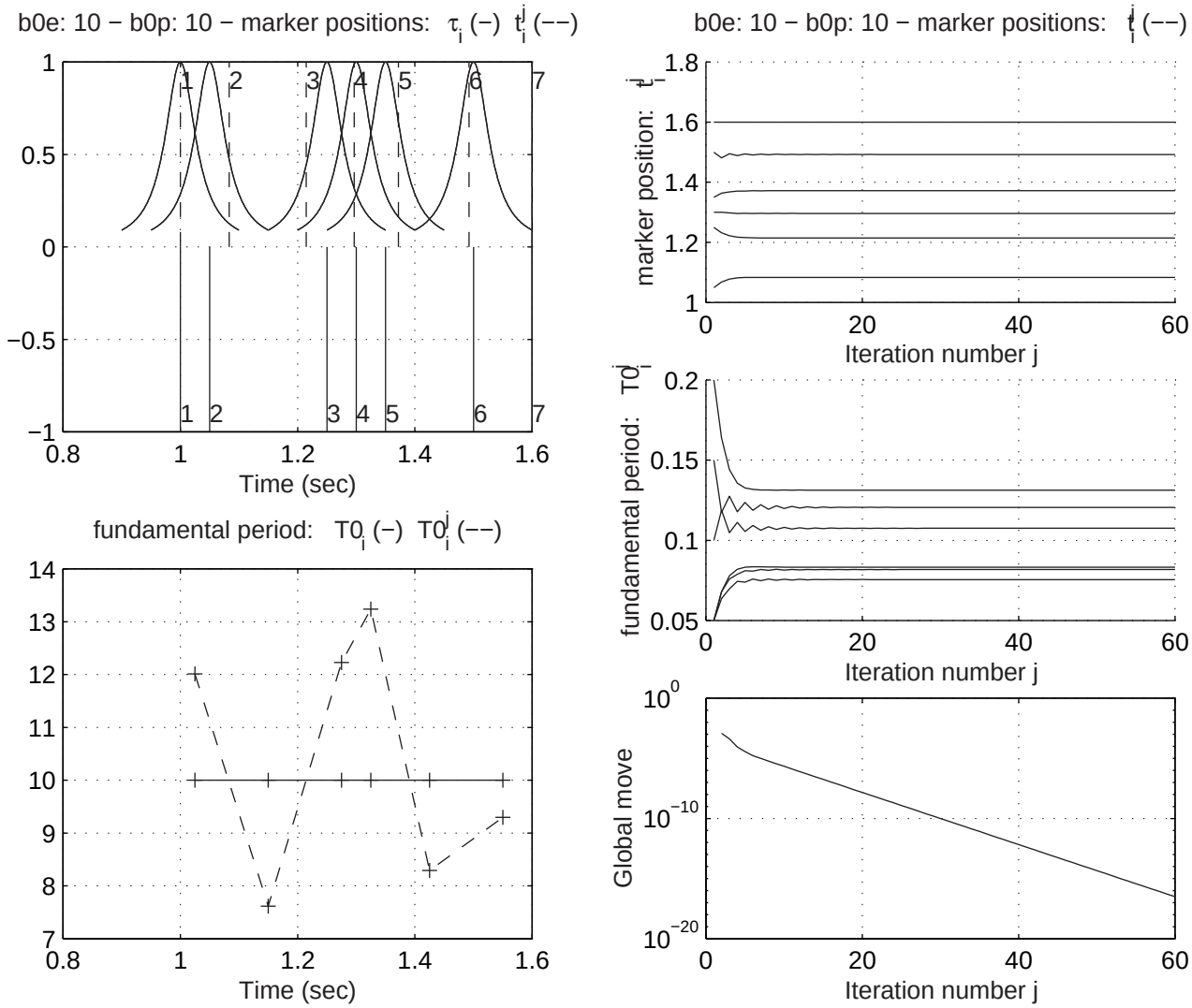


FIG. 6.12 – Algorithme itératif, formulation 1, étude de la convergence pour $b0 = 10$

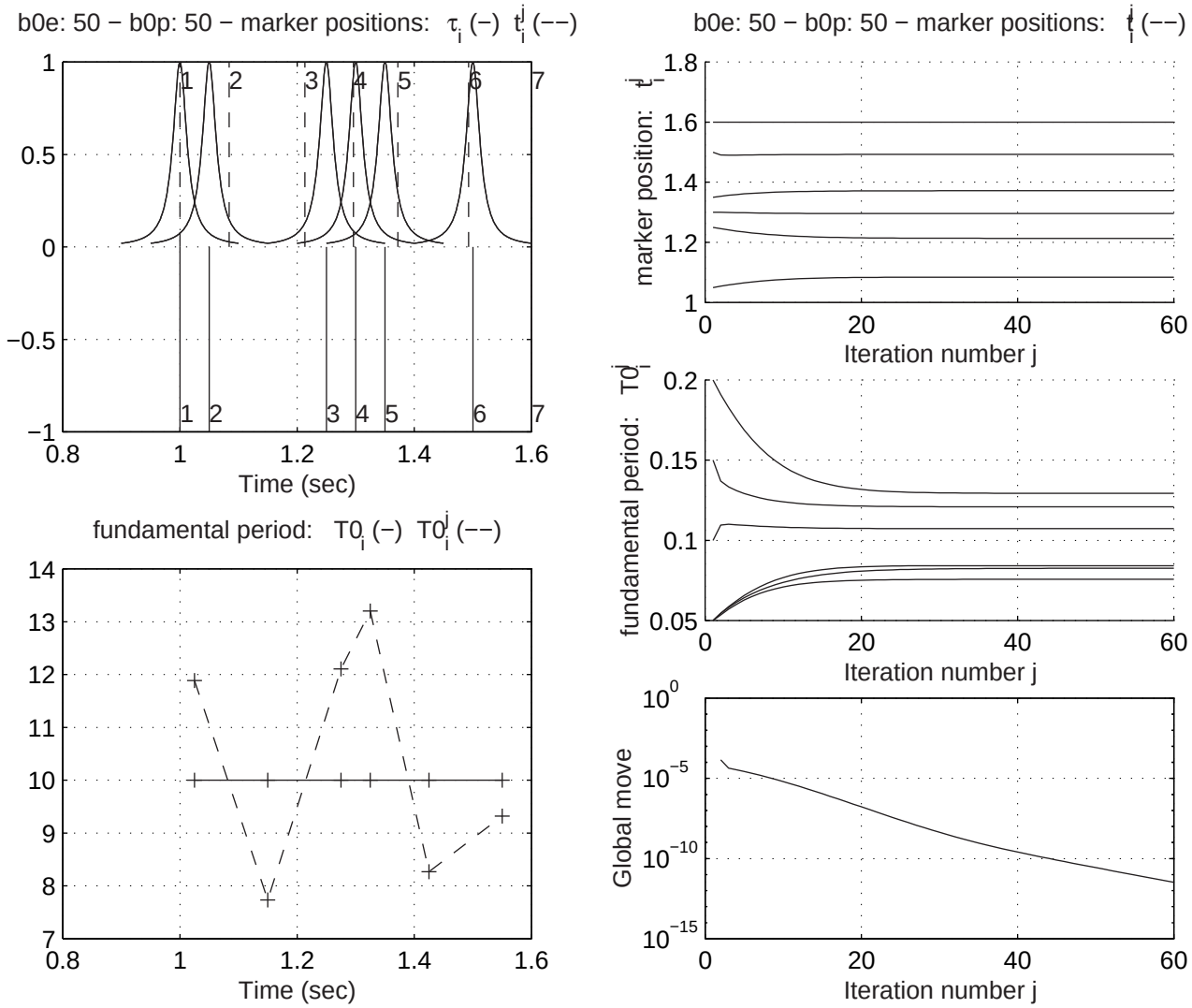


FIG. 6.13 – Algorithme itératif, formulation 1, étude de la convergence pour $b0 = 50$

$b0p = b0e = b0$	nombre d'itérations nécessaires	ϵ_p	ϵ_e
1	13	0.0099	6.5249e-04
6	24	0.0046	0.0020
10	32	0.0038	0.0023
13	37	0.0035	0.0025
20	51	0.0032	0.0027
50	105	0.0027	0.0029
100	188	0.0026	0.0030
...			

Observations : Cet algorithme converge très rapidement (de manière inversement proportionnelle à $b0$). Cependant, de par sa formulation, l'algorithme favorise la contrainte d'énergie au dépend de la contrainte de périodicité. Ceci s'observe particulièrement pour les valeurs faibles de $b0$: dans ce cas ϵ_e est inférieur à ϵ_p d'un ordre de grandeur 10.

6.3.2 Minimisation d'une erreur quadratique énergie/ périodicité

Le deuxième algorithme que nous proposons [SP00] minimise une erreur quadratique définie comme la somme des carrés des erreurs dues à l'éloignement des maxima locaux et des erreurs dues à la non-périodicité du marquage.

6.3.2.1 Notations (voir FIG. 6.14)

Nous notons (voir FIG. 6.14)
- Soit τ_i $i \in I$ les positions des maxima locaux de l'énergie.
- $T0_i$ la période fondamentale (supposée localement constante) au temps τ_i.
- m_i $i \in I$ les positions des marques que nous cherchons

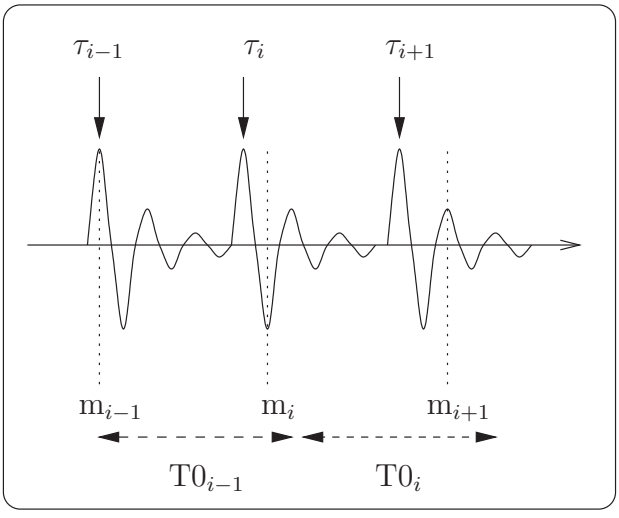


FIG. 6.14 – Algorithme des moindres carrés

6.3.2.2 Formulation du problème

Pour chaque marque m_i, nous voulons satisfaire les contraintes suivantes :
- contrainte de périodicité droite : $\mathrm{m}_{i+1} - \mathrm{m}_i = \mathrm{T0}_i$
- contrainte de périodicité gauche : $\mathrm{m}_i - \mathrm{m}_{i-1} = \mathrm{T0}_{i-1}$
- contrainte d'énergie : $\mathrm{m}_i = \tau_i$

$$\begin{cases} \mathrm{m}_i - \mathrm{m}_{i-1} = \mathrm{T0}_{i-1} \\ \mathrm{m}_{i+1} - \mathrm{m}_i = \mathrm{T0}_i \\ \mathrm{m}_i = \tau_i \end{cases} \tag{6.18}$$

Nous cherchons à minimiser la somme des carrés des erreurs, ϵ, définit comme

$$\epsilon = \sum_{i \in I} \left[((\mathrm{m}_i - \mathrm{m}_{i-1}) - \mathrm{T0}_{i-1})^2 + \beta(\mathrm{m}_i - \tau_i)^2 \right] \tag{6.19}$$

Pondération β et γ β sert à pondérer l'influence de chacune des deux contraintes :
- $\beta \to 0$: l'algorithme prend en compte uniquement la contrainte de périodicité,
- $\beta \to \infty$: l'algorithme prend en compte uniquement la contrainte d'énergie,
- $\beta = 1$: l'algorithme prend en compte de manière équivalente les deux contraintes.

β peut être remplacé par γ aux extrémités du segment. Ceci permet de pondérer différemment les critères aux extrémités (exemple : le choix de $\gamma > \beta$ permet de favoriser le respect des maxima d'énergie aux extrémités du segment, ainsi que de «bloquer» les marques extrémités s'il s'agit de transitoires).

Interprétation : La pondération de l'erreur d'énergie par le facteur β peut s'interpréter comme une modification de la fonction d'éloignement de la marque par rapport au maximum d'énergie. En effet, $\beta(\mathrm{m}_i - \tau_i)^2$ peut s'interpréter comme la valeur de la fonction $f(x) = \beta(x - \tau_i)^2$ évaluée à la position m_i. Le coefficient β règle l'étroitesse de la fonction. La fonction $f(x)$ est représentée à la Fig. 6.15 pour différentes valeurs de β.

La minimisation de ϵ s'obtient par annulation de sa dérivée par rapport aux m_i :

$$\frac{\partial \epsilon}{\partial \mathrm{m}_i} = 0 \qquad \forall i \in I \tag{6.20}$$

Notons $\mathbf{m} = [\mathrm{m}_0 \mathrm{m}_1 ... \mathrm{m}_i ... \mathrm{m}_I]$ le vecteur des marques que nous cherchons. La position optimale des marques est alors donnée par

$$\mathbf{m} = \mathbf{M^{-1}} \cdot \begin{pmatrix} 0 & -\mathrm{T0}_0 & +\gamma\tau_0 \\ \mathrm{T0}_0 & -\mathrm{T0}_1 & +\beta\tau_1 \\ \mathrm{T0}_1 & -\mathrm{T0}_2 & +\beta\tau_2 \\ \vdots & \vdots & \vdots \\ \mathrm{T0}_{i-1} & -\mathrm{T0}_i & +\beta\tau_i \\ \vdots & \vdots & \vdots \\ \mathrm{T0}_{I-2} & -\mathrm{T0}_{N-1} & +\beta\tau_{I-1} \\ \mathrm{T0}_{I-1} & 0 & +\gamma\tau_I \end{pmatrix} \tag{6.21}$$

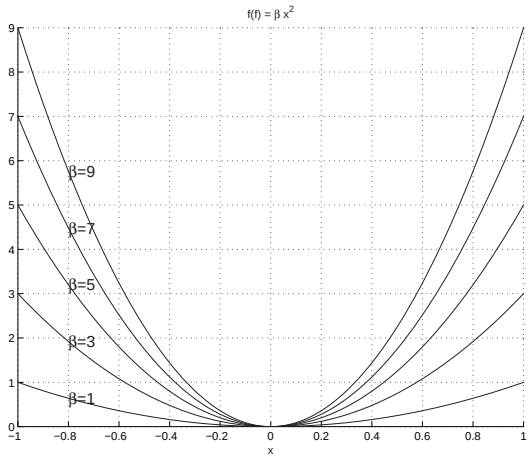


FIG. 6.15 – Interprétation de la pondération β comme fonction d'éloignement de la marque m_i du maximum local d'énergie τ_i

où $\mathbf{M}$ est défini comme

$$\mathbf{M} = \begin{pmatrix} 1+\gamma & -1 & 0 & \dots & & & & \\ -1 & 2+\beta & -1 & 0 & \dots & & & \\ 0 & -1 & 2+\beta & -1 & 0 & \dots & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & \dots & 0 & -1 & 2+\beta & -1 & 0 & \dots \\ & & & \ddots & \ddots & \ddots & \ddots & \\ & & & & 0 & -1 & 2+\beta & -1 \\ & & & & \dots & 0 & -1 & 1+\gamma \end{pmatrix} \tag{6.22}$$

6.3.2.3 Observation des résultats

Dans l'exemple des FIG. 6.11, FIG. 6.12 et FIG. 6.13, nous obtenons avec l'algorithme de minimisation de l'erreur quadratique $\epsilon_p = 0.0015$ et $\epsilon_e = 0.0038$.

Notes de bas de page relatives à la partie 6

1. Afin de localiser les variations brusques de $v(t)$, nous utilisons une conséquence de cette variation : la modification de la forme d'onde locale. Pour détecter cela, un artefact de l'algorithme d'estimation de la fréquence fondamentale f0 par auto-corrélation du signal, la sensibilité aux variations de la forme d'onde , est utilisé. L'estimation de f0 obtenue par la méthode d'auto-corrélation est comparée à chaque instant à l'estimation obtenue par une méthode fréquentielle d'estimation de f0. La méthode fréquentielle utilisée est celle de [Dov94]. Une divergence importante entre les deux estimations nous indique une modification importante de la forme d'onde à cet endroit. Ceci est illustré aux figures Fig. 6.3. La partie de gauche de la figure illustre le cas d'une transition rapide de formant, celle de droite celui d'une transition lente.

Résumé de la partie caractérisation

A la figure FIG. 6.16, nous avons représenté sous forme de diagramme les différentes étapes de la caractérisation du signal utilisée dans cette recherche. Les traits continus indiquent la communication des données entre les différents blocs d'analyse. Les traits pointillés indiquent les données optionnelles.

Fréquence fondamentale : estimation de la fréquence fondamentale soit par la méthode de l'auto-corrélation (dans ce cas un filtrage passe-bas, de fréquence de coupure de 1000 Hz, est appliqué préalablement au signal), soit par la méthode du maximum de vraisemblance.

Voisement, inharmonicité : estimation des coefficients de voisement et d'inharmonicité globaux en fréquence $vois(t), inharm(t)$ et locaux en fréquence $vois(t, w_h), inharm(t, w_h)$.

Singularités : détection des singularités dans le signal par calcul des fonctions globales en fréquence $\gamma_n(t)$, $\sigma_n(t)$ (dans ce cas un filtrage passe-haut, de fréquence de coupure de 600 Hz, est appliqué préalablement au signal) et locales en fréquence (par bande d'octave W) $\gamma_n(t, W)$, $\sigma_n(t, W)$; ainsi que des valeurs limites $\gamma_{h,n}(t)$ et $\sigma_{h,n}(t)$. L'estimation peut s'effectuer soit sur le signal (estimateur GDS), soit sur le signal résiduel (estimateur GDR). Le même algorithme est utilisé pour la détection des transitoires en utilisant un horizon plus large. Une connaissance approximative de la fréquence fondamentale est souhaitée mais pas indispensable.

Placement des marques PSOLA : dans les régions de coefficient $vois(t)$ supérieur à un seuil, régions dites voisées, les marques PSOLA tm sont positionnées en satisfaisant aux contraintes de périodicité ($f0$) et de maxima locaux d'énergie (maxima locaux de γ_n) à l'aide de l'algorithme de minimisation de l'erreur quadratique pour les segments courts, de l'algorithme itératif pour les segments longs.

Sinusoïdalité : l'estimation des paramètres du modèle sinusoïdal d'ordre 1 $(\omega_h, \Delta_h, a_{0,h}, a_{1,h}, \phi_{0,h})$ est effectuée par mesure de la distorsion du spectre complexe. Une connaissance approximative de la fréquence fondamentale ou l'utilisation des marques PSOLA (pour une analyse synchrone) bénéficie à l'analyse mais n'est pas indispensable.

Création de trajets : les trajets du modèle sinusoïdal sont créés à partir des paramètres $(\omega_h, \Delta_h, a_{0,h}, a_{1,h}, \phi_{0,h})$ et de l'algorithme de courbure polynomial/ distance euclidienne.

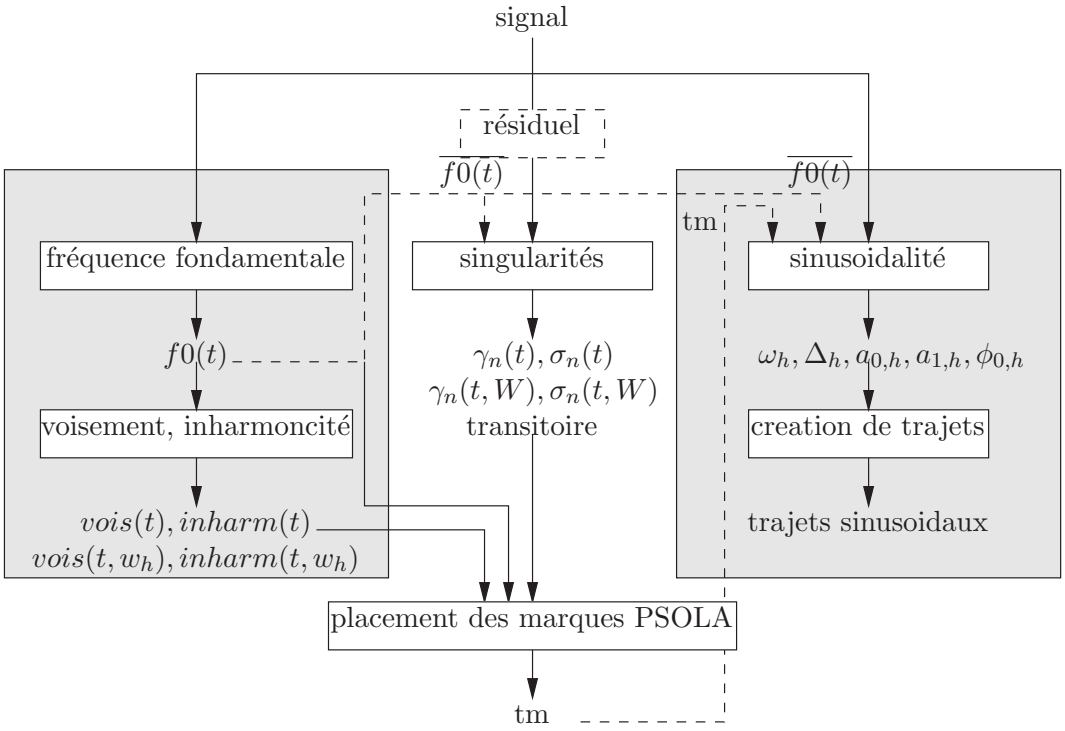


FIG. 6.16 – Diagramme de la partie caractérisation du signal

Nous pouvons valider maintenant le deuxième point de notre thèse
«la phase contient une information pertinente pour l'analyse des signaux sonores (localisation temporelle, localisation fréquentielle, synchronie des évènements fréquentiels)»
Dans cette première partie de notre recherche, l'information de phase du signal a été utilisée pour la localisation temporelle des singularités du signal, pour l'estimation des paramètres du modèle sinusoïdal, ainsi que pour la création de trajets temporels de sinusoïdes.

Deuxième partie

Modifications du signal

Chapitre 7

Modifications du signal par la méthode PSOLA

7.1 Introduction

Les principes de base de la méthode PSOLA ont été décrit dans la partie 2.1. L'estimation des paramètres nécessaire à son utilisation ont été étudiés dans les parties 3.4 et 6. Ce sont ces estimations que nous utilisons ici. Dans cette partie, nous détaillons la méthode PSOLA étudiée ainsi que les améliorations pouvant être apportées.

7.2 Algorithmes de PSOLA étudiés

7.2.1 Découpage du signal en formes d'onde élémentaires

Le signal $s(t)$ est découpé en formes d'onde élémentaires par un fenêtrage $h(t)$ exactement centré sur les périodes fondamentales m_i. Les marques de lecture m_i telles que déterminées lors de l'analyse (voir partie 6) déterminent le centre des fenêtres de découpage. Chaque fenêtre est définie sur une longueur égale à 2 périodes fondamentales locales (du signal original ou du signal de synthèse voulu).

$$s_i(t) = h_i(\mathrm{m}_i - t)s(t) \tag{7.1}$$

$$h_i = h\left(\frac{t}{2\mathrm{T0}(\mathrm{m}_i)}\right) \tag{7.2}$$

7.2.1.1 Positionnement des marques de découpage

Les marques m_i sont issues de l'algorithme de satisfaction de contraintes périodicité/énergie (voir partie 6)

7.2.1.2 Fenêtre de découpage du signal

Choix du type de la fenêtre : La fenêtre de pondération utilisée est une fenêtre de type Hann. Ceci pour les raisons suivantes :

1. Nous souhaitons une fenêtre telle que, en l'absence de modification du signal, la somme des contributions des fenêtres soit égale à 1 [1] .

$$\sum_i h_i(n - \mathrm{m}_i) = 1 \, \forall n \tag{7.3}$$

   Pour un facteur de recouvrement de 50 % entre fenêtres adjacentes (cas du PSOLA-WB), cette condition est satisfaite par les fenêtres de Hann et triangulaires.
2. Nous souhaitons - et ceci est un corollaire de notre algorithme de marquage reposant sur un critère de concentration d'énergie -, une fenêtre conservant au maximum l'énergie locale du signal.

Symétrie/Dissymétrie de la fenêtre : Puisque les variations de fréquence fondamentale peuvent être rapides (en particulier dans le cas de la voix), la question se pose quant au choix entre

- une fenêtre symétrique (i.e. définie sur l'intervalle $[\mathrm{m}_i - \mathrm{T0}_i, \mathrm{m}_i + \mathrm{T0}_i]$) ou
- une fenêtre dissymétrique (i.e. définie sur l'intervalle $([\mathrm{m}_i - \mathrm{T0}_{i-1}, \mathrm{m}_i + \mathrm{T0}_i]$ et de valeur maximale en m_i [2]).

En l'absence de modifications du signal, le choix d'une fenêtre dissymétrique présente l'avantage de permettre la reconstruction exacte du signal original. En présence de modifications dans le domaine fréquentiel, le choix d'une fenêtre symétrique permet de garantir la nullité du spectre de phase de la fenêtre. Notre choix se porte sur une fenêtre symétrique.

Choix de la longueur de la fenêtre : La longueur de la fenêtre de découpage peut être prise proportionnelle

- à la période du signal original
- à la période du signal de synthèse

Ce dernier choix permet de garantir un facteur de recouvrement constant égal à un pour le signal de synthèse, et donc ne nécessite pas de normalisation.

Dans [HMC89], la fenêtre est prise proportionnelle à la période du signal original lors d'une diminution de hauteur, à la période du signal de synthèse lors d'une augmentation de hauteur. Nous adoptons également ce choix.

7.2.2 Modification des formes d'onde élémentaires

Avant reconstruction du signal, chaque forme d'onde élémentaire peut subir certaines modifications dans le domaine temporel ou fréquentiel.

$$\tilde{s}_j(t) = F(s_i(t + \mathrm{m}_i)) \tag{7.4}$$

Les modifications étudiées dans le cadre de cette recherche sont : l'interpolation temporelle entre forme d'onde élémentaire adjacentes (TDI-PSOLA), l'interpolation fréquentielle (FDI-PSOLA), la dilatation spectrale (FD-PSOLA), la transposition spectrale (FS-PSOLA), et le traitement des régions fréquentielles non-harmoniques (VUV-PSOLA).

7.2.3 Reconstruction du signal

Le signal modifié (signal de synthèse) est construit par superposition/addition des formes d'onde élémentaires placées en de nouvelles positions $\tilde{\mathrm{m}}_j$ appelées marques d'écriture. Ces marques d'écriture sont déterminées par les modifications voulues du signal : modification de hauteur, modification du déroulement de l'axe temporel

$$\tilde{s}(t) = \sum_j \tilde{s}_j(t - \tilde{\mathrm{m}}_j) \tag{7.5}$$

7.2.3.1 Positionnement des marques d'écriture

Les marques d'écriture $\tilde{\mathrm{m}}_j$ déterminent les périodes fondamentales du signal de synthèse. Leur positionnement dépend donc de la fréquence fondamentale $f(t)$ voulue pour le signal de synthèse (où t est le temps référencé par rapport au signal original).

Nous introduisons un temps dit de «correspondance» : $\hat{\mathrm{co}}_j$. Celui-ci détermine le temps sur le signal original correspondant à $\tilde{\mathrm{m}}_j$ (voir FIG. 7.1). $\hat{\mathrm{co}}_j$ dépend non seulement de la fréquence fondamentale voulue pour le signal de synthèse $f(t)$, mais également de la modification voulue de l'axe temporel. Soit $\beta(t)$ la fonction de déroulement de l'échelle temporelle (correspondance entre l'axe temporel du signal original et l'axe temporel du signal de synthèse voulu).

$$\beta(t) = \int_0^t D(\tau)d\tau \tag{7.6}$$

dans lequel $D(t)$ désigne le facteur de dilatation.

Nous pouvons maintenant écrire

$$\tilde{\mathrm{m}}_{j+1} = \tilde{\mathrm{m}}_j + \frac{1}{\hat{\mathrm{co}}_{j+1} - \hat{\mathrm{co}}_j} \int_{\hat{\mathrm{co}}_j}^{\hat{\mathrm{co}}_{j+1}} \frac{1}{f(t)}\, dt \tag{7.7}$$

$$\hat{\mathrm{co}}_{j+1} = \hat{\mathrm{co}}_j + \frac{1}{\hat{\mathrm{co}}_{j+1} - \hat{\mathrm{co}}_j} \int_{\hat{\mathrm{co}}_j}^{\hat{\mathrm{co}}_{j+1}} \frac{1}{f(t) \cdot D(t)}\, dt \tag{7.8}$$

Le calcul de $\tilde{\mathrm{m}}_{j+1}$ nécessite la connaissance de $\hat{\mathrm{co}}_{j+1}$. De même, le calcul de $\hat{\mathrm{co}}_{j+1}$ nécessite la connaissance des bornes de l'intégration, c'est-à-dire également de $\hat{\mathrm{co}}_{j+1}$. Deux solutions se présentent :

Approximation locale constante : Nous considérons $f(t)$ et $D(t)$ comme variant lentement sur l'intervalle $[\hat{\mathrm{co}}_j, \hat{\mathrm{co}}_{j+1}]$. Dans ce cas, nous pouvons approximer leur valeur sur l'intervalle $[\hat{\mathrm{co}}_j, \hat{\mathrm{co}}_{j+1}]$ par une valeur constante $f(\hat{\mathrm{co}}_j)$ et $D(\hat{\mathrm{co}}_j)$. (7.7) et (7.8) se réécrivent

$$\tilde{\mathrm{m}}_{j+1} = \tilde{\mathrm{m}}_j + \frac{1}{f(\hat{\mathrm{co}}_j)} \qquad \hat{\mathrm{co}}_{j+1} = \hat{\mathrm{co}}_j + \frac{1}{f(\hat{\mathrm{co}}_j) \cdot D(\hat{\mathrm{co}}_j)} \tag{7.9}$$

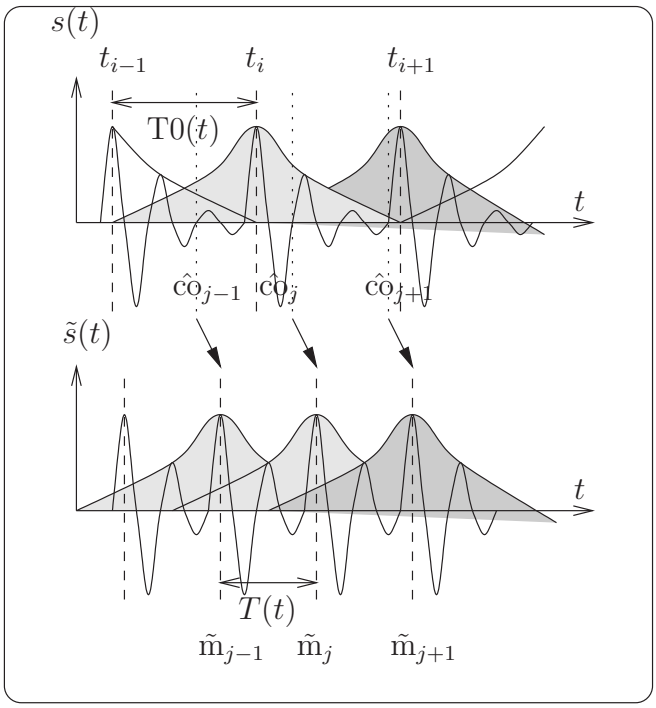


FIG. 7.1 – Positionnement des marques d'écriture $\tilde{m}_j$ et des temps de correspondance $\hat{t}_j$

Algorithme itératif [Sty96] (voir annexe H) : Nous procédons de manière itérative. Partant d'une première approximation de $\hat{co}_{j+1}$ (par exemple celle donnée par l'approximation constante $f(\hat{co}_j)$ et $D(\hat{co}_j)$), nous calculons la valeur de l'intégrale (7.8) ce qui nous donne une nouvelle approximation de $\hat{co}_{j+1}$ qui est alors utilisée pour une nouvelle évaluation de l'intégrale (7.8). L'itération est arrêtée quand $\hat{co}_{j+1}$ ne change plus entre deux itérations.

Le choix entre les deux algorithmes s'effectue selon les modifications voulues du signal : pour un facteur de compression élevé (dans ce cas un intervalle de temps important peut séparer deux $\hat{co}_j$ et donc la fréquence fondamentale est susceptible de varier de manière non-négligeable entre deux $\hat{co}_j$) l'algorithme itératif est utilisé. Pour le reste des transformations, l'approximation locale constante donne des résultats satisfaisants.

En annexe I, nous proposons une méthode de correction du spectre de phase des formes d'onde élémentaires permettant, dans le cas de fréquences d'échantillonnage basses, de tenir compte de la troncature des positions des marques obtenues en temps continu et appliquées à un signal en temps discret.

7.2.3.2 Sélection des formes d'onde élémentaires

Dans la méthode PSOLA standard, aucune modification n'est apportée aux formes d'onde élémentaires

$$\tilde{s}_j(t) = s_l(t + m_i) \tag{7.10}$$

La forme d'onde élémentaire choisie est celle dont le m_i est le plus proche de $\hat{\mathrm{co}}_j$

$$l = \arg\min_i |\hat{\mathrm{co}}_j - \mathrm{m}_i| \tag{7.11}$$

Un allongement du signal, ou une augmentation de hauteur sans modification de l'axe temporel, est obtenu par duplication d'une même forme d'onde élémentaire (partie gauche de la Fig. 7.2). Un raccourcissement du signal, ou un abaissement de la hauteur sans modification de l'axe temporel, est obtenu par élimination de certaine formes d'onde élémentaires (partie droite de la Fig. 7.2). Nous verrons dans la partie suivante des améliorations pouvant être apportées à cet algorithme.

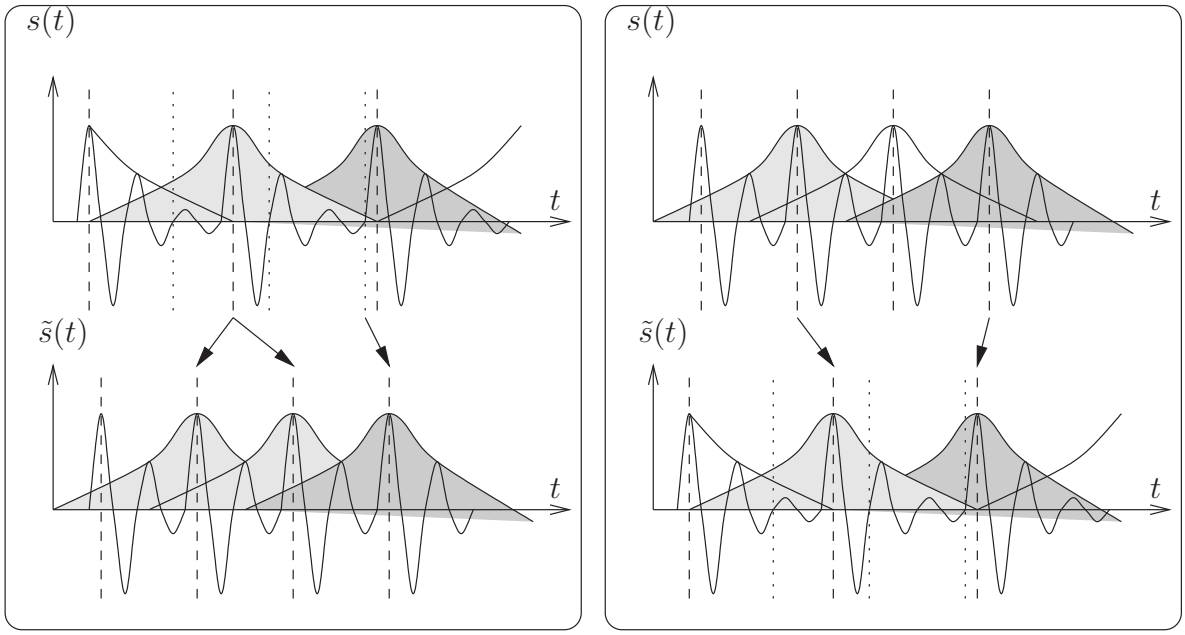


Fig. 7.2 – [G] Augmentation de f0 : duplication de forme d'onde élémentaire , [D] Diminution de f0 : élimination de forme d'onde élémentaire

7.2.3.3 Addition/Recouvrement

L'étape de superposition/addition est suivie par une procédure de normalisation, de manière à tenir compte du facteur de recouvrement éventuellement variable au cours du temps. Les deux types de normalisation utilisées sont :

Méthode de Allen [All77] : Il s'agit de la procédure normale de normalisation faisant suite à une méthode de superposition/addition

$$\tilde{s}(n) = \frac{\sum_j s_j(n)}{\sum_j f_j(\mathrm{m}_j - n)} \tag{7.12}$$

dans lequel $s_j(n)$ désigne les formes d'onde élémentaires successives du signal de synthèse, et $f_j(n)$ la fenêtre de synthèse utilisée.

Méthode de Griffin [GL84] : La méthode de Griffin est obtenue par minimisation de l'erreur énergétique entre spectre du signal original et spectre modifié correspondant à y_j. Cette méthode n'est utile que lorsque des modifications fréquentielles sont appliquées.

$$\tilde{s}(n) = \frac{\sum_m f_j(\mathrm{m}_j - n) y_j(n)}{\sum_m f_j^2(\mathrm{m}_j - n)} \tag{7.13}$$

Nous n'appliquons pas l'algorithme de normalisation du signal pour les facteurs de transposition inférieurs à 0.8 (abaissement de la fréquence fondamentale). Ceci pour des raisons de déformation trop importante des forme d'onde élémentaire censées représenter la RI du filtre du système.

7.2.3.4 Traitement PSOLA sur le signal résiduel

Le diagramme de la méthode LP-PSOLA que nous avons utilisée est illustré à la Fig. 7.3. L'estimation des filtres est effectuée de manière synchrone à la période fondamentale autour des marques m_i à l'aide de la méthode de Burg [Kay88]. L'ordre du modèle est choisi de manière à obtenir deux pôles par bande de 1000 Hz. Le traitement PSOLA est effectué sur le signal résiduel ; le signal transformé est ensuite re-filtré. Ce filtrage s'effectue de manière synchrone autour des instants $\tilde{\mathrm{m}}_j$. Les filtres sont obtenus par interpolation des coefficients log-area ratio ($\frac{1-k_i}{1+k_i}$), de manière à garantir la stabilité des filtres

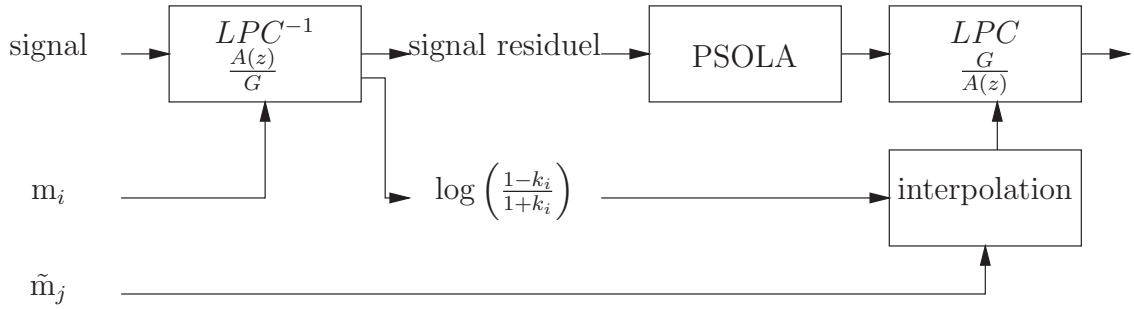


Fig. 7.3 – Diagramme de la méthode LP-PSOLA

7.3 Améliorations de la synthèse PSOLA

7.3.1 Interpolation des formes d'onde élémentaires

Lorsque la modification de l'axe temporel du signal ou de sa hauteur devient trop importante, l'algorithme de duplication/élimination introduit des discontinuités dans le signal :

- dans le cas de la duplication, par la reproduction d'une même forme d'onde puis passage à une autre,
- dans le cas de l'élimination, par le passage d'une forme d'onde à une autre forme d'onde distante donc a priori non similaire.

Ceci se traduit perceptivement par une certaine «rugosité» du signal de synthèse.

Une solution face à cela est de créer de nouvelles formes d'onde élémentaires réduisant ces discontinuités par «lissage». Ces formes d'onde élémentaires nouvelles sont obtenues par interpolation des formes d'onde associées aux marques m_i et m_{i+1} voisines de $\hat{\mathrm{co}}_j$. L'interpolation peut s'effectuer

- soit dans le domaine temporel (Time Domain Interpolation PSOLA ou TDI-PSOLA) :
- soit dans le domaine fréquentiel (Frequency Domain Interpolation PSOLA ou FDI-PSOLA) :

Dans les deux cas, le signal de synthèse est formé par superposition/addition des formes d'onde élémentaires interpolées $\tilde{s}_j(t)$: $\tilde{s}(t) = \sum_j \tilde{s}_j(t - \tilde{\mathrm{m}}_j)$.

7.3.1.1 Interpolation temporelle des formes d'onde élémentaires (TDI-PSOLA)

La forme d'onde élémentaire $\tilde{s}_j(t)$ est obtenue par interpolation temporelle des formes d'onde élémentaires $s_i(t)$ les plus proches de $\hat{\mathrm{co}}_j$ (voir Fig. 7.5 [M]).

$$\begin{cases} l = \arg\min_i |\hat{\mathrm{co}}_j - \mathrm{m}_i| & \text{tel que } \hat{\mathrm{co}}_j > \mathrm{m}_i \\ l' = \arg\min_i |\hat{\mathrm{co}}_j - \mathrm{m}_i| & \text{tel que } \mathrm{m}_i > \hat{\mathrm{co}}_j \end{cases} \tag{7.14}$$

$$\begin{cases} \tilde{s}_j(t) = (1-\alpha) s_l(t + \mathrm{m}_l) + \alpha s_{l'}(t + \mathrm{m}_{l'}) \\ \alpha = \dfrac{\hat{\mathrm{co}}_j - \mathrm{m}_l}{\mathrm{m}_{l'} - \mathrm{m}_l} \\ \tilde{s}(t) = \displaystyle\sum_j \tilde{s}_j(t - \tilde{\mathrm{m}}_j) \end{cases} \tag{7.15}$$

dans lequel α est le coefficient d'interpolation et m_l, $\mathrm{m}_{l'}$ les marques de lecture entourant $\hat{\mathrm{co}}_j$

TDI-PSOLA permet d'éviter les discontinuités du PSOLA standard. Cependant, du fait d'une interpolation dans le domaine temporel de formes d'onde élémentaires proches mais non nécessairement alignées (soit du fait d'un mauvais positionnement des m_i, soit du fait de l'erreur de troncature des m_i vers leur équivalent en échantillon), l'algorithme TDI-PSOLA peut produire l'effet d'un filtrage passe-bas des formes d'onde élémentaires [3] .

7.3.1.2 Interpolation fréquentielle des formes d'onde élémentaires (FDI-PSOLA)

La forme d'onde élémentaire $\tilde{s}_j(t)$ est obtenue par interpolation des spectres d'amplitude et de phase [4] des formes d'onde élémentaires $s_i(t)$ les plus proches de $\hat{\text{co}}_j$ (voir FIG. 7.5 [B]).

$$\begin{cases} l = \arg\min_i |\hat{\text{co}}_j - \text{m}_i| & \text{tel que } \hat{\text{co}}_j > \text{m}_i \\ l' = \arg\min_i |\hat{\text{co}}_j - \text{m}_i| & \text{tel que } \text{m}_i > \hat{\text{co}}_j \end{cases} \tag{7.17}$$

Soient, en notant $S_l(\omega) = A_l(\omega)e^{j\phi_l(\omega)}$ et $S_{l'}(\omega) = A_{l'}(\omega)e^{j\phi_{l'}(\omega)}$, les TF de $s_l(t + \text{m}_l)$ et $s_{l'}(t + \text{m}_{l'})$:

$$\forall\, k \in [0, N] \qquad \begin{cases} \tilde{A}_j(k) = (1-\alpha)A_l(k) + \alpha A_{l'}(k) \\ \tilde{\phi}_j(k) = (1-\alpha)\phi_l(k) + \alpha\phi_{l'}(k) \\ \alpha = \dfrac{\hat{\text{co}}_j - \text{m}_l}{\text{m}_{l'} - t_l} \end{cases} \tag{7.18}$$

La forme d'onde élémentaire interpolée $\tilde{s}(t)$ est obtenue par TF inverse des spectres interpolés $\tilde{S}_j(\omega) = \tilde{A}_j(\omega)e^{j\tilde{\phi}_j(k)}$. Une attention particulière doit être portée afin d'éviter le repliement temporel du signal lors de la TF inverse. L'utilisation d'un facteur de prolongement par zéro important permet cela. Une fenêtre de synthèse est alors appliquée et la normalisation du signal $\tilde{s}(t)$ utilise la formule (7.13) de minimisation de l'erreur de Griffin.

◊ *Interpolation des spectres d'amplitude*

Les spectres d'amplitude des formes d'onde élémentaires $s_i(t)$ constituent une approximation de l'enveloppe spectrale du signal. L'enveloppe spectrale est supposée varier lentement sur la durée d'une période fondamentale. De ce fait, les spectres d'amplitude A_l et $A_{l'}$ sont supposés proches (au sens d'une distance spectrale).

$$\tilde{A}_j(k) = (1-\alpha)A_l(k) + \alpha A_{l'}(k) \tag{7.19}$$

◊ *Interpolation des spectres de phase*

A l'inverse du spectre d'amplitude, le spectre de phase ne peut être considéré comme un signal à variation lente. Même dans le cas de deux formes d'onde élémentaires quasi-similaires, le positionnement des marques par rapport à chacune de ces forme d'onde élémentaire peut conduire à des spectres de phase très différents. [5] Rajoutons à ce problème le fait que la phase n'est définie qu'en valeur principale ($[-\pi, \pi]$). Une interpolation aveugle des spectres de phase peut conduire à la création d'une forme d'onde élémentaire interpolée $\tilde{s}_j(t)$ très différente de ses voisines [6], produisant une discontinuité dans le signal de synthèse, au lieu de l'effet de «lissage» recherché.

Afin de simplifier le problème, nous faisons dans la suite l'hypothèse que les spectres de phase sont quasi-similaires à une composante linéaire τ près, composante due au décalage de l'observation des formes d'onde élémentaires . L'algorithme que nous proposons repose sur une re-synchronisation des spectres de phase avant interpolation.

L'algorithme proposé est le suivant :

1. Déroulement des deux spectres de phase $\phi(\mathrm{m}_l)$ et $\phi(\mathrm{m}_{l'})$ de manière à minimiser la discontinuité (retard de groupe) fréquentielle des spectres $|\phi(\omega_k) - \phi(\omega_{k+1})| \leq \pi \ \ \forall k$
2. Synchronisation :
   L'indice de synchronisation utilisé est la pente moyenne Δ (calculée par régression linéaire) du spectre de phase sur l'intervalle $\omega = [-\pi, \pi]$ (base de la méthode GDP [SY95] étudiée dans la partie 3.4).
   Cet indice est calculé pour les deux formes d'onde élémentaires et sert au ré-alignement des deux formes d'onde élémentaires l'une par rapport à l'autre. La correction apportée au spectre de chaque forme d'onde élémentaire est proportionnelle au coefficient d'interpolation. Ceci permet d'éviter une discontinuité aux bords des couples de formes d'onde élémentaires interpolées [7] .
   Soient Δ_1 et Δ_2 les indices de synchronisation associés aux formes d'onde élémentaires $s_l(t)$ et $s_{l'}(t)$. La correction apportée à leur spectre de phase est de
   $$\begin{aligned} \hat{\phi}(\omega_k, \mathrm{m}_l) &= \phi(\omega_k, \mathrm{m}_l) \cdot \alpha \cdot \omega_k(\Delta_1 - \Delta_2) \\ \hat{\phi}(\omega_k, \mathrm{m}_{l'}) &= \phi(\omega_k, \mathrm{m}_{l'}) \cdot (1-\alpha) \cdot \omega_k(\Delta_2 - \Delta_1) \end{aligned} \tag{7.20}$$
3. Déroulement relatifs des spectres de phase
   Même après re-synchronisation des spectres de phase, rien ne garantit que $\hat{\phi}(\omega_k, \mathrm{m}_l)$ et $\hat{\phi}(\omega_k, \mathrm{m}_{l'})$ soient définis dans le même domaine en chaque fréquence. Un déroulement relatif de la phase de $\hat{\phi}(\omega_k, \mathrm{m}_{l'})$ par rapport à $\hat{\phi}(\omega_k, \mathrm{m}_l)$ est effectué de manière à minimiser $|\hat{\phi}(\omega_k, \mathrm{m}_l) - \hat{\phi}(\omega_k, \mathrm{m}_{l'})|$ en chaque fréquence.
4. Interpolation
   $$\tilde{\phi}_j(k) = (1-\alpha)\hat{\phi}_l(k) + \alpha\hat{\phi}_{l'}(k) \tag{7.21}$$

7.3.1.3 Comparaison des méthodes TD-PSOLA, TDI-PSOLA, FDI-PSOLA

Lorsque le facteur de modification de l'échelle temporelle devient important, la différence de qualité du signal de synthèse entre les méthodes TD-PSOLA et les méthodes d'interpolation TDI et FDI-PSOLA devient très perceptible. La première procédant par duplication de forme d'onde produit un signal variant par pallier perceptivement «rugueux».

La différence entre les méthodes TDI et FDI-PSOLA est plus fine. Elle se situe au niveau d'un filtrage fréquentiel dû à l'addition en TDI-PSOLA de composantes non nécessairement en phase. Ce phénomène de filtrage est évité en FDI-PSOLA du fait de l'interpolation sur les spectres d'amplitude et de phase et du fait du ré-alignement.

A la FIG. 7.4, nous illustrons dans le plan des phases l'interpolation temporelle et fréquentielle dans le cas simple de deux composantes fréquentielles S_l et $S_{l'}$ en opposition de phase. Sur ces diagrammes, il est aisé de comprendre que l'interpolation temporelle (avec un facteur $\alpha = 0.5$) de ces composantes produit une composante $\hat{S}_j$ d'amplitude quasi nulle, alors que l'interpolation fréquentielle résulte en une composante d'amplitude moyenne et de signe inversé.

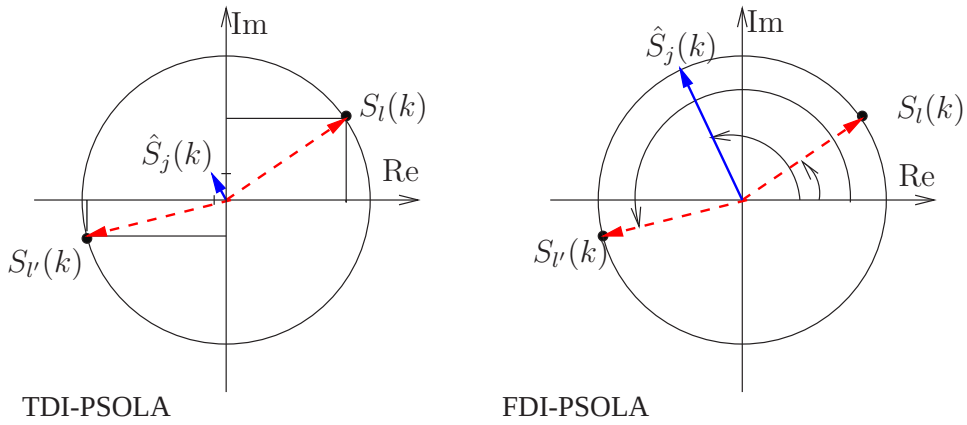


FIG. 7.4 – Comparaison dans le plan des phases de l'interpolation temporelle (TDI-PSOLA) et fréquentielle (FDI-PSOLA)

A la FIG. 7.6, nous comparons les spectres d'amplitude du signal original, du même signal traité par l'algorithme TDI-PSOLA et par l'algorithme FDI-PSOLA (facteur de dilatation de 30) . Le filtrage introduit par l'algorithme TDI-PSOLA apparaît clairement sur les composantes autour de 500 Hz, 1000 Hz et 3500 Hz.

Tout ceci nuance quelque peu [CM88], qui conclut à une différence perceptive faible.

7.3.2 Modification du spectre des formes d'onde élémentaires

Afin d'étendre les modifications du signal possibles par la méthode PSOLA-WB, nous étudions ici deux types de modifications spectrales des formes d'onde élémentaires .

7.3.2.1 Dilatation/Compression du spectre des formes d'onde élémentaires (FD-PSOLA)

L'algorithme FD-PSOLA [CS86] a été initialement proposé dans le cadre de la synthèse PSOLA-NB afin de corriger (mettre en adéquation) le spectre des formes d'onde élémentaires (spectre à structure fine, harmoniques résolues) vis-à-vis d'une modification voulue de hauteur du signal.

En PSOLA-WB, chacune des formes d'onde élémentaires $s_i(t)$ constitue une approximation de l'enveloppe spectrale autour de l'instant m_i. Nous utilisons l'algorithme FD-PSOLA pour modifier cette enveloppe spectrale et permettre un changement de timbre du signal.

Deux méthodes FD-PSOLA ont été initialement proposées : une méthode de dilatation/compression du spectre, une méthode d'élimination/répétition des régions du spectre. Pour FDI-PSOLA, nous avons seulement appliqué la méthode de dilatation/compression du spectre. La dilatation/compression du spectre d'une forme d'onde élémentaire s'obtient dans notre algorithme par ré-échantillonnage du spectre d'une forme d'onde élémentaire . Deux implémentations du ré-échantillonnage ont été testées : l'interpolation linéaire, et l'interpolation par filtrage à bande limitée (sinus cardinal ; voir annexe F) avec des résultats comparables mais un coût de calcul beaucoup plus élevé pour la seconde méthode.

7.3.2.2 Déplacement du spectre des formes d'onde élémentaires (FS-PSOLA)

Un deuxième type de modification spectrale a été étudié : le déplacement spectral. Cet effet, généralement connu sous le nom de modulateur en anneau, consiste à transposer le spectre d'un signal de manière uniforme en fréquence.

$$x(t) \cdot e^{j\Omega t} \rightleftharpoons X(\omega - \Omega) \tag{7.22}$$

Ceci a pour conséquence, lorsqu'il est appliqué dans des conditions d'analyse à bande étroite, de rendre inharmonique un signal initialement harmonique (voir FIG. 7.7 partie du haut).

A l'inverse, lorsque (7.22) est appliqué à chacune des formes d'onde élémentaires dans des conditions d'analyse à bande large, (7.22) décale le spectre d'un facteur Ω en fréquence. De par l'interprétation en termes de ré-échantillonnage aux multiples de hf0 de PSOLA-WB, ceci ne modifie pas le caractère harmonique du signal (voir FIG. 7.7 partie du bas).

Le déplacement spectral de la forme d'onde élémentaire $s_i(t)$ est obtenu par

$$\tilde{s}_j(t) = \Re\left\{[s_a(t) \cdot h_{L_i}(t - \mathrm{m}_i)] \exp(-j\Omega t)\right\} \tag{7.23}$$

dans lequel $s_a(t)$ est le signal analytique correspondant à $s_i(t)$, L_i est la taille de $s_i(t)$ et Ω est le facteur de transposition.

Selon le signe de la transposition, un filtrage passe-bas ou passe-haut est appliqué afin de prévenir tout repliement fréquentiel. $\tilde{s}_j(t)$ est ensuite traité par l'algorithme PSOLA-WB en vue d'obtenir les modifications de hauteur et de durée voulues.

Dans le cas particulier où nous désirons faire coïncider la modification d'enveloppe spectrale et celle de hauteur (par exemple en vue de garder la correspondance entre le premier formant et la fréquence fondamentale [8]), nous effectuons le choix $\Omega = \omega - \omega_0$ et $L = 2\max\{T, T_0\}$; dans lequel ω (T) est la période fondamentale requise et ω_0 (T_0) la période fondamentale du signal original. Dans le cas $T < T_0$, une fenêtre de synthèse de taille $L = 2T$ est utilisée.

7.3.3 Traitement des régions non-périodiques

L'algorithme PSOLA tel que présenté jusqu'à présent s'applique uniquement aux régions périodiques du signal. Les deux autres types de signaux pris en compte par notre algorithme OLA sont les signaux de caractères «singularité non-périodique» et «non-singularité non-périodique non-sinusoïdale».

Singularités non-périodiques : Les singularités non-périodiques, appelées précédemment «transitoires» (voir partie 3.5), ne font l'objet d'aucun traitement spécifique dans la version actuelle des algorithmes. Les transitoires détectées sont fenêtrées et recopiées dans le signal de synthèse sans modification. L'algorithme prend en compte leur spécificité en empêchant :

- la réutilisation d'une forme d'onde élémentaire renfermant une transitoire (cas d'une dilatation de l'axe temporel),
- le saut d'une forme d'onde élémentaire renfermant une transitoire (cas d'une compression de l'axe temporel)

ceci de manière à préserver le naturel de la production et de la perception d'un signal sonore (les contractions/dilatations s'opérant en majorité dans les parties tenues du signal). Dans la version actuelle de l'algorithme, les transitoires ne font pas l'objet

d'un traitement dans le domaine fréquentiel. Si bien qu'une portion du signal renfermant une transitoire sera traitée de manière uniforme, quelle que soit la localisation fréquentielle du bruit transitoire et quel que soit le contenu du reste du spectre. Ce point est évidemment à améliorer.

Régions non-périodiques absentes de singularités : Les régions non-périodiques absentes de singularités, appelées communément régions «bruitées» ou régions «non-voisées» dans le cas de la parole, font l'objet d'un traitement spécifique.

Le traitement PSOLA des zones non-périodiques doit permettre la prise en compte des spécificités de ces régions tout en autorisant la transformation du signal. La transformation considérée ici (qui est la transformation la plus communément utilisée) est la dilatation des régions non-périodiques. Dans ce cas, l'algorithme de dilatation doit permettre l'allongement/raccourcissement de ces régions sans l'introduction de périodicité artificielle qui résulterait de l'application de l'algorithme PSOLA tel qu'étudié précédemment.

La caractérisation dans le temps (fonction $vois(t)$) et dans le plan temps/fréquence (fonction $vois(t, w_h)$), obtenue dans la partie 5, est utilisée dans deux algorithmes différents.

7.3.3.1 Traitement dans le domaine temporel

Dans ([CM89]), une méthode de dilatation du bruit est proposée reposant sur l'inversion alternative de l'axe du temps d'une forme d'onde lorsque cette forme d'onde est utilisée plusieurs fois successivement. Cette méthode permet de préserver le spectre d'amplitude du signal, tout en inversant le spectre de phase, empêchant ainsi (dans une certaine mesure) la corrélation des formes d'onde et donc l'apparition de l'effet «tunnel» (effet «flanger») dans le signal de synthèse. D'après [CM89], cette méthode permettrait d'obtenir des dilatations jusqu'à un facteur 2.

La méthode que nous proposons inclut la méthode précédente tout en permettant l'utilisation de facteurs de dilatation plus importants et en diminuant encore d'avantage l'effet tunnel (effet «flanger»).

Méthodes proposées :

- Décalage aléatoire de la forme d'onde élémentaire : la forme d'onde élémentaire utilisée pour la synthèse est une portion du signal décalée aléatoirement par rapport au marqueur m_i (voir FIG. 7.8). L'ordre de grandeur de ce décalage est de 5 % de la taille de la fenêtre.
- Inversion alternative de l'axe du temps lorsqu'une forme d'onde est réutilisée
- Interpolation temporelle (TDI-PSOLA) : les deux points précédents sont appliqués aux deux formes d'onde entourant $\hat{co}_j$. La forme d'onde $\hat{s}_j(t)$ utilisée dans le signal de synthèse résulte de l'interpolation temporelle de ces deux formes d'onde .

Positionnement des marques : Dans le cas d'une région temporelle dont la totalité des bandes de fréquences est considérée comme non-voisée, l'algorithme est appliqué sur la globalité du signal. Dans ce cas, les marques d'écriture $\tilde{m}_j$ sont placées de manière équidistante. La distance choisie est égale à la moyenne des périodes fondamentales contenues dans les régions périodiques avoisinantes, notée $\overline{T0}$. Les temps de «correspondance» $\hat{co}_j$ sont calculés

de la même manière que dans les régions périodiques, à la différence près que la période fondamentale est remplacée par $\overline{\text{T0}}$.

Dans le cas d'une région temporelle dont seule une partie des fréquences est considérée comme non-voisée (partie supérieure à la fréquence de coupure voisé/non-voisé), le signal est séparé en deux parties par filtrage. L'algorithme de dilatation des régions non-périodiques est appliqué à la partie non-voisée. Dans ce cas, afin de garder la correspondance des enveloppes d'énergie des parties périodique et non-périodique du signal [9], les marques $\tilde{\text{m}}_j$ et $\hat{\text{co}}_j$ utilisées dans la partie non-voisée sont identiques à celle de la partie voisée.

7.3.3.2 Traitement dans le domaine fréquentiel (VUV-PSOLA)

Le but de l'algorithme fréquentiel est de permettre directement (sans séparation fréquentielle des signaux) une prise en compte de l'aspect périodique/non-périodique dans l'algorithme PSOLA. La technique proposée (VUV-PSOLA) repose sur l'introduction d'une composante aléatoire dans le spectre de phase de chaque forme d'onde élémentaire . Dans le cadre de la synthèse sinusoïdale, une approche similaire est proposée par [MC97].

L'importance de cette composante aléatoire est proportionnelle à la non-périodicité du signal (voir FIG. 7.9). En notant $\phi_i(\omega_k)$ $\tilde{\phi}_j(\omega_k)$ les spectres de phase du signal original et modifié,

$$\tilde{\phi}_j(\omega_k) = \phi_i(\omega_k) + \alpha(\omega_k) \cdot randn \cdot \pi \tag{7.24}$$

dans lequel $randn$ est un générateur de nombres aléatoires selon une loi gaussienne de moyenne 0 et de variance 1. La valeur de $\alpha(\omega_k)$ est déterminée par le voisement à la fréquence ω_k : $\alpha(\omega_k) = 1 - v(\omega_k)$, dans lequel $v(k)$ est obtenu par interpolation de $vois(\omega_h)$.

Comme indiqué dans la partie 5, l'utilisation de $vois(\omega_h)$ est rendue difficile du fait de sa grande variabilité au cours du temps. Le traitement étudié dans la partie 5 consiste en l'estimation d'une fréquence de coupure séparant une région dite majoritairement voisée d'une région majoritairement non-voisée . Cette fréquence de coupure ne rend cependant pas compte de l'évolution fréquentielle du coefficient de voisement.

Dans la version finale de l'algorithme VUV-PSOLA, la fonction $v(k)$ est remplacée par une fonction affine par morceaux constituée des trois segments (voir FIG. 7.9) suivants :

- $[(\text{f} = 0, a), (f = f_c, b)]$,
- $[(f = f_c, b), (f = f_c + \Delta, c)]$ et
- $[(f = \text{f}_c + \Delta, c), (f = Fe/2, d)]$.

Le choix de $a = 1$, $b = 1$, $c = 0.5$ $d = 0$ nous a conduits à de bons résultats d'un point de vue perceptif.

La forme d'onde élémentaire $\tilde{s}_j(t)$ correspondant à $\tilde{\phi}_j(\omega_k)$ est finalement obtenue par IFFT après re-symétrisation du spectre complexe. Une fenêtre de synthèse est ensuite appliquée sur chaque forme d'onde élémentaire $\tilde{s}_j(t)$.

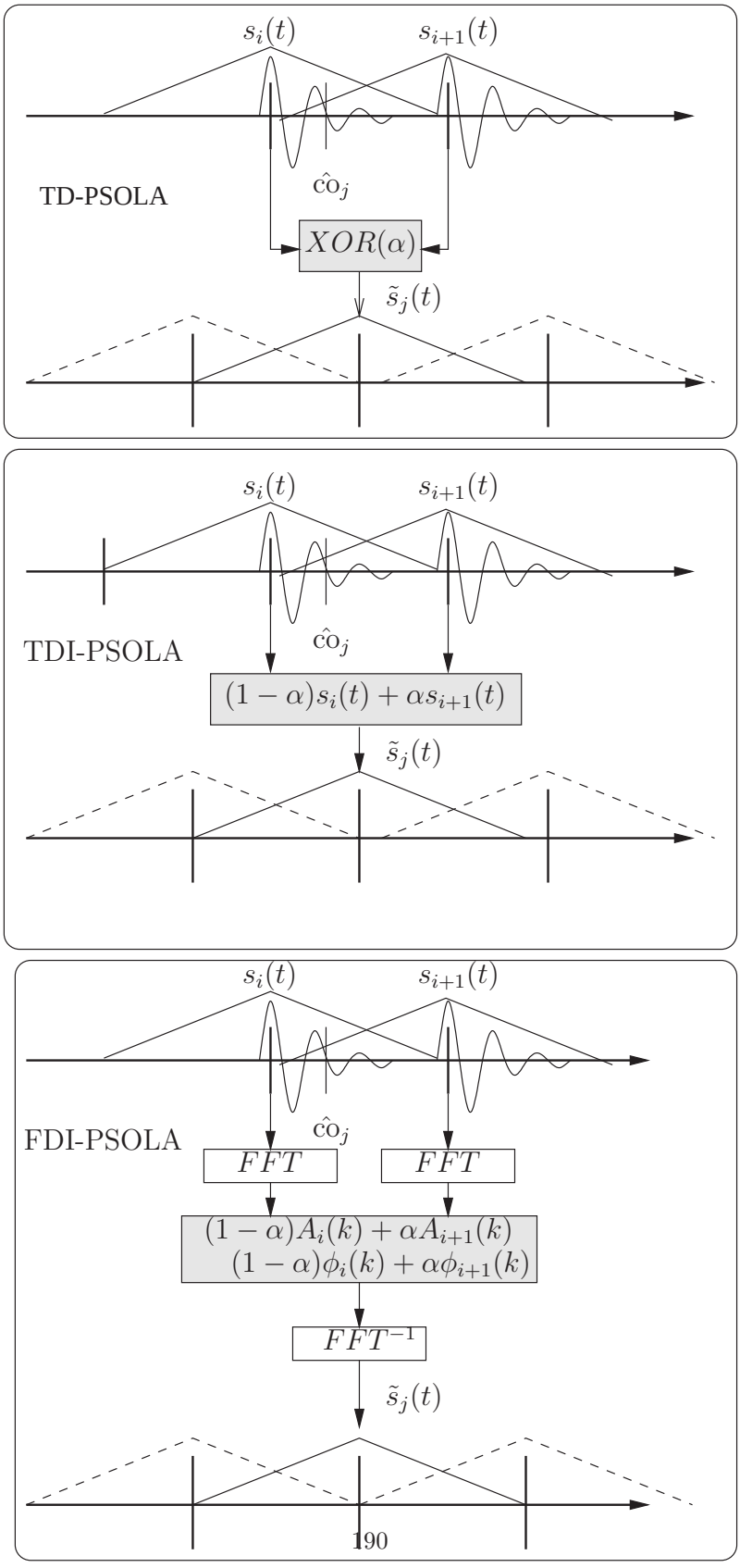


FIG. 7.5 – Algorithmes [H] TD-PSOLA, [M] TDI-PSOLA et [B] FDI-PSOLA

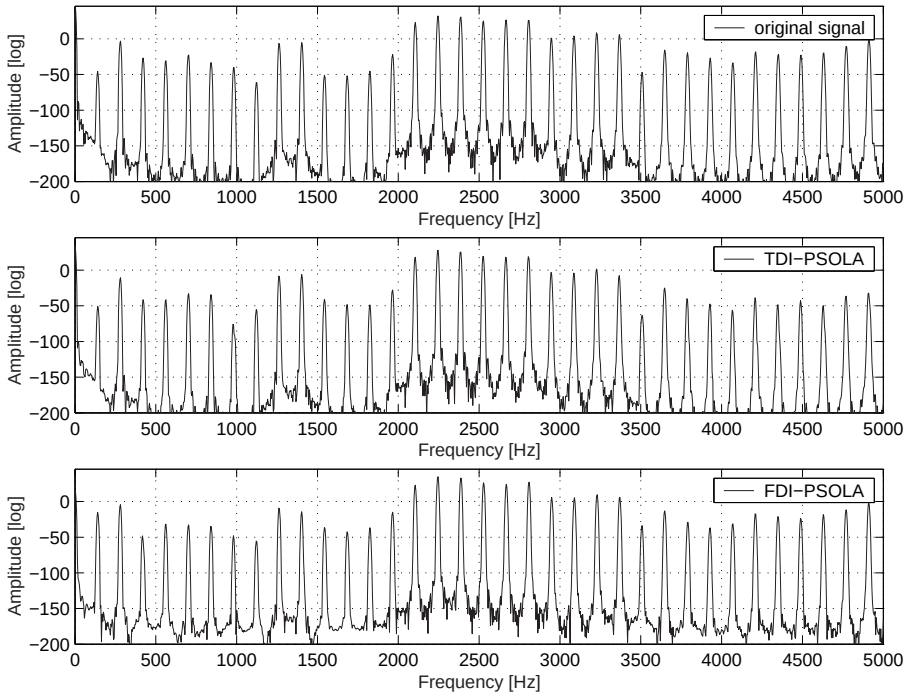


FIG. 7.6 – Comparaison TDI-PSOLA/FDI-PSOLA : [H] Spectre d'amplitude du signal original [M] Spectre d'amplitude du signal traité par l'algorithme TDI-PSOLA [B] Spectre d'amplitude du signal traité par l'algorithme FDI-PSOLA. Signal= speech

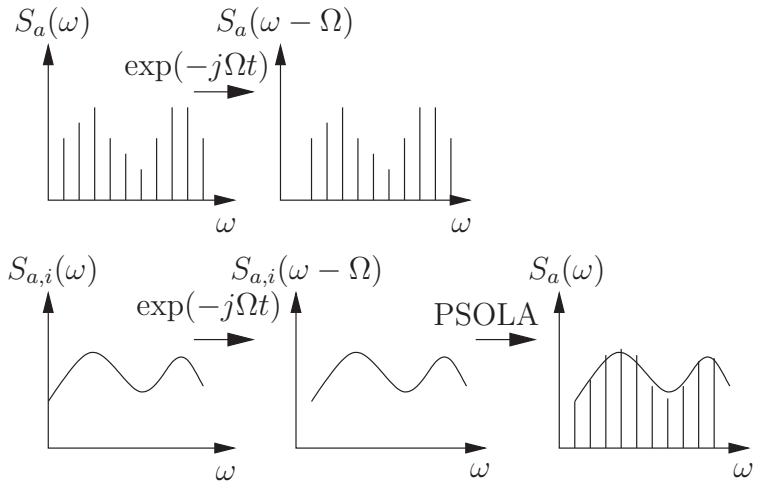


FIG. 7.7 – Algorithme FS-PSOLA : [H] transposition de fréquence dans des conditions d'analyse à bande étroite, [B] transposition de fréquence dans des conditions d'analyse à bande large suivie du ré-échantillonnage PSOLA-WB.

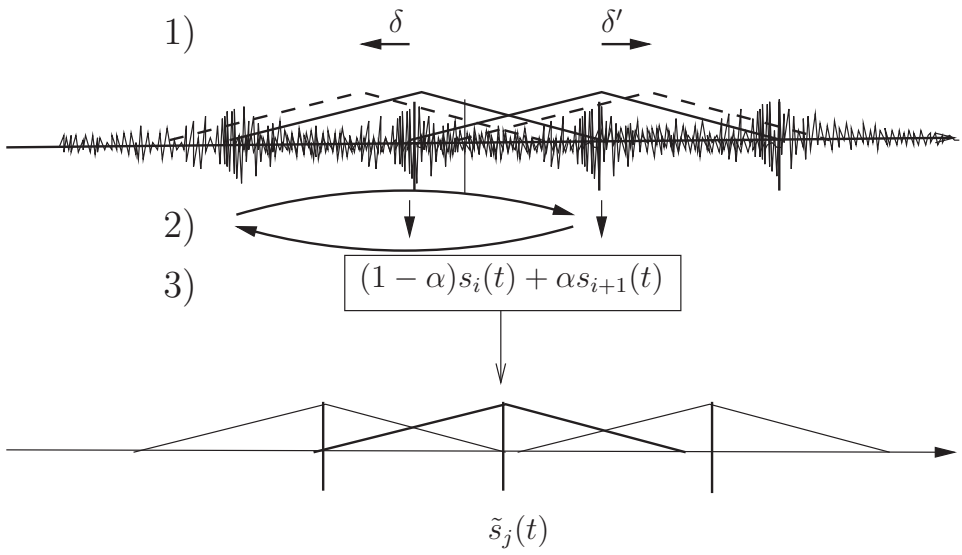


FIG. 7.8 – Algorithme temporel de traitement des régions non-périodiques : 1) décalage aléatoire, 2) inversion de l'axe du temps, 3) interpolation temporelle

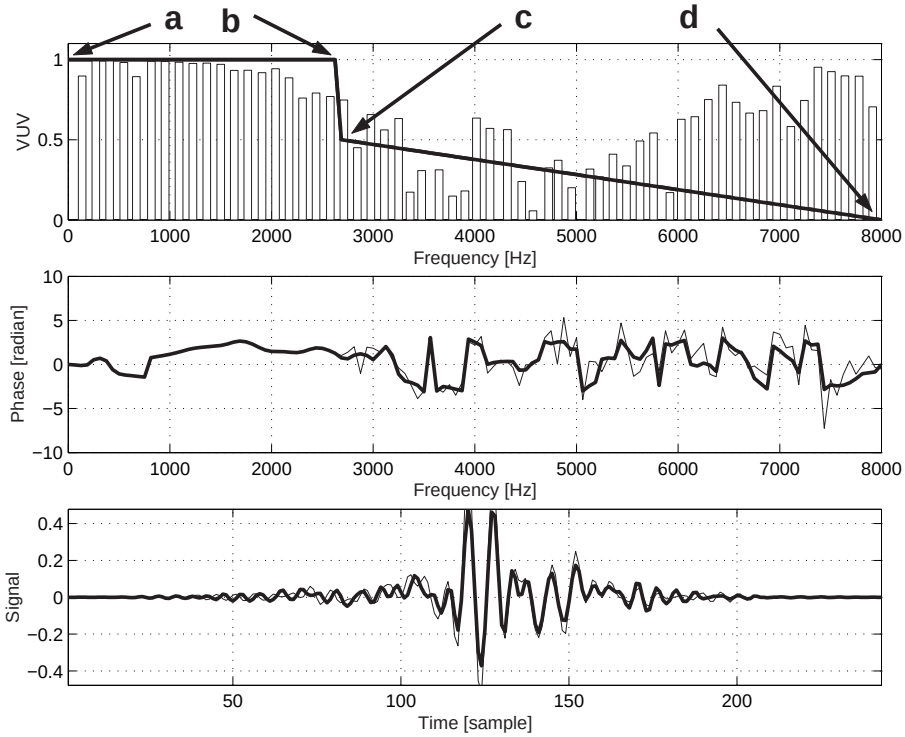


FIG. 7.9 – Algorithme fréquentiel de traitement des zones non-périodiques : [H] Fonction de voisement $vois(\omega_h)$, fréquence de coupure et fonction affine par morceau [M] spectre de phase de la forme d'onde élémentaire (trait gras), spectre de phase avec ajout de composante aléatoire en fonction du voisement (trait léger) [B] forme d'onde élémentaire originale (trait gras), forme d'onde élémentaire modifiée (trait léger)

7.4 Résumé

A la figure FIG. 7.10, nous avons représenté sous forme de diagramme les différentes étapes de l'étape de modification du signal par la méthode PSOLA.

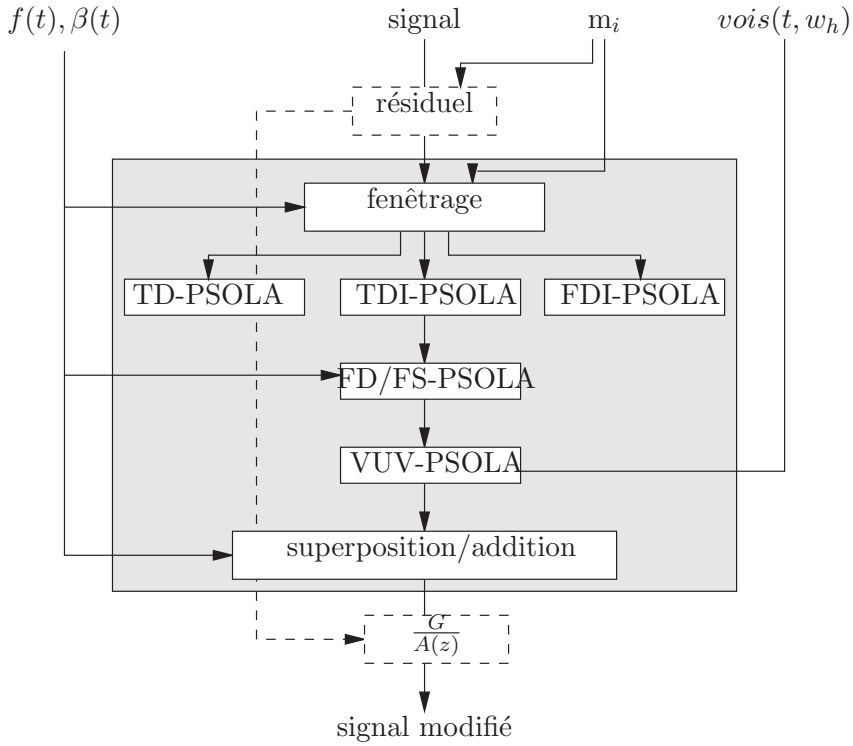


FIG. 7.10 – Diagramme de la partie modification du signal par l'algorithme PSOLA

Notes de bas de page relatives à la partie 7

1. Dans ce cas, il n'y a pas lieu de normaliser le signal puisque le dénominateur de normalisation de Allen (7.12) se réduit à 1
2. une telle fenêtre est obtenue par concaténation en m_i de deux demi-fenêtres de largeurs $T0_{i-1}$ et $T0_i$
3. Soit un signal constitué de la répétition d'une forme d'onde identique $x_{T_0}(t)$;
si $m_l = kT_0$ et $m_{l'} = (k+1)T_0 + \Delta$ désignent les deux marques de lecture des formes d'onde élémentaires à interpoler,
alors $\tilde{s}_j(t) = (1-\alpha)x_{T_0}(t) + \alpha x_{T_0}(t)e^{j\omega\Delta}$.
Dans le cas $\alpha = 0.5$ et $\Delta = 1$, nous retrouvons l'expression d'un filtrage passe-bas

$$\tilde{s}_j(t) = 0.5x_{T_0}(t)(1+e^{j\omega}) = 0.5x_{T_0}(t)(1+z^{-1}) \tag{7.16}$$

4. Il ne s'agit donc pas d'une interpolation sur les parties réelle et imaginaire du spectre, ce qui par linéarité de la Transformée de Fourier reviendrait à faire une interpolation temporelle
5. Ceci peut paraître étonnant pour le lecteur, étant donné que l'algorithme de marquage proposé repose précisément sur le spectre de phase. Rappelons néanmoins que cet algorithme de marquage sur la phase est suivi d'un algorithme de contrainte périodicité/énergie à la suite duquel nous ne pouvons plus garantir la similitude des spectres de phase des formes d'onde élémentaires successives. Pour s'en convaincre, considérons le cas d'un signal de fréquence fondamentale constante mais d'enveloppe spectrale variant au cours du temps. Dans ce cas il n'y a pas de raison apparente d'observer des spectres de phase identiques d'une période à la suivante. Un autre facteur expliquant le décalage des spectres de phase provient de l'erreur de troncature résultant du passage d'un marquage en temps continu à son équivalent en nombre d'échantillons du signal discret.
6. Un phénomène souvent rencontré lors de l'interpolation aveugle des spectres de phase est l'inversion de l'axe temporel de la forme d'onde élémentaire interpolée
7. Explication : Lorsque α passe de 0 à 1 (passage progressif de la forme d'onde élémentaire l à la forme d'onde élémentaire l'), la resynchronisation est d'abord prise en charge par l' (alignement de l' par rapport à l) puis progressivement par l (alignement de l par rapport à l'). De la sorte, lorsque le couple de formes d'onde élémentaires à interpoler change (passage du couple (l, l') au couple $(l', l")$), la forme d'onde élémentaire l' se trouve à sa position exacte dans le signal de synthèse tant à la fin de l'interpolation du couple (l, l') qu'au début de l'interpolation du couple $(l', l")$. Ceci est indispensable afin d'éviter des discontinuités dans le signal de synthèse.
8. On constate par exemple que dans le cas de la voyelle «i», un formant très accentué qui se situe aux alentours de la fréquence fondamentale. Une transposition du signal par PSOLA-WB sans correction du spectre résulte en une perte importante d'énergie de la fréquence fondamentale. L'algorithme FS-PSOLA peut être utilisé dans ce cas en vue de garder la localisation du premier formant proche de celle de la fréquence fondamentale.
9. La synchronie des enveloppes d'énergie entre les parties voisée et non-voisée est un paramètre perceptivement important, faute de quoi le signal résultant de la superposition des deux parties est perçu comme issu de sources séparées. Les signaux ne fusionnent plus [Sty96]

Chapitre 8

Modifications du signal en synthèse par addition de sinusoïdes

8.1 Introduction

Les principes de base du modèle sinusoïdal ont été décrit dans la partie 2.2.

L'estimation des paramètres d'un modèle sinusoïdal ainsi que la création de trajets sinusoïdaux ont été étudiés dans les parties 4 et 5. Ce sont ces estimations que nous utilisons ici dans le but de :

re-synthétiser le signal original à partir de sa modélisation en somme de sinusoïdes ; dans ce cas, les paramètres utilisés pour la synthèse sont ceux estimés lors de l'analyse ;

synthétiser un signal modifié : dans ce cas, les paramètres utilisés sont une modification de ceux estimés lors de l'analyse.

Les deux algorithmes de synthèse les plus communément utilisés pour la synthèse sinusoïdale sont la synthèse par :

superposition/addition : dans ce cas, le spectre correspondant au modèle sinusoïdal estimé est reconstitué à chaque instant et le signal reconstruit par IFFT et superposition/addition ([RD92] et [Dut93]) : cet algorithme extrêmement rapide ne permet cependant qu'un nombre limité de modifications du signal ;

banc d'oscillateurs : dans ce cas, chaque trajet de sinusoïdes est généré de manière indépendante au cours du temps par un oscillateur ; le signal de synthèse est obtenu par l'addition des différents trajets.

L'algorithme utilisé dans ce travail est du type banc d'oscillateurs.

8.2 Re-synthèse du signal original

Création de trajets continus de fréquence, d'amplitude et de phase : L'analyse sinusoïdale nous donne une estimation de l'amplitude $A_{h,m}$, de la fréquence $\omega_{h,m}$ et de

la phase $\phi_{h,m}$ de chacune des H sinusoïdes à chaque trame d'analyse t_m. L'algorithme de synthèse par banc d'oscillateurs nécessite la création de trajets continus («continu» désigne ici «à chaque échantillon» n par opposition à «à chaque trame» m) à partir de ces informations discrètes.

La création des trajets continus d'**amplitude** est généralement obtenue par interpolation linéaire entre les estimations en les instants t_m. Les **fréquences** étant liées aux **phases** par la relation $\omega_h(t) = \frac{\partial \phi_h(t)}{\partial t}$, les trajets continus de fréquence sont donc contraints par ceux de phase. Vis-à-vis de cela, différentes stratégies peuvent être adoptées :

Ignorance des phases : Les phases estimées lors de l'analyse ne sont pas prises en compte. Les phases utilisées pour le trajet sont obtenues par intégration temporelle des fréquences : $\phi_h(t) = \int_{\tau=t_m}^{t} \omega_h(\tau) d\tau + \phi(t_m)$ dans lequel $\omega_h(\tau)$ est obtenu par interpolation linéaire des estimations $\omega_h(t_m)/\omega_h(t_{m+1})$ en τ.

Fréquence instantanée : Les fréquences estimées lors de l'analyse ne sont pas prises en compte. Les fréquences utilisées pour le trajet sont les fréquences instantanées obtenues par différenciation des phases entre deux instants t_m et t_{m+1} : $\omega_h\left(\frac{t_m+t_{m+1}}{2}\right) = \frac{\phi_{h,m+1}-\phi_{h,m}}{t_{m+1}-t_m}$

Satisfaction des contraintes fréquence/phase : Dans ce cas, l'algorithme le plus communément utilisé est l'algorithme dit du polynôme cubique de phase [MQ86b]. Cet algorithme, rappelé en annexe O, cherche sur l'intervalle $[t_m, t_{m+1}]$ le trajet de phase satisfaisant les contraintes de fréquence et de phase aux extrémités de l'intervalle. D'autres algorithmes ont également été proposés, comme l'utilisation de polynômes quadratiques de phase estimés par B-Spline [DQ97].

Notre choix se porte sur l'algorithme du polynôme cubique de phase [MQ86b]. Ceci parce que nous cherchons une re-synthèse sinusoïdale pouvant être soustraite du signal, donc une re-synthèse tenant compte de la localisation temporelle de l'information. Le choix de l'algorithme de [MQ86b] parmi les autres algorithmes est également fait en raison de sa simplicité. Cette partie de notre méthode devra toutefois être perfectionnée dans le futur.

8.3 Modification du signal

Les modifications possibles en synthèse sinusoïdale sont nombreuses, du fait d'un contrôle individuel de chacun des paramètres : transformations classiques de type modification de l'axe temporel et transposition, transformations plus fines de type modification de l'espacement fréquentiel entre sinusoïdes (modification de l'harmonicité d'un son), tremolo, vibrato, ...

Nous nous intéressons ici aux deux types principaux de modifications - la modification de l'axe temporel et la transposition - et étudions les améliorations possibles de l'algorithme de synthèse dans ce cas.

Dans la suite, nous notons :

- β la fonction de correspondance entre temps sur le signal original t_m (m) et temps sur le signal de synthèse t'_m (m') ; nous notons également D la dérivée locale de $\beta(t)$ (appelée usuellement facteur de dilatation),
- T la fonction de correspondance (appelée usuellement facteur de transposition) entre fréquence sur le signal original $\omega_{h,m}$ et fréquence sur le signal de synthèse $\hat{\omega}_{h,m'}$.

Transposition : Une transposition du signal d'un facteur T est obtenue par multiplication

des fréquences $\omega_{h,m}$ par un facteur T.

$$\omega_{h,m} \rightarrow \hat{\omega}_{h,m'} = T \cdot \omega_{h,m} \tag{8.1}$$

Compression/dilatation : Une compression/dilatation du signal d'un facteur D est obtenu par une modification de l'intervalle temporel séparant deux instants t_m d'analyse :

$$t_m - t_{m-1} \rightarrow t'_m - t'_{m-1} = D \cdot (t_m - t_{m-1}) \tag{8.2}$$

Différentes stratégies sont adoptées pour la reconstruction des trajets de phase.

Ignorance des phases : Les phases sont obtenues par intégration des fréquences transposées sur l'intervalle de temps modifié.

$$\hat{\phi}_{h,m'} = \hat{\phi}_{h,m'-1} + D \cdot (t_m - t_{m-1}) \cdot T \cdot \frac{\omega_{h,m-1} + \omega_{h,m}}{2} \tag{8.3}$$

Fréquence instantanée : Cette méthode est similaire à celle utilisée dans le vocodeur de phase. L'incrément de phase mesuré à une fréquence donnée et sur un intervalle donné est utilisé pour le calcul de la fréquence instantanée. Celle-ci, après transposition, est utilisée pour la détermination de l'avancement de phase sur l'intervalle de temps modifié.

$$\hat{\phi}_{h,m'} = \hat{\phi}_{h,m'-1} + D \cdot T \cdot (\phi_{h,m} + M2\pi - \phi_{h,m-1}) \tag{8.4}$$

Le déroulement des phases est nécessaire pour les valeurs $D \cdot T \notin \mathbb{N}$. Pour cela, nous déterminons M tel que

$$\phi_{h,m} + M2\pi - \left(\phi_{h,m-1} + (t_m - t_{m-1})\frac{\omega_{h,m-1} + \omega_{h,m}}{2}\right) \leq \pi \tag{8.5}$$

8.3.1 Améliorations

8.3.1.1 Introduction

La qualité des transformations du son obtenue par la méthode PSOLA-WB tient principalement à deux points : 1) le respect de l'enveloppe spectrale (voir partie 2.1.1) ; 2) le respect des relations de phase entre les composantes fréquentielles lors d'une transformation de l'axe temporel (ceci tient au fait de l'absence de modification des formes d'onde dans le PSOLA normal et de la modification de l'axe temporel obtenu par superposition/addition).

Les améliorations que nous étudions dans la suite visent à rapprocher la synthèse sinusoïdale de la synthèse PSOLA,

- tant au niveau des principes sous-jacents aux modifications (-respect de l'enveloppe spectrale en amplitude et en phase lors de transposition, - respect des relations de phase et respect de la forme d'onde en dilatation)
- qu'au niveau de la qualité du son produit.

Ces améliorations s'appliquent principalement au modèle sinusoïdal harmonique, puisque les relations utilisées reposent pour l'essentiel sur l'hypothèse d'un son harmonique ou quasi-harmonique.

8.3.1.2 Transposition : préservation de l'enveloppe spectrale

Une transposition en synthèse sinusoïdale ne préserve pas l'enveloppe spectrale. Afin de permettre la préservation de l'enveloppe spectrale du signal (lorsque celle-ci est définie, c'est-à-dire généralement dans le cas d'un signal harmonique), une correction des amplitudes $A_{h,m}$ doit être appliquée [Sch98]. Cette conservation de l'enveloppe spectrale correspond à l'invariance du filtre d'une décomposition source/filtre. La réponse fréquentielle du filtre peut donc être soustraite du signal avant transposition et rajoutée après transposition. Dans le cas d'une modélisation sinusoïdale, il est toutefois avantageux de tirer parti de la nature des données, et une estimation de l'enveloppe spectrale par cepstre discret [GR90], voire par cepstre discret régularisé [CLM95], est bien adaptée. Les amplitudes corrigées $A_{h,m}$ sont obtenues par échantillonnage de l'enveloppe spectrale aux nouvelles fréquences $\hat{\omega}_{h,m'}$.

De la même manière, les composantes de phase $\phi_{h,m}$ peuvent être corrigées par ré-échantillonnage du spectre de phase du signal original [Oud98].

8.3.1.3 Compression/Dilatation : préservation de la forme d'onde

Dans ce paragraphe, nous illustrons trois méthodes permettant la préservation de la forme d'onde (respect des relations de phase) lors d'une transformation du signal. La première méthode retrouve comme cas particulier l'analyse/synthèse synchrone à la période fondamentale. La deuxième méthode permet d'obtenir le même bénéfice qu'une analyse/synthèse synchrone sans contrainte sur le pas d'analyse. La troisième méthode est une méthode que nous proposons reposant sur le respect du retard de groupe lors d'une transformation du son.

Séparation source/filtre Dans [QM92], une méthode permettant de conserver la forme d'onde («shape invariant») du signal de parole est proposée. Le signal est séparé en contributions d'un signal source $e(t)$ et d'un filtre $v(t)$:

$$s(t_m) = \sum_h A_h(t_m)\cos(\phi_h(t_m)) = A_{e,h}(t_m)\cdot A_{v,h}(t_m)\cos(\phi_{e,h}(t_m)+\phi_{v,h}(t_m)) \quad (8.6)$$

Les composantes h du signal source sont supposées être toutes en phase aux instants t_0 de fermeture de la glotte. La phase du signal source à l'instant d'analyse m peut donc s'exprimer en terme de décalage par rapport à ces instants t_0 :

$$\phi_{e,h}(t_m) = (t_m - t_0)\cdot\omega_h(t_m) \quad (8.7)$$

Lors d'une dilatation du signal d'un facteur D, de nouveaux instants t'_0 de fermeture de la glotte sont définis d'une manière similaire à la définition des marques de synthèse dans la méthode PSOLA (voir partie 7.2.3.1). La préservation de la forme d'onde est obtenue est gardant les phases du signal source égales entre elles aux instants t'_0. La phase du signal source à un instant t'_m s'exprime donc également en terme de décalage par rapport à t'_0.

$$\hat{\phi}_{e,h}(t'_m) = (t'_m - t'_0)\cdot\hat{\omega}_h(t'_m) \quad (8.8)$$

La contribution de phase du filtre $\phi_{v,h}(t)$ est gardée inchangée et la phase du signal de synthèse s'exprime donc

$$\hat{\phi}_h(t'_m) = \phi_{v,h}(t'_m) + (t'_m - t'_0)\cdot\hat{\omega}_h(t'_m) \quad (8.9)$$

$$\boxed{\hat{\phi}_h(t'_m) = \phi_{v,h}(t'_m) + D \cdot (t_m - t_0) \cdot T \cdot \omega_h(t_m)} \tag{8.10}$$

dans lequel T est un facteur de transposition.

Analyse et re-synthèse synchrone à la période fondamentale Dans le cas d'une analyse et d'une re-synthèse synchrone à la période fondamentale, nous avons $t_m = t_0$ et donc, d'après (8.10), la correction de phase à apporter est nulle. L'analyse/re-synthèse sinusoïdale synchrone à la période fondamentale respecte ainsi les relations de phase entre les composantes.

Une autre interprétation est de considérer que lors d'une analyse synchrone à la période fondamentale, la différence de phase entre deux instants d'analyse ne comprend pas l'incrément de phase linéaire ($\phi_{h,m} - \phi_{h,m-1} = T0 \cdot h\omega_0 = h2\pi$), mais uniquement la variation due au filtre du système.

Remarquons que ceci ne constitue pas le seul avantage d'une analyse synchrone. Un autre avantage tient dans le fait que, dans ce cas, l'estimation des paramètres du modèle sinusoïdal est effectuée en des endroits où les relations de phase entre les composantes sont quasi-identiques d'une trame à l'autre, et donc la résolution spectrale est quasi-identique d'une trame à la suivante. De cette manière, le phénomène d'oscillation de la résolution spectrale (observé pour des fenêtres de courtes durée) peut être évité (voir partie 4.2.1.2 FIG. 4.1).

Utilisation du retard de phase relatif L'algorithme proposé par [Fed98] repose sur la conservation du retard de phase relatif des composantes du signal lors de transformations (voir FIG. 8.1 [G]). Le retard de phase de la composante h à la trame t_m est défini comme

$$\tau_{\phi,h,m} = -\frac{\phi_{h,m}}{\omega_{h,m}} \tag{8.11}$$

Le retard de phase relatif est défini comme la différence entre le retard de phase de la composante h et celui de la première composante du modèle (non forcément égale à la fréquence fondamentale) :

$$\Delta\tau_{\phi,h,m} = \tau_{\phi,h,m} - \tau_{\phi,0,m} = -\frac{\phi_{h,m} + M2\pi}{\omega_{h,m}} + \frac{\phi_{0,m} + M'2\pi}{\omega_{0,m}} \tag{8.12}$$

Les deux retards sont indéterminés à $\pm\frac{M2\pi}{\omega_{h,m}}$. La détermination de M et/ou de M' s'effectue de manière à minimiser $\Delta\tau_{\phi,h,m}$. $\Delta\tau_{\phi,h,m}$ représente le décalage temporel minimal entre les maxima de la cosinusoïde de fréquence $\omega_{h,m}$ et de celle de fréquence $\omega_{0,m}$.

La modification du signal est effectuée de manière à préserver le retard de phase relatif lors de la modification du signal (voir FIG. 8.1 [G]). En définissant, de manière similaire, un retard de phase à la synthèse

$$\hat{\tau}_{\phi,h,m'} = -\frac{\hat{\phi}_{h,m'}}{\hat{\omega}_{h,m'}} \tag{8.13}$$

nous imposons donc l'égalité des retards de phase relatifs à l'analyse et à la synthèse

$$\begin{aligned} \Delta\tau_{\phi,h,m} &= \hat{\Delta}\tau_{\phi,h,m'} \\ \frac{\hat{\phi}_{h,m'}}{\hat{\omega}_{h,m'}} - \frac{\hat{\phi}_{0,m'}}{\hat{\omega}_{0,m'}} &= \frac{\phi_{h,m}}{\omega_{h,m}} - \frac{\phi_{0,m}}{\omega_{0,m}} \end{aligned} \tag{8.14}$$

La nouvelle phase $\hat{\phi}_{h,m'}$ est finalement donnée par

$$\boxed{\hat{\phi}_{h,m'} = -\left(\hat{\tau}_{\phi,0,m'} + \left(\tau_{\phi,h,m} - \tau_{\phi,0,m}\right)\right)\hat{\omega}_{h,m'}} \tag{8.15}$$

L'évolution de la phase de la première composante du spectre $\hat{\phi}_{0,m'}$ est obtenue par l'algorithme de fréquence instantanée (8.4) vu précédemment.

Outre le respect des relations de phase entre les composantes, cet algorithme présente l'avantage de n'utiliser le modèle d'évolution de la phase du modèle sinusoïdal (évolution temporelle de la phase correspondant au cas d'une sinusoïde pure) que pour la composante de plus basse fréquence. Ce choix est adéquat étant donné la variation plus lente et le plus faible niveau de bruit dans cette région du spectre, donnant par conséquence une meilleure estimation de la phase. Les phases des harmoniques supérieures sont construites non pas temporellement mais fréquentiellement.

Il est facile de montrer l'équivalence entre la méthode dite «shape invariant» [QM92] et la méthode du retard de phase relatif [Fed98] (voir annexe P). L'intérêt de la méthode du retard de phase réside cependant dans la non-nécessité d'une analyse synchrone à la période fondamentale.

Il est également facile de montrer que, dans le cas d'un signal composé de sinusoïdes pures de fréquences constantes, la correction apportée aux phases par la méthode du retard de phase est identique à celle apportée par la méthode de la fréquence instantanée (8.4).

Utilisation du retard de groupe relatif Nous avons développé un algorithme proche de celui de [Fed98], mais reposant sur le respect du retard de groupe relatif. L'objectif de cet algorithme est de remplacer la mesure du retard de phase par une mesure de la localisation temporelle de l'énergie des fréquences du signal, mesure fournie par le retard de groupe. L'objectif est donc de préserver, non pas les relations de phase entre les composantes du modèle, mais les relations de localisation d'énergie entre les composantes fréquentielles (voir FIG. 8.1 [D]).

Nous définissons le retard de groupe $\tau_{g,h,m}$ entre les composantes h et $h-1$ à l'instant t_m

$$\tau_{g,h,m} = -\frac{\phi_{h,m} + M2\pi - \phi_{h-1,m}}{\omega_{h,m} - \omega_{h-1,m}} \tag{8.16}$$

M est calculé de manière à minimiser $\tau_{g,h,m}$ Nous définissons, de la même manière, le retard de groupe à la synthèse $\hat{\tau}_{g,h,m'}$

$$\hat{\tau}_{g,h,m'} = -\frac{\hat{\phi}_{h,m'} - \hat{\phi}_{h-1,m'}}{\hat{\omega}_{h,m'} - \hat{\omega}_{h-1,m'}} \tag{8.17}$$

L'évolution de phase des deux premières composantes ϕ_0 et ϕ_1 est déterminée par l'algorithme de fréquence instantanée (8.4). La phase des composantes de la bande h est calculée de manière à respecter la distance temporelle, appelée «retard de groupe relatif» $\Delta\tau_{g,h,m}$, entre son retard de groupe $\hat{\tau}_{g,h,m'}$ et celui de la bande de base $\hat{\tau}_{g,1,m'}$ (voir FIG. 8.1 [D]) :

$$\Delta\tau_{g,h,m} = \hat{\tau}_{g,h,m'} - \hat{\tau}_{g,1,m'} = \tau_{g,h,m} - \tau_{g,1,m} \tag{8.18}$$

La phase de synthèse $\hat{\phi}_{h,m'}$ est alors donnée par

$$\boxed{\hat{\phi}_{h,m'} = \hat{\phi}_{h-1,m'} - \left(\hat{\tau}_{g,1,m'} + \left(\tau_{g,h,m} - \tau_{g,1,m}\right)\right)\left(\hat{\omega}_{h,m'} - \hat{\omega}_{h-1,m'}\right)} \tag{8.19}$$

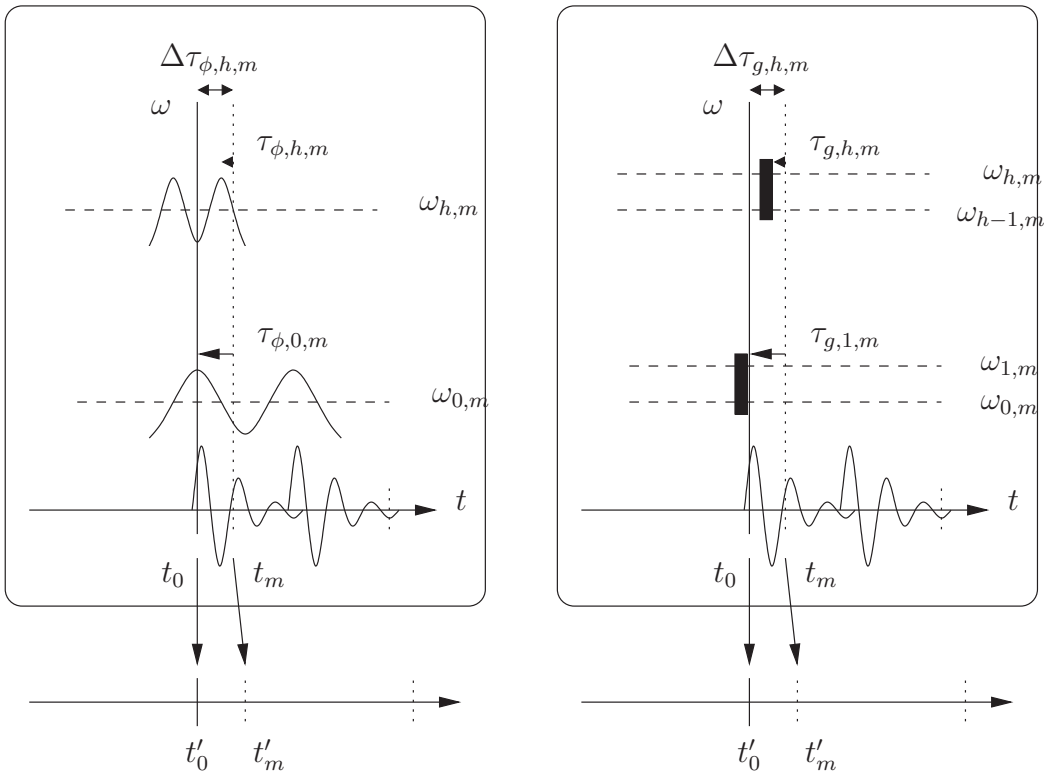


FIG. 8.1 – [G] Méthode du retard de phase relatif [D] Méthode du retard de groupe relatif

8.3.2 Observations

Les algorithmes de modification de la phase étudiés ci-dessus ont été comparés sur des signaux de voix parlées. Nous avons constaté une légère amélioration, d'un point de vue perceptif, lorsque la fréquence instantanée est utilisée. Nous avons constaté une nette amélioration, d'un point de vue perceptif, lorsque la synchronie ou les retards de phase/groupe sont utilisés. D'un point de vue perceptif les deux méthodes de correction par retard sont quasi-indiscernable. Ces résultats sont illustrés à la FIG. 8.2.

Ceci permet de valider le troisième point de notre thèse :
«la prise en compte des relations de phase lors de modifications d'un signal permet d'atteindre un haut niveau de qualité sonore»
Nous avons montré, dans cette partie, que la prise en compte des relations de phase entre composantes fréquentielles permet d'améliorer significativement la qualité du signal transformé. Ceci était connu dans le cas de méthode d'analyse/synthèse synchrone à la période fondamentale. Nous l'avons montré dans le cas général d'une analyse/synthèse non-synchrone, par respect du retard de phase relatif et du retard de groupe relatif.

Dans les deux méthodes utilisant le retard relatif (de phase et de groupe), les deux premiers trajets (resp. le premier et les deux premiers) du modèle sont supposés continus et sont supposés correspondre aux fréquences les plus basses. Ceci suppose un ordonnancement des trajets de la fréquence la plus basse à la fréquence la plus élevées. Ceci suppose également qu'une composante de même ordre harmonique soit représentée au cours du temps par un trajet de numéro identique. Ceci favorise l'utilisation d'algorithmes de création de trajets basés sur l'hypothèse d'harmonicité tel qu'expliqué dans la partie 5.

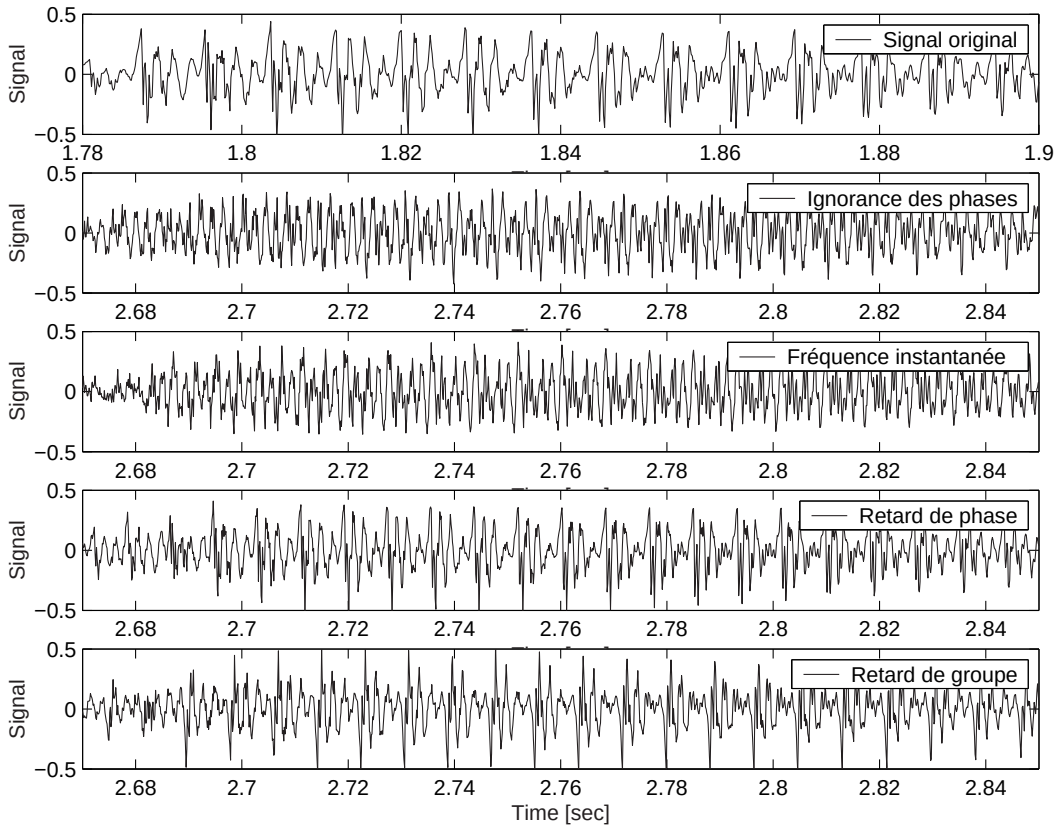


FIG. 8.2 – Dilatation d'un signal en synthèse sinusoïdale [1] Signal original [2] Ignorance des phases [3] Algorithme de la fréquence instantanée [4] Algorithme du retard de phase relatif, [5] Algorithme du retard de groupe relatif. Signal= speech. Facteur de dilatation : 1.5

8.4 Résumé

A la figure FIG. 8.3, nous avons représenté sous forme de diagramme les différentes étapes de la phase de modification du signal en synthèse sinusoïdale.

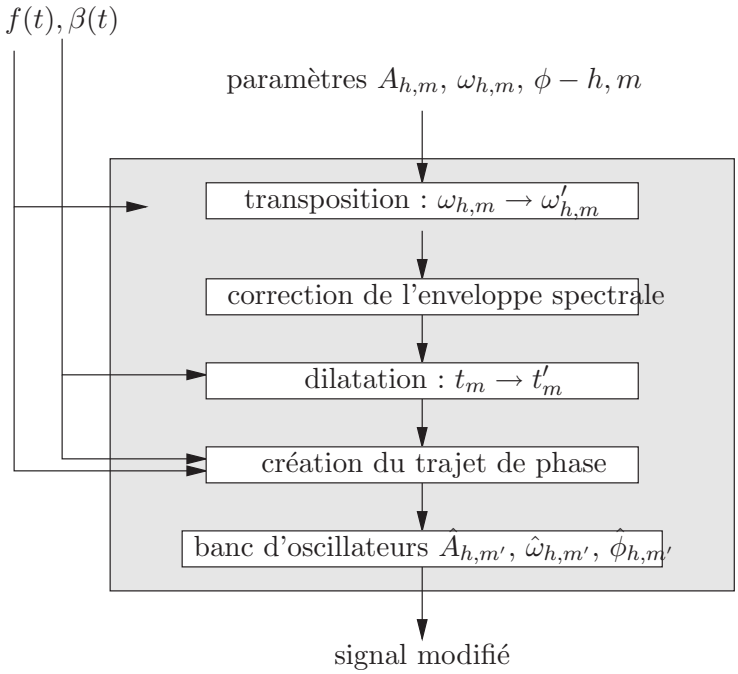


FIG. 8.3 – Diagramme de l'algorithme de modification (transposition et dilatation) en synthèse sinusoïdale

Notes de bas de page relatives à la partie 8

Chapitre 9

Modification du signal par synthèse hybride

9.1 Introduction

Dans cette partie, nous proposons une méthode hybride de modification du signal permettant, parmi les modèles étudiés, le choix de celui le plus adapté aux caractéristiques locales du signal. De nombreuses méthodes de synthèse hybride ont été proposées. La plus connue est sans doute la méthode Harmonique plus Bruit ([GL88], [SS90], [AMT91] [Sty96]), étendue aux transitoires Harmonique plus Bruit plus Transitoires [Lev98] .

Notre objectif est de proposer une méthode de modification du signal (synthèse AVEC modification [1]) utilisant simultanément les avantages de la méthode PSOLA (ainsi que sa version adaptée aux traitement des régions non-voisées du signal) et ceux du modèle sinusoïdal.

9.1.1 Avantage et inconvénient des méthodes PSOLA et du modèle sinusoïdal

La **méthode PSOLA** est particulièrement bien adaptée et robuste pour certaines classes de signaux sonores : les signaux harmoniques ou pseudo-harmoniques et présentant une forme d'onde dont la Réponse Impulsionnelle du filtre est courte par rapport à la période fondamentale du signal. Nous désignons cette dernière caractéristique comme le caractère «singularité» (carcactère «impulsionnel») d'une forme d'onde . Un contre-exemple est illustré à la Fig. 9.1 panneau de gauche. Dans cet exemple, la voyelle «i» de premier formant aligné sur la fréquence fondamentale présente une forme d'onde sans localisation temporelle à l'intérieur de la période. La méthode PSOLA présente l'avantage, par rapport à d'autres méthodes de transformation du son comme la synthèse sinusoïdale, de ne pas nécessiter l'estimation d'un modèle. Même si la méthode PSOLA repose sur l'hypothèse d'un signal harmonique, le modèle harmonique n'est pas estimé. Ceci favorise non seulement la robustesse des transformations (par exemple l'absence de bruit musical dans les hautes fréquences) mais également la préservation de caractéristiques difficilement modélisables comme les variations fines du signal au cours du temps et le jeu de relations de phase entre les composantes, la préservation

de l'enveloppe spectrale, la possibilité de modifier les signaux pseudo-périodiques (voir FIG. 9.1 panneau de droite) difficiles à modéliser en synthèse sinusoïdale. De plus, son extension, VUV-PSOLA permet également de l'utiliser pour les signaux non-voisés ou mixtes voisés/non-voisés.

Le **modèle sinusoïdal** permet de représenter une large classe de signaux sonores harmoniques ou inharmoniques ; il n'est par contraint par le caractère «singularité» du signal. Le passage par un modèle permet un contrôle total sur l'ensemble des paramètres et donc un nombre important de transformations de types différents. Se pose cependant le problème de son estimation, difficile en dehors de l'hypothèse d'harmonicité, et difficile dans le cas de signaux pseudo-périodiques (voir FIG. 9.1 panneau de droite). De plus, son estimation ainsi que la création de trajets réguliers sont rendus difficiles en haute fréquence du fait d'une diminution du rapport signal sinusoïdal sur bruit.

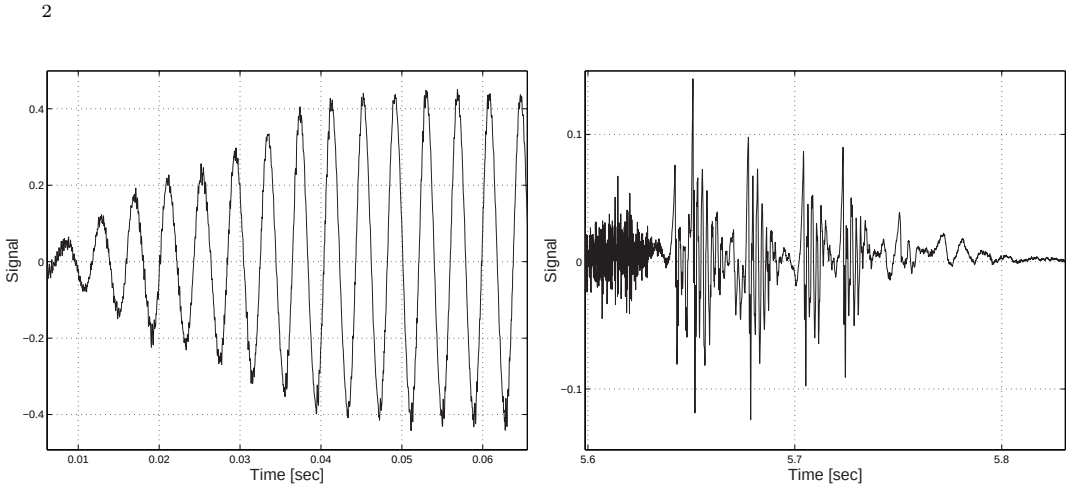


FIG. 9.1 – forme d'onde correspondant [G] à un «i», [D] à une «vocal fry» de la langue anglaise

9.1.2 Positionnement des méthodes dans un espace de caractéristiques

A la FIG. 9.2 partie du haut, nous représentons, dans un espace de caractéristiques, les différentes catégories de signaux que nous considérons et les modèles que nous y associons. Nous y distinguons trois types de caractéristiques :

- le caractère «singularité» du signal (au sens de la partie 3.4 de ce rapport),
- le caractère réducteur du modèle où l'«erreur de modélisation» (au sens de la partie 4.3 de ce document) dépend non seulement de la qualité de l'estimation du modèle mais également de la présence de bruit,
- le caractère harmonique ou non du signal (au sens de la partie 5 de ce document)

Les caractéristiques voisée/non-voisée font référence à l'erreur de modélisation dans le cas d'un modèle harmonique. L'erreur de modélisation dans le cas général d'un modèle non nécessairement harmonique est plus généralement appelée bruit.

Cette représentation permet de localiser les différentes méthodes de modification considérées par adéquation aux catégories de signal :

- la méthode PSOLA s'applique aux signaux quasi-harmoniques dont la forme d'onde présente des singularités et le signal renferme une quantité faible de bruit,
- la modélisation sinusoïdale correspond au cas d'un signal harmonique ou inharmonique mais exempt de bruit,
- les transitoires correspondent au cas d'un signal non-harmonique dont la forme d'onde présente des singularités.

Sous l'angle de cette répartition dans notre espace de caractéristiques, la synthèse sinusoïdale permet de représenter la plus grande classe de signaux.

A la FIG. 9.2 partie du bas, nous représentons les méthodes, non pas dans les catégories de signaux considérées mais dans les estimations que nous avons des modèles. L'axe de l'erreur de modélisation est remplacé par l'erreur de spécification (théoriquement indépendante du bruit). Dans ce cas, la synthèse sinusoïdale n'est plus le signal moins le bruit mais la partie du signal pour lequel le modèle sinusoïdal est bien spécifié selon un critère de régularité temporelle. Sous cet angle, la synthèse sinusoïdale n'est plus limitée par le bruit mais par la qualité et l'adéquation de l'estimation. A l'inverse, sous cet angle, la méthode PSOLA est indépendante de toute spécification de modèle puisque ne reposant pas sur un modèle.

Les deux espaces de la FIG. 9.2 correspondent à deux visions différentes, la première est une vision en termes d'adéquation de modèle par rapport à des classes ; la deuxième est un point de vue d'ingénieur prenant également en compte l'applicabilité du modèle, i.e. la qualité de son estimation dans chaque cas.

9.2 Méthode SINOLA (SINusoidal OverLap-Add)

L'idée de notre méthode hybride, appelée SINOLA (**SIN**usoidal **O**verLap-**A**dd) est de combiner simultanément les deux méthodes de modification du signal.

9.2.1 Première formulation

La première formulation de la méthode SINOLA, proposée dans [PR99b], tend à favoriser la modélisation sinusoïdale.

En voici les principales étapes :

1. Dans un premier temps, le signal $s(t)$ est modélisé par une somme de sinusoïdes $\hat{s}(t)$.
2. La re-synthèse du signal par le modèle sinusoïdal est soustraite du signal original : $b(t) = s(t) - \hat{s}(t)$. Le signal $b(t)$ renferme toutes les composantes du signal n'ayant pas pu être modélisées par le modèle sinusoïdal : bruit, transitoire, pulses peu périodiques
3. Le signal $b(t)$ est ensuite modifié par l'algorithme PSOLA (VUV-PSOLA) donnant le signal transformé $b'(t)$.
4. La synthèse sinusoïdale avec modifications est effectuée : $\hat{s}'(t)$.
5. Les deux signaux $b'(t)$ et $\hat{s}'(t)$ sont ré-additionnés, donnant le signal modifié $s'(t)$

Ainsi, même si une région du signal ne rentrait dans aucun des deux modèles, elle serait malgré tout (par défaut) prise en compte lors de la modification du signal.

Ceci est illustré à la FIG. 9.3

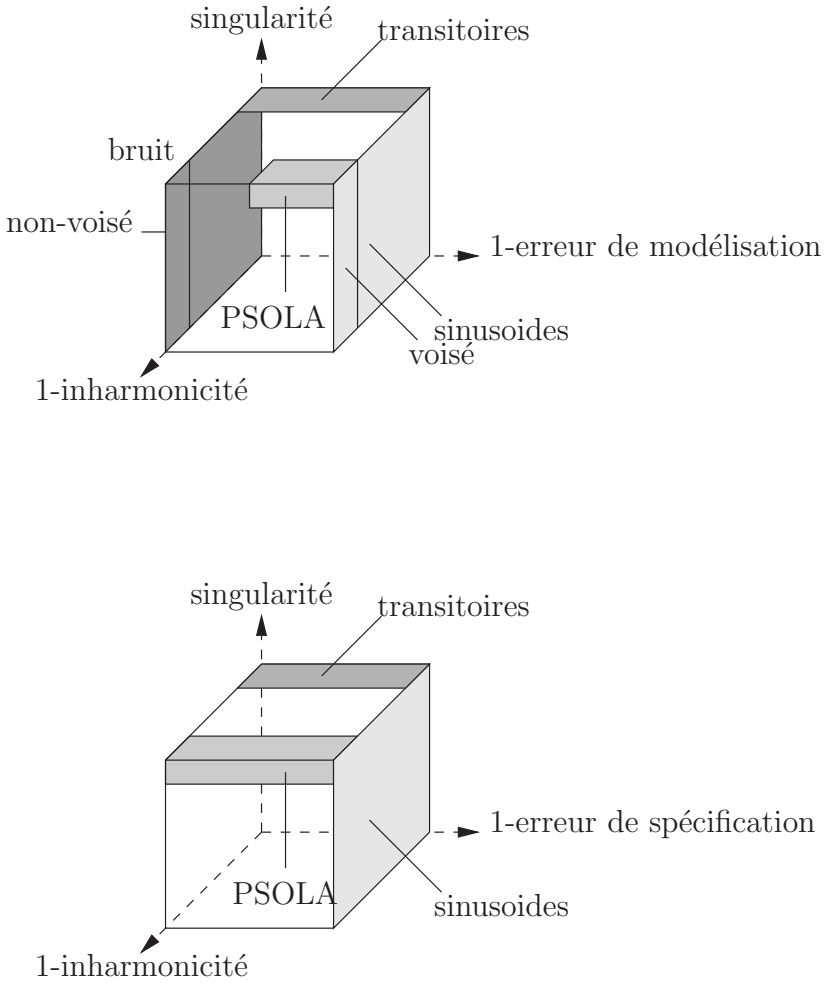


FIG. 9.2 – Positionnement des méthodes dans un espace de caractéristiques

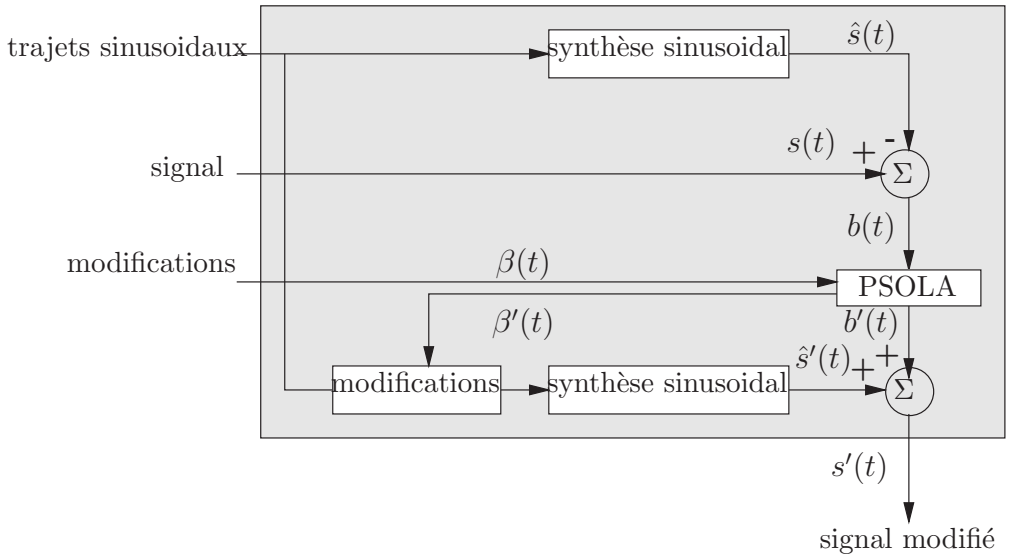


FIG. 9.3 – Diagramme de la partie séparation (soustraction) et modification du signal de SINOLA

Un point important consiste à assurer la synchronie des deux transformations. Une transformation de l'axe temporel en PSOLA s'effectue en nombre entier de périodes fondamentales. Pour un facteur de modification du signal donné, il n'est donc pas possible de garantir la synchronie des phases de la partie du signal modifiée par PSOLA et par la synthèse sinusoïdale. La solution choisie est d'asservir la synthèse sinusoïdale à la méthode PSOLA. Pour une fonction voulue de modification de l'axe temporel $\beta(t)$, nous obtenons en sortie de l'algorithme PSOLA une fonction $\beta'(t)$ qui correspond à la modification réellement apportée. C'est cette fonction $\beta'(t)$ qui est utilisée pour la modification de l'axe temporel de la partie sinusoïdale (voir FIG. 9.3).

Remarquons que la fonction de voisement $v(t, \omega)$ utilisée pour PSOLA (VUV-PSOLA) est la fonction calculée avant soustraction de la partie sinusoïdale. D'un point de vue perceptif, ceci n'est pas gênant. Cela s'explique par le fait que dans les régions voisées (régions fréquentielles voisées), l'essentiel de la contribution de l'énergie provient de la périodicité du signal.

- Dans le cas où la modélisation sinusoïdale représenterait l'ensemble des composantes périodiques, nous appliquerions donc un algorithme périodique (PSOLA) à une région non-périodique d'énergie faible. Après addition des deux contributions, cette modification périodique du bruit n'est pas (ou difficilement) perçue.
- A l'inverse, dans le cas où la modélisation sinusoïdale ne représenterait qu'une partie des composantes périodiques, l'application de l'algorithme périodique (PSOLA) dans cette région du signal différence serait souhaitable.

Observations : Cette formulation de la synthèse hybride tend à favoriser un type de synthèse, la synthèse sinusoïdale. Cette formulation ne permet pas de tirer profit de la méthode PSOLA dans les régions auxquelles celle-ci s'applique. Dans cette formulation, la méthode PSOLA est bien souvent réduite à un rôle de traitement des parties non-voisées (VUV-PSOLA) et des transitoires du signal.

9.2.2 Deuxième formulation

La deuxième formulation de la méthode SINOLA repose sur le choix, à chaque instant et pour chaque bande de fréquence, de la méthode la plus adaptée au signal, selon un critère de présence ou d'absence de trajets sinusoïdaux, de concentration de la forme d'onde dans cette bande et d'inharmonicité du signal.

A l'inverse de la première formulation, l'objectif ici est d'utiliser au maximum l'algorithme PSOLA.

En voici les principales étapes.

1. **sélection :**
   De la même manière que précédemment, la formulation est inversée, c'est-à-dire que nous ne cherchons pas les régions qui peuvent être modélisées par PSOLA, mais bien celles parmi l'ensemble des régions qui peuvent être représentées par des sinusoïdes, celles ne pouvant pas être modélisées par PSOLA. Ceci permet, comme précédemment, de prendre en compte (par défaut) les régions du signal ne rentrant dans aucun des deux modèles.
2. re-synthèse du signal par le modèle sinusoïdal et calcul du signal différence : $b(t) = s(t) - \hat{s}(t)$. Le signal $b(t)$ renferme maintenant également des composantes pouvant être représentées par des sinusoïdes, mais sélectionnées comme répondant au modèle PSOLA
3. modification du signal $b(t)$ par l'algorithme PSOLA (VUV-PSOLA) : $b'(t)$.
4. synthèse sinusoïdale avec modifications : $\hat{s}'(t)$
5. addition des deux signaux modifiés : $s'(t) = b'(t) + \hat{s}'(t)$

9.2.2.1 Choix d'une méthode pour chaque région du signal

Le choix de la méthode la plus appropriée s'effectue selon les critères de :
- caractère "singularité" de la forme d'onde ,
- présence ou absence de sinusoïdes,
- inharmonicité du signal.

♢ ***Caractérisation en «singularité» par la fonction*** σ_n

Dans la partie 3.4, nous avons proposé l'utilisation des fonctions γ_n de localisation temporelle et σ_n de mesure de concentration d'énergie locale. La fonction de localisation nous a servi au chapitre 6 pour le positionnement des marques PSOLA. Nous utilisons ici la fonction de mesure de concentration σ_n afin de déterminer le caractère non-destructif du fenêtrage PSOLA et donc sa propension à être représenté par des formes d'onde élémentaires . Dans la partie 3.4, cette fonction à été illustrée pour une mesure en bande formantique et en bande d'octave. Le choix des bandes de fréquences se porte ici sur :

- W_1 : 0-1000 Hz ; la fréquence de 1000 Hz est choisie de manière à correspondre grossièrement à la fréquence à partir de laquelle la perception de l'oreille devient logarithmique. Cette bande renferme également, dans la majorité des cas, la fréquence du premier formant ;

- W_2 : 1000-5000 Hz ; la fréquence de 5000 Hz est choisie de manière à se trouver au-dessus de la fréquence des deux premiers formants de la voix ;
- W_3 : 5000-Fe/2 Hz ; dans lequel Fe représente la fréquence d'échantillonnage.

Notre approche est donc comparable à celle prise par [d'A89] dans laquelle la synthèse CHANT (modélisation par Forme d'Onde Formantique) est combinée à la synthèse sinusoïdale. La première sert à modéliser la partie supérieure du spectre, la seconde la bande de base du signal (bande de fréquence inférieure à 800 Hz). [3] Dans notre cas, le choix de l'algorithme utilisé pour la modification du signal dans la bande de base n'est pas figée et dépend de la valeur de σ_n dans cette bande.

◊ *Critère utilisé*

La mesure du caractère «singularité» de la forme d'onde est théoriquement faite en comparant la valeur de $\sigma_n(t_m, W_j)$ (valeur de σ_n évaluée à la position t_m des marques PSOLA et dans la bande de fréquence W_j) à sa valeur limite $\sigma_{h,n}$. ($\sigma_{h,n}$ dépend uniquement du type de la fenêtre de pondération utilisée et est considéré comme correspondant au cas d'une forme d'onde sans concentration temporelle. Une valeur de $\sigma_n(t_m, W_j) > \sigma_n$ désigne une concentration de l'énergie locale de la forme d'onde . Le calcul est effectué pour la globalité d'un segment T_i et la valeur moyenne est calculée. Le passage d'un modèle à un autre dépendant de la valeur de σ_n, nous avons jugé préférable de prendre une décision sur l'ensemble d'un segment, afin d'éviter le passage intempestif d'un modèle à l'autre. Les segments choisis correspondent aux segments temporels voisés du signal. Nous définissons T_i comme le $i^{\text{ème}}$ segment temporel obtenu par découpage du signal en régions voiséss/ non-voisées (voir partie 5 FIG. 5.4). Un meilleur choix résulterait cependant d'une procédure de segmentation en notes ou phonèmes [Ros99].

Afin de rendre l'algorithme plus robuste et de prévenir le phénomène de double périodicité de σ_n (voirFIG. 9.4), nous évaluons également σ_n aux positions t_{int} entre deux marqueurs t_m. Nous notons cette valeur $\sigma_n(t_{int}, W_j)$. Les valeurs moyennes sur l'ensemble du segment T_i, notées $\overline{\sigma_n(t_m, W)}_{T_i}$ et $\overline{\sigma_n(t_{int}, W)}_{T_i}$, sont ensuite comparées à la valeur limite $\sigma_{h,n}$. Le caractère «singularité» est attribué à un segment T_i dans la bande W_j si

$$\frac{\overline{\sigma_n(t_m, W_j)}_{T_i} - \sigma_{h,n}}{1 - \sigma_{h,n}} > \alpha_1 \qquad \text{et} \qquad \frac{\overline{\sigma_n(t_{int}, W_j)}_{T_i} - \sigma_{h,n}}{1 - \sigma_{h,n}} < \alpha_2 \tag{9.1}$$

Les valeurs $\alpha_1 = 0.2$ et $\alpha_2 = 0$ nous ont conduit à des résultats satisfaisants.

Pour chaque segment temporel voisé T_i et chaque bande de fréquence W_j, une prise de décision quant au caractère «singularité» de la région est effectuée. Nous notons P les régions (T_i, W_j) caractérisées comme «singularité» et NP les autres régions.

Ceci est illustré à la FIG. 9.6 panneau de gauche.

◊ *Sélection d'une sinusoïde*

La sélection s'effectue par segments. Un trajet sinusoïdal h est annulé sur le segment T_i (c'est-à-dire ne sera pas utilisé pour la re-synthèse sinusoïdale sur le segment T_i) uniquement si son énergie sur le segment T_i se trouve majoritairement dans les régions de type P.

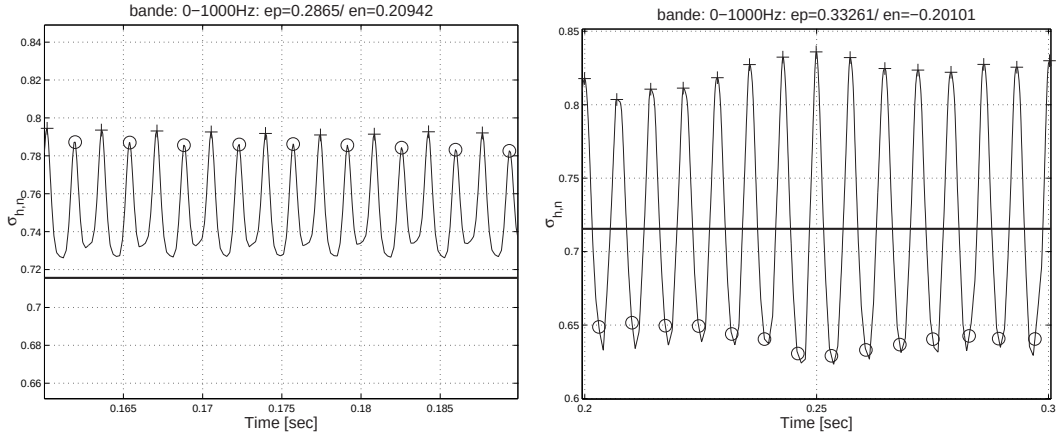


FIG. 9.4 – Valeurs de σ_n à t_m (+) à t_{int} (O) [G] illustration du phénomène de double périodicité de σ_n, Signal= vie [D] simple périodicité de σ_n, Signal= speech

Ce choix est effectué pour les raisons suivantes :

- L'annulation d'un trajet sur l'ensemble du segment, lorsque l'énergie est répartie de manière égale entre régions P et NP, conduirait à la création d'un trou d'énergie dans les régions de type NP.
- L'annulation du trajet uniquement dans les régions P lorsque la composante traverse une région P, se ferait au risque de la création d'un bruit musical (apparition/disparition du trajet).

Pour un segment T_i donné, nous cherchons toutes les sinusoïdes h traversant les régions $(T_i, W_*) = P$. Pour chacune de ces sinusoïdes, nous calculons le rapport de l'énergie de la sinusoïde contenue dans les régions $(T_i, W_*) = P$ à celui du trajet total de la sinusoïde dans le segment T_i.

$$r(h, T_i) = \frac{\sum_{j \text{ tel que } (T_i, W_j)=P} \; \sum_{t \in T_i \text{ et } \omega_h(t) \in W_j} A_h^2(t)}{\sum_{t \in T_i} A_h^2(t)} \tag{9.2}$$

Si ce rapport est important (supérieur à un seuil), l'énergie de la sinusoïde sur T_i est majoritairement contenue dans les régions de type P et le trajet est annulé sur l'intervalle T_i. Le seuil choisi est de 75%.

Ceci est illustré à la FIG. 9.5. Le panneau de gauche représente un trajet à mi-chemin entre une région $(T_i, W_j) = P$ et une région $(T_i, W_j) = NP$. Le panneau de droite représente l'énergie du trajet au cours du temps. La partie en trait gras représente l'énergie présente dans la région $(T_i, W_j) = P$, celle en trait fin représente l'énergie sur le segment T_i.

Cet algorithme est appliqué à l'ensemble des trajets h. Le résultat final de l'algorithme de sélection est illustré à la FIG. 9.6. Le panneau de gauche représente les trajets initiaux, celui de droite les trajets sélectionnés.

◊ ***Prise en compte de l'inharmonicité du signal***

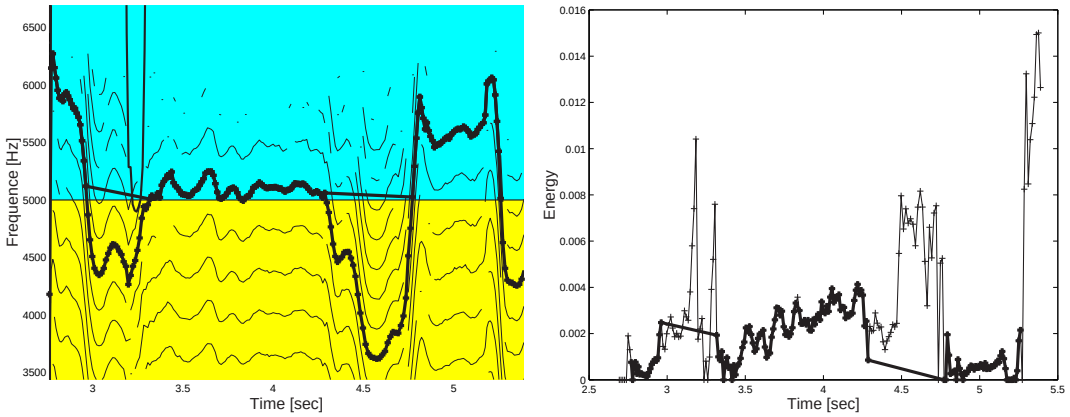


FIG. 9.5 – Algorithme de sélection de sinusoïdes. [G] Région P (zone grisée) et NP (zone grisée claire), trajets sinusoïdaux traversant successivement les régions P et NP (trait gras) autres trajets sinusoïdaux (traits légers) Signal=tibet-extract [D] Evolution temporelle de l'énergie sur la région T_i (trait léger), évolution temporelle de l'énergie sur les régions $(T_i, W_*) = P$

L'algorithme PSOLA ne s'applique en principe qu'aux sons harmoniques. A l'inverse la synthèse sinusoïdale, sous sa formulation générale, n'est pas contrainte par l'harmonicité et peut représenter, lorsque la régularité des trajets temporels le permet, des composantes inharmoniques.

Nous nous intéressons dans ce modèle aux taux faibles d'inharmonicité (cas de la dilatation des partiels sinusoïdaux). Dans ce cas, comme indiqué dans la partie 5, l'utilisation d'une valeur petite de Δ (coefficient de tolérance à l'inharmonicité) du coefficient de voisement permet de mesurer le caractère non harmonique du signal.

L'utilisation de cette mesure permet de favoriser l'utilisation du modèle sinusoïdal, dans la méthode SINOLA, pour la représentation des composantes inharmoniques. Les composantes inharmoniques mais présentant une régularité temporelle élevée et une énergie importante (ces deux points sont voulus afin de distinguer les composantes inharmoniques des composantes bruitées) sont gardées dans les régions P.

9.2.2.2 Résumé

A la FIG. 9.7, nous avons représenté sous forme de diagramme les différentes étapes de la caractérisation du signal utilisées pour la séparation du signal entres modèle sinusoïdal et modèle PSOLA. La sortie de ce diagramme constitue l'entrée du diagramme 9.3.

9.2.2.3 Résultats

L'application de la méthode SINOLA est montrée à la FIG. 9.8. Le panneau du haut représente le spectrogramme du signal original. Le signal est ensuite décomposé en parties à modéliser respectivement par le modèle sinusoïdal (partie de gauche) et par l'algorithme PSOLA. Chacune des parties est ensuite modifiée selon son algorithme respectif. Le signal

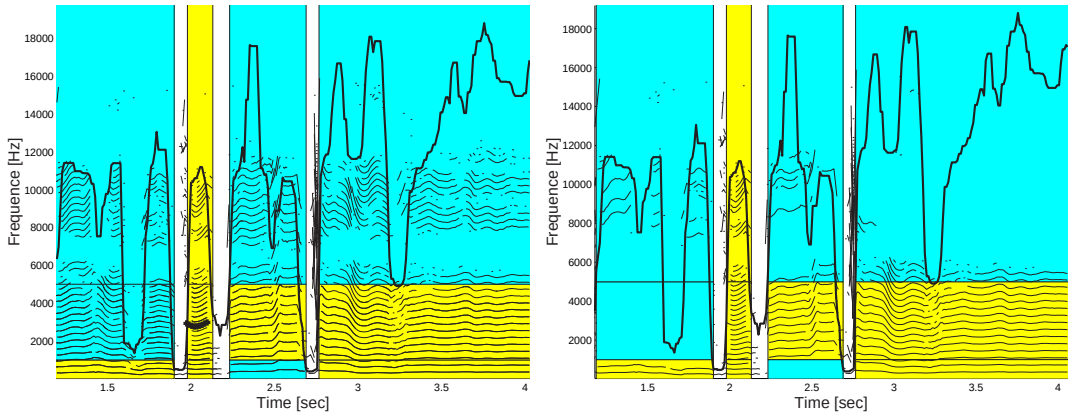


FIG. 9.6 – Région P (zone grisée) et NP (zone grisée claire), région non-voisée (zone blanche), fréquence de coupure voisé/non-voisé (trait gras), [G] trajets de sinusoïdes initiaux [D] trajets de sinusoïdes sélectionnés Signal= tibet-extract

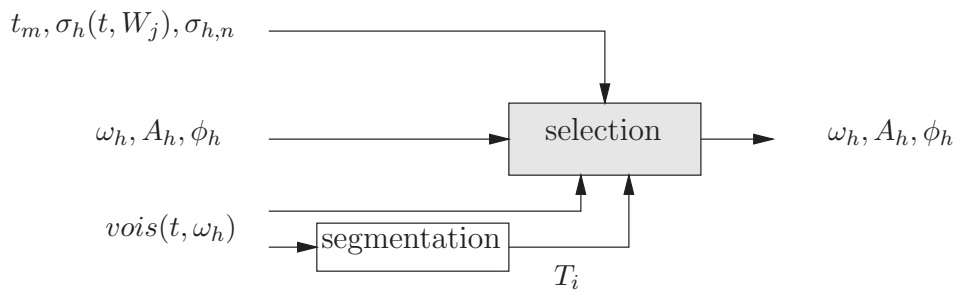


FIG. 9.7 – Diagramme de la partie sélection d'une méthode de SINOLA

est enfin recomposé par superposition des deux signaux (panneau inférieur).

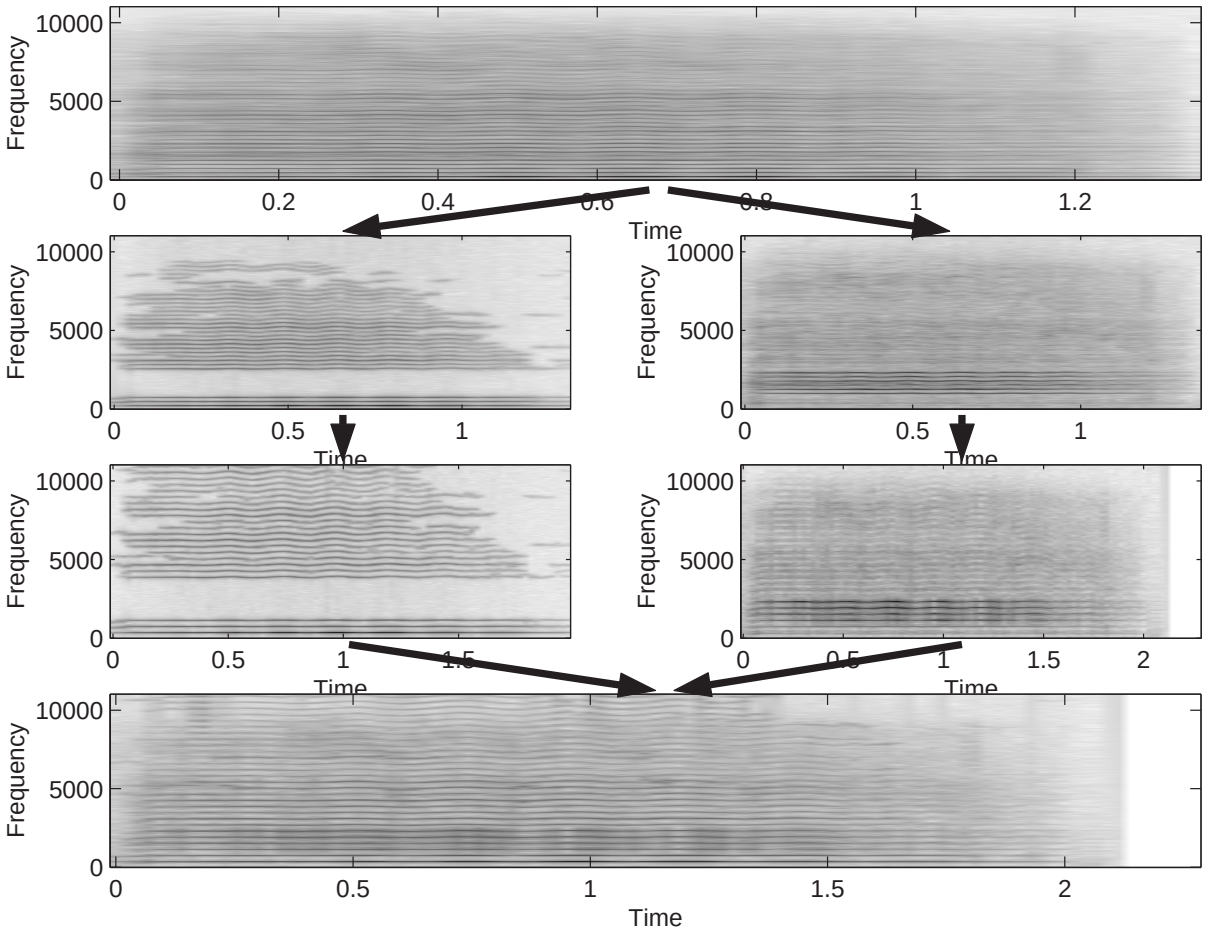


FIG. 9.8 – Illustration de l'application de la méthode SINOLA à un signal de violon (voir texte)

Notes de bas de page relatives à la partie 9

1. La synthèse sans modification est un non-sens en PSOLA, puisqu'en PSOLA le signal non modifié est le signal original
2. Le lecteur intéressé par une comparaison des algorithmes sinusoïdal harmonique et TD-PSOLA dans le cadre de la synthèse concaténative pourra se référer à [Dut94].
3. Dans [d'A89], ce choix repose sur le modèle fonctionnel du conduit auditif périphérique dans lequel les harmoniques du signal sont résolues indépendamment pour les fréquences inférieures à 800 Hz. [d'A89] propose de remplacer la synthèse CHANT en bande de base par une synthèse plus fine de fréquences indépendantes.

Dans le cas de la méthode PSOLA, un point de vue fondé sur la RI des formants basses fréquences généralement plus longue que celle des formants hautes fréquences [Fla72] favorise l'utilisation de PSOLA pour les hautes fréquences. A l'opposé, la modélisation sinusoïdale est rendue difficile en hautes fréquences du fait de la diminution du rapport signal sinusoïdal sur bruit et de la variation plus importante des composantes du signal.

Chapitre 10

Conclusion générale et Perspectives

L'objectif de cette recherche était l'étude des algorithmes de modifications du signal sonore par décomposition du signal en formes d'onde élémentaires et modification du signal par la méthode PSOLA d'une part, et par modélisation en sinusoïdes et synthèse par addition de sinusoïdes d'autre part.

Un aspect sous-jacent à l'ensemble de la recherche est l'utilisation de la phase du signal.

Première partie : Caractérisation du signal

La première partie de cette recherche a été consacrée à l'étude des caractéristiques permettant la mise en oeuvre des deux modèles, ainsi que celle permettant le choix du modèle le plus adapté aux caractéristiques locales du signal.

Chapitre 3 : Détection de singularités dans le signal

Un certain nombres de choix ont été effectués dans cette recherche. Le premier concernait le choix, parmi les différentes méthodes PSOLA existantes, de la méthode PSOLA à bande large. Ceci afin de mieux tenir compte du paradigme de forme d'onde élémentaire , et en raison de la complémentarité de cette méthode par rapport à la modélisation sinusoïdale. A la différence du PSOLA à bande étroite, le PSOLA bande large nécessite un positionnement des marques (nécessaires au découpage du signal en formes d'onde élémentaires) non seulement synchrone à la période fondamentale mais également proche des maxima locaux d'énergie. Nous avons constaté le peu de littérature existante dans ce domaine, comparativement à l'algorithme de modification du signal PSOLA. Ceci nous a conduit à étudier les approches prises en traitement de la parole pour la détection des instants de fermeture de la glotte (IFGS). Les méthodes fondées sur la rupture du modèle auto-régressif du signal de parole, ainsi que les propriétés phase minimale du signal glottal, ont été étudiées dans le but de déterminer leur applicabilité au marquage d'un signal non nécessairement de parole. Nous avons proposé une méthode utilisant le retard de groupe du signal ou du signal résiduel, permettant la caractérisation d'instants, appelés singularités, en termes de localisation et de largeur temporelle. Ces deux caractéristiques permettent le positionnement des marques

PSOLA ainsi qu'une mesure de la détérioration du signal engendrée par le fenêtrage de l'algorithme PSOLA à bande large. Pour une observation du signal sur un horizon plus large, cette méthode du retard de groupe peut également être utilisée pour la localisation des transitoires dans le signal, définis ici comme des singularités non-périodiques.

Chapitre 4 : Sinusoïdalité

La modélisation du signal en sinusoïdes a été étudiée sous l'angle de l'estimation des paramètres du modèle et de la détection de sinusoïdes ; les deux étant intimement liées au travers du choix d'un modèle sinusoïdal. Les estimateurs les plus communément utilisés ont été étudiés et divisés en plusieurs classes, selon que l'estimation s'effectue de manière locale ou globale en fréquence, en utilisant la forme du spectre - estimateurs morphologiques - ou en minimisant un critère d'erreur de modélisation - estimateur des moindres carrés , que l'estimation permette la prise en compte de variations du signal ou non. Nous avons proposé un modèle sinusoïdal fondé sur la mesure de la distorsion du spectre, et dont les paramètres d'amplitude et de fréquence varient de manière linéaire sur la durée de l'observation. Ce nouveau modèle a été comparé aux autres estimateurs en termes de biais, variance et erreur quadratique moyenne pour différents signaux tests représentant différentes conditions d'analyse : échelle du spectre, résolution spectrale, finesse spectrale, présence de bruit, signaux non-stationnaires. Cette expérience nous a permis de conclure à deux voies d'estimation envisageable :

- l'estimation sur un intervalle temporel court de manière à justifier une hypothèse de stationnarité locale du signal ; dans ce cas le meilleur estimateur est du type moindre carré, permettant la prise en compte de la faible résolution spectrale ;
- l'estimation sur un intervalle temporel de durée plus longue à l'aide d'un estimateur permettant la prise en compte des non-stationnarités ; dans ce cas le meilleur estimateur, parmi ceux considérés, est celui fondé sur la mesure de distorsion du spectre.

Le deuxième choix que nous avons effectué est l'estimation des paramètres du modèle sinusoïdal sur un intervalle temporel plus long à l'aide de notre estimateur non-stationnaire.

La détection des composantes sinusoïdales a ensuite été étudiée sous l'angle d'une erreur de modélisation et d'une erreur de spécification. Nous avons montré que la valeur de l'erreur de modélisation dépend non seulement du contenu du signal (présence ou absence d'une sinusoïde dans la région considérée, rapport signal à bruit, variations du signal), mais également de la définition d'une largeur temporelle et fréquentielle d'observation. De ce fait, son utilisation en dehors d'hypothèses sur localisation des composantes est rendue difficile.

Nous avons proposé l'utilisation d'une erreur de spécification, mesurant à quel point la spécification d'un modèle sinusoïdal correspondant à un ensemble d'observations (estimations à un instant donné) peut être étendu à un autre ensemble d'observations (estimations à l'instant suivant). Cette erreur est proposée sous l'angle du problème de création de trajets sinusoïdaux. Un algorithme reposant sur une double contrainte de régularité - critère de courbure du trajet d'amplitude et de fréquence et comparaison des trajets de phase - a été proposé.

Chapitre 5 : Caractérisation des signaux périodiques/harmoniques

La caractérisation des signaux en périodicité/harmonicité repose sur une mesure de la répétition temporelle d'une forme d'onde ou sur une mesure de la répartition fréquentielle des composantes sinusoïdales. A ces deux interprétations correspondent la méthode PSOLA et la synthèse sinusoïdale harmonique. De manière similaire, deux algorithmes d'estimation de la période/fréquence fondamentale ont été utilisés : la méthode de l'auto-corrélation et la méthode du maximum de vraisemblance. Nous définissons un coefficient de voisement

mesurant l'erreur de modélisation dans le cas d'un signal harmonique, et un coefficient d'inharmonicité mesurant le décalage des fréquences des composantes du signal par rapport à un modèle purement harmonique.

Chapitre 6 : Marquage des singularités périodiques

La localisation des singularités et la caractérisation des signaux en périodicité/harmonicité nous permet le positionnement de marqueurs PSOLA utilisés pour le découpage du signal en forme d'onde élémentaire . Deux algorithmes sont proposés, afin de satisfaire aux contraintes de périodicité et de localisation d'énergie :

- un algorithme procédant de manière itérative par satisfaction de chacune des contraintes successivement,
- un algorithme prenant en compte simultanément les deux contraintes par minimisation d'un critère d'erreur quadratique.

Deuxième partie : Modification du signal

La deuxième partie de cette recherche a été consacrée à l'étude des méthodes de modification du signal et des améliorations et extensions possibles de ces méthodes.

Chapitre 7 : Modification du signal par la méthode PSOLA

La modification du signal par la méthode PSOLA ne s'applique qu'aux parties périodiques du signal. Dans ce cas, les modifications applicables au signal sont du type modification de l'axe temporel et transposition du signal. Les modifications de l'axe temporel du signal peuvent être améliorées par l'utilisation de techniques d'interpolation des formes d'onde élémentaires . Nous proposons un algorithme d'interpolation fréquentielle des formes d'onde amenant à de meilleurs résultats que l'algorithme d'interpolation temporelle habituellement utilisé. La méthode PSOLA préserve l'enveloppe spectrale du signal. Ceci peut être vu comme un avantage, mais il peut être intéressant de rendre possible la modification de cette enveloppe de manière simple. Nous appliquons l'algorithme FD-PSOLA et proposons l'algorithme FS-PSOLA de manière à permettre une dilatation et une transposition de l'enveloppe spectrale. L'algorithme PSOLA peut être étendu de manière à permettre la modifications des régions non-périodiques du signal ou des régions mixtes. Dans ce cas, une composante de phase aléatoire est ajoutée au spectre de manière proportionnelle à une fonction de voisement définie.

Chapitre 8 : Modification du signal en synthèse par addition de sinusoïdes

Nous étudions les améliorations apportées à la synthèse sinusoïdale dans le cas de signaux harmoniques. En particulier, nous étudions les améliorations permettant de préserver l'enveloppe spectrale du signal et sa forme d'onde. Dans ce dernier cas, nous étudions l'analyse/synthèse synchrone à la période fondamentale, la décomposition du modèle en contribution de la source et du filtre, et l'utilisation du retard de phase relatif. Nous rapprochons ces différentes méthodes et proposons une méthode utilisant le retard de groupe relatif permettant le rapprochement avec la méthode PSOLA.

Chapitre 9 : Synthèse hybride

Dans ce dernier chapitre, nous proposons une méthode hybride de modification du signal, baptisée SINOLA, permettant le choix du modèle (parmi les deux modèles étudiés) le plus adapté aux caractéristiques locales du signal. Nous définissons un ensemble de régions du signal correspondant aux régions voisées, chacune d'elles étant décomposée en trois bandes

de fréquence. Dans chaque région, une mesure du caractère “singularité” de la forme d'onde et du caractère inharmonique du signal est utilisée, afin de déterminer l'adéquation de l'algorithme PSOLA à la modification du signal. La sinusoïdalité des composantes du signal est déterminée par l'algorithme de création de trajets fondé sur la mesure de l'erreur de spécification. Une méthode de modification du signal est alors attribuée à chaque région. Le signal est séparé en deux parties par soustraction de la contribution des régions attribuées au modèle sinusoïdal. Le signal différence est modifié par l'algorithme PSOLA et VUV-PSOLA selon les caractéristiques de voisement local. Le signal différence modifié ainsi que la re-synthèse sinusoïdale avec modification sont recombinés pour former le signal modifié.

Revenons à notre thèse :

▷ *L'utilisation conjointe de plusieurs modèles de signaux, par choix d'un modèle adapté aux caractéristiques locales du signal sonore, plutôt que le perfectionnement indéfini d'un même modèle, permet d'atteindre des modifications de grande qualité.*
Ce premier point à été montré dans la partie 9 pour l'utilisation de deux modèles de signaux fondés l'un sur une décomposition temporelle du signal, l'autre sur un décomposition fréquentielle.

▷ *La phase contient une information pertinente pour l'analyse des signaux sonores (localisation temporelle, localisation fréquentielle, synchronie des évènements fréquentiels).*
Ce deuxième point a été montré dans les parties 3 et 4 de cette recherche, pour la localisation temporelle des singularités du signal, pour l'estimation des paramètres du modèle sinusoïdal, ainsi que pour la création de trajets temporels de sinusoïdes.

▷ *La prise en compte des relations de phase lors de modifications d'un signal permet d'atteindre un haut niveau de qualité sonore.*
Ce troisième point a été montré dans la partie 8 dans le cadre d'une analyse/synthèse non-synchrone à la période fondamentale, par respect du retard de phase relatif et du retard de groupe relatif.

Perspectives

Les perspectives sont généralement nombreuses en fin de thèse et nous ne dérogeons pas à la règle.

Ces perspectives concernent autant les méthodes de caractérisation que celles de synthèse.

Prédiction linéaire : la méthode de prédiction linéaire utilisée dans ce travail est une méthode de Burg. Dans le cas de la parole, il est cependant possible de tirer parti de la détection des Instants de Fermeture de la Glotte, puisque ceux-ci correspondent théoriquement aux instants de rupture du modèle auto-régressif. L'estimation s'effectue alors sur les segments disjoints correspondant aux segments de meilleure prédiction. Des premiers pas ont été effectués dans cette direction, mais nous ne sommes pas encore parvenus à exploiter pleinement cet avantage.

Estimation des paramètres de modulation du modèle sinusoïdal : la méthode d'estimation des modulations du paramètre du modèle sinusoïdal proposée repose sur une observation locale du spectre. En ce sens, cette méthode n'est pas robuste aux faibles résolutions spectrales. La reformulation du modèle sous forme d'un problème de minimi-

sation d'erreur globale pourrait la rendre robuste. Cependant, du fait de l'utilisation de la fonction logarithme et arc tangente, cette résolution n'est pas linéaire. Une solution itérative pourrait donc être envisagée.

Algorithme de création de trajets : L'algorithme de création de trajets proposé repose sur une erreur de spécification. Cette erreur mesure l'erreur commise en prolongeant le modèle, évalué à un instant donné, à l'instant suivant. Dans notre algorithme, les trajets sont créés par maximisation de probabilités locales et non globales. Nous justifions cela par le fait que la probabilité cumulée d'un trajet est mal définie dans le cas d'un signal où les différentes composantes fréquentielles n'apparaissent pas simultanément. Une solution consisterait toutefois à appliquer les probablitiés de non-existence de trajets de [Gar92].

Description des transitoires : la description des transitoires est assez sommaire dans notre recherche. Ceci provient du fait que nous n'avons pas considéré la modification des transitoires comme pertinente, puisque non naturelle dans le cas de la parole. Il serait cependant intéressant d'introduire dans notre méthode une modélisation de ces transitoires afin d'étendre la représentation des classes de sons.

Synthèse hybride : la méthode SINOLA proposée ne correspond qu'à une partie de nos ambitions initiales. Elle répond aux décompositions en forme d'onde élémentaire et en sinusoïde. Une décomposition non pas en forme d'onde élémentaire mais en Formes d'Ondes Formantiques (FOF) devrait permettre un meilleur contrôle du modèle. Plutôt qu'une estimation des paramètres des FOFs, une localisation temporelle et fréquentielle des FOFs devrait permettre l'application d'un découpage non pas seulement temporel (décomposition en forme d'onde élémentaire) mais également fréquentiel (décomposition en bandes d'onde formantique). Le signal est reconstruit dans ce cas par superposition/addition tant sur l'axe temporel que sur l'axe fréquentiel, donnant la possibilité de modifier la position des formants du signal. La méthode SINOLA est dans ce cas utilisée afin de déterminer le modèle le plus approprié pour représenter chacune des FOFs. Des premiers pas ont été effectués dans cette direction, mais se pose cependant le problème de la robustesse du modèle du fait de la nécessité de déterminer des trajets de formants se croisant et disparaissant dans le temps.

Ondelette : Cette recherche repose sur la représentation temps/fréquence de la Transformée de Fourier à Court Terme. Cette représentation, choisie dés le départ, n'est cependant pas la plus adaptée à la localisation temporelle et fréquentielle conjointe utilisée lors de notre recherche. L'utilisation des représentation temps/échelle [Mal99] [Ld89] reste certainement un domaine à explorer pour l'évolution de notre modèle.

Voilà autant de perspectives constituant des domaines à parcourir encore plus vaste que le chemin déjà parcouru dans le but de décrire cet objet complexe que constitue le signal sonore.

Troisième partie

Annexes

Annexe A

Applications de la recherche de thèse

Quatre applications utilisant la recherche effectuée dans cette thèse ont été réalisées.

- Les deux premières ont été réalisées grâce aux programmes d'analyse et de synthèse que nous avons développés dans le cadre de nos recherches.
- Les deux suivantes n'utilisent que la partie analyse de nos programmes. La synthèse est réalisée en temps-réel dans l'environnement JMax. L'objet «PAGS» développé par Norbert Schnell est une implémentation de l'algorithme PSOLA sur JMax.

A.1 Post-production pour le film «Vatel» de Roland Joffé

Fiche technique du film :

Réalisation	Roland Joffé
Production	Légende Entreprises, Gaumont
Date de sortie	10/05/2000
Durée	01 :57 :00
Distribution	Gérard Depardieu, Uma Thurman, Tim Roth, ...
Musique	Ennio Morricone

L'objectif du travail est de corriger l'accent anglais de l'acteur Gérard Depardieu.

La méthode de synthèse utilisée est du type FDI-PSOLA-WD. La méthode d'analyse est la méthode GDR (voir partie 3) suivie de l'algorithme de satisfaction de contrainte par minimisation de l'erreur quadratique (voir partie 6).

analyse	marquage sur l'énergie du signal résiduel (GDR) algorithme moindre carré f0/énergie
synthèse	TDI-PSOLA, FDI-PSOLA
modifications	modification de durée, de hauteur et d'énergie
manipulation	Les paramètres sont changés manuellement.

A.2 Post-production pour le film «Vercingétorix» de Jacques Dorfmann

Fiche technique du film :

Réalisation	acques Dorfmann
Production	Tesson
Date de sortie	24/01/2001
Durée	02 :02 :00
Distribution	Christophe Lambert, Klaus Maria Brandauer, Max Von Sydow, ...
Musique	Pierre Charvet

Le travail consiste à transformer l'accent germanophone de Klaus Maria Brandauer en un accent français. Pour cela, la voix d'un locuteur français est utilisée comme cible. Un ensemble de programme permettant de modifier la prosodie (hauteur, durée, intensité, timbre) de manière semi-automatique afin de l'ajuster à celle d'un locuteur français a été réalisé.

La synthèse utilisée est du type LP-FD-FDI-PSOLA-WB. La méthode d'analyse est la méthode GDR suivie de l'algorithme de satisfaction de contrainte par minimisation de l'erreur.

analyse	marquage sur l'énergie du signal résiduel (GDR) algorithme moindre carré f0/énergie
synthèse	TDI-PSOLA, FDI-PSOLA, LP-PSOLA, filtrage croisé
modifications	modification de durée, de hauteur, d'énergie, de spectre, filtrage croisé, dé-voisement
manipulation	création de programmes permettant l'ajustement des paramètres d'un fichier source vers un fichier cible (ajustement de la durée, de la hauteur, de l'enveloppe spectrale)

A.2.1 Interfaces d'ajustement de la prosodie

A.2.1.1 Ajustement des durées

Le programme prend en entrée deux fichiers sons : le fichier original qui est à modifier (abscisse), le fichier cible (ordonnée) Le programme nécessite également deux fichiers de segmentation. Chacun des deux fichiers est segmenté en unités dont la signification au sein de la phrase est semblable (les deux locuteurs n'utilisent pas le même vocabulaire). La segmentation est manuelle.

Les paramètres de contrôle du programme sont

- alpha [0,1] : permet le passage progressif entre la durée du segment original et celle du segment cible
- dsmooth : paramètre de lissage permettant d'éviter les transitions brusques entre segments du à des facteurs de dilatation différents à chaque segments. Le programme génère des fonctions par points (breakpoint functions ou encore bpf). Le programme de synthèse interpole cette bpf aux instants de synthèse.
  - Lorsque dsmooth=0, les temps de dilatation de la bpf sont placés aux extrémités inférieure et supérieure du segment (sinf+0.005 et ssup-0.005 ou sinf et ssup sont les temps correspondant aux extrémités inférieure et supérieure d'un segment).

– Lorsque dsmooth¿0, les temps sont éloignés des extrémités du segment de manière à permettre une interpolation (de la bpf de dilatation) lors des transitions. Dans ce cas, le facteur de dilatation est recalculé de manière à obtenir la modification correcte de temps :

```
dc1 = segment(l-1).dilat_bp(2,2);
dc2 = segment(l).dilatation;
m1  = segment(l).initdebut;
m2  = segment(l).initfin;
d = (dc2*(m2-m1)-dc1*(x-m1)/2)/((x-m1)/2+(m2-x));
```

A.2.1.2 Ajustement des hauteurs

L'ajustement de la fréquence fondamentale est plus compliqué puisque l'application, telle quelle, de la fréquence fondamentale du fichier cible sur le fichier source conduit à introduire les variations fines (variations d'une période à l'autre) de f0 du fichier cible dans le fichier source. Ceci produit un son rugueux. A l'inverse, notre but est d'appliquer le "contour" musical du fichier cible au fichier source. Pour cela nous calculons au sein de chaque segment le contour mélodique du segment par une composante continue M de fréquence et un polynôme P d'ordre q. Ces deux paramètres sont ajustés par régression et calculés pour le fichier source (Ms et Ps) et le fichier cible (Mc et Pc). L'ajustement s'effectue ensuite à partir des paramètres Ms, Ps, Mc et Pc et donne lieu à la création d'une bpf de transposition du fichier source.

paramètre de contrôle :

- Q : ordre du polynôme pour Ps et Pc (en pratique nous utilisons un polynôme d'ordre 2),
- beta [0,1] : règle le passage de la moyenne de fréquence Ms vers Mc,
- gamma [0,1] : règle le passage du polynôme Ps vers Pc,
- tsmooth : paramètre de lissage. Comme dans le cas de la dilatation, le lissage par éloignement des temps de la bpf des extrémités du segment.

A.2.2 Ajustement de l'enveloppe spectrale

L'objectif de cette transformations est de changer la sonorité de certaines voyelles. La modification est effectuée par un algorithme de type LP-PSOLA. Comme pour le traitement de la prosodie, deux fichiers sont utilisés : un fichier source et un fichier cible. La succession de filtre AR (modélisation AR par la méthode de Burg) est estimée pour les deux fichiers. Un ensemble de segments correspondants aux voyelles à modifier est déterminé sur le fichier source. Les segments équivalents sont déterminés sur le fichier cible. Une bpf, de valeur comprise entre 0 et 1 permet le passage pour chaque voyelle du filtre source vers le filtre cible. L'interpolation des filtres (tant l'interpolation entre filtre temporellement que le passage d'un filtre source vers un filtre cible) est effectué sur les coefficients Log-Area Ratio du filtre.

A.3 Création d'un choeur virtuel pour l'opéra «K ...» de Philippe Manoury

Le projet de création du choeur virtuel a fait l'objet de la publication suivante [SP00]. Le lecteur intéressé par un complément d'information pourra s'y référer.

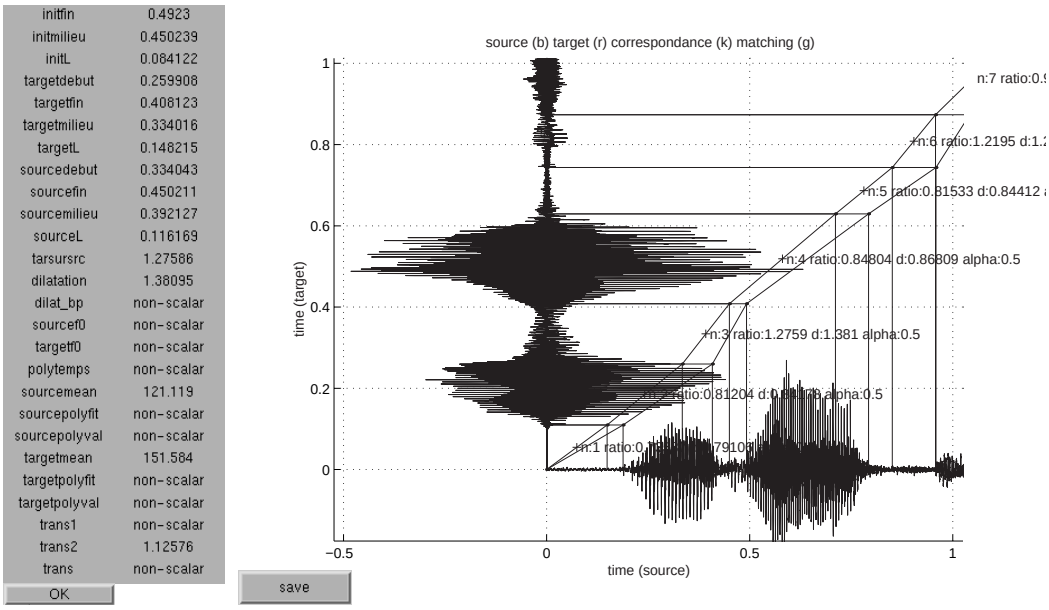


FIG. A.1 – Interface de modification de durée utilisé pour le film "Vercingétorix" : abscisse signal source, ordonnée signal cible, correspondance des temps des segments

Le projet s'insère dans la création d'un nouvel opéra de Philippe Manoury (commande de l'opéra Bastille). L'opéra est fondé sur l'oeuvre de Kafka «Le procès». L'opéra s'intitule «K ...». Le compositeur désire intégrer dans son opéra un choeur virtuel permettant la synthèse de l'effet d'un choeur naturel mais également la construction d'un choeur de sonorité inouïe.

Pour la première tâche, un enregistrement du choeur de l'opéra Bastille est effectué. Les voix sont enregistrées par groupes (basses, sopranes, ...) ainsi que par voix individuelles afin de voir l'évolution de l'effet de choeur en fonction du nombre de voix. Différentes techniques d'analyse/synthèse sont essayées afin de permettre la restitution de l'effet de choeur. Finalement, le choix se porte sur l'utilisation d'une multitude de synthétiseur PSOLA prenant en charge, chacun, la synthèse d'une voix. Les paramètres des synthétiseurs sont contrôlés par un interface de plus haut niveau permettant la modification du choeur en terme de désynchronie des voix, variations de hauteurs, variations de vitesse de vibrato, changement de timbre, dé-voisement, ...

Ce travail est effectué en collaboration avec Norbert Schnell et Serge LeMouton.

analyse	marquage sur l'énergie du signal (GDS) (voix chantée de fréquence élevée) algorithme moindre carré f0/énergie
synthèse	objet «PAGS» sous JMAX : TD-PSOLA, TDI-PSOLA, séparation partie voisée/non-voisée
manipulation	paramètres de haut niveaux

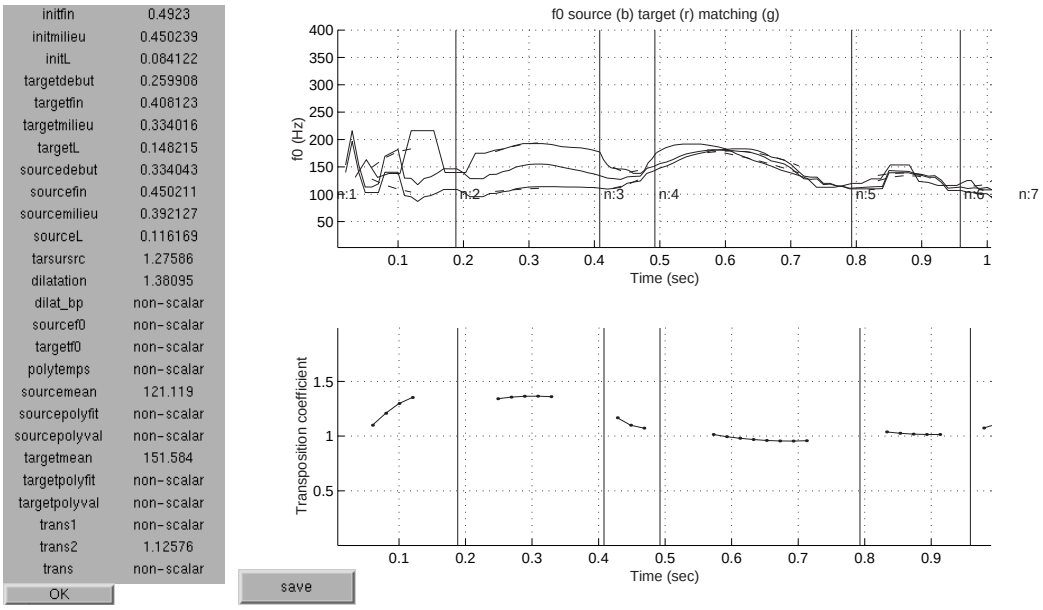


FIG. A.2 – Interface de modification de hauteur pour le film "Vercingétorix" : [H] tracé des fréquences fondamentales des fichiers source et cible, fréquence fondamentale interpolée. [B] fonction de transposition crée

A.4 Réalité virtuelle "Elle et la voix" de Catherine Ikam et Louis-François Fléri, musique de Pierre Charvet

«Rencontre avec un personnage virtuel «Elle», qui n'existe que dans la mémoire de l'ordinateur. «Elle» est une réplique numérisée dotée de modèles de comportement. La présence et le déplacement des visiteurs dans la pièce sont analysés en temps réel et permettent une interaction avec «Elle». Une étrange rencontre à la fois sonore et visuelle, une ébauche de dialogue. En l'absence de visiteur, «Elle» développe une vie propre, aléatoire. Entre plusieurs visiteurs «Elle» choisit celui auquel elle sourit. Les visiteurs de l'installation sont invités à interagir avec «Elle». Captée en temps réel, leur voix va modifier celle de "Elle", à l'aide de nouvelles techniques de synthèse de la voix et de spatialisation développées à l'Ircam.»

L'objectif de l'artiste est de permettre l'interaction du spectateur avec un personnage virtuel «Elle» représenté visuellement en image de synthèse et auditivement par la méthode PSOLA (objet «PAGS» sous JMax). Le spectateur interagit au travers d'un microphone et d'un capteur de position. L'ordinateur enregistre la position, le signal, caractérise le signal par l'intensité, la vitesse et la hauteur et répond auditivement au spectateur par déclenchement et transformation d'une voix pré-enregistrée. L'analyse des voix pré-enregistrées est effectuée de manière préalable par la méthode GDS suivie de l'algorithme de satisfaction de contrainte par minimisation de l'erreur.

analyse	marquage sur le signal (GDS) (voix parlée de fréquence élevée) algorithme moindre carré f0/énergie
synthèse	objet «PAGS» sous JMAX : TD-PSOLA, TDI-PSOLA
mdofications	modification de durée, de hauteur
manipulation	Les paramètres sont changés par l'interaction du spectateur dans un microphone : détection d'énergie et de fréquence fondamentale

Annexe B

Articles Opera-K

Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)

Norbert Schnell, Geoffroy Peeters, Serge Lemouton, Philippe Manoury, Xavier Rodet
{schnell,peeters,lemouton,manoury,rod}@ircam.fr
IRCAM - CENTRE GEORGES-POMPIDOU
1, pl. Igor Stravinsky, F-75004 Paris, France
http://www.ircam.fr

ABSTRACT

The paper presents a method to synthesize a choir in real-time and its application in the framework of an opera production. It intentionally integrates artistic considerations with research and engineering matters, thus giving a complete picture of a concrete collaboration in the context of the creation of electronic music.
The synthesis of the "virtual choir" is implemented for the jMax real-time sound processing system using the Pitch Synchronous Overlap Add (PSOLA) technique. The synthesis algorithm derives multiple voices of a same group from a single recording of a real choir singer. The first stage of the analysis segments harmonic, non harmonic and transient parts of the signal. The second stage places PSOLA markers in the harmonic parts by a novel two-steps algorithm.
The synthesis algorithm allows various transformations of the analysed sound of a single voice by the introduction of stochastic as well as deterministic variations. It is controlled by an extended set of parameters and results in a wide range of different timbres and textures in addition to those of a realistic choir sound.
The last section of the paper is dedicated to the application of the algorithm in the context of the composition and its integration into the rest of the environment of the opera production. It describes the experiments with the recordings of a choir and the work in the production studio using the jMax environment. Finally a set of commented examples is associated with the paper, which will be presented during the paper session.

1 INTRODUCTION

The opera "K..." and the concept of the virtual choir

Since spring 1998 Philippe Manoury is working on the composition of the opera "K..." based on Franz Kafka's novel "Der Prozess" which will have its premiere in march 2001 at the Opera Bastille in Paris. The work has an important electro-acoustic part, which is entirely implemented in *jMax* [Déchelle et al., 1998] [Déchelle et al., 1999a] and realized at IRCAM with the musical assistance of Serge Lemouton.

For several scenes of this Opera (such as the trial) Manoury has expressed the need for choral voices evoking the notion of crowd. This led to the concept of a *virtual choir*.

The goal was to create an algorithm which is able to realistically reproduce the sound of a choir, permitting sounds unusual or impossible for a real choir. It was decided to evaluate several technical possibilities. Although there is a lot of research on synthesis methods for a single voice [Sundberg, 1987] [Ternstr'on, 1989], the domain of vocal ensemble synthesis is not much explored .

After some unsatisfying trials to obtain a choir sound with various techniques such as granular synthesis, modified additive synthesis or various chorus effects it was found that the only way to obtain the realistic notion of a choir would be by superposition of multiple well enough distinguishable solo voices.

This assumption leads to the following two questions:

1. How to efficiently synthesize a single voice allowing a wide range of transformations?
2. Which individual variations should be attributed to each voice in order to obtain a chorus effect when superposing them?

The answer to the first question was found in the *PSOLA* technique described in the first part of this paper. The second part part of the paper explains the real-time algorithm implemented for the synthesis of a group of voices proposing an answer to the second question. The paper concludes with the experiments made during the research on the virtual choir and its integration into the opera.

2 PSOLA

PSOLA (**P**itch **S**ynchronous **O**ver**L**ap-**A**dd [Charpentier, 1988] [Moulines and Charpentier, 1990]) is a method based on the decomposition of a signal into a series of elementary waveforms in such a way that each waveform represents one of the successive pitch periods of the signal and the sum (overlap-add) of them reconstitues the signal.

PSOLA works directly on the signal waveform without any sort of model and therefore does not lose any detail of the signal. But in opposition to usual sampling, PSOLA allows independent control of pitch, duration and formants of the signal.

One of the main advantages of the PSOLA method is the preservation of the spectral envelope (formant positions) when pitch shifting is used. High-quality transformations of signals can be obtained by time manipulation only, therefore with very low computational cost. For a simultaneous modification of pitch and spectral envelope, a **F**requency **S**hifting (*FS*-PSOLA [Peeters and Rodet, 1999]) method has been proposed.

PSOLA is very popular for speech transformation because of the properties of the speech signal. Indeed, PSOLA requires the signal to be harmonic and well-suited for a decomposition into elementary waveforms by windowing, which means that the signal energy must be concentrated around one instant inside each period.

The PSOLA method can be understood as

- granular synthesis in which each "grain" corresponds to one pitch period
- synthesis based on a source/filter model like *CHANT* [d'Alessandro and Rodet, 1989]: the elementary waveforms can be considered as an approximation of the *CHANT Formant Waveforms* but without explicit estimation of source and filter parameters

G. Peeters has developed a PSOLA analysis and synthesis package described in the following.

2.1 Time/Frequency signal characterization

By its definition, the PSOLA method allows only modification of the periodic parts of the signal. It is therefore important to estimate which parts of the signal are periodic, which are non-periodic and which are transient. In the case of the voice, the periodic part of the signal is produced by the vibration of the vocal chords and is called "voiced".

At each time instant t, a "voicing" coefficient $v(t)$ is estimated. This coefficient is obtained by use of the "Phase Derived Sinusoidality measure" from *SINOLA* [Peeters and Rodet, 1999]. For each time/frequency region, the instantaneous frequency is compared to the frequency measured from spectrum peaks. If they match, the time/frequency region is said to be "sinusoidal". If for a specific time most regions of the spectrum are sinusoidal, this time frame is said to be "voiced" and is therefore processed by the PSOLA algorithm.

2.2 PSOLA analysis

PSOLA analysis consists of decomposing a signal $s(t)$ into a series of elementary waveforms $s_i(t)$. This decomposition is obtained by applying analysis windows $h(t)$ centered on times m_i:

$$s_i(t) = h(t - \mathrm{m}_i)s(t) \tag{1}$$

The m_i, called "markers", are positioned [Peeters, 1998]

- pitch-synchronously, i.e. the difference $\mathrm{m}_i - \mathrm{m}_{i-1}$ is close to the local fundamental period [Kortekaas, 1997],
- close to the local maxima of the signal energy. This last condition is required in order to avoid deterioration of the waveform due to the windowing.

After estimating the signal pitch period $\mathrm{T0}(t)$ and the signal energy function $e(t)$, the markers m_i are positionned using the following two-step algorithm.

Step 1: Estimation of the local maxima of the energy function

Because PSOLA markers m_i must be close to the local maxima of the energy function, the first step is the estimation of these maxima.

Let us define a vector of pitch instants $\Theta_l = [\theta_{l,0}, \theta_{l,1}, ..., \theta_{l,i}, ...]$ such that $\theta_{l,i} - \theta_{l,i-1} = \mathrm{T0}_{i-1}$ (see Figure 1). Around each instant $\theta_{l,i}$ let us define an interval $I_{l,i} = \left[\theta_{l,i} - \frac{\mathrm{T0}_{i-1}}{\alpha}, \theta_{l,i} + \frac{\mathrm{T0}_i}{\alpha}\right]$, where α controls the extent of the interval. Inside each interval $I_{l,i}$, the maximum of the energy is estimated and noted $t_{l,i}$. For each vector Θ_l, i.e. for each choice of starting time $\theta_{l,0}$, the sum of the values of the energy function at the times $t_{l,i}$, $\sigma_l = \sum_i e(t_{l,i})$, is computed. Finally the selected maxima τ_i are those of the vector Θ_l which maximize σ_l: $\tau_i = t_{l',i}$ with $l' = \arg\max_L \sigma_l$.

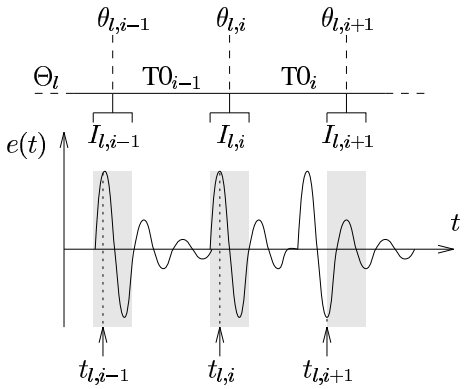


Figure 1: Estimation of the local maxima of the energy function

Step 2: Optimization of periodicity and energy criterions

Because PSOLA markers m_i must be placed pitch-synchronously and close to the local maxima, the two criteria have to be minimized simultaneously.

A novel least-squares resolution is proposed, as follows:

Let m_i denote the markers we are looking for, τ_i the time locations of the local maxima of the energy function estimated at the previous stage, $\mathrm{T0}_i$ the fundamental period at time τ_i. A least-squares resolution is used in order to minimize the periodicity criterion (distance between two markers close to the fundamental period: $\mathrm{m}_i - \mathrm{m}_{i-1} \simeq \mathrm{T0}_{i-1}$) and energy criterion (markers close to the local maxima of energy: $\mathrm{m}_i \simeq \tau_i$). The quantity to be minimized is $\epsilon = \sum_i((\mathrm{m}_i - \mathrm{m}_{i-1}) - \mathrm{T0}_{i-1})^2 + \beta(\mathrm{m}_i - \tau_i)^2$. β is used to weigh the criteria: $\beta < 1$ favours periodicity while $\beta > 1$ favours energy.

If the vector of markers is $\overline{\mathrm{m}} = [\mathrm{m}_0\ \mathrm{m}_1\ ...\ \mathrm{m}_i\ ...\ \mathrm{m}_{N-1}\ \mathrm{m}_N]^T$, the optimal marker positions are obtained by

$$\overline{\mathrm{m}} = M^{-1} \begin{pmatrix} 0 & -\mathrm{T0}_0 & +\gamma\tau_0 \\ \mathrm{T0}_0 & -\mathrm{T0}_1 & +\beta\tau_1 \\ & \vdots & \\ \mathrm{T0}_{i-1} & -\mathrm{T0}_i & +\beta\tau_i \\ & \vdots & \\ \mathrm{T0}_{N-2} & -\mathrm{T0}_{N-1} & +\beta\tau_{N-1} \\ \mathrm{T0}_{N-1} & 0 & +\gamma\tau_N \end{pmatrix} \tag{2}$$

where M is a tri-diagonal matrix, with main diagonal $[1+\gamma\ 2+\beta\ ...2+\beta\ ...\ 2+\beta\ 1+\gamma]$ and lower and upper diagonal $[-1\ -1\ ...\ -1\ ...\ -1\ -1]$ where γ is used for specific border weighting.

2.3 PSOLA Synthesis

2.3.1 Voiced parts

For the voiced parts, PSOLA synthesis proceeds by overlap-add of the waveforms $s_i(t)$ re-positionned on time instants $\tilde{\mathrm{m}}_j$ (see Figure 2):

$$\begin{cases} \tilde{s}_j(t) = s_i(t + \mathrm{m}_i) \\ \tilde{s}(t) = \sum_j \tilde{s}_j(t - \tilde{\mathrm{m}}_j) \end{cases} \tag{3}$$

where m_i are the PSOLA markers which are the closest to the current time in the input sound file.

A modification of the pitch of the signal from T0(t) to T(t) is obtained by changing the distance between the successive waveforms: $\tilde{m}_j - \tilde{m}_{j-1} = T(t)$. In the usual PSOLA, time stretching/compression is obtained by repeating/skipping waveforms.

However, in case of strong time-stretching, the repetition process produces signal discontinuities. This is the reason why a *TDI*-PSOLA (**T**ime **D**omain **I**nterpolation PSOLA) has been proposed [Peeters, 1998]. TDI-PSOLA proceeds by overlap-add of continuously interpolated waveforms:

$$\begin{cases} \tilde{s}_j(t) = \alpha s_i(t + m_i) + (1-\alpha) s_{i-1}(t + m_{i-1}) \\ \alpha = (\hat{m} - m_{i-1})/(m_i - m_{i-1}) \\ \tilde{s}(t) = \sum_j \tilde{s}_j(t - \tilde{m}_j) \end{cases} \quad (4)$$

where m_{i-1} and m_i are the PSOLA markers which frames the current time, $\hat{m}$, in the input sound file.

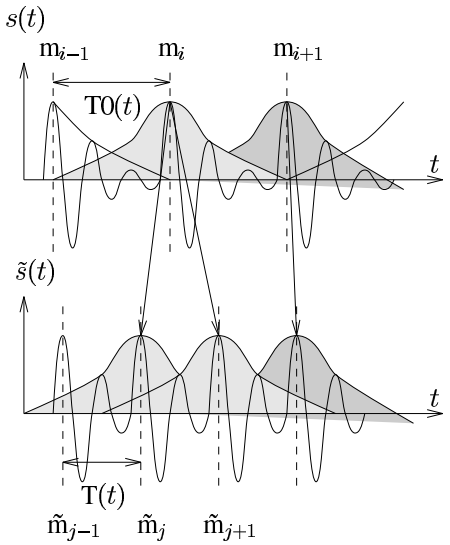


Figure 2: Example of pitch-shifting and time stretching using PSOLA

2.3.2 Unvoiced parts

Unvoiced parts of signals are characterized by a relatively weak long-term correlation (no pitch period) while a short-term correlation is due to the (anti)resonances of the vocal tract.

Special care has to be taken in order to avoid introducing artificial correlations in these parts, which would be perceived as artificial tones ("flanging effect").

Several methods [Moulines and Charpentier, 1990] [Peeters and Rodet, 1999] has been proposed in order to process the unvoiced part while keeping the low computational-cost advantage of the OLA framework. These methods use various techniques to randomize the phase, in order to reduce the inter-frame correlation.

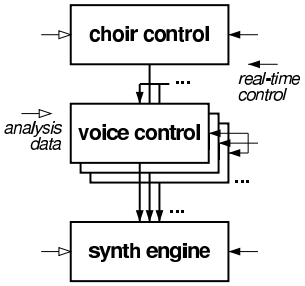


Figure 3: Stages of the voice group synthesis module

3 SYNTHESIZING A GROUP OF VOICES IN REAL-TIME

It was decided to apply a PSOLA resynthesis on recordings of entire phrases of singing solo voices.

In addition to the PSOLA markers determined by the analysis stage two levels of segmentation were manually applied to the recorded phrases:

- pitched notes according to the original score
- segments of musical interest for the process of resynthesis such as phonemes, words and phrases

A synthesis module for *j*Max [Déchelle et al., 1999b] [IRCAM, 2000] was designed, which reads the output of the analysis stage as well as the original sound file and performs the synthesis of a group of individual voices. It was decided to "clone" a whole group of voices from the same sound and analysis data file.

The chosen implementation of the voice group synthesis module shown in figure 3 divides the involved processes into three stages. The first stage determines the parameters, which are common to a group of voices derived from the same analysis data. The parameters are the common pitch and the onset position within an analyzed phrase.

The second stage contains for each voice a process applying individual modulations to the output of the first stage, which causes the voices not to be synchronous and assures that each voice is distinguished from the others. The third stage is a synthesis engine common to all voices performing an optimized construction of the resulting sound from the parameter streams generated by the voice processes of the second stage.

3.1 A PSOLA real-time synthesis algorithm

In the simplest case, the output of the analysis stage is a vector of increasing time values m_i each of them marking the middle of an elementary wave form. For simplicity non-periodic segments are marked using a constant period.

The real-time synthesis algorithm reads a marker file as well as the original sound file. It copies an elementary waveform from a given onset time m_i defined by a marker, applies a windowing function and adds it to the output periodically according to the desired frequency. The fundamental frequency can be either taken from the analysis data as $f0 = \frac{1}{m_{i+1} - m_i}$ or determined as a synthesis parameter of arbitrary value[1].

[1] It is evident that the higher the frequency - or better, the ratio between the orig-

An analysis file can be understood as a pool of available synthesis spectra linearly ordered by their appearance in a recorded phrase[2]. The onset time determines the synthesized spectrum.

In general the onset time and the pitch are independent synthesis parameters so that time-stretching/compression can be easily obtained by moving through the onset times with an arbitrary speed. Modifications of the pitch can be performed simultaneously. The variable increment of the onset time (i.e. speed) represents an interesting synthesis parameter as an alternative to the absolute onset time. The *TDI*-PSOLA (see 2.3.1) interpolation produces a smooth development of timbre for a wide range of onset speeds including extremely slow stretching.

3.2 Resynthesis of unvoiced segments

A first extension of the synthesis algorithm described in the previous section uses the voicing coefficient $v(t)$ output from the analysis stage. The coefficient $v(t)$ indicates whether the sound signal at time t is voiced or unvoiced.

PSOLA synthesis is used for voiced sound segments only. For the synthesis of unvoiced segments a simple granular synthesis algorithm is used [Schnell, 1994]. Grains of constant duration are randomly taken from a limited region around the current onset time. The amount of the onset variation and an overlapping factor are parameters which can be controlled in real-time.
Signal transients are treated in the same way as unvoiced segments.

In order to amplify and attenuate either the voiced or the unvoiced parts, the output of the synthesis stage can be weighted with an amplitude coefficient $c(t)$ calculated from the voicing coefficients by a clipped linear function:

$$c(t) = \begin{cases} 0 & : \ \frac{v(t)-a}{b-a} \leq 0 \\ 1 & : \ \frac{v(t)-a}{b-a} \geq 1 \\ \frac{v(t)-a}{b-a} & : \ else \end{cases} \qquad (5)$$

Giving adequate values for a and b for example the voiced parts can be attenuated or even suppressed so that only the consonants of a phrase are synthesized.

PSOLA synthesis as well as the synthesis of unvoiced segments can be performed by a single granular synthesis engine applying different constraints for either case. Figure 4 shows an overview of the implemented voice resynthesis engine and its control parameters.

The pitch and the onset are computed by a previous synthesis control stage which will be described below.

3.3 Original pitch modulation

Experiments with the implemented synthesis engine for a single voice like other algorithms performing time-stretching on recordings containing vibrato show undesired effects. Blind time-stretching slows down the vibrato frequency and often leads to the perception of an annoying pitch bend in the resulting sound. It is desirable to change the duration of a musical gesture while leaving the vibrato frequency untouched.

inal frequency and the synthesized frequency - the more the elementary waveforms overlap. Since the computation load of a typical synthesis algorithm depends of the number of simultaneously calculated overlapping waveforms, it increases with the synthesized frequency.

[2] Although this is convenient for the resynthesis of entire words and phrases for further applications, it could be interesting to construct differently structured feature spaces from the same analysis data.

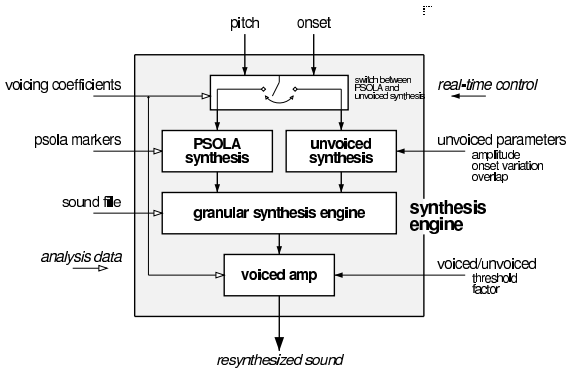


Figure 4: Synthesis engine combining PSOLA and unvoiced synthesis

For the implemented algorithm, the original pitch modulation is removed from the analysis data in two steps:

1. segmentation of the recorded singing voice into notes for voiced segments
2. determination of an averaged (note) frequency $\bar{f}0$ for each segment

An example of the segmentation of a singing voice phrase derived from the voicing coefficient, and the assignment of the note frequency according to the score is shown in figure 5.

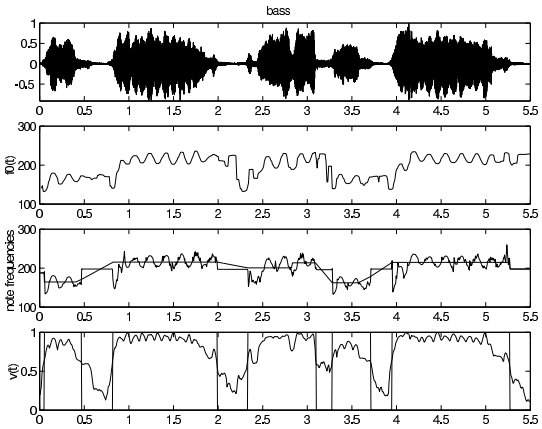


Figure 5: Note segmentation and pitch of a singing voice phrase

The note frequency is integrated into the analysis data by assigning it to each marker within a given segment representing a note. In addition, a modulation coefficient $k(t)$ is stored with each marker which contains the original pitch modulation of a note:

$$k(t) = \frac{f0(t) - \bar{f}0(t)}{\bar{f}0(t)} \qquad (6)$$

The original instantaneous frequency can be recalculated as $f0(t) = \bar{f}0(t)(1 + M \cdot k(t))$. The modulation index M determines the amount of original re-synthesized pitch modulation. This technique allows a preservation of the musical expression contained in the pitch modulation of a note when the absolute original frequency is replaced. For a modulation index of $M = 0$ the modulation is removed and can be replaced by a synthesized modulation independent of the applied time-stretching/compression. With $M > 1$ an exaggerated modulation can be achieved.

3.4 Controlling a group of voices

Figure 6 shows the control stage determining pitch and onset for the synthesis of a single voice as well as for a group of voices.

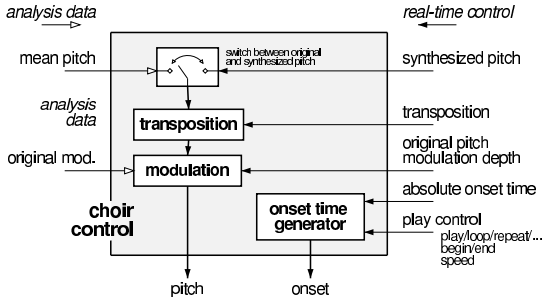


Figure 6: Pitch and onset control for a group of voices

The pitch is input from the analysis data or as real-time control parameter and a transposition (given in cent) is calculated before the original modulation. The onset time is generated by a module, which advances the onset time according to an arbitrary segmentation. A segment is specified by its begin and end time, its reading mode (play forward/backward, loop back and forth, repeat looping forward, ...) and the speed at which the onset time is advancing.

3.5 Individual variations of the voices

A major concern designing the algorithm was the variations of timbre and pitch performed by each voice in order to obtain a realistic impression of a choir by the superposition of multiple voices re-synthesized from the same analysis data.

In intensive experiments comparing synthesized groups of voices with recordings of real choir groups the following variations where found important:

- pitch variations
- timing (onset) variations
- vibrato frequency variations

The pitch and timing variations are mainly corresponding to the individual imprecision of a singer in a choir making that never two singers sing exactly the same pitch and start and end the same note at the same time. The onset variations lead as well to a diversity of the spectrum of the voices at each moment. A synthesized vibrato of an individual frequency can be added to each voice.

It was considered to give individual formant characters to each synthesis voice in order to create additional individuality close to reality. However the experiments have shown that in the context of the accompanying sound and spatialization effects, the additional computation was found to be too costly in comparison with the produced effect[3].

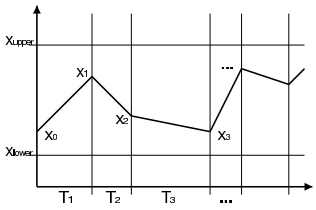


Figure 7: Example of a random break point function

The variations for each voice are performed by *random break point functions (rbpf)*. In the synthesis cycle of the algorithm an *rbpf* computes for each synthesized waveform a new value $x(t)$ on a line segment between two break-points x_i guaranteeing a smooth development of the synthesized sound (see figure 7). A new target value x_i as well as a new interpolation time T_i are randomly chosen inside the boundaries each time a target value x_{i-1} is reached.

The parameters of a general *rbpf* generator are the boundaries for the generated values (x_{lower}/x_{upper}) and for the duration (T_{lower}/T_{upper}) between two successive break-points. As an alternative to its duration as well the slope of a line segment can be randomly chosen taking in this case the minimum and maximum slope as parameters.

Using these generators a constantly changing pitch transposition, onset time and vibrato frequency can be performed. Depending on the chosen parameters this can result either in a realistic chorus effect or, when exaggerating the parameter values, a completely different impression.

A schematic overview of the modulations for each voice acting on the pitch and onset produced by the choir control module is shown in figure 8. The produced pitch and onset parameters are directly fed into the synthesis engine.

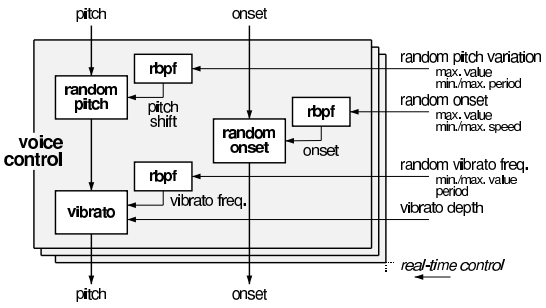


Figure 8: Individual pitch and onset variations performed for each voice

[3]The computation load for a synthesis voice using a simple re-sampling technique in order to modify its formants must be estimated as about three times as costly as a straight forward PSOLA synthesis with the same transposition or overlap ratio.

4 CONSTRUCTING THE VIRTUAL CHOIR

The implementation of the voice group synthesis module was accompanied by intensive experiments in order to adjust the synthesis algorithm and parameter values corresponding to a realistic choral sound.

The sound sources for the PSOLA analysis and further choral sounds for comparative tests were obtained in a special recording session with the choir of the Opera Bastille Paris in the *Espace de Projection* at IRCAM configured for a dry acoustic. The same musical phrases written by Manoury based on a Czech text were sung individually by the four choir sections (soprano, alto, tenor and bass) in unison. For each choir section several takes of 2, 4, 6 and 10 singers as well as a solo singer were recorded.

Various analysis tools have been tested in the research of the choir sound as a phenomenon of the superposition of single voices and their individualities as well as its particularities of the signal level.
Classical signal models (such as those used for the estimation of pitch period or spectral peaks) are difficult to apply in the case of a choir signal. The signal is composed of several sources of slightly shifted frequencies spreading and shifting the lines of the spectrum and preventing usual sinusoidal analysis methods from working properly. The de-synchronization of the signal sources prevents most usual temporal method from working with the mixed signal.

The nature and amount of variation between one singer and another in terms of timbre and intonation[4] have been considered as well as the amount of synchronization between the singers at different points of a phrase and the synchronization of their vibrato. For example it was found that plosive consonants correspond to stronger synchronization points than than vowels.

Only the recordings of solo singers have been analyzed and segmented. The re-synthesized sound of a group of voices by the implemented module was perceptually compared with the original recording of multiple singers singing the same musical phrase. The experiments have shown that about 7 well differentiated synthetic voices gave the same impression as a group of 10 real voices. A pitch variation in the range of 25 cents and a uncertainty of 20 ms for the onset position have been found to give a realistic impression of a choir.

4.1 Segmentation

In addition to the segmentation into elementary waveforms (by the PSOLA markers), voiced and unvoiced segments as well as pitched notes (manually, see 3.3), a fourth level of segmentation was applied to the analysis data. It cuts the musical phrases into segments of musical interest like phonemes, words and entire phrases.

With this segmentation, the recorded phrases can be used as a data base for a wide range of different synthesis processes. The sequence of timbre and pitch of the original phrases can be completely re-composed. In order to reconstitute an entire virtual choir, phrases of different voice groups, based on different analysis files, can be re-synchronized word by word.

Interesting effects can be obtained controlling the synthesis by a function of the voicing coefficients. For example, the voiced segments of the signal can be more stretched than unvoiced segments. Similarly, vowels and consonants can be independently processed and spatialized.

[4]Expressed by Sundberg's *degree of unison* [Sundberg, 1987].

4.2 Spatialization

The realization of the piece "Vertigo Apocalypsis" by Philippe Schoeller at IRCAM [Nouno, 1999] showed the importance of spatialization for a realistic impression of a choir. In this work multiple solo recorded singers were precisely placed in the acoustic space. For "K...", each re-synthesized voice or voice section will be processed by IRCAM's *Spatializateur* [Jot and Warusfel, 1995] allowing the composer to control the spatial placement and extent of the virtual choir.

In the general context of the electro-acoustic orchestration of "K...", an important role will be given to the Spatializateur taking into account the architectural and acoustic specificities of the opera house.

4.3 Conclusions

The implemented system reveals itself to be very versatile and flexible. The choir impression obtained with it is much more interesting and realistic than any classical chorus effect.

The used synthesis technique produces an excellent audio quality, close to the choir recordings. The quality of transformation achieved with PSOLA is better than the usual techniques based on re-sampling.
The application of an individual vibrato for each synthesis voice after having canceled the recorded vibrato turned out to be extremely effective for the perception of the choral effect.

The efficiency of the algorithm allows polyphony of a large number of voices. The virtual choir is embedded into a rich environment of various synthesis and transformation techniques such as phase-aligned formants synthesis, sampling and classical sound transformations like harmonizing and frequency-shifting. The virtual choir will be constituted of 32 simultaneous synthesis voices grouped into 8 sections.

During the experiments it appeared clearly that vocal vibrato does not affect only the fundamental frequency. It is accompanied by synchronized amplitude and spectral modulations. Canceling the vibrato by smoothing the pitch leaves an effect of unwanted roughness in the resulting sound.

Another limitation of the system appears for the processing of very high soprano notes (above 1000 Hz). For these frequencies the impulse response of the vocal tract extends over more than one signal period and can not be isolated by simple windowing of the time domain signal.

4.4 Future extensions

While the used analysis algorithm performs signal characterization into voiced and unvoiced parts in the time/frequency domain, in the context of "K..." it has only been applied for segmentation in the time domain. Separation into both time and frequency domains would certainly benefit the system, especially for mixed voiced/unvoiced signals (voiced consonants).

In order to produce timbre differences between individual voices, several techniques are currently being evaluated. They rely on an efficient modification of the spectral envelope (i.e. formants) of the vocal signal.

An interesting potential of the paradigm of superposing simple solo voices can be seen in its application to non-vocal sounds. The synthesis of groups of musical instruments could be obtained in the same way as the virtual choir, i.e. deriving the violin section of an orchestra from a single violin recording.

REFERENCES

[Charpentier, 1988] Charpentier, F. (1988). *Traitement de la parole par Analyse/Synthèse de Fourier application à la synthèse par diphones.* PhD thesis, ENST, Paris, France.

[d'Alessandro and Rodet, 1989] d'Alessandro, C. and Rodet, X. (1989). Synthèse et analyse-synthèse par fonctions d'ondes formantiques. *J. Acoustique*, (2):163–169.

[Déchelle et al., 1998] Déchelle, F., Borghesi, R., Cecco, M. D., Maggi, E., Rovan, B., and Schnell, N. (1998). jMax: A new JAVA-based Editing and Control System for Real-time Musical Applications. In *Proceedings of the International Computer Music Conference*, San Francisco. International Computer Music Association.

[Déchelle et al., 1999a] Déchelle, F., Borghesi, R., Cecco, M. D., Maggi, E., Rovan, B., and Schnell, N. (1999a). jMax: An Environment for Real-Time Musical Applications. *Computer Music Journal*, 23(3):50–58.

[Déchelle et al., 1999b] Déchelle, F., Cecco, M. D., Maggi, E., and Schnell, N. (1999b). jMax Recent Developments. In *Proceedings of the 1999 International Computer Music Conference*, San Francisco. International Computer Music Association.

[IRCAM, 2000] IRCAM (2000). *jMax home page*. IRCAM, http://www.ircam.fr/jmax.

[Jot and Warusfel, 1995] Jot, J.-M. and Warusfel, O. (1995). A real-time spatial sound processor for music and virtual reality applications. In *Proceedings of the International Computer Music Conference*, Banff. International Computer Music Association.

[Kortekaas, 1997] Kortekaas, R. (1997). *Physiological and psychoacoustical correlates of perceiving natural and modified speech*. PhD thesis, TU, Eindhoven, Holland.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones. *Speech Communication*, (9):453–467.

[Nouno, 1999] Nouno, G. (1999). Vertigo apocalypsis. *Internal Report IRCAM*.

[Peeters, 1998] Peeters, G. (1998). Analyse-Synthèse des sons musicaux par la mèthode PSOLA. In *Journées Informatique Musicale*, Agelonde, France.

[Peeters and Rodet, 1999] Peeters, G. and Rodet, X. (1999). Non-Stationary Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum. In *ICSPAT*, Orlando, USA.

[Schnell, 1994] Schnell, N. (1994). GRAINY - Granularynthese in Echtzeit. *Beiträge zur Elektronischen Musik*, (4).

[Sundberg, 1987] Sundberg, J. (1987). *The Science of Singing Voice*. University Press, Stocholm.

[Ternström, 1989] Ternström, S. (1989). *Acoustical Aspects of Choir Singing*. Royal Institute of Technology, Northern Illinois.

Annexe C

Programmation

Les algorithmes développés lors de cette thèse sont nombreux. Une partie des algorithmes a été développé en langage C en utilisant la librairie de calcul vectoriel UDI, le reste des algorithmes a été développé en MATLAB également sous forme de librairie (`function`) plutôt que de scripts.

Les programmes de l'Ircam préexistants utilisés lors de cette thèse sont

- ["f0"] : algorithme d'estimation de la fréquence fondamentale par maximum de vraisemblance développé par Boris Doval
- ["syntadd"] : algorithme de synthèse sinusoïdal par banc d'oscillateurs

Dans un premier temps, les programmes développés reposaient sur des programmes préexistants de l'Ircam. Cependant il est apparu très vite que des erreurs d'allocation temporelle de ces programmes empêchaient un positionnement précis de l'information. Une partie du temps a été consacrée au débogage de ces programmes mais il s'est avéré plus judicieux à ce stade d'entreprendre le développement en MATLAB et de porter ensuite ce développement en C. Ceci permettait de plus un contrôle total de l'exécution des algorithmes grâce aux fonctionnalités graphiques de MATLAB. Le développement sous MATLAB a été effectué sous forme de librairies de manière à permettre facilement le portage de ces librairies en C.

L'essentiel des algorithmes se présente sous forme de fonctions emboîtées. De plus, un soin particulier a été apporté quand à l'uniformisation non seulement des APIs mais également de la nomenclature interne à chaque fonction ceci de manière à faciliter la lecture du code.

La nomenclature suivante a été utilisée et est proposée comme base d'une standardisation de la dénomination de variable.

C.1 Nomenclature MATLAB

La nomenclature suivante a été utilisée lors de cette recherche : `codecontenu_codeechelle_codetype`

Chaque nom de variable se termine par un `code type` représentant le type de données renfermé par la variable :

- `_v` vecteur
- `_iv` vecteur de nombres complexes

- `_m` matrice
- `_im` matrice de nombre complexe
- `_bp` fonction par point (matrice de valeur x/y)
- `_struct` structure

Le `code echelle` détermine l'échelle dans laquelle sont exprimées les données :

- `_hz` Hertz
- `_sec` seconde
- `_k` pour les fréquences discretes
- `_n` pour une échelle normalisée
- `_pos` pour une position dans un vecteur

Le `code contenu` n'est pas standardisé. Nous avons cependant généralement utilisé les codes suivants :

- `data` donnée
- `fenetre` fenêtre d'analyse
- `am` spectre d'amplitude
- `ph` spectre de phase
- `unph` spectre de phase déroulé
- `amh` amplitude de la composante h du spectre
- `phh` phase de la composante h du spectre

Exemples :

– une variable contenant le spectre de phase s'exprime `unph_rad_v`

– une variable breakpoint fonction de fréquence fondamentale `f0_sec_hz_bp` .

Cette standardisation de la dénomination des variables, contraignante dans un premier temps, c'est avérée très utile dans la suite : relecture très facile des fichiers, non-ambiguïté des variables, débogage instantané.

De même, la dénomination des noms des fichiers détermine si il s'agit

- `F*` une fonction
- `G*` une méta-fonction
- `P*` un programme (script)

L'en-tête de TOUTES fonctions DOIT contenir les champs suivants

- `DESCRIPTION:` la description de l'action accomplie par la fonction
- `INPUTS:` le nom, la description, le type ainsi que l'échelle des variables d'entrée
- `OUTPUTS:` le nom, la description, le type ainsi que l'échelle des variables de sortie
- `LAST EDIT:` la date de la dernière modification apportée à la fonction

C.2 Formats

Les programmes développés dans le cadre de ce travail lisent et écrivent les fichiers répondant aux standard SDIF. Le standard SDIF a été développé par l'Ircam, le CNMAT de Berkeley et l'IUA/ UPF de Barcelone et est actuellement utilisé par un nombre de plus en plus important de centre de recherche de la communauté musicale. Ce standard définit un format de données ainsi qu'un ensemble de format de descriptions du son. Il s'agit d'un standard ouvert dans la mesure où de nouveaux formats de descriptions peuvent être proposés.

Annexe D

Propriétés générales : Retard de groupe

D.1 Définition

Le retard de groupe est défini comme la dérivée fréquentielle de la phase :

Retard de groupe :

$$\tau_g(\omega) = -\frac{\partial\phi(\omega)}{\partial\omega} \tag{D.1}$$

$\tau_g(\omega)$ est le temps moyen d'arrivée de la fréquence ω.

Le retard de groupe contient l'information temporelle du signal, information absente du module de la Transformée de Fourier. Il peut être montré que le retard de groupe d'une bande de fréquence est égal au centre de gravité temporel de l´énergie contenue dans cette bande [Fla93] [AF95]. En utilisant la convention passe-bas de la TFCT [1]

$$\tau_g(s, t_m, \omega) = \Re\left\{\frac{\int_t tx(t)h^*(t_m - t)e^{j\omega t}dt}{\int_t x(t)h^*(t_m - t)e^{j\omega t}dt}\right\} \tag{D.2}$$

Le centre de gravité temporel de l'énergie est égal à la moyenne du retard de groupe pondéré par l'énergie des bandes de fréquence [Coh95], en notant $s(t_m, t) = x(t)h(t_m - t)$

$$\langle s(t_m)\rangle = \int t|s(t_m, t - t_m)|^2 dt \tag{D.3}$$

$$\langle s(t_m)\rangle = \int \tau_g(s, t_m, \omega)|S(t_m, \omega)|^2 d\omega \tag{D.4}$$

[1] Convention passe-bas de la TFCT : L'origine temporelle de la transformée de Fourier est fixe et est référencée par rapport au début du signal

Une application particulièrement intéressante du retard de groupe réside dans son utilisation afin d'améliorer la lecture des représentations temps/fréquence et en particulier du spectrogramme [AF95]. Ceci constitue d'ailleurs notre première motivation à son utilisation pour la détection des singularités.

D.2 Estimation

Différenciation : Le retard de groupe peut s'obtenir par différenciation des valeurs consécutives du spectre de phase ramenées à la taille du pas fréquentiel du spectre

$$\tau_g\left(\frac{\omega_k+\omega_{k+1}}{2}\right) = -\frac{\phi(\omega_k+1)-\phi(\omega_k)}{\omega_{k+1}-\omega_k} \tag{D.5}$$

L'approximation de la fonction de phase sera d'autant meilleure que la résolution du spectre sera grande. L'obtention d'une résolution suffisante est obtenue par prolongement par zéro . Le spectre $S(z)$ du signal $s(n)$ devant être analytique dans son domaine de définition, la phase de $S(z)$ doit être une fonction continue de z. Cette condition se traduit sur $S(\omega_k)$ par la nécessité d'un déroulement de la phase : $|S(\omega_k) - S(\omega_{k-1})| \leq \pi \ \forall k$.

Rapport de TF : [AF95] proposent de calculer le retard de groupe par le rapport de deux TFs (voir annexe G) :

- la TF du signal $x(t)$ pondéré par la fonction produit $t \cdot h(t)$, dans lequel t est la variable temps et $h(t)$ la fonction fenêtre de pondération
- la TF du signal $x(t)$ pondérée par la fenêtre $h(t)$.

En utilisant la notation passe-bande (BP) de la TFCT et en notant $s(t) = x(t)h(t-t_m)$:

$$\tau_g(s,t_m,\omega) = \Re\left\{\frac{STFT^{BP}_{(t-t_m)h}(x,t_m,\omega)}{STFT^{BP}_{h}(x,t_m,\omega)}\right\} \tag{D.6}$$

La division complexe peut être évitée en utilisant la formulation en terme d'énergie

$$\tau_g(s,t_m,\omega) = \Re\left\{\frac{STFT^{BP}_{(t-t_m)h}(x,t,\omega)\cdot STFT^{BP,*}_{h}(x,t,\omega)}{|STFT^{BP}_{h}(x,t,\omega)|^2}\right\} \tag{D.7}$$

qui se réécrit

$$\tau_g(s,t_m,\omega) = \frac{TH_{\Re}\cdot H_{\Re} + TH_{\Im}\cdot H_{\Im}}{H^2_{\Re}+H^2_{\Im}} \tag{D.8}$$

dans lequel nous avons noté $H = STFT^{BP}_{h}(x,t_m,\omega)$ et $TH = STFT^{BP}_{(t-t_m)h}(x,t_m,\omega)$; $\Re$ et$\Im$ désigne les parties réelles et imaginaires de la TF.

Annexe E

Propriétés générales : Signaux à phase minimale

E.1 Définition

Les filtres à phase minimale ou, d'une manière plus générale, les séquences ou signaux à phase minimale sont définis par la localisation des pôles et des zéros de leur Transformée en Z à l'intérieur du cercle unité. Nous nous plaçons dans le cas de séquences causales et réelles. Les séquences à phase minimale sont caractérisées par les propriétés suivantes :

1. la séquence à phase minimale $h_{min}(n)$ est la séquence, parmi les séquences ayant une réponse fréquentielle d'amplitude $|H(\omega)|$ donnée, qui possède la concentration d'énergie la plus importante

$$\sum_{n=0}^{N} h_{min}^2(n) \geq \sum_{n=0}^{N} h^2(n) \quad \forall N \tag{E.1}$$

2. Pour une séquence réelle, causale et à phase minimale, $\log(|X(\omega)|)$ et $\arg(H(\omega))$ sont transformées de Hilbert réciproques [OS75][1].
3. L'absence de zéro dans la région de convergence de la Transformée en Z de h_{min} (la région de convergence étant extérieure au cercle unité pour une séquence causale) a pour conséquence un spectre de phase de pente moyenne nulle [RD86].
4. Toute séquence peut se décomposer en une séquence à phase minimum et une séquence de type passe-tout.
5. L'inverse d'un filtre à phase minimale est encore un filtre à phase minimale.
6. La cascade de deux filtres à phase minimale est un filtre à phase minimale.

[1] Pour une séquence réelle et causale, la partie réelle et imaginaire du spectre sont transformées de Hilbert réciproques ($X_{\Im}(\omega)$ peut se déduire de $X_{\Re}(\omega)$, et $X_{\Re}(\omega)$ peut se déduire de $X_{\Im}(\omega)$ à une constante près).

E.2 Filtrage Homomorphique et signaux à phase minimale

Le cepstre complexe $cc(n)$ est défini comme la Transformée de Fourier inverse du logarithme complexe du spectre :

$$cc(n) \overset{\text{TF}}{\rightleftharpoons} \log(X(\omega)) = \log(|X(\omega)|) + j \arg(X(\omega)) \tag{E.2}$$

La partie du cepstre complexe proche de l'origine représente l'évolution lente du spectre et donc sa TF fournit le spectre de log-amplitude et le spectre de phase d'une approximation de l'enveloppe spectrale voire du filtre du système. Cependant l'utilisation du cepstre complexe pose le problème de l'indétermination de la phase $\arg(X(\omega))$.

Le cepstre réel $c(n)$ est défini comme la Transformée de Fourier inverse du spectre de log-amplitude.

$$c(n) \overset{\text{TF}}{\rightleftharpoons} \log(|X(\omega)|) \tag{E.3}$$

Pour un signal réel et causal, le spectre réel $X_{\Re}(\omega)$ et imaginaire $X_{\Im}(\omega)$ sont reliés par relation de Hilbert dans le sens que l'un peut se déduire de l'autre (du moins à une constante près). De manière équivalente à la relation liant $X_{\Re}(\omega)$ et $X_{\Im}(\omega)$, si $c(n)$ est causal (et donc $x(n)$ à phase minimale), le spectre de log-amplitude $\log(|X(\omega)|)$ et le spectre de phase $\arg(X(\omega))$ sont reliés par relation de Hilbert. Dans ce cas $\arg(X(\omega))$ peut être obtenu à partir de $\log(|X(\omega)|)$ sans nécessité le calcul du logarithme complexe et donc sans problème d'indétermination de la phase.

Un cepstre réel causal correspond à une séquence à phase minimale. De même un cepstre réel anti-causal correspond à une séquence à phase maximale. Une séquence peut donc se décomposer en une contribution à phase minimale et une contribution à phase maximale dont les spectres de log-amplitude et de phase sont obtenus par relation de Hilbert. La TF de la partie proche de l'origine d'un cepstre réel causal fournit donc une approximation d'un filtre à phase minimal, celle d'un cepstre réel anti-causal d'un filtre à phase maximale.

En considérant donc la partie réelle du cepstre proche de l'origine, nous obtenons le spectre de phase du filtre du signal à phase minimale correspondant à la séquence $x(n)$.

Annexe F

Propriétés générales : Ré-échantillonnage

Nous rappelons les bases théoriques du ré-échantillonnage temporel, dans le cas du sur-échantillonage et du sous-échantillonnage, ainsi que du ré-échantillonnage correspondant au zéro-padding ou au prolongement du signal par zéro.

F.1 Ré-échantillonnage temporel

F.1.1 Théorie

Le ré-échantillonnage d'un signal peut être vu de deux manière différentes ([Smi98]) :

1. Il s'agit de calculer les valeurs du signal en des temps qui n'existent pas au taux d'échantillonnage initial. Nous pouvons calculer ces points par interpolation (sinus cardinal ou autre) des échantillons adjacents.
2. Il s'agit de reconstruire le signal continu correspondant au signal discret initial et de le re-discrétiser au nouveau taux d'échantillonnage.

Dans la suite cette partie nous détaillerons la deuxième interprétation.

Soit $X_F(f)$ le spectre associé au signal continu $x(t)$ (voir Fig. F.1). Soit $\tilde{X}_F(f)$ le spectre associé au signal discret $\tilde{x}_F(t)$ de pas d'échantillonnage $1/F$. $\tilde{X}_F(f)$ est donc périodique de période F. $X_F(f)$ est égale à $\tilde{X}_F(f)$ après filtrage entre $-\frac{F}{2}$ et $\frac{F}{2}$.

Soit $x_F(t)$ la version échantillonnée de $x(t)$:

$$x_F(t) = \sum_n x(t)\delta\left(t - \frac{n}{F}\right) \tag{F.1}$$

Soit $\tilde{x}_F(t)$ le signal discret observé

$$\begin{aligned} \tilde{x}_F(t) &= \frac{1}{F} \cdot x_F(t) \\ &= \frac{1}{F} \sum_n x(t)\delta\left(t - \frac{n}{F}\right) \end{aligned} \tag{F.2}$$

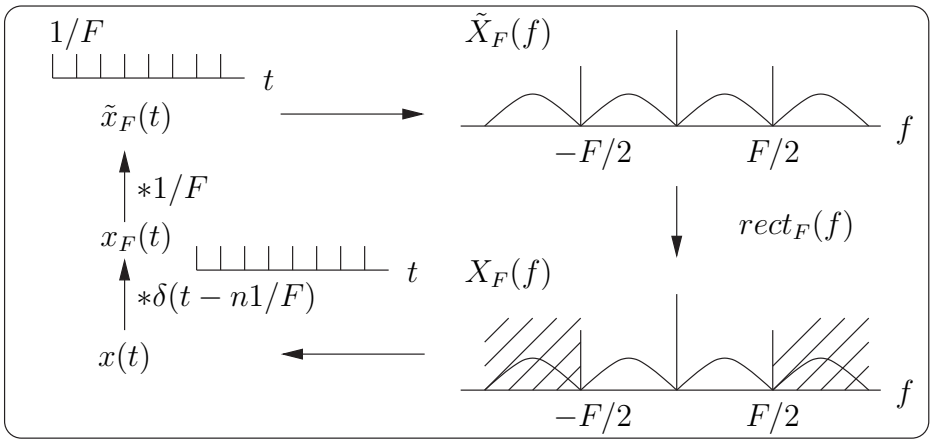


FIG. F.1 – Ré-échantillonnage : illustration de l'échantillonnage d'un signal et de la reconstruction d'un signal continu à partir d'un signal discret

F.1.1.1 Reconstruction du signal continu

Nous cherchons le signal continu $x(t)$ correspondant à $X_F(f)$:

$$X_F(f) = \tilde{X}_F(f) \cdot rect_F(f) \tag{F.3}$$

ce qui se réécrit dans le domaine temporel

$$\begin{aligned} x(t) &= \tilde{x}_F(t) \otimes F \ \mathrm{sinc}(\pi F t) \\ &= \left[\frac{1}{F} \sum_n x(t) \delta \left(t - \frac{n}{F} \right) \right] \otimes F \ \mathrm{sinc}(\pi F t) \\ &= \sum_n x \left(\frac{n}{F} \right) \ \mathrm{sinc} \left(\pi F \left(t - \frac{n}{F} \right) \right) \end{aligned} \tag{F.4}$$

Remarque : *$x\left(\frac{n}{F}\right)$ est égale à $\tilde{x}_F\left(\frac{n}{F}\right)$*

La reconstruction du signal continu $x(t)$ peut donc être vu
- soit comme la transformée de Fourier inverse du spectre du signal discret filtré à $\frac{F}{2}$,
- soit comme l'interpolation à l'aide d'un sinus cardinal des points intermédiaires entre les échantillons.

F.1.1.2 Ré-échantillonnage du signal continu

Ré-échantillonnage du signal continu aux points n'/F' (nouveau taux d'échantillonnage F') :

$$x_{F'}(n') = \sum_{t=-\infty}^{+\infty} x(t) \cdot \delta \left(t - \frac{n'}{F'} \right) \tag{F.5}$$

$$\boxed{x_{F'}(n') = \sum_{n=-\infty}^{+\infty} x\left(\frac{n}{F}\right) \cdot \ \mathrm{sinc}\left(\pi F\left(\frac{n'}{F'} - \frac{n}{F}\right)\right)} \tag{F.6}$$

Afin d'éviter le repliement de spectre (aliasing) lors d'un **sous-échantillonnage** ($F' < F$), nous devons filtrer le signal de manière passe-bas à $f_c \leq \frac{F'}{2}$ avant ré-échantillonnage. En terme d'interprétation par «transformée de Fourier inverse du spectre filtré du signal discret», ceci revient à filtrer le spectre non pas par $\frac{F}{2}$ mais par $\frac{F'}{2}$.

$$\boxed{x_{F'}(n') = \sum_{n=-\infty}^{+\infty} x\left(\frac{n}{F}\right) \cdot \frac{F'}{F} \cdot \ \mathrm{sinc}\left(\pi F'\left(\frac{n'}{F'} - \frac{n}{F}\right)\right)} \tag{F.7}$$

F.1.2 Implémentation

Le ré-échantillonnage s'effectue par interpolation à l'aide d'un sinus cardinal. Le sinus cardinal est stocké dans une table et sur-échantillonné afin d'obtenir avec une grande précision ses valeurs entre deux passages par zéro («zero-crossing»). La longueur de cette table dépend donc du nombre de valeurs entre deux passages par zéro («number of samples per zero crossing», nspzc) ainsi que du nombre de passages par zéro du sinus cardinal. Les points intermédiaires de la table seront calculés par interpolation linéaire.

La même table est utilisée pour le sur-échantillonnage et le sous-échantillonnage mais elle y est interprétée différemment.

F.1.2.1 Sur-échantillonnage (Fig. F.2) :

Pour calculer le signal en un point $x(n')$, nous convoluons le vecteur signal $x(n)$ par le vecteur de points du sinus cardinal

- commençant en : $(\lceil n \rceil_{n'} \cdot 1/F - n' \cdot 1/F') \cdot F \cdot \mathrm{nspzc}$
  où $\lceil n \rceil_{n'}$ désigne l'échantillon directement supérieur à n'.
- de pas d'avancement : nspzc

Une interpolation linéaire entre les points de la table est effectuée si les échantillons du signal $x(n)$ ne «tombent» pas sur les échantillons de la table du sinus cardinal.

F.1.2.2 Sous-échantillonnage (voir Fig. F.3) :

Dans le cas d'un sous-échantillonnage,

- commençant en : $(\lceil n \rceil_{n'} \cdot 1/F - n' \cdot 1/F') \cdot F_e \cdot \frac{F'}{F} \cdot \mathrm{nspzc}$
- de pas d'avancement : $\frac{F'}{F} \cdot \mathrm{nspzc}$

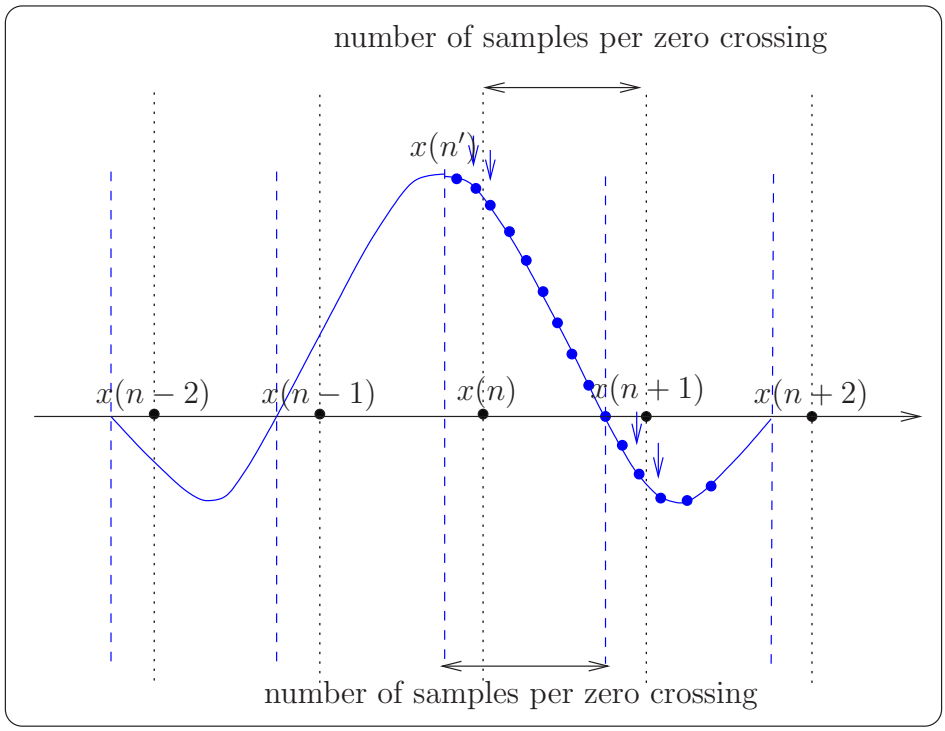


FIG. F.2 – Sur-échantillonnage ($F' > F$)

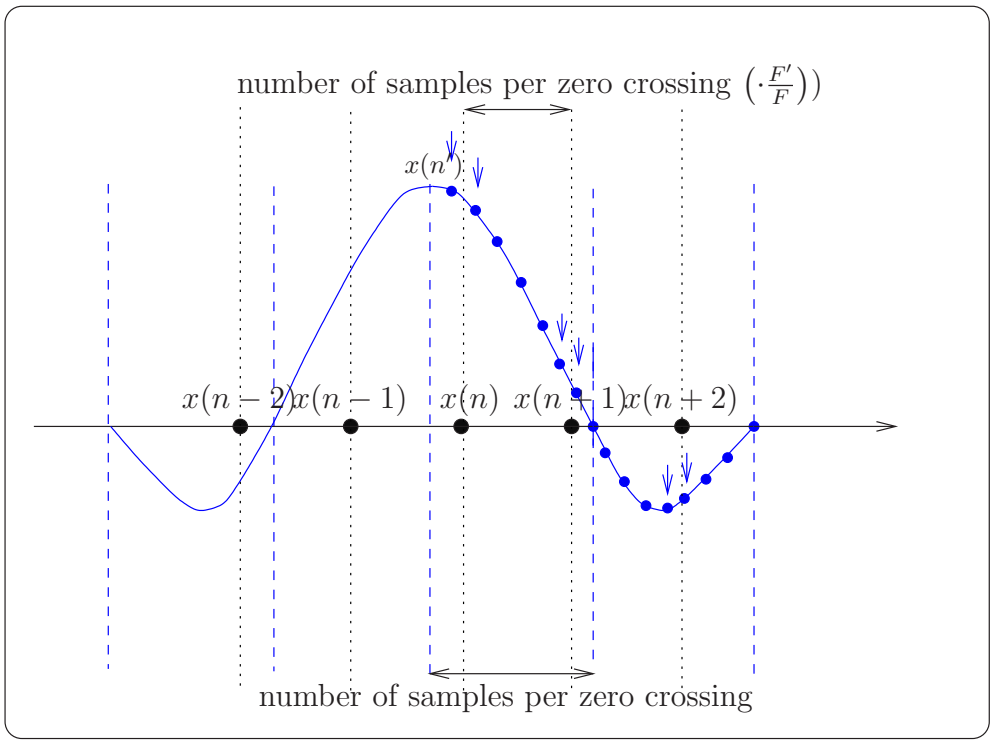


FIG. F.3 – Sous-échantillonnage ($F' < F$)

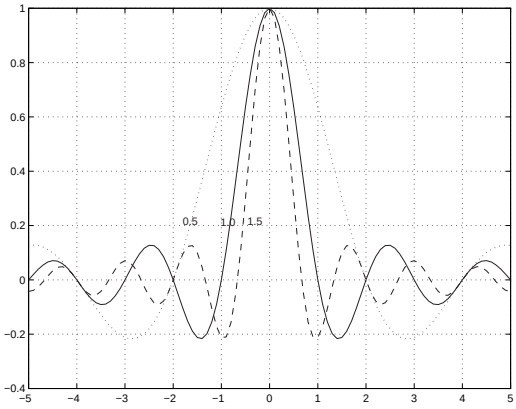


FIG. F.4 – $\text{sinc}(x) = \frac{sin(\pi x)}{\pi x}$ pour différentes valeurs de x

F.2 Ré-échantillonnage fréquentiel (Zéro-padding ou prolongement par zéro)

Nous montrons dans la suite comment la même formulation peut être utilisée pour expliquer le prolongement par zéro en terme de ré-échantillonnage du spectre.

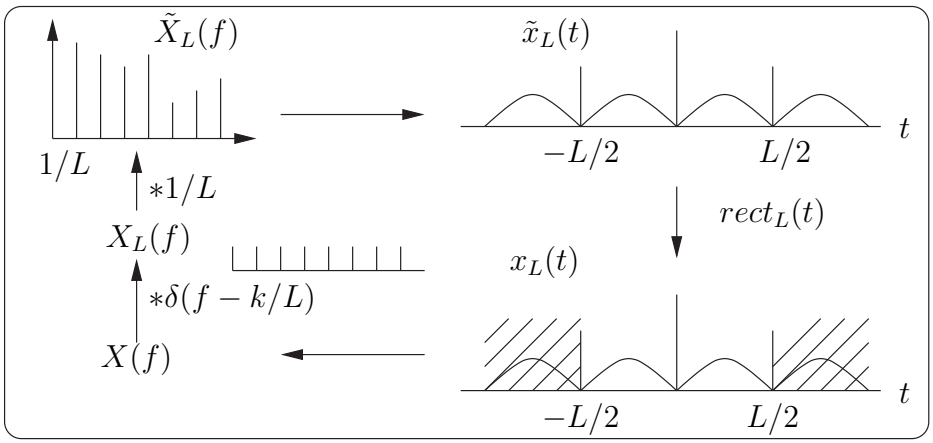


FIG. F.5 – Prolognement par zéro : illustration de l'échantillonnage d'un spectre et de la reconstruction d'un spectre continu à partir d'un spectre discret

Soit $X(f)$ le spectre continu recherché, $X_L(f)$ le spectre discret de pas d'echantillonnage $1/L$:

$$X_L(f) = \sum_k X(f)\delta\left(f - \frac{k}{L}\right) \tag{F.8}$$

Soit $\tilde{X}_L(f)$ le spectre discret observé

$$\begin{aligned} \tilde{X}_L(f) &= \frac{1}{L} X_L(f) \\ &= \frac{1}{L} \sum_k X(f)\delta\left(f - \frac{k}{L}\right) \end{aligned} \tag{F.9}$$

F.2.0.3 Reconstruction du spectre continu

Nous cherchons le spectre continu $X(f)$ correspondant à $x_L(t)$:

$$x_L(t) = \tilde{x}_L(t) \cdot rect_L(t) \tag{F.10}$$

ce qui se réécrit dans le domaine fréquentiel

$$\begin{aligned} X(f) &= \tilde{X}_L(f) \otimes L \ \mathrm{sinc}(\pi L f) \\ &= \left[\frac{1}{L} \sum_k X(f) \delta \left(f - \frac{k}{L} \right) \right] \otimes L \ \mathrm{sinc}(\pi L f) \\ &= \sum_k X \left(\frac{k}{L} \right) \ \mathrm{sinc} \left(\pi L \left(f - \frac{k}{L} \right) \right) \end{aligned} \tag{F.11}$$

Remarque : $X\left(\frac{k}{L}\right)$ est égale à $\tilde{X}_L\left(\frac{k}{L}\right)$:

F.2.0.4 Ré-échantillonnage du spectre continu

Ré-échantillonnage du spectre continu aux points k'/N

$$\begin{aligned} X_N \left(\frac{k'}{N} \right) &= \sum_f X(f) \delta \left(f - \frac{k'}{N} \right) \\ &= \sum_n X \left(\frac{k}{L} \right) \ \mathrm{sinc} \left(\pi L \left(\frac{k'}{N} - \frac{k}{L} \right) \right) \\ &= X \left(\frac{k'}{N} \right) \otimes \ \mathrm{sinc} \left(\pi L \frac{k'}{N} \right) \\ &= x_N(t) \cdot rect_L(t) \end{aligned} \tag{F.12}$$

La dernière formulation est bien la formulation du prolongement par zéro d'un signal. Deux interprétations :

- signal de longueur L allongé par des zéros jusqu'à N,
- signal de longueur N pondérée par une fenêtre rectangulaire de longueur L.

Annexe G

Propriétés générales : Ré-assignement

Nous rappelons la formulation mathématique du ré-assignement temporel et fréquentiel [AF95] , ainsi que le passage de son interprétation en tant que centre de gravité temporel et fréquentiel de l'énergie aux expressions en terme de dérivée fréquentielle et temporelle du spectre de phase. Ces formulations sont indiquées dans le cas de la convention passe-bas et passe-bande de la Transformée de Fourier à Court Terme (TFCT).

Nous notons $x(t)$ le signal, t le temps, $h(t)$ la fenêtre de pondération utilisée, ω la pulsation et $STFT_y$ la TFCT obtenue par fenêtrage du signal à l'aide de la fonction y.

G.1 Ré-assignement temporel

G.1.1 Définition en tant que centre de gravité temporelle de l'énergie

convention passe-bas

$$t_r(x;t,\omega) = t + \Re\left\{\frac{\int_s s x(s) h^*(t-s) e^{-j\omega s} ds}{STFT_h^{LP}(x;t;\omega)}\right\} \tag{G.1}$$

convention passe-bande

$$t_r(x;t,\omega) = t + \Re\left\{\frac{\int_s (s-t) x(s) h^*(t-s) e^{j\omega(t-s)} ds}{STFT_h^{BP}(x;t;\omega)}\right\} \tag{G.2}$$

G.1.2 Réécriture en terme de dérivée du spectre

Nous pouvons réécrire (G.1) et (G.2) comme

$$t_r(x;t,\omega) = t + \Re\left\{\frac{j\frac{\partial}{\partial\omega} STFT_h(x;t,\omega)}{STFT_h(x;t,\omega)}\right\} \tag{G.3}$$

G.1.3 Réécriture en terme de dérivée de la phase

Si nous notons

$$STFT_h(x;t,\omega) = M(x;t,\omega)e^{j\phi(x;t,\omega)} \tag{G.4}$$

nous pouvons réécrire (G.3) comme

$$\begin{aligned} t_r(x;t,\omega) &= t + \Re\left\{\frac{j(M'e^{j\phi} + Mj\phi'e^{j\phi})}{Me^{j\phi}}\right\} \\ &= t - \frac{\partial}{\partial\omega}\phi(x;t,\omega) \end{aligned} \tag{G.5}$$

Définition du retard de groupe

$$\tau_g(\omega) = -\frac{\partial\phi}{\partial\omega} \tag{G.6}$$

$\tau_g(\omega)$ est le temps moyen d'arrivé de la fréquence ω

G.1.4 Calcul

Nous pouvons également écrire (G.1) et (G.2) comme

convention passe-bas

$$t_r(x;t,\omega) = t + \Re\left\{\frac{STFT^{LP}_{sh}(x;t,\omega)}{STFT^{LP}_{h}(x;t,\omega)}\right\} \tag{G.7}$$

convention passe-bande

$$t_r(x;t,\omega) = t + \Re\left\{\frac{STFT^{BP}_{(s-t)h}(x;t,\omega)}{STFT^{BP}_{h}(x;t,\omega)}\right\} \tag{G.8}$$

Une **autre formulation** ne nécessitant pas de division complexe est possible :

$$t_r(x,t,\omega) = t + \Re\left\{\frac{STFT_{sh}(x,t,\omega)STFT^*_h(x,t,\omega)}{|STFT_h(x,t,\omega)|^2}\right\} \tag{G.9}$$

Ceci s'obtient aisément en remplaçant $STFT^*_h(x,t,w)$ par $Me^{-j\phi}$ et $STFT_{sh}(x,t,w)$ par $j(M'e^{j\phi} + Mj\phi'e^{j\phi})$:

$$t_r(x,t,\omega) = t + \Re\left\{\frac{j(M'e^{j\phi} + Mj\phi'e^{j\phi})Me^{-j\phi}}{M^2}\right\} = -\phi' \tag{G.10}$$

Finalement

$$t_r(x,t,\omega) = \frac{TH_{\Re}H_{\Re} + TH_{\Im}H_{\Im}}{H^2_{\Re} + H^2_{\Im}} \tag{G.11}$$

dans lequel nous avons noté $H = STFT_h(x,t,\omega)$ et $TH = STFT_{sh}(x,t,\omega)$

G.2 Ré-assignement fréquentiel

G.2.1 Définition en tant que centre de gravité fréquentiel de l'énergie

$$\omega_r(x;t,\omega) = \Re\left\{\frac{\int_\xi \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi}{STFT_h(x;t;\omega)}\right\} \tag{G.12}$$

G.2.2 Passage de la formule de la TFCT en terme de convolution des TF

Convention passe-bas :

$$STFT_h^{LP}(x;t,\omega) = \int_s x(s)h^*(t-s)e^{-j\omega s}ds = \int_\xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi\cdot e^{-j\omega t} \tag{G.13}$$

Démonstration :

$$X_h^{LP}(t,\omega) = \int_s x(s)h^*(t-s)e^{-j\omega s}ds \tag{G.14}$$

équivalent en terme de convolution fréquentielle à

$$\begin{aligned} X_h^{LP}(t,\omega) &= X(\omega)\otimes[H^*(-\omega)e^{-j\omega t}] \\ &= \int X(\xi)H^*(\xi-\omega)e^{j(\xi-\omega)t}d\xi \\ &= \int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi\cdot e^{-j\omega t} \end{aligned} \tag{G.15}$$

Convention passe-bande :

$$STFT_h^{BP}(x;t,\omega) = \int_s x(s)h^*(t-s)e^{j\omega(t-s)}ds = \int_\xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi \tag{G.16}$$

Démonstration :

$$\begin{aligned} X_h^{BP}(t,\omega) &= \int_s x(s)h^*(t-s)e^{j\omega(t-s)}ds \\ &= \int_s x(s)h^*(t-s)e^{-j\omega s}ds\cdot e^{j\omega t} \end{aligned} \tag{G.17}$$

$$\boxed{X_h^{BP}(t,\omega) = X_h^{LP}(t,\omega)\cdot e^{j\omega t}} \tag{G.18}$$

(G.17) est donc équivalent en terme de convolution fréquentielle à

$$\begin{aligned} X_h^{BP}(t,\omega) &= \int X(\xi)H^*(\xi-\omega)e^{j(\xi-\omega)t}d\xi \cdot e^{j\omega t} \\ &= \int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi \end{aligned} \tag{G.19}$$

G.2.3 Réécriture de (G.12) en terme de dérivée du spectre

Convention passe-bas La dérivée de (G.13) s'écrit

$$\frac{\partial}{\partial t}X_h^{LP}(t,\omega) = j\int \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi \cdot e^{-j\omega t} - j\omega\int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi \cdot e^{-j\omega t} \tag{G.20}$$

et donc

$$\frac{\frac{\partial}{\partial t}X_h^{LP}(t,\omega)}{X_h^{LP}(\omega)} = -j\omega + j\frac{\int \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi}{\int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi} \tag{G.21}$$

et donc aussi

$$\frac{\int \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi}{\int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi} = \omega - j\frac{\frac{\partial}{\partial t}X_h^{LP}(t,\omega)}{X_h^{LP}(\omega)} \tag{G.22}$$

Convention passe-bande La dérivée de (G.16) s'écrit

$$\frac{\partial}{\partial t}X_h^{BP}(t,\omega) = j\int \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi \tag{G.23}$$

et donc

$$\frac{\int \xi X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi}{\int X(\xi)H^*(\xi-\omega)e^{j\xi t}d\xi} = -j\frac{\frac{\partial}{\partial t}X_h^{BP}(t,\omega)}{X_h^{BP}(\omega)} \tag{G.24}$$

Étant donné (G.22) et (G.24), nous pouvons réécrire (G.12) comme

convention passe-bas :

$$\omega_r(x;t;\omega) = \omega - \Re\left\{\frac{j\frac{\partial}{\partial t}STFT_h^{LP}(x;t,\omega)}{STFT_h^{LP}(x;t,\omega)}\right\} \tag{G.25}$$

convention passe-bande :

$$\omega_r(x;t;\omega) = -\Re\left\{\frac{j\frac{\partial}{\partial t}STFT_h^{BP}(x;t,\omega)}{STFT_h^{BP}(x;t,\omega)}\right\} \tag{G.26}$$

G.2.4 Réécriture de (G.12) en terme de dérivée de la phase

Si nous notons

$$STFT_h(x;t,\omega) = M(x;t,\omega)e^{j\phi(x;t,\omega)} \tag{G.27}$$

nous pouvons réécrire (G.25) et (G.26) comme

convention passe-bas :

$$\begin{aligned}\omega_r &= \omega - \Re\left\{\frac{j(M'e^{j\phi} + Mj\phi'e^{j\phi})}{Me^{j\phi}}\right\} \\ &= \omega + \frac{\partial}{\partial t}\phi^{LP}(x;t,\omega)\end{aligned} \quad \text{(G.28)}$$

convention passe-bande :

$$\begin{aligned}\omega_r &= -\Re\left\{\frac{j(M'e^{j\phi} + Mj\phi'e^{j\phi})}{Me^{j\phi}}\right\} \\ &= \frac{\partial}{\partial t}\phi^{BP}(x;t,\omega)\end{aligned} \quad \text{(G.29)}$$

Définition de la fréquence instantanée

$$\omega_\phi(\omega) = \frac{\partial \phi}{\partial t} \quad \text{(G.30)}$$

$\omega_\phi(\omega)$ est la fréquence moyenne du temps t

G.2.5 Réécriture de (G.12) en terme de dérivée de la fenêtre d'analyse

Repartons de (G.25) et (G.26). Nous pouvons calculer

convention passe-bas :

$$\frac{\partial}{\partial t}STFT_h^{LP}(x;t,\omega) = STFT_{dh}^{LP}(x) \quad \text{(G.31)}$$

convention passe-bande :

$$\frac{\partial}{\partial t}STFT_h^{BP}(x;t,\omega) = STFT_{dh}^{BP}(x) + j\omega STFT_h^{BP}(x) \quad \text{(G.32)}$$

et donc obtenir

convention passe-bas :

$$\omega_r = \omega - \Re\left\{\frac{jSTFT_{dh}^{LP}(x)}{STFT_h^{LP}(x)}\right\} \quad \text{(G.33)}$$

$$\omega_r = \omega - \Im\left\{\frac{STFT_{dh}^{LP}(x)}{STFT_h^{LP}(x)}\right\} \quad \text{(G.34)}$$

convention passe-bande :

$$\omega_r = -\Re\left\{\frac{jSTFT_{dh}^{BP}(x) - \omega STFT_h^{BP}(x)}{STFT_h^{BP}(x)}\right\} \quad \text{(G.35)}$$

$$\boxed{\omega_r = \omega - \Im\left\{\frac{STFT_{dh}^{BP}(x)}{STFT_h^{BP}(x)}\right\}} \quad \text{(G.36)}$$

Une **autre formulation** ne nécessitant pas de division complexe est possible :

$$\omega_r(x,t,\omega) = \omega - \Im\left\{\frac{STFT_{dh}(x,t,\omega)STFT_h^*(x,t,w)}{|STFT_h(x,t,\omega)|^2}\right\} \quad \text{(G.37)}$$

Ceci s'obtient aisément en remplaçant $STFT_h^*(x,t,w)$ par $M_h e^{-j\phi_h}$ et $STFT_{dh}(x,t,\omega)$ par $M_{dh}e^{-j\phi_{dh}}$. Dans ce cas

$$\frac{STFT_{dh}(x,t,\omega)STFT_h^*(x,t,w)}{|STFT_h(x,t,\omega)|^2} = \frac{M_{dh}e^{j\phi_{dh}}M_h e^{j\phi_h}}{M_h^2} = \frac{M_{dh}e^{j\phi_{dh}}}{M_h e^{j\phi_h}} = \frac{STFT_{dh}^{BP}(x)}{STFT_h^{BP}(x)} \quad \text{(G.38)}$$

G.3 Résumé

Ré-assignement temporel

$$\begin{cases} t_r(x,t,\omega) = & t + \Re\left\{\dfrac{STFT_{th}(x,t,\omega)}{STFT_h(x;t,\omega)}\right\} \\ t_r(x,t,\omega) = & t + \Re\left\{\dfrac{STFT_{th}(x,t,\omega)\overline{STFT}_h^*(x,t,w)}{|STFT_h(x,t,\omega)|^2}\right\} \\ t_r(x,t,\omega) = & t - \dfrac{\partial}{\partial\omega}\phi(x,t,\omega) \end{cases} \tag{G.39}$$

où $-\frac{\partial}{\partial\omega}\phi(x,t,\omega)$ est le retard de groupe.

Ré-assignement fréquentiel

$$\begin{cases} \omega_r = & \omega - \Im\left\{\dfrac{STFT_{dh}(x,t,\omega)}{STFT_h(x,t,\omega)}\right\} \\ \omega_r(x,t,\omega) = & \omega - \Im\left\{\dfrac{STFT_{dh}(x,t,\omega)\overline{STFT}_h^*(x,t,w)}{|STFT_h(x,t,\omega)|^2}\right\} \\ \omega_r(x,t,\omega) = & \dfrac{\partial}{\partial t}\phi(x,t,\omega) \end{cases} \tag{G.40}$$

où $\frac{\partial}{\partial\tau}\phi(x,t,\omega)$ est la fréquence instantanée.

Annexe H

PSOLA : Algorithme itératif de positionnement des marques de «correspondance» et de synthèse PSOLA

Nous rappelons l'algorithme itératif de positionnement des marques de «correspondance» et de synthèse PSOLA proposé par [Sty96].

Fonction de déroulement de l'échelle temporelle : Nous pouvons définir une fonction de déroulement de l'échelle temporelle $\beta(t)$ liée au facteur de dilatation $D(t)$ (voir FIG. H.1) :

$$\beta(t) = \int_0^t D(\tau) d\tau \tag{H.1}$$

telle que

$$\begin{cases} \tilde{\mathrm{m}}_j = \beta(\hat{\mathrm{co}}_j) \\ \hat{\mathrm{co}}_j = \beta^{-1}(\tilde{\mathrm{m}}_j) \end{cases} \tag{H.2}$$

Rappel des notations :

- m_i : marques d'analyse synchrones à la période fondamentale du signal original (distance entre marque égale à la période fondamentale locale)
- $\tilde{\mathrm{m}}_j$: marques de synthèse synchrones à la période fondamentale $P(t)$ du signal voulu
- $\hat{\mathrm{co}}_j$: marques de correpondance (temps sur le signal original correspondant au marques sur le signal de synthèse)

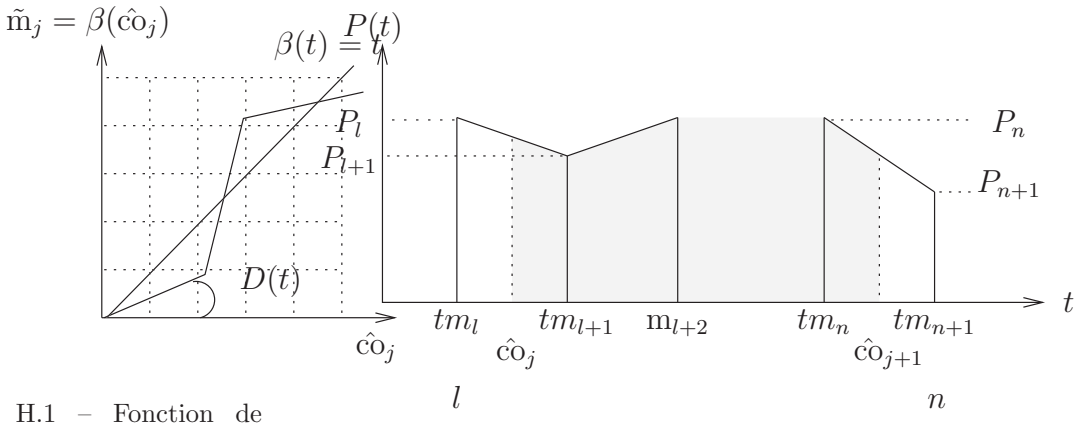


FIG. H.1 – Fonction de déroulement de l'échelle temporelle $\beta(t)$ et facteur de dilatation $D(t)$

FIG. H.2 – Intégration par morceau de $\int_{\hat{\mathrm{co}}_j}^{\hat{\mathrm{co}}_{j+1}} P(t)dt$

Présentation du problème : Connaissant $\hat{\mathrm{co}}_j$, $\tilde{\mathrm{m}}_j$, $f(t)$ et $\beta(t)$, nous voulons trouver $\hat{\mathrm{co}}_{j+1}$ et $\tilde{\mathrm{m}}_{j+1}$. Comme il peut y avoir plusieurs marques de lecture t entre $\hat{\mathrm{co}}_j$ et $\hat{\mathrm{co}}_{j+1}$ (et que par conséquent $f(t)$ et $\beta(t)$ ne peuvent plus être considérés comme constant sur l'intervalle $[\hat{\mathrm{co}}_j, \hat{\mathrm{co}}_{j+1}]$) nous ne pouvons plus faire l'approximation de valeur constante précédente $f(t) = f(\hat{\mathrm{co}}_j)$, $D(t) = D(\hat{\mathrm{co}}_j)$.

Nous supposons $D(t)$ constant par morceaux,

$$D(t) = D_i \text{ pour } \mathrm{m}_i \leq t < \mathrm{m}_{i+1} \tag{H.3}$$

donc par intégration $\beta(t)$ est linéaire par morceaux

$$\beta(t) = \beta(\mathrm{m}_i) + (t - \mathrm{m}_i)D_i \text{ pour } \mathrm{m}_i \leq t < \mathrm{m}_{i+1} \tag{H.4}$$

Soit

$$\begin{cases} \tilde{\mathrm{m}}_j = \beta(\hat{\mathrm{co}}_j) \\ \tilde{\mathrm{m}}_{j+1} = \beta(\hat{\mathrm{co}}_{j+1}) \end{cases} \tag{H.5}$$

Soit m_l la marque de lecture directement inférieure à $\hat{\mathrm{co}}_j$ et m_n la marque de lecture directement inférieure à $\hat{\mathrm{co}}_{j+1}$ (voir FIG. H.2).

$$\begin{cases} \tilde{\mathrm{m}}_j = \beta(\hat{\mathrm{co}}_j) = \beta(\mathrm{m}_l) + (\hat{\mathrm{co}}_j - \mathrm{m}_l)D_l \\ \tilde{\mathrm{m}}_{j+1} = \beta(\hat{\mathrm{co}}_{j+1}) = \beta(\mathrm{m}_n) + (\hat{\mathrm{co}}_{j+1} - \mathrm{m}_n)D_n \end{cases} \tag{H.6}$$

Nous pouvons réécrire l'équation (7.7) en utilisant (H.6), Nous notons $P(t) = \frac{1}{f(t)}$:

$$\tilde{\mathrm{m}}_{j+1} - \tilde{\mathrm{m}}_j = \beta(\mathrm{m}_n) + (\hat{\mathrm{co}}_{j+1} - \mathrm{m}_n)D_n - \beta(\mathrm{m}_l) - (\hat{\mathrm{co}}_j - \mathrm{m}_l)D_l \tag{H.7}$$

D'autre part, la distance entre marque de synthèse $\tilde{\mathrm{m}}_{j+1} - \tilde{\mathrm{m}}_j$ doit être égale à la période fondamentale moyenne voulue sur l'intervalle $[\hat{\mathrm{co}}_j, \hat{\mathrm{co}}_{j+1}]$:

$$\tilde{\mathrm{m}}_{j+1} - \tilde{\mathrm{m}}_j = \frac{1}{\hat{\mathrm{co}}_{j+1} - \hat{\mathrm{co}}_j} \int_{\hat{\mathrm{co}}_j}^{\hat{\mathrm{co}}_{j+1}} P(t)dt \tag{H.8}$$

L'intégrale exprimée sous forme de trapèze , en notant $P_l = P(\mathrm{m}_l)$ (voir FIG. H.2), s'exprime :

$$\begin{aligned}\int_{\hat{\mathrm{co}}_j}^{\hat{\mathrm{co}}_{j+1}} P(t)dt = &\frac{\left[\left(\frac{P_{l+1}-P_l}{\mathrm{m}_{l+1}-\mathrm{m}_l}\right)(\hat{\mathrm{co}}_j - \mathrm{m}_l) + P_l\right] + P_{l+1}}{2}(\mathrm{m}_{l+1} - \hat{\mathrm{co}}_j) \\ &+ \mathrm{cste}(\mathrm{m}_{l+1}, \mathrm{m}_n) \\ &+ \frac{P_n + \left[\left(\frac{P_{n+1}-P_n}{\mathrm{m}_{n+1}-\mathrm{m}_n}\right)(\hat{\mathrm{co}}_{j+1} - \mathrm{m}_n) + P_n\right]}{2}(\hat{\mathrm{co}}_{j+1} - \mathrm{m}_n)\end{aligned} \tag{H.9}$$

où «cste» représente les termes d'intégration du milieu $(\mathrm{m}_{l+1}, \mathrm{m}_{l+2}, ..., \mathrm{m}_n)$ dont le nombre dépend de la distance $\mathrm{m}_n - \mathrm{m}_{l+1}$ et donc de $\hat{\mathrm{co}}_{j+1}$ et $\hat{\mathrm{co}}_j$. Il ne s'agit donc pas à proprement parler d'une constante puisque cette distance est à priori inconnue. Ceci implique l'utilisation d'une méthode récursive. Nous partons d'une première approximation de la distance $\hat{\mathrm{co}}_{j+1} - \hat{\mathrm{co}}_j$ où $\hat{\mathrm{co}}_{j+1}$ est donnée par l'approximation constante de $f(t)$ et $D(t)$.

Après développement nous obtenons l'équation en $\hat{\mathrm{co}}_{j+1}$ (H.10) permettant de calculer la nouvelle valeur de $\hat{\mathrm{co}}_{j+1}$ qui sera réintroduite dans (H.10) afin de calculer la valeur suivante de la récursion et ainsi de suite.

Celle-ci donnera donc aussi la valeur de $\tilde{\mathrm{m}}_{j+1}$ à partir de $\tilde{\mathrm{m}}_j$ (moyennant $\beta(t)$).

$$A(\hat{\mathrm{co}}_{j+1})^2 + B(\hat{\mathrm{co}}_{j+1}) + C = 0 \tag{H.10}$$

où

$$A = D_n - \frac{1}{2}\frac{P_{n+1} - P_n}{\mathrm{m}_{n+1} - \mathrm{m}_n} \tag{H.11}$$

$$B = \beta(\mathrm{m}_n) - \beta(\mathrm{m}_l) - D_n\mathrm{m}_n + D_l\mathrm{m}_l - \hat{\mathrm{co}}_j(D_n + D_l) - \frac{P_n\mathrm{m}_{n+1} - P_{n+1}\mathrm{m}_n}{\mathrm{m}_{n+1} - \mathrm{m}_n} \tag{H.12}$$

$$\begin{aligned}C = &(\hat{\mathrm{co}}_j)^2\left(D_l + \frac{1}{2}\frac{P_{l+1} - P_l}{\mathrm{m}_{l+1} - \mathrm{m}_l}\right) \\ &+ \hat{\mathrm{co}}_j\left(-\beta(\mathrm{m}_n) + \beta(\mathrm{m}_l) + D_n\mathrm{m}_n - D_l\mathrm{m}_l + \frac{P_l\mathrm{m}_{l+1} - P_{l+1}\mathrm{m}_l}{\mathrm{m}_{l+1} - \mathrm{m}_l}\right) \\ &+ \frac{1}{2}\mathrm{m}_n\left(2P_n - \mathrm{m}_n\frac{P_{n+1} - P_n}{\mathrm{m}_{n+1} - \mathrm{m}_n}\right) - \frac{1}{2}\mathrm{m}_{l+1}\left(-\mathrm{m}_l\frac{P_{l+1} - P_l}{\mathrm{m}_{l+1} - \mathrm{m}_l} + P_l + P_{l+1}\right) \\ &- \mathrm{cste}(\mathrm{m}_{l+1}, \mathrm{m}_n)\end{aligned} \tag{H.13}$$

Annexe I

PSOLA : Du marquage PSOLA en temps continu au signal en temps discret

Nous proposons une méthode simple permettant de tenir compte du marquage PSOLA en temps continu appliqué au signal en temps discret.

Le marquage tm_i obtenu lors de l'analyse exprime un temps (non discret), il en est de même des marques de synthèse $\tilde{\mathrm{tm}}_j$. Ces marques sont utilisées pour le traitement d'un signal en temps discret $s(n)$.

Le passage de valeurs continues tm_i et $\tilde{\mathrm{tm}}_j$ aux valeurs discrètes m_i et $\tilde{\mathrm{m}}_j$ introduit une erreur sur le signal de synthèse. Cette erreur est d'autant plus importante que le rapport $\frac{\mathrm{f0}}{Fe}$ croît. Pour une période fondamentale définie en temps continu comme

$$\mathrm{T0} = \mathrm{tm}_{i+1} - \mathrm{tm}_i \tag{I.1}$$

cette période fondamentale devient en temps discret

$$\begin{aligned} \mathrm{T0}_m &= Te(\mathrm{m}_{i+1} - \mathrm{m}_i) \\ &= Te([\mathrm{tm}_{i+1}/Te] - [\mathrm{tm}_i/Te]) \end{aligned} \tag{I.2}$$

dans lequel Te désigne le pas d'échantillonnage (inverse de la fréquence d'échantillonnage).

Il va de soit que l'imprécision croit à mesure que Te croît. Dans le cas le plus critique, l'erreur est de un échantillon. Dans ce cas, l'erreur relative, définie comme $\Delta = \frac{\mathrm{T0}_m - \mathrm{T0}}{\mathrm{T0}}$ est égale à f0 et croît donc avec la fréquence fondamentale.

Cet effet de troncature peut être évitée de deux manières

- par sur-échantillonnage du signal (voir annexe F). La forme d'onde est dans ce cas calculée à la position exacte de la marque tm_i. Le coût du sur-échantillonnage est cependant élevé si celui-ci doit être effectué pour chaque forme d'onde .
- par modification du spectre. Étant donné la nature fréquentielle des traitements qui seront appliquées à ces formes d'onde , une solution basée sur la modification du spectre de phase est ici proposée (cette solution est cependant incomplète puisque le spectre de phase n'est pas corrigé).

Soit l'équation de synthèse PSOLA

$$\tilde{s}(t) = \sum_j s(t + \mathrm{tm}_i - \tilde{\mathrm{tm}}_j) \tag{I.3}$$

et sa ré-écriture en temps discret

$$\tilde{s}(nTe) = \sum_j s(nTe + \mathrm{m}_i Te - \tilde{\mathrm{m}}_j Te) \tag{I.4}$$

Décrivons les deux étapes du processus

$$\begin{aligned} s_1(t) &= s(t + \mathrm{tm}_i) \\ s'_1(nTe) &= s(nTe + Te\left[\frac{\mathrm{tm}_i}{Te}\right]) \\ s_1(t) &= s'_1(t + (\mathrm{tm}_i - Te\left[\frac{\mathrm{tm}_i}{Te}\right])) \\ S_1(\omega) &= \underbrace{S'_1(\omega) \cdot e^{j\omega(\mathrm{tm}_i - Te\left[\frac{\mathrm{tm}_i}{Te}\right])}}_{S_1>} \end{aligned} \tag{I.5}$$

$$\begin{aligned} \tilde{s}(t) &= s_1(t - \tilde{\mathrm{tm}}_j) \\ \tilde{s}'(nTe) &= s_1>(nTe - Te\left[\frac{\tilde{\mathrm{tm}}_j}{Te}\right]) \\ \tilde{s}(t) &= \tilde{s}'(nTe + (Te\left[\frac{\tilde{\mathrm{tm}}_j}{Te}\right] - \tilde{\mathrm{tm}}_j)) \\ \tilde{S}(\omega) &= \tilde{S}' \cdot e^{j\omega(Te\left[\frac{\tilde{\mathrm{tm}}_j}{Te}\right] - \tilde{\mathrm{tm}}_j)} \end{aligned} \tag{I.6}$$

En réunissant (I.5) et (I.6) nous trouvons la correction à apporter

$$\tilde{S}(\omega) = S'_1(\omega, \mathrm{m}_i) \cdot \boxed{\exp\left(j\omega(\mathrm{tm}_i - \tilde{\mathrm{tm}}_j - \left(Te\left[\frac{\mathrm{tm}_i}{Te}\right] - Te\left[\frac{\tilde{\mathrm{tm}}_j}{Te}\right]\right)\right)} \tag{I.7}$$

Le terme encadré est la correction de phase à apporter à la forme d'onde élémentaire obtenue en temps discret à l'échantillon m_i et placée à l'échantillon $\tilde{\mathrm{m}}_j$ pour obtenir l'équivalent de la forme d'onde au temps tm_i placée au temps $\tilde{\mathrm{tm}}_j$.

Annexe J

Modèle sinusoïdal : Détermination des paramètres s (parabole) et σ (gaussienne) pour les fenêtres cosinusoïdales

Dans cette annexe, nous cherchons à déterminer les valeurs du paramètre s d'une fonction parabole et σ d'une fonction gaussienne telle que, avec ces valeurs, ces fonctions approximent «le mieux possible» la forme de la transformée de Fourier d'une fenêtre cosinusoïdale.

J.1 Détermination du paramètre s (parabole) pour les fenêtres cosinusoïdales

Soit la parabole $P(f)$ d'expression

$$P(f) = P_h - \frac{1}{4s}(f - f_h)^2 \tag{J.1}$$

dans laquelle f désigne les fréquences.

La forme de la parabole dépend du paramètre s. Le paramètre s étant une fonction non-linéaire de la largeur temporelle L de la fenêtre, nous définissons donc le paramètre τ, linéaire en L, tel que

$$\boxed{\tau \triangleq 2\sqrt{s}} \tag{J.2}$$

et réécrivons l'équation de la parabole

$$P(f) = P_h - \frac{1}{\tau^2}(f - f_h)^2 \tag{J.3}$$

τ étant inversement proportionnel à la largeur temporelle L de la fenêtre, nous définissons le paramètre normalisé α tel que

$$\boxed{\alpha \triangleq \tau \cdot L} \tag{J.4}$$

Nous cherchons maintenant à déterminer les valeurs du paramètre α permettant d'approximer «le mieux possible» la forme de la transformée de Fourier $H(f)$ d'une fenêtre cosinusoïdale. Selon l'échelle dans laquelle est exprimée $H(f)$ (échelle de puissance ou de log-amplitude), nous devons déterminer la valeur de α permettant d'approximer par une parabole le mieux possible la forme du

- spectre de puissance d'une fenêtre cosinusoïdale : $P = A^2$
- spectre de log-amplitude : $P = \log(A)$

Puisque ces valeurs de α doivent servir à l'estimation des fréquences et des amplitude d'un modèle sinusoïdale par régression parabolique (voir partie 4.2.1.1 85), les deux critères à minimiser pour déterminer α sont

critère 1 : minimisation de l'erreur d'estimation de fréquence

critère 2 : minimisation de la différence d'énergie parabole/fenêtre cosinusoïdale

Pourquoi minimiser deux critères ? Réponse : Une minimisation d'énergie ne conduit pas à une minimisation d'estimation de fréquence

La minimisation s'effectue sur un nombre $2M + 1$ de valeurs discrètes autour de la fréquence centrale de la réponse en fréquence de la fenêtre étudiée. La valeur de α est fonction du type de la fenêtre (hanning, hamming, blackman, gauss) et de M. Du fait de la dépendance en M, α dépend également de la résolution du spectre [1].

Pour les besoins de l'estimation de fréquence et d'amplitude par régression parabolique (voir partie 4.2.1.1 85), $M = 1$, i.e. l'ajustement s'effectue sur l'intervalle $[k_h - 1, k_h, k_h + 1]$, k_h étant la fréquence discrète la plus proche de la fréquence centrale de la fenêtre.

Dans le cas de l'estimation sur le spectre de puissance, α dépend également de la puissance du signal considéré. Dans ce cas, nous calculerons α pour la valeur normalisée de puissance $A_h^2 = 1$ et en déduisons les autres valeurs de α par $\alpha_{|A|^2} = \frac{\alpha_0}{\sqrt{|A|^2}}$.

Si nous définissons δ comme le décalage de la fréquence centrale ω_h de la fenêtre par rapport aux fréquences discrètes ω_k de la TFDCT : $\omega_h = \omega_k + \delta 2\pi \frac{Fe}{N}$; pourquoi doit-on calculer une valeur de τ pour chaque valeur de δ ? Réponse : La première chose que nous avons essayé est une minimisation globale de l'énergie en considérant une largeur importante du lobe de manière justement à rendre le plus possible τ indépendant de δ. Nous avons constaté que les résultats étaient nettement meilleurs si l'on réduisait la taille de l'intervalle (ceci est puisque l'approximation par un polynôme d'ordre 2 est d'autant plus exacte que l'on réduit l'intervalle, cnfr développement en série). Mais dés lors, la valeur de τ dépend de δ. Néanmoins nous indiquons les valeurs «clefs» de δ dont le τ permet d'annuler le biais sur l'intervalle le plus important.

Nous indiquons dans les tables suivantes, la valeur de α pour $H(f)$ exprimé en échelle de puissance et de log-amplitude, pour différents facteurs de prolongement par zéro et pour

[1] Rappel : nous définissons la résolution comme le rapport prolongement par zéro défini comme $ZP = N/L$)

différentes valeurs decalage δ. Ces valeurs ont été obtenues en minimisant le critère 1 ou 2 entre la parabole ($P = P_h - \frac{1}{4s}(f - f_h)^2$) et le spectre de puissance, ou de log-amplitude, de la fenêtre cosinusoïdale.

J.1.1 Minimisation de l'erreur d'estimation de fréquence

Conseil : Si nous définissons $biais(\delta)$ comme le biais obtenu pour un décalage de δ bin, et définissons $\mu(biais)$ comme la moyenne de biais pour les différents décalages, la valeur obtenue à 0.4 bin permet de minimiser $\mu(biais)$ c'est à dire de minimiser la moyenne des biais.

Valeurs de α :

Spectre de puissance (N=1024)

- fenêtre rectangulaire

α	δ=0.3 bin	0.4 bin	0.5 bin
ZP=N/L=1	3.515	2.766	2.351
2	0.800	0.814	0.831
4	0.603	0.606	0.610
8	0.563	0.564	0.565

- fenêtre de hanning

α	0.3 bin	0.4 bin	0.5 bin
1	1.6537	1.6727	1.6988
2	1.0296	1.0385	1.0500
4	0.9125	0.9147	0.9176
8	0.8823	0.8829	0.8836

- fenêtre de hamming

α	0.3 bin	0.4 bin	0.5 bin
1	1.7264	1.7324	1.7445
2	0.9676	0.9774	0.9902
4	0.8359	0.8384	0.8416
8	0.8076	0.8082	0.8090

- fenêtre de blackman

α	0.3 bin	0.4 bin	0.5 bin
1	1.648	1.670	1.701
2	1.134	1.143	1.153
4	1.033	1.036	1.038
8	1.013	1.014	1.014

- fenêtre de gauss

α	0.3 bin	0.4 bin	0.5 bin
1	2.7673	2.7808	2.7981
2	2.6002	2.6036	2.6080
4	2.5598	2.5606	2.5617
8	2.5498	2.5500	2.5503

Spectre de log-amplitude (N=1024)

- fenêtre rectangulaire

α	0.3 bin	0.4 bin	0.5 bin
1	1.386	1.371	1.347
2	0.696	0.687	0.675
4	0.761	0.760	0.758
8	0.775	0.775	0.774

- fenêtre de hanning

α	0.3 bin	0.4 bin	0.5 bin
1	1.1401	1.1290	1.1137
2	1.2203	1.2187	1.2166
4	1.2349	1.2346	1.2341
8	1.2341	1.2340	1.2339

- fenêtre de hamming

α	0.3 bin	0.4 bin	0.5 bin
1	1.0300	1.0204	1.0073
2	1.1037	1.1022	1.1003
4	1.1206	1.1203	1.1199
8	1.1271	1.1270	1.1270

- fenêtre de blackman (log-amplitude)

α	0.3 bin	0.4 bin	0.5 bin
1	1.361	1.357	1.352
2	1.401	1.401	1.400
4	1.413	1.413	1.413
8	1.421	1.421	1.421

- fenêtre de gauss

α	0.3 bin	0.4 bin	0.5 bin
1	$\frac{\sqrt{2}}{2\pi\sigma}L$		
2			
4			
8			

J.1.2 Minimisation de l'énergie :

Conseil : Si nous définissons $biais(\delta)$ comme le biais obtenu pour un décalage de δ bin, et définissons $\mu(biais)$ comme la moyenne des biais pour les différents décalages, la valeur obtenue à 0 bin permet de minimiser $\mu(biais)$ c'est à dire de minimiser la moyenne des biais.

Valeurs de α :

Spectre de puissance (N=1024)

- fenêtre rectangulaire

α	0 bin	0.5 bin
1	0.999	1.455
2	0.647	0.777
4	0.574	0.602
8	0.556	0.563

- fenêtre de hanning

α	0 bin	0.5 bin
1	1.1536	1.4923
2	0.9439	1.0247
4	0.8931	0.9125
8	0.8776	0.8824

- fenêtre de hamming

α	0 bin	0.5 bin
1	1.1050	1.4730
2	0.8685	0.9600
4	0.8139	0.8358
8	0.8023	0.8076

- fenêtre de blackman

α	0 bin	0.5 bin
1	1.244	1.546
2	1.059	1.131
4	1.016	1.034
8	1.009	1.013

- fenêtre de gauss

α	0 bin	0.5 bin
1	2.6451	2.7645
2	2.5710	2.6005
4	2.5526	2.5599
8	2.5480	2.5498

Spectre de log-amplitude (N=1024)

- fenêtre rectangulaire

α	0 bin	0.5 bin
1	0.379	1.181
2	0.744	0.685
4	0.771	0.761
8	0.777	0.775

- fenêtre de hanning

α	0 bin	0.5 bin
1	1.2000	1.1276
2	1.2326	1.2197
4	1.2379	1.2348
8	1.2349	1.2341

- fenêtre de hamming

α	0 bin	0.5 bin
1	1.0834	1.0198
2	1.1149	1.1031
4	1.1233	1.1206
8	1.1278	1.1271

- fenêtre de blackman

α	0 bin	0.5 bin
1	1.389	1.359
2	1.408	1.401
4	1.415	1.413
8	1.422	1.421

- fenêtre de gauss

α	0 bin	0.5 bin
1	$\frac{\sqrt{2}}{2\pi\sigma}L$	
2		
4		
8		

J.2 Détermination du paramètre σ (gaussienne) pour les fenêtres cosinusoïdales

En l'absence d'effet de troncature de la fenêtre gaussienne, σ s'obtient directement à partir de τ :

$$\sigma = \frac{\sqrt{2}}{2\pi\tau} \tag{J.5}$$

dans lequel τ est défini comme $\tau = \frac{\alpha}{L}$. Les valeurs de σ se déduisent donc directement de celles de s trouvées dans la partie précédente.

Annexe K

Modèle sinusoïdal : Modèle sinusoïdal non-stationnaire

Nous rappelons le modèle sinusoïdal de fréquence linéaire et d'amplitude gaussienne proposé par [MA89]. Nous montrons ensuite l'équivalence entre les solutions de ce modèle et de notre modèle de mesure de distorsion du spectre complexe [PR99a].

K.1 Modèle de fréquence linéaire et d'amplitude gaussienne [MA89]

Soit le modèle

$$\hat{x}(t) = A(e^{-\mu t^2} e^{\lambda t}) e^{j(\phi_0 + \omega_0 t + \Delta t^2)} \tag{K.1}$$

La Transformée de Fourier du modèle sinusoïdal fenêtré $\hat{s}(t) = \hat{x}(t) \cdot h(t)$ peut s'écrire en notant en terme de module/argument [MA89] :

$$\hat{S}(\omega) = A e^{j\phi_0} \frac{e^{\frac{\sigma^2}{2} \frac{(\lambda - j(\omega - \omega_0))^2}{(E - j2\Delta\sigma^2)}}}{\sqrt{E - j\Delta 2\sigma^2}} \tag{K.2}$$

dans lequel E est égale à $E \triangleq (1 + 2\mu\sigma^2)$

K.2 Equivalence des solutions du modèle de [MA89] et du modèle de [PR99a]

Nous avons obtenus les solutions suivantes pour le modèle [MA89].

Remarque : *Les solutions sont exprimées sous une forme similaire à celles du modèle de fréquence linéaire et d'amplitude linéaire afin de faciliter la comparaison.*

$S(\omega)$ est un polynôme d'ordre 2 en ω. En égalisant (K.2) avec les paramètres des polynômes d'ordre 2 du spectre de log-amplitude $P_{\log|S|}(\omega) = a_{\log|S|}\omega^2 + b_{\log|S|}\omega + c_{\log|S|}$ et

de phase $P_{\phi(S)}(\omega) = a_{\phi(S)}\omega^2 + b_{\phi(S)}\omega + c_{\phi(S)}$ nous obtenons :

$$\begin{cases} \mu = -\frac{1}{4}\left(\frac{2}{\sigma^2} + \frac{a_{\log|S|}}{G}\right) \\ \lambda = \frac{1}{2}\frac{b_{\phi(S)}a_{\log|S|} - a_{\phi(S)}b_{\log|S|}}{G} \\ \omega_0 = -\frac{1}{2}\frac{b_{\log|S|}a_{\log|S|} + a_{\phi(S)}b_{\phi(S)}}{G} \\ \Delta = -\frac{1}{4}\frac{a_{\phi(S)}}{G} \end{cases} \tag{K.3}$$

Explication de l'égalité $\sigma^2 = -2\frac{G}{a_{\log|S|}}$ utilisé dans la partie 7 Lorsque $\mu = 0$, i.e. en l'absence de modulation symétrique par rapport à $t = 0$, nous obtenons l´égalité $\sigma^2 = -2\frac{G}{a_{\log|S|}}$ dans lequel $G = a^2_{\log|S|} + a^2_{\phi(S)}$.

Également lorsque $\mu = 0$, et si nous approximons $e^{\lambda t}$ par $a_1/a0$ (développement de e^t autour de $t = 0$ égale à t pour l'ordre 1) alors (K.3) est équivalent à (4.41). (K.3) étant obtenu sans développement en série limitée, constitue donc une solution exacte. L'équivalence de (K.3) et de (4.41) justifie à posteriori le développement en série limitée à l'ordre 1 de $\log(\alpha'(\omega))$ et de $\beta'(\omega)$ effectué en dans la partie 7.

Annexe L

Modèle sinusoïdal : Comparaison des estimateurs des paramètres des modèles sinusoïdaux

Nous indiquons ici les conditions dans lesquelles nous avons comparé les méthodes d'estimation de fréquence et d'amplitude des composantes d'un modèle sinusoïdal, présentée dans la partie 4.2.3.

Les estimateurs de fréquence ω_h et d'amplitude A_h que nous avons testés sont :

i	interpolation du spectre sur $[\omega_{k-1}\omega_k\omega_{k+1}]$ (voir équation (4.3))
r	régression du spectre sur $[\omega_{k-2}\omega_k\omega_{k+2}]$
R	régression du spectre sur $[\omega_{k-1}\omega_k\omega_{k+1}]$ en imposant la forme du spectre autour de ω_h (voir équation (4.4) et (4.6))
d	mesure de distorsion du spectre complexe (voir équation (4.41))
p	fréquence instantanée (voir équation (4.9))
h	moindres carrés itératif (voir équation (4.25))
l	moindres carrés polynôme d'amplitude complexe (voir équation (4.48))

Nous comparons également les estimations de modulations linéaires de fréquence et d'amplitude lorsque les estimateurs le permettent : estimateur **d** et **h** .

Les facteurs influençant l'estimation que nous avons testés sont :

Le décalage δ de la fréquence ω_h de la sinusoïde test par rapport aux fréquences discrètes ω_k de la TFDCT : $\omega_h = \omega_k + \delta 2\pi \frac{Fe}{N}$.
L'échelle de représentation du spectre d'amplitude : échelle de puissance, échelle logarithmique
Le facteur de prolongement par zéro appliqué : 1, 2
Le type de la fenêtre de pondération : blackman, gauss
Le rapport signal à bruit SNR (Signal to Noise Ratio)
Les modulations de fréquence et d'amplitude
La proximité de composantes voisines de ω_h et la relation de phase de ces composantes avec la phase ϕ_h de la sinusoïde test

Création des signaux tests La partie déterministe $x(n)$ du signal test de fréquence ω_h, d'amplitude A_h, et de longueur L échantillons s'exprime : $x(n) = A_h \cos(\omega_h[0 : L-1]/Fe)$.

L'amplitude de A_h est prise égale à 1 dans l'ensemble des expériences sauf dans l'expérience testant l'effet des modulations de fréquence et d'amplitude. Dans ce cas elle vaut 1 au centre de l'observation.

La fréquence ω_h est choisie de manière à observer 8 périodes fondamentales sur la durée de la fenêtre ($\omega_h = 2\pi 8Fe/L$), sauf dans la dernière expérience concernant les basses résolutions où elle est choisie de manière à observer 3 périodes fondamentales.

Décalage de la fréquence de la sinusoïde Soit δ le décalage de la fréquence ω_h de la sinusoïde test par rapport à la position de la fréquence discrète ω_k de la TFDCT la plus proche de ω_h : $\omega_h = \omega_k + \delta 2\pi \frac{Fe}{N}$. En l'absence de bruit, les valeurs suivantes sont testées $\delta = [-0.5, -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.4]$. En présence de bruit, le nombre de valeur est réduit aux trois valeurs les plus représentatives $\delta = [-0.5, -0.25, 0]$ ou uniquement à $\delta = [-0.25]$. La fréquence d'échantillonnage Fe est prise égale au nombre N de fréquences discrètes de la TFDCT. De cette manière δ exprime donc également un décalage en Hz.

Définition du rapport signal à bruit Soit $b(n)$ la partie aléatoire de type bruit blanc gaussien de variance σ_b^2. Le signal test total s'exprime $s(n) = x(n) + b(n)$. Pour une fenêtre de longueur L, nous définissons le rapport signal à bruit[1] comme

$$SNR = \log_{10} \frac{\sqrt{\frac{1}{L}(\sum_n x^2(n))}}{\sigma_b} \tag{L.1}$$

Pour chaque valeur de SNR, 1000 réalisations sont effectuées. Les valeurs de biais, variance et "erreur quadratique moyenne" ("mean square error" ou mse) indiquées sont obtenus sur cet ensemble de 1000 réalisations.

Calcul du biais, variance et erreur quadratique moyenne Les définitions suivantes ont été utilisées

- Biais : $biais(\hat{\theta}) = E\{\hat{\theta}\} - \theta$
- Variance : $variance(\hat{\theta}) = \{\hat{\theta}^2\} - E\{\hat{\theta}\}^2$
- Erreur quadratique moyenne ("Mean Square Error" ou mse) : $mse = E\{(\hat{\theta} - \theta)^2\}$ [2]. Les valeurs indiquées par le terme mse dans la suite correspondent en fait à la *racine carré de la valeur mse*. De même, les variances sont indiquées sous forme d'écart-types. Ceci de manière à permettre une comparaison plus aisée avec les valeurs de biais.

Moyenne des biais, variance et erreur quadratique moyenne Dans la suite nous parlerons de biais-moyen, de variance-moyenne et de valeur mse-moyenne. Il s'agit des moyennes des valeurs absolues de biais, variance et valeur mse pour l'ensemble des décalages δ.

[1] Une définition du SNR utilisant, non pas la variance du bruit, mais l'intégrale de la Densité Spectrale de Puissance dans la largeur de bande effective de l'estimation (bande fréquentielle réellement utilisée pour l'estimation) serait plus réaliste. Cependant ceci impliquerait la définition d'un SNR différent pour chaque estimateur puisque les estimateurs étudiés utilise des plages fréquentielles différentes, qui de plus peuvent varier (itérations successives de l'estimateur h). Pour cette raison la définition du SNR est choisie comme le rapport de l'énergie du signal sur la variance du bruit. Nous gardons cependant en tête que cette définition surestime l'importance du bruit par rapport au bruit influençant réellement l'estimation.

[2] Rappelons que l'erreur quadratique moyenne tient simultanément compte du biais et de la variance : $mse(\hat{\theta}) = biais^2(\hat{\theta}) + variance(\hat{\theta})$

Initialisation des algorithmes L'initialisation des algorithmes (i.e. le point milieu des vecteurs utilisés pour l'interpolation/régression, le point du vecteur de fréquence instantanée, la fréquence initiale des algorithmes de moindre carrés) est pris égale à ω_k. Ceci constitue une hypothèse raisonnable dans la mesure où, dans le cas de niveau de bruit pas trop élevés, il s'agit du maximum local d'amplitude.

À propos de l'utilisation d'un signal analytique : Étant donné que ω_h se trouve proche de l'origine des fréquences, afin d'éviter l'influence de l'axe négatif des fréquences, nous utilisons le signal analytique correspondant au signal $s(n)$ pour les estimateurs basés sur la forme du spectre (**i** ,**r** ,**R** ,**d**).

À propos du SNR effectif et du prolongement par zéro : La finesse fréquentielle est gardée constante dans toutes les expériences. Pour une TFDCT sur N points, et une fréquence d'échantillonnage notée Fe, nous prenons $N = Fe = 1024$. Ceci permet de garder une finesse fréquentielle identique entre les expériences et autorise les comparaisons entre expériences. Le prolongement par zéro du signal est de ce fait effectué, non pas par augmentation du nombre de point fréquentiel N, mais par diminution de la taille de la fenêtre. La fréquence ω_h est augmentée de manière à observer toujours un nombre de période du signal égal à 8. Dans une bande de fréquence donnée, la diminution de la taille de la fenêtre engendre une augmentation de la puissance du bruit $b(t)$ relativement à celle de la composante sinusoïdale $x(n)$ (L'espérance de la valeur d'un point du spectre de puissance de $b(n)$ est égale à $\overline{|X(k)|^2} = \sigma_b^2 \sum_n h^2(n)$. La valeur du spectre de puissance de $s(n)$ pour $\omega_k = \omega_h$ est égale à $|X(k)|^2 = \frac{A^2}{4}(\sum_n h(n))^2$.) Dans nos expériences, pour corriger cela, la variance du bruit est divisée par deux pour un facteur de prolongement par zéro de 2.

À propos de l'estimateur h : Pour l'estimateur **h** , 10 itérations ont été effectuées. La fréquence d'initialisation de l'algorithme étant proche de la valeur cherchée ω_h, la stabilité des résultats est obtenue en dessous de 10 itérations.

À propos de l'estimation de l'amplitude pour p : L'estimateur **p** est un estimateur uniquement de fréquence. Étant donné ses bonnes propriétés (mse faible constatées a posteriori), nous l'utilisons afin de tester un estimateur d'amplitude. Cette estimateur d'amplitude est l'estimateur de minimisation de l'erreur quadratique de modélisation **locale** en fréquence :

$$A_h = \frac{\sum_{\omega_k \in W_h} S(\omega_k) H(\omega_h - \omega_k)}{\sum_{\omega_k \in W_h} |H(\omega_k)|^2} \tag{L.2}$$

À propos du choix de s pour l'estimateur R et d : Les estimateurs **R** et **d** reposent sur un critère de forme du spectre (forme du lobe). Le paramètre **s** de la parabole est utilisé dans l'estimateur **R** afin d'imposer la forme du lobe et est utilisé dans l'estimateur **d** comme paramètre équivalent au σ de la fenêtre gaussienne. La forme du lobe est modélisée dans les deux cas par une parabole.

Dans le cas d'une fenêtre de pondération cosinusoïdale (hanning, hamming, blackman), la forme du spectre ne correspond plus à celle d'une parabole mais nous pouvons cependant minimiser la différence de forme par un choix adéquat du paramètre **s** de la parabole. Comme nous le montrons en annexe J, deux valeurs de **s** différentes peuvent être obtenues pour une fenêtre cosinusoïdale donnée : l'une correspond à (critère A) la minimisation d'un critère d'énergie entre la fenêtre cosinusoïdale et la parabole de paramètre s, l'autre correspond à (critère B) la minimisation du biais de la fréquence obtenue par l'estimateur **R** en utilisant la fenêtre cosinusoïdale. Dans le cas d'une fenêtre de pondération cosinusoïdale, l'estimation se fait en deux étapes : une étape pour l'estimation de fréquence (en utilisant **s** résultant du critère B) et une étape pour l'estimation d'amplitude (en utilisant **s** résultant du critère A). Ceci permet d'utiliser la valeur du paramètre **s** la plus appropriée à chaque cas.

À propos du type de fenêtre pour les estimateurs de forme i ,r ,R ,d : Pour les estimateurs **i** ,**r** ,**R** ,**d** seules les fenêtres de blackman (0.42, 0.5, 0.08) et gaussiennes ont été testées. L'utilisation d'une fenêtre gaussienne est justifiée par l'expression théorique (4.5) obtenue. La fenêtre de blackman est choisie parmi l'ensemble des fenêtres cosinusoïdales puisqu'il s'agit de celle dont la forme est la plus proche de la fenêtre gaussienne (selon un critère d'erreur quadratique de modélisation ϵ : rectangulaire : ϵ =0.46, hanning : ϵ =0.0575, hamming : ϵ =0.0956, blackman : ϵ =0.0288).

À propos du type de fenêtre pour l'estimateur p : La fenêtre de pondération utilisée pour l'estimateur **p** est la fenêtre de hamming. Le choix de cette fenêtre est effectué en raison du biais de fréquence obtenue avec cette fenêtre qui est inférieure à celui obtenu avec les autres fenêtres. En effet, nous obtenons les valeurs suivantes de mse de fréquence pour un ensemble de 1000 réalisations avec un niveau de bruit de variance égale à 2 (amplitude de la composante égale à 1) : rectangulaire : mse=0.09, hanning : mse=0.035, hamming : mse=0.03, blackman : mse=0.055. Dans [Har78], la fenêtre de hamming n'est cependant pas la fenêtre d'ENBW (Equivalent Noise Band Width) minimal. La fenêtre d'ENBW minimal est la fenêtre rectangulaire (ENBW-rectangle : 1, ENBW-hanning : 1.5, ENBW-hamming : 1.36, ENBW-blackman : 1.73). Cependant, l'utilisation d'une fenêtre discontinue aux bords, telle la fenêtre rectangulaire, pour l'estimation de la fréquence instantanée dont le calcul repose sur la dérivé de la fenêtre introduit un biais supplémentaire.

À propos du type de fenêtre pour les algorithmes moindres carrés h ,l : La fenêtre d'analyse utilisée pour les estimateurs **h** et **l** est la fenêtre rectangulaire. Dans la mesure où l'initialisation des algorithmes s'effectue près de la fréquence exacte et sur le lobe principale, l'utilisation d'une fenêtre rectangulaire permet une meilleur résistance au bruit (valeur ENBW la plus faible).

Pour l'algorithme itératif de **h** , l'utilisation de la fenêtre rectangulaire présente un autre avantage, celui d'une convergence plus rapide (dérivée plus importante de la fenêtre rectangulaire autour de ω_h due à l'étroitesse de son lobe principal).

Remarque à propos du prolongement par zéro et de la troncature de la fenêtre de gauss : Nous définissons la durée effective de la fenêtre de gauss, L_{eff}, comme la durée renfermant 99% de son énergie (L_{eff} est égale à 6 σ) et L comme la durée totale de la fenêtre.

1) Pour $L = 6\sigma$, Leff et L sont égaux.

2) Pour $L = 12\sigma$, L_{eff} n'est plus égale à L. Comme nous désirons garder N constant (finesse fréquentielle constante), L doit rester constant, et le passage de 6 à 12 σ implique

la diminution de L_{eff} de moitié. Il est évident que puisque $L/(6\sigma) > L/(12\sigma)$ le passage de 6σ à 12σ diminue la durée effective du signal et donc diminue le rapport signal à bruit.

3) Nous voulons garder N constant, donc L constant et augmenter σ mais sans changer le SNR. Comment faire ? Théoriquement nous pourrions diminuer la valeur de la variance du bruit s_b. La solution que avons choisie est de considérer que dans le cas de l'utilisation d'un facteur 12σ, les parties du signal fenêtrées au delà de $t = \pm 3\sigma$ n'apporte que 1% d'énergie et constitue donc une contribution très faible. Dés lors nous considérons que cette contribution peut s'apparenter à un facteur de prolongement par zéro =2. Nous considérons donc l'équivalence entre un facteur prolongement par zéro =1 appliquées à une fenêtre de taille $L = 6\sigma = Leff$ et un facteur prolongement par zéro =2 appliqué à une fenêtre de taille $L = 12\sigma = 2Leff$.

Annexe M

Modèle sinusoïdal : Mesure de l'erreur de modélisation d'un modèle sinusoïdal

Nous rappelons la méthode d'estimation des amplitudes et des phases d'un modèle sinusoïdal par minimisation de l'erreur quadratique, erreur de modélisation. Nous montrons ensuite l'équivalence entre cette erreur de modélisation et la corrélation complexe entre le spectre du signal et la TF de la fenêtre d'analyse utilisée pour la TFDCT.

M.1 Estimation de l'amplitude A_h du modèle sinusoïdal par minimisation de l'erreur quadratique de modélisation locale en fréquence

Soit $\hat{S}_h = A_h e^{j\omega_h t+\phi_h}$ [1] la $h^{\text{ème}}$ composante du modèle sinusoïdal, $H(\omega)$ la réponse fréquentielle de la fenêtre d'observation, $S(\omega_k)$ la TFDCT du signal observé et W_h la largeur fréquentielle considérée.

- L'amplitude A_h est obtenue par minimisation de l'erreur quadratique de modélisation ϵ_h par rapport à A_h :

$$\begin{aligned}\epsilon_h &= \sum_{\omega_k \in W_h} |S(\omega_k) - A_h H(\omega_h - \omega_k)|^2 \\ &= \sum_{\omega_k \in W_h} S^2(\omega_k) + A_h^2 H^2(\omega_h - \omega_k) - 2S(\omega_k) A_h H(\omega_h - \omega_k)\end{aligned} \tag{M.1}$$

  La valeur de A_h est celle permettant de minimiser l'erreur quadratique

$$\frac{\partial \epsilon_h}{\partial A_h} = \sum_{\omega_k \in W_h} 2A_h H^2(\omega_h - \omega_k) - 2S(\omega_k) H(\omega_h - \omega_k) = 0 \tag{M.2}$$

[1] Nous considérons uniquement l'axe positif des fréquences.

$$\boxed{A_h = \frac{\sum\limits_{\omega_k \in W_h} S(\omega_k) H(\omega_h - \omega_k)}{\sum\limits_{\omega_k \in W_h} |H(\omega_k)|^2}} \tag{M.3}$$

- L'obtention de l'amplitude complexe $A_h^* = |A_h| e^{j \arg(A_h)}$ s'obtient de la même manière par minimisation de l'erreur quadratique de modélisation ϵ_h par rapport à A_h. Nous notons $X_\Re$ et $X_\Im$ ma partie réelle et imaginaire de X.

$$\begin{aligned}
\epsilon_h &= \sum |S_\Re + jS_\Im - (A_\Re + jA_\Im)H|^2 \\
&= \sum |S_\Re - A_\Re H + j(S_\Im - A_\Im H)|^2 \\
&= \sum (S_\Re - A_\Re H)^2 + (S_\Im - A_\Im H)^2 \\
&= \sum S_\Re^2 + S_\Im^2 + (A_\Re^2 + A_\Im^2)H^2 - 2S_\Re A_\Re H - 2S_\Im A_\Im H
\end{aligned} \tag{M.4}$$

La valeur de A_h est celle permettant de minimiser l'erreur quadratique

$$\begin{aligned}
\frac{\partial \epsilon}{\partial A_h} &= \frac{\partial \epsilon}{\partial A_{h,\Re}} + j \frac{\partial \epsilon}{\partial A_{h,\Im}} \\
&= \sum 2A_\Re H^2 - 2S_\Re H + j\left(2A_\Im H^2 - 2S_\Im H\right) \\
&= \sum 2AH^2 - 2SH
\end{aligned} \tag{M.5}$$

$$\boxed{A = \frac{\sum SH}{\sum H^2}} \tag{M.6}$$

M.2 Équivalence entre l'erreur quadratique de modélisation et la corrélation complexe

Nous montrons l'équivalence entre
- l'erreur quadratique normalisée de modélisation ϵ_h du signal par une composante sinusoïdale de fréquence w_h
- et la corrélation complexe $|c_h|$ entre la TFDCT $S(\omega_k)$ du signal autour de la fréquence et la réponse fréquentielle de la fenêtre d'observation transposée à la fréquence ω_h et échantillonnée aux fréquences ω_k de la TFDCT.

L'erreur quadratique de modélisation s'exprime

$$\epsilon_h = \frac{\sum\limits_{\omega_k \in W_h} |S(\omega_k) - \hat{S}_h(\omega_k)|^2}{\sum\limits_{\omega_k \in W_h} |S(\omega_k)|^2} \tag{M.7}$$

Pour un modèle sinusoïdal de paramètre constant par morceaux $\hat{S}_h = A_h e^{j\omega_h t}$ [2] et ϵ_h peut se réécrire

$$\epsilon_h = \frac{\sum |S - A_h H|^2}{\sum |S|^2} \tag{M.8}$$

En notant $\|S\|^2$ et $\|H\|^2$ les normes L^2 de $S(\omega_k)$ et de $H(\omega_h - \omega_k)$ sur l'intervalle W_h, en notant S^0 et H^0 les versions normées de S et H, et en utilisant le résultat obtenu dans la section précédente $A = (\sum SH)/(\sum H^2)$ nous pouvons réécrire

$$\begin{aligned}
\epsilon_h &= \sum \left| \frac{S}{\|S\|} - A_h \frac{H}{\|S\|} \right|^2 \\
&= \sum \left| \frac{S}{\|S\|} - \frac{\sum SH}{\|H\|^2} \cdot \frac{H}{\|S\|} \right|^2 \\
&= \sum \left| S^0 - \underbrace{(\sum S^0 H^0)}_{c_h} \cdot H^0 \right|^2
\end{aligned} \tag{M.9}$$

dans lequel c_h est la valeur complexe du coefficient de corrélation.

$$\begin{aligned}
\epsilon_h &= \sum \left| S^0_{\Re} + j S^0_{\Im} - (c_{\Re} + j c_{\Im})\, H^0 \right|^2 \\
&= \sum \left| S^0_{\Re} - c_{\Re} H^0 + j \left(S^0_{\Im} - c_{\Im} H^0 \right) \right|^2 \\
&= \sum \left[(S^0_{\Re})^2 + (c_{\Re} H^0)^2 - 2 S^0_{\Re} c_{\Re} H^0 + (S^0_{\Im})^2 + (c_{\Im} H^0)^2 - 2 S^0_{\Im} c_{\Im} H^0 \right] \\
&= \sum \left[|S^0|^2 + |c|^2 (H^0)^2 \right] - 2 \sum \Re \left[S^0 H^0 c^* \right] \\
&= 1 + |c|^2 - 2\Re \left[\sum \underbrace{S^0 H^0}_{c} c^* \right] \\
&= 1 + |c|^2 - 2|c|^2 \\
&= 1 - |c|^2
\end{aligned} \tag{M.10}$$

$$\boxed{\epsilon_h = 1 - |c_h|^2} \tag{M.11}$$

[2] Nous considérons uniquement l'axe positif des fréquences.

Annexe N

Modèle sinusoïdal : Discrimination sinusoïde/bruit par calcul de l'erreur de modélisation

Nous étudions l'influence des paramètres d'analyse suivant
- durée L de l'observation temporelle
- largeur W_h de l'observation fréquentielle, (nous utiliserons la notation W_ω dans la suite à la place de W_h pour bien préciser qu'il s'agit d'une largeur exprimée en fréquence et non en «bins»)
- rapport signal à bruit,

sur la discrimination sinusoïde/bruit obtenue par le calcul de l'erreur de modélisation.

Les développements sont effectués pour l'erreur de modélisation en énergie plutôt que pour l'erreur de modélisation complexe[1] étant donné l'indétermination de l'espérance de la phase d'un bruit. La formule d'erreur de modélisation d'énergie utilisée s'exprime :

$$\epsilon = 1 - \frac{\sum_k |S(\omega)|^2}{\sum_k |\hat{S}(\omega_k)|^2} \tag{N.4}$$

[1] Rappel : Nous pouvons définir trois erreurs de modélisation

complexe

$$\epsilon_1 = \sum_k |S(\omega_k) - \hat{S}(\omega_k)|^2 \tag{N.1}$$

en énergie (formulation quadratique)

$$\epsilon_2 = \sum_k \left(|S(\omega_k)| - |\hat{S}(\omega_k)|\right)^2 \tag{N.2}$$

en énergie (formulation linéaire)

$$\epsilon_3 = \sum_k |S(\omega_k)|^2 - \sum_k |\hat{S}(\omega_k)|^2 \tag{N.3}$$

La discrimination que nous entendons est la différence de valeur entre l'erreur de modélisation obtenue pour une composante sinusoïdale, notée ϵ_S, et celle obtenue en région bruitée, notée ϵ_{NS}.

N.1 Rappels :

Energie d'un signal $x(n)$ de durée L et de TFCT sur N points $X(k)$.

Dans le cas d'un signal pondéré par une fenêtre $h(n)$, en notant $s(n) = x(n)h(n)$,

$$E = \sum_n (x(n)h(n))^2 = \frac{1}{N}\sum_k |S(k)|^2 \tag{N.5}$$

Puissance d'un signal de durée L et d'énergie E.

Dans le cas d'un signal pondéré par une fenêtre $h(n)$ cette expression devient

$$P = \frac{1}{\sum_n h^2(n)} E \tag{N.6}$$

On notera dans la suite
- $L_{\sum^2} = \sum_n h^2(n)$
- $L_{\sum} = \sum_n h(n)$

Remarque : *Dans le cas particulier d'un signal non pondéré ($h(n) = 1$ pour $n \in [1, L]$), $L_{\sum^2} = L_{\sum} = L$.*

Puissance d'un bruit blanc gaussien centré de variance σ_b^2

$$P = \sigma_b^2 \tag{N.7}$$

Puissance d'une sinusoïde d'amplitude A pour une durée tendant vers l'infini

$$P = \frac{A^2}{2} \tag{N.8}$$

N.1.1 Observation sur un spectre

pour une sinusoïde d'amplitude A : (en considérant que la fréquence de la sinusoïde correspond à une fréquence discrète (bin) $k0$)

$$|X(k0)| = \frac{A}{2} L_{\sum} \tag{N.9}$$

on en déduit

$$|X(k0)|^2 = \frac{A^2}{4} L_{\sum}^2 \tag{N.10}$$

Si l'énergie total du signal ($\frac{A^2}{2} L_{\sum^2}$) était concentrée en un seul point (un point sur l'axe positif des fréquences et un point sur l'axe négatif), on aurait $|X(k0)|^2 = |X(-k0)|^2 = \frac{A^2}{4} L_{\sum^2} N$. La différence entre $\frac{A^2}{4} L_{\sum}^2$ et $\frac{A^2}{4} L_{\sum^2} N$ provient de l'étalement fréquentiel dû au fenêtrage.

pour un bruit blanc gaussien centré de variance σ^2 : Espérance de $|X(k)|^2 \qquad \forall k$

$$\overline{|X(k)|^2} = \sigma_b^2 L_{\sum^2} \qquad \forall k \tag{N.11}$$

L'énergie totale E est égale à l'énergie en chaque bin du spectre puisque le bruit a une espérance constante en fréquence

On en déduit

$$\overline{|X(k)|} = \sigma_b \sqrt{L_{\sum^2}} \tag{N.12}$$

Récapitulatif :

Sinusoïde	Amplitude	Puissance	Energie	$\lvert X(k0)\rvert$	$\lvert X(k0)\rvert^2$
	A	$\frac{A^2}{2}$	$\frac{A^2}{2} L_{\sum^2}$	$\frac{A}{2} L_{\sum}$	$\frac{A^2}{4} L_{\sum}^2$
Bruit	Variance	Puissance	Energie	$\overline{\lvert X(k)\rvert^2}$	$\overline{\lvert X(k)\rvert}$
	σ^2	σ^2	$\sigma^2 L_{\sum^2}$	$\sigma^2 L_{\sum^2}$	$\sigma \sqrt{L_{\sum^2}}$

N.1.2 Energie dans une bande de fréquence

Energie d'un bruit blanc gaussien centré de variance σ_b^2 bruit dans une bande de fréquence : L'énergie totale d'un bruit de variance σ_b^2 pour une observation de durée L s'exprime

$$E_{NS} = PL_{\sum^2} = \sigma_b^2 L_{\sum^2} = \frac{1}{N}\sum_k |X(k)|^2 \tag{N.13}$$

- L'espérance de $|X(k)|^2 \quad \forall k$ s'exprime : $\overline{|X(k)|^2} = \sigma_b^2 L_{\sum^2} \quad \forall k$
- L'espérance de l'énergie dans une bande de fréquence de largeur W_k bin ($W_k \in [0, \frac{N}{2}]$ et $\frac{W_k}{N} = \frac{W_\omega}{Fe}$) :

$$\boxed{E_{NS,W_k} = 2\sigma_b^2 L_{\sum^2} \frac{W_k}{N}} \tag{N.14}$$

Energie d'une sinusoïde d'amplitude A dans une bande de fréquence : L'énergie totale d'une sinusoïde d'amplitude A pour une observation de durée L s'exprime

$$E_S = PL_{\sum^2} = \frac{A^2}{2} L_{\sum^2} = \frac{1}{N}\sum_k |X(k)|^2 \tag{N.15}$$

Bande passante équivalente B_k : Nous définissons une «bande passante équivalente» B_k tel que $2B_k|X(k0)|^2 = \sum_k |X(k)|^2$ dans lequel $|X(k0)|$ est l'amplitude de la composante du spectre discret la plus proche de la fréquence de la sinusoïde.

$$B_k = \frac{\sum_k |X(k)|^2}{2|X(k0)|^2} \tag{N.16}$$

En supposant $k0$ proche de la fréquence de la sinusoïde et donc $|X(k0)|^2$ proche de $\frac{A^2}{4} L_{\sum}^2$, B_k peut se réécrire :

$$B_k = N \frac{L_{\sum^2}}{L_{\sum}^2} \tag{N.17}$$

Nous reconnaissons dans le deuxième terme l'Equivalent Noise Bandwidth (ENBW) de [Har78].

Cette approximation permet de considérer l'étalement fréquentiel de l'énergie dû au fenêtrage du signal tout en gardant une grande simplicité des expressions (nous considérons une fenêtre de largeur variable en fonction de L mais d'amplitude constante en fréquence). Nous définissons également une bande passante équivalente en Hz : $B_\omega = B_k \frac{Fe}{N}$. B_k et B_ω sont inversement proportionnel à L. [2]

L'énergie totale de la sinusoïde se réécrit en utilisant B_k comme

$$E_S = \frac{A^2}{2} L_{\sum^2} = \frac{1}{N} 2B_k |X(k0)|^2 \tag{N.18}$$

[2] Pour $L \to \infty$, $B_k \to 1$ et $B_\omega \to \frac{Fe}{N}$. Dans ce cas toute l'énergie du signal est concentrée en un point et $|X(k0)|^2 = \frac{A^2}{4} L_{\sum^2} N$.

- La valeur de $|X(k_0)|^2$ s'exprime en fonction de B_k

$$|X(k0)|^2 = \frac{A^2}{4} L_{\Sigma^2} \frac{N}{B_k} \tag{N.19}$$

- L'énergie dans une bande de fréquence de largeur W_k bin dépend du rapport W_k sur B_k.
  Si $W_k < B_k$, alors

$$E_{S,W_k} = \frac{1}{N} 2W_k |X(k0)|^2 \tag{N.20}$$

$$\boxed{E_{S,W_k} = \frac{A^2}{2} L_{\Sigma}^2 \frac{W_k}{N}} \tag{N.21}$$

  Si $W_k > B_k$ alors

$$\boxed{E_{S,W_k} = \frac{A^2}{2} L_{\Sigma^2}} \tag{N.22}$$

N.2 Erreur de modélisation pour une région fréquentielle contenant une sinusoïde

L'erreur de modélisation sur une largeur fréquentielle W_ω et pour une région contenant une sinusoïde d'amplitude A et de fréquence $k0$ s'exprime comme le rapport de l'énergie de la sinusoïde à celle du signal. L'énergie du signal est égale à la somme de l'énergie de la sinusoïde et à celle du bruit [3] :

$$\epsilon_S = 1 - \frac{E_{S,W_\omega}}{E_{S,W_\omega} + E_{NS,W_\omega}} \tag{N.23}$$

Remarque Hypothèse indépendance statistique: *Nous n'avons pas tenu compte de l'influence du bruit dans l'estimation de l'amplitude de la sinusoïde. Une formulation exacte devrait tenir compte de cette influence supplémentaire.*

Si $W_k < B_k$ (largeur d'observation fréquentielle plus faible que la bande passante équivalente de la sinusoïde)]

$$\begin{aligned} \epsilon_S &= 1 - \frac{\frac{A^2}{2} L_{\Sigma}^2 \frac{W_k}{N}}{\frac{A^2}{2} L_{\Sigma}^2 \frac{W_k}{N} + 2\sigma_b^2 L_{\Sigma^2} \frac{W_k}{N}} \\ &= 1 - \frac{\frac{A^2}{2}}{\frac{A^2}{2} + 2\sigma_b^2 \frac{L_{\Sigma^2}}{L_{\Sigma}^2}} \end{aligned} \tag{N.24}$$

[3] Nous supposons l'indépendance entre l'énergie de la sinusoïde et celle du bruit : $E_{S+NS} = E_S + E_{NS}$.

Soit en définissant un rapport signal à bruit $SNR = \frac{A^2/2}{\sigma_b^2}$ et en utilisant la notation $ENBW = \frac{L_{\sum^2}}{L_{\sum}^2}$

$$\boxed{\epsilon_s = 1 - \frac{SNR}{SNR + 2ENBW}} \tag{N.25}$$

Interprétation : Pour un SNR tendant vers l'infini, l'erreur de modélisation ϵ tend vers 0 et est indépendante de $ENBW$. Pour les autres valeurs de SNR, ϵ dépend de $ENBW$. Puisque $ENBW$ diminue quand la durée temporelle de l'observation L augmente, ϵ diminue quand L augmente. Pour un SNR tendant vers 0, ϵ tend vers 1.

Si $W_k > B_k$ (largeur d'observation fréquentielle plus grande que la bande passante équivalente de la sinusoïde)]

$$\epsilon_S = 1 - \frac{\frac{A^2}{2} L_{\sum^2}}{\frac{A^2}{2} L_{\sum^2} + 2\sigma_b^2 L_{\sum^2} \frac{W_k}{N}} \tag{N.26}$$

Soit en définissant un rapport signal à bruit $SNR = \frac{A^2/2}{\sigma_b^2}$.

$$\boxed{\epsilon_s = 1 - \frac{SNR}{SNR + 2\frac{W_\omega}{Fe}}} \tag{N.27}$$

Interprétation : Pour un SNR tendant vers l'infini, l'erreur de modélisation ϵ tend vers 0 et est indépendante de la largeur fréquentielle de l'observation W_ω. Pour les autres valeurs de SNR, ϵ dépend de W_ω. Si W_ω augmente alors l'erreur augmente également. Pour un SNR tendant vers 0, ϵ tend vers 1.

N.3 Erreur de modélisation pour une région fréquentielle ne contenant pas de sinusoïde

L'erreur de modélisation s'exprime comme le rapport de l'énergie de la sinusoïde modélisée à partir de l'observation du bruit, énergie notée $E_{S(NS)}$, à l'énergie du signal, i.e. du bruit, énergie notée E_{NS}.

$$\epsilon_{NS} = 1 - \frac{E_{S(NS),W_\omega}}{E_{NS,W_\omega}} \tag{N.28}$$

N.3.1 Expression de $E_{S(NS)}$

Dans le cas d'une sinusoïde modélisée à partir de l'observation d'une sinusoïde, nous observons dans le bin $k0$ une valeur égale à $|X(k0)|^2 = \frac{A^2}{4} L_{\sum}^2$. L'énergie totale de la sinusoïde vaut $\frac{A^2}{2} L_{\sum^2}$, soit un facteur de correction de $2\frac{L_{\sum^2}}{L_{\sum}^2}$ par rapport à l'observation.

N.3.1.1 Cas 1) modélisation de l'amplitude de la sinusoïde à partir de l'observation du bruit en $|X(k0)|^2$

Dans le cas du bruit, nous observons $|X(k0)|^2 = \overline{|X(k)|^2} = \sigma_b^2 L_{\sum^2}$. L'énergie totale de la sinusoïde modélisée à partir de l'observation du bruit s'exprime, en appliquant le même facteur de correction que pour une sinusoïde, $E_{S(NS)} = 2\sigma^2 \frac{(L_{\sum^2})^2}{L_\sum^2}$.

Nous pouvons également définir une bande passante équivalente B_k tel que

$$2\frac{1}{N} B_k \sigma^2 L_{\sum^2} = 2\sigma^2 (L_{\sum^2})^2 / L_\sum^2 \tag{N.29}$$

et nous trouvons la même valeur que précédemment : $B_k = N\frac{L_{\sum^2}}{L_\sum^2}$.

L'énergie dans une bande de fréquence de largeur W_k bin dépend du rapport W_k sur B_k

si $W_k < B_k$, nous trouvons

$$\begin{aligned} E_{S(NS),W_k} &= \frac{1}{N} 2 W_k \sigma_b^2 L_{\sum^2} \\ &= 2\sigma_b^2 L_{\sum^2} \frac{W_k}{N} \end{aligned} \tag{N.30}$$

et donc l'erreur de modélisation s'exprime

$$\epsilon_{NS,W_k} = 1 - \frac{2\sigma_b^2 L_{\sum^2} \frac{W_k}{N}}{2\sigma_b^2 L_{\sum^2} \frac{W_k}{N}} = 1 \tag{N.31}$$

ce qui est logique puisque si $W_k < B_k$ les deux rectangles sont égaux (le rectangle équivalent à la sinusoïde et le rectangle W_k considérée du signal) voir FIG. N.1 FIGURE 2

Si $W_k > B_k$, nous trouvons

$$\begin{aligned} E_{S(NS)} &= \frac{1}{N} 2 B_k \sigma^2 L_{\sum^2} \\ &= 2\sigma^2 \frac{(L_{\sum^2})^2}{L_\sum^2} \end{aligned} \tag{N.32}$$

et donc l'erreur de modélisation s'exprime

$$\epsilon_{NS,W_k} = 1 - \frac{2\sigma^2 \frac{(L_{\sum^2})^2}{L_\sum^2}}{2\sigma^2 L_{\sum^2} \frac{W_k}{N}} \tag{N.33}$$

$$\boxed{\epsilon_{NS,W_k} = 1 - \frac{ENBW}{\frac{W_\omega}{Fe}}} \tag{N.34}$$

Interprétation :

La valeur de ϵ est indépendante de la variance du bruit σ_b^2. Au plus la durée de l'observation temporelle est grande L, au plus $ENBW$ est petit et au plus l'erreur ϵ tend vers 1. Pour un L donné, l'erreur décroît en fonction de W_ω.

N.3.1.2 Cas 2) modélisation de l'amplitude de la sinusoïde à partir de l'observation du bruit par minimisation de l'erreur quadratique

Dans ce cas, l'amplitude A de la sinusoïde est modélisée de manière à minimiser

$$\epsilon = \sum_{k \in W_k} (|S(\omega_k)| - AH(\omega_h - \omega_k))^2 \tag{N.35}$$

La minimisation conduit à

$$A = \frac{\sum_{k \in W_k} H|S|}{\sum_{k \in W_k} |H|^2} \tag{N.36}$$

En considérant l'espérance de $|X(k)|^2$ comme $\sigma^2 L_{\sum^2}$ nous trouvons l'espérance de $|S| = \sigma\sqrt{L_{\sum^2}}$ et donc celle de A

$$E(A) = \sigma\sqrt{L_{\sum^2}} \underbrace{\frac{\sum_{k \in W_k} H(k)}{\sum_{k \in W_k} H^2(k)}}_{D} \tag{N.37}$$

Nous notons D le dernier terme. D dépend de L et de W_k (D est inversement proportionnel à L, mais si on normalise la fenêtre, $H \rightarrow H/L_{\sum}$, alors D est proportionnel à L. D est proportionnel à W_k. , $D = 1$ pour $L = 1$).

Dans le cas d'une estimation par moindre carré, nous devons donc remplacer l'observation $\sigma_b^2 L_{\sum^2}$ par $\sigma_b^2 L_{\sum^2} D^2$ dans $E_{S(NS)}$ et $\epsilon_{S(NS)}$.

N.4 Discrimination entre sinusoïde et bruit

La discrimination entre sinusoïde et bruit est finalement définie comme la différence de valeur entre l'erreur de modélisation obtenue pour une région contenant une sinusoïde et celle pour une région ne contenant pas de sinusoïde :

Si $W_k < B_k$

$$\boxed{d = \epsilon_S - \epsilon_{NS} = \left(1 - \frac{SNR}{SNR + 2ENBW}\right) - \left(1 - \frac{ENBW}{\frac{W_\omega}{Fe}}\right)} \tag{N.38}$$

Si $W_k > B_k$

$$\boxed{d = \epsilon_S - \epsilon_{NS} = \left(1 - \frac{SNR}{SNR + 2\frac{W_\omega}{Fe}}\right) - \left(1 - \frac{ENBW}{\frac{W_\omega}{Fe}}\right)} \tag{N.39}$$

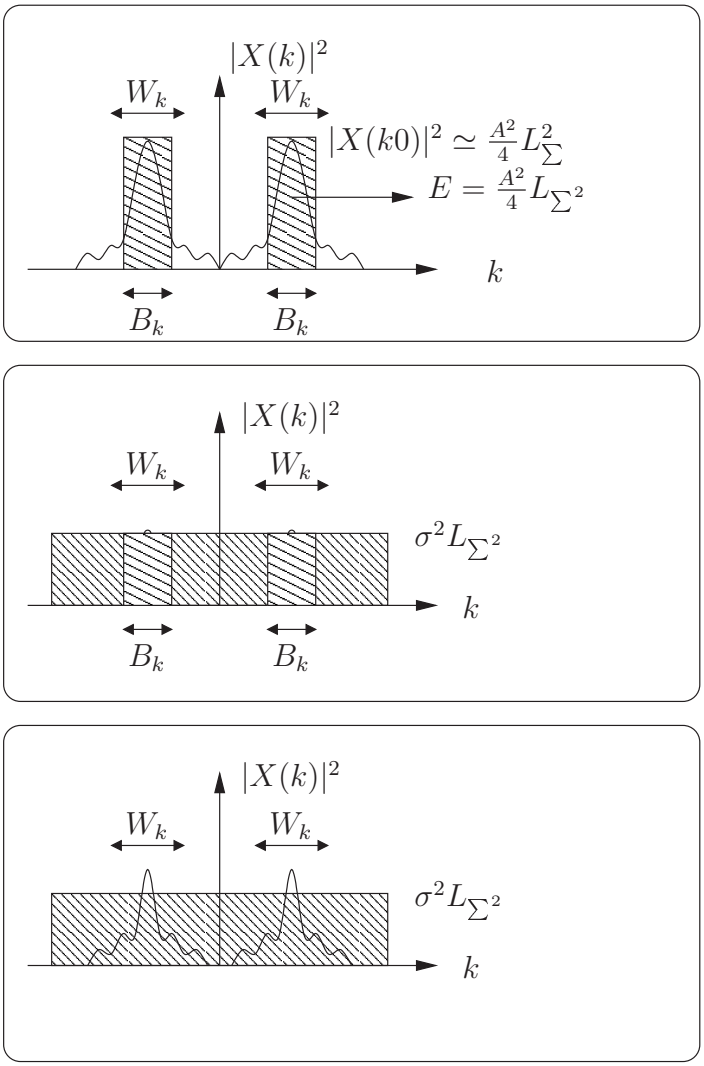

$|X(k)|^2$
W_k
W_k
$|X(k0)|^2 \simeq \frac{A^2}{4} L_{\sum}^2$
$E = \frac{A^2}{4} L_{\sum^2}$
k
B_k
B_k
$|X(k)|^2$
W_k
W_k
$\sigma^2 L_{\sum^2}$
k
B_k
B_k
$|X(k)|^2$
W_k
W_k
$\sigma^2 L_{\sum^2}$
k

FIG. N.1 –

Annexe O

Modèle sinusoïdal : Méthode du trajet de phase par polynôme cubique

Nous rappelons l'algorithme de création de trajets de phase en synthèse sinusoïdale par utilisation d'un polynôme cubique de phase.

L'algorithme du polynôme cubique de phase [MQ86b] cherche le trajet de phase sur l'intervalle $t = [rI, (r+1)I]$ satisfaisant aux contraintes suivantes, de fréquence et de phase, aux extrémités de l'intervalle :

$$\begin{cases} \varphi_h(rI) = \phi_{0,h,rI} \\ \varphi_h((r+1)I) = \phi_{0,h,(r+1)I} \\ \dfrac{\partial \varphi_h}{\partial t}(rI) = \omega_{h,rI} \\ \dfrac{\partial \varphi_h}{\partial t}((r+1)I) = \omega_{h,(r+1)I} \end{cases} \tag{O.1}$$

dans lequel I dénote le pas d'avancement de l'analyse, $\phi_{0,h,rI}/\phi_{0,h,(r+1)I}$ et $\omega_{h,rI}/\omega_{h,(r+1)I}$ dénotent les estimations de phases et de fréquences aux bords de l'intervalle.

Afin de résoudre le système de manière directe, [MQ86b] propose l'utilisation d'un polynôme $\varphi_h(t)$ d'ordre 3 : $\varphi_h(t) = \alpha + \beta t + \gamma t^2 + \delta t^3$. Dans ce cas la solution est donnée par

$$\begin{cases} \alpha = \phi_{0,h,m} \\ \beta = \omega_{h,m} \\ \gamma(M) = \dfrac{3}{I^2}(\phi_{0,h,m+1} - \phi_{0,h,m} - \omega_{h,m} I + 2\pi M) - \dfrac{1}{I}(\omega_{h,m+1} - \omega_m) \\ \delta(M) = \dfrac{-2}{I^3}(\phi_{0,h,m+1} - \phi_{0,h,m} - \omega_{h,m} I + 2\pi M) + \dfrac{1}{I^2}(\omega_{h,m+1} - \omega_{h,m}) \end{cases} \tag{O.2}$$

M est le facteur d'indétermination de phase de $\phi_{0,h,m+1}$. Il est déterminé de manière à rendre le trajet de $\varphi(t)$ le plus régulier (lisse) possible sur l'intervalle $[rI, (r+1)I]$. Cette régularité

est obtenue en cherchant la valeur de M qui minimise $f(M)$

$$f(M) = \int_0^I \left(\frac{\partial^2 \varphi(t)}{\partial t^2} \right)^2 dt \tag{O.3}$$

La solution est finalement donnée par

$$M = \left[\frac{1}{2\pi} \left((\phi_{0,h,m} + \omega_{h,m} I - \phi_{0,h,m+1}) + (\omega_{h,m+1} - \omega_{h,m}) \frac{I}{2} \right) \right] \tag{O.4}$$

dans lequel $[x]$ désigne l'entier le plus proche de x.

Annexe P

Modèle sinusoïdal : Equivalence de la méthode dite «shape invariant» et de celle du retard de phase relatif

Nous montrons l'équivalence entre la méthode de synthèse sinusoïdale dite «shape invariance» [QM92] et celle du retard de phase relatif [Fed98].

Dans la méthode shape-invariant, la phase $\phi_{v,h}(t_m)$ du filtre v du système est gardée invariante lors d'une dilatation ou d'une transposition :

$$\begin{aligned}\phi_{v,h}(t_m) &= \phi_h(t_m) - (t_m - t_0)\omega_h(t_m) \\ &= \hat{\phi}_h(t'_m) - (t'_m - t'_0)T\omega_h(t_m)\end{aligned} \tag{P.1}$$

Si nous supposons l'instant de fermeture de la glotte comme l'instant pour lequel la phase de la cosinusoïde à ω_0 est maximum, nous pouvons écrire $(t_m - t_0) = \frac{\phi_0(t_m)}{\omega_0(t_m)}$, et donc

$$\frac{\phi_{v,h,m}}{\omega_{h,m}} = \frac{1}{\omega_{h,m}}\left(\phi_{h,m} - \frac{\phi_{0,m}}{\omega_{0,m}}\omega_{h,m}\right) = \frac{\phi_{h,m}}{\omega_{h,m}} - \frac{\phi_{0,m}}{\omega_{0,m}} \tag{P.2}$$

Si nous faisons la même hypothèse sur le signal de synthèse, nous écrivons également $(t'_m - t'_0) = \frac{\hat{\phi}_0(t'_m)}{T\omega_0(t_m)}$, et donc

$$\frac{\phi_{v,h,m}}{T\omega_{h,m}} = \frac{1}{T\omega_{h,m}}\left(\hat{\phi}_{h,m'} - \frac{\hat{\phi}_{0,m'}}{T\omega_{0,m}}T\omega_{h,m}\right) = \frac{\hat{\phi}_{h,m'}}{T\omega_{h,m}} - \frac{\hat{\phi}_{0,m'}}{T\omega_{0,m}} \tag{P.3}$$

Soit

$$\boxed{\frac{\hat{\phi}_{h,m'}}{T\omega_{h,m}} - \frac{\hat{\phi}_{0,m'}}{T\omega_{0,m}} = \frac{\phi_{h,m}}{\omega_{h,m}} - \frac{\phi_{0,m}}{\omega_{0,m}}} \tag{P.4}$$

qui est la méthode du retard de phase relatif normalisé

Notations employées

Notation	Signification
t	temps
Fe	fréquence d'échantillonage
$n = t \cdot Fe$	échantillon
$x(t), x(n)$	signal
$s(n) = h(n) \cdot x(n)$	signal fenêtré
$s(t) = h(t) \cdot x(t)$	
$I_\#$, I_{sec}	pas d'avancement de l'analyse en échantillons, en secondes
$L_\#$, L_{sec}	taille de la fenêtre d'analyse en échantillons, en secondes
$N_\#$	taille de la FFT
x	valeur mesurée
$\hat{x}$	estimateur de x
$h \in [1, H]$	indice des composantes d'un modèle sinusoïdal
m, t_m	instants d'analyse ou milieu d'une fenêtre d'analyse
$\nu \in [0, 1]$	fréquence continue normalisée
$f \in [0, Fe]$	fréquence continue
$\omega \in [0, 2\pi Fe]$	pulsation continue
$k \in [0, N]$	indice vecteur FFT
$\nu_k = \frac{k}{N}$	fréquence normalisée correspondant à k
$f_k = \frac{k}{N} Fe$	fréquence correspondant à k
$\omega_k = 2\pi \frac{k}{N} Fe$	pulsation correspondant à k
TF	Transformée de Fourier
TFTD	Transformée de Fourier à Temps Discret
TFD	Transformée de Fourier Discrète
TFDCT	Transformée de Fourier Discrète à Court Terme

Bibliographie

[AF95] F. Auger and P. Flandrin, *Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method*, IEEE Trans. Signal Processing **43** (1995), no. 5, 1068–1089. 32, 86, 19, 20, 31

[AKI95] T. Abe, T. Kobayashi, and S. Imai, *Harmonics tracking and pitch extraction based on instantaneous frequency*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1995. 82, 86

[All77] J.B. Allen, *Short-term spectral analysis, synthesis and modification by discrete Fourier Transform*, 235–238. 181

[AMT91] A. Abrantes, J. Marques, and I. Transcoso, *Hybrid sinusoidal modeling of speech without voicing decision*, Proc. EUROSPEECH, 1991. 205

[AR77] J. B. Allen and L. R. Rabiner, *A unified approach to short-time Fourier analysis and synthesis*, Proceedings of the IEEE (1977). 15

[Bag] P. Bagshaw, *Speech files and laryngeal frequency contours*, http ://www.cstr.ed.ac.uk/∼pcb/fda_eval.tar.gz. 23, 30

[Bas89] M. Basseville, *Distance Measures for Signal Processing and Pattern Recognition*, Signal Processing **18** (1989), 349–369. 53

[Bas95] P. Bastien, *Adaptation de la technique PSOLA aux sons musicaux*, Master's thesis, DEA ATIAM, Ircam, France, 1995. 157

[BB83] M. Basseville and A. Benveniste, *Sequential detection of abrupt changes in spectral changes of digital signals*, IEEE Trans. Info. Theory **29** (1983), 709–724. 53

[Cha88] F. Charpentier, *Traitement de la parole par Analyse/Synthèse de Fourier application à la synthèse par diphones*, Ph.D. thesis, École Nationale Supérieur des Télécommunications (Paris), 1988. 9, 12, 22, 48, 86

[CLM95] O. Cappé, J. Laroche, and E. Moulines, *Regularized estimation of cepstrum envelope from discrete frequency points*, IEEE Trans. Acoust., Speech and Signal Processing (1995). 198

[CM88] F. Charpentier and E. Moulines, *Text-To-Speech Algorithms Based on FFT Synthesis*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1988. 186

[CM89] _____, *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones*, Proc. EUROSPEECH, 1989. 9, 22, 48, 188

[CO89] Y. Cheng and D. O'Shaughnessy, *Automatic and Reliable Estimation of Glottal Closure Instant and Period*, IEEE Trans. Acoust., Speech and Signal Processing (1989). 27

[Coh95] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, 1995. 38, 19

[Cor99] F. Corson, *Étude de la décomposition des sons en sinusoïdes et bruit*, Master's thesis, DEA ATIAM, Ircam, France, 1999. 82, 86

[CS86] F. Charpentier and M. Stella, *Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation*, Proc. Int. Conf. on Audio, Speech and Signal Proc. (Tokyo), 1986. 9, 186

[d'A89] C. d'Alessandro, *Représentation du signal de parole par une somme de fonctions élémentaires*, Ph.D. thesis, Laforia (Université Paris VI), 1989. 10, 45, 211

[DH97] P. Depalle and T. Helie, *Extraction of Spectral Peak Parameters using a Short-Time Fourier Transform Modeling and no-sidelobe Windows*, IEEE ASSP workshop on App. of Sig. Proc. to Audio and Acoust., 1997. 89

[Dov94] B. Doval, *Estimation de la fréquence fondamentale des signaux sonores*, Ph.D. thesis, Laforia (Université Paris VI), 1994. 143

[DQ97] Y. Ding and X. Qian, *Sinusoidal and Residual Decomposition and Residual Modeling of Musical Tones Using the QUASAR Signal Model*, Proc. Int. Computer Music Conf., 1997. 196

[dR89] C. d'Alessandro and X. Rodet, *Synthèse et analyse-synthèse par fonctions d'ondes formantiques*, J. Acoustique (1989). 10, 45

[DT96] P. Depalle and L. Tromp, *An improved additive analysis method using parametric modeling of the short-time Fourier transform*, Proc. Int. Computer Music Conf., 1996. 89

[Dut93] T. Dutoit, *High Quality Text-To-Speech Synthesis of the French Langage*, Ph.D. thesis, Polytechnique de Mons, 1993. 195

[Dut94] _____, *High Quality Text-To-Speech Synthesis : a comparison of four candidate algorithms*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1994. 88

[DYM89] G. Ducan, B. Yegnanarayana, and H. Murthy, *A Non-Parametric Method of Formant Estimation using Group Delay Spectra*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1989, pp. 572–575. 40

[EL96] M. Edgington and A. Lowry, *Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis*, International Conference on Spoken Langage Processing, ICSLP, vol. 3, 1996, pp. 1425–1428. 14

[Fed98] R. Di Federico, *Waveform Preserving Time Stretching and Pitch Shifting for Sinusoidal Models of Sound*, Proc. Digital Audio Effects, 1998. 199, 200, 73

[Fit99] K. R. Fitz, *The Re-assigned Bandwidth-Enhanced Method of Additive Synthesis*, Ph.D. thesis, University of Illinois, 1999. 86

[Fla72] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, 1972.

[Fla93] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, San Diego, California, Paris : Hermes, 1993. 19

[Gar92] G. Garcia, *Analyse des signaux sonores en termes de partiels et de bruit : Extraction automatique des trajets fréquentiels par des modèles de Markov cachés*, Master's thesis, DEA Automatique et Traitement des signaux, Orsay, France, 1992. 126, 127, 131, 221

[GBM+96] R. Gribonval, E. Bacry, S. Mallat, P. Depalle, and X. Rodet, *Analysis of Sound Signals with High Resolution Matching Pursuit*, Proc. IEEE Symp. Time-Freq. and Time-Scale Anal., June 1996, pp. 125–128. 92

[GL84] D. Griffin and J. Lim, *Signal Estimation from Modified Short-Time Fourier Transform*, IEEE Trans. Acoust., Speech and Signal Processing **32** (1984), no. 2, 236–243. 9, 182

[GL88] ———, *Multiband Excitation Vocoder*, IEEE Trans. Acoust., Speech and Signal Processing **36** (1988), no. 8, 1223–1235. 88, 118, 143, 145, 205

[GR90] T. Gallas and X. Rodet, *An Improved Cepstral Method for Deconvolution of Source-Filter Systems with Discrete Spectra*, Proc. Int. Computer Music Conf., 1990. 198

[Gri99] R. Gribonval, *Approximations non-linéaires pour l'analyse des signaux musicaux sonores*, Ph.D. thesis, Université Paris IX Dauphine, 1999. 92

[GS97] E. George and M. Smith, *Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model*, IEEE Trans. Speech and Audio Processing (1997). 143

[Har78] F. Harris, *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*, Proceedings of the IEEE **66** (1978), no. 1. 121, 54, 64

[HE75] H. Helmholtz and A. Ellis, *On the sensations of tone as a physiological basis for the theory of music*, Dover Publications, 1875. 15

[Hel97] T. Helie, *Estimation des paramètres de partiels par modélisation de la transformée de Fourier à court-terme utilisant des fenêtres spectrales sans lobes secondaires*, Master's thesis, DEA ATIAM, Ircam, France, 1997. 90

[HMC89] C. Hamon, E. Moulines, and F. Charpentier, *A Diphone Synthesis System Based on Time-Domain Prosodic Modifications*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1989. 178

[Jeh97] T. Jehan, *Musical Signal Parameter Estimation*, Master's thesis, CNMAT, Berkeley - IRISA, Rennes, 1997. 22

[Kaw00] H. Kawahara, *Vocal Fold and speech event detection using group delay*, Acoustical Society of Japan (2000). 41, 42

[Kay88] S. Kay, *Modern Spectral Estimation (Theory and Application)*, Prentice Hall, 1988. 40, 182

[KK97] R. Kortekaas and A. Kohlrausch, *Psychoacoustical evaluation of the pitch synchronous overlap and add speech-waveform manipulation technique using single-formant stimuli*, J. Acoust. Soc. of America (1997). 151

[Lar89] J. Laroche, *Étude d'un système d'analyse et de synthèse utilisant la méthode de Prony*, Ph.D. thesis, École Nationale Supérieur des Télécommunications (Paris), 1989. 97

[Ld89] J. Lienard and C. d'Alessandro, *Time-Frequency methods and phase space*, ch. Wavelet Transform and Granular Analysis of Speech, Springer Verlag, Berlin, 1989. 221

[Lev98] S. N. Levine, *Audio Representations for Data Compression and Compressed Domain Processing*, Ph.D. thesis, Stanford University, 1998. 53, 205

[LG96] T-H Li and J. Gibson, *Speech Analysis and Segmentation by Parametric Filtering*, IEEE Trans. Speech and Audio Processing (1996). 53

[MA86] J. Marques and L. Almeida, *A Background for Sinusoid Based Representation of Voiced Speech*, Proc. Int. Conf. on Audio, Speech and Signal Proc. (Tokyo), 1986. 15, 92, 94

[MA89] ———, *Frequency-Varying Sinusoidal Modeling of Speech*, IEEE Trans. Speech and Audio Processing (1989). iv, 49, 50

[MA94] J. Marques and A. Abrantes, *Hybrid harmonic coding of speech at low-bit-rates*, Speech Communication (1994), no. 14, 231–247. 88, 89

[Mal79] D. Mallah, *Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals*, no. 2, 121–133. 9

[Mal99] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 1999. 221

[Mas96] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*, Ph.D. thesis, University of Bristol, 1996. 92

[MC90] E. Moulines and F. Charpentier, *Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis using Diphones*, Speech Communication (1990). 9, 10, 11, 14

[MC97] M. Macon and M. Clements, *Sinusoidal Modeling and Modification of Unvoiced Speech*, IEEE Trans. Speech and Audio Processing (1997). 189

[MD92] C. McIntyre and D. Dermott, *A New Fine-Frequency Estimation Algorithm based on Parabolic Regression*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1992. 84, 85, 86, 103, 115

[MF90] E. Moulines and R. Di Francesco, *Detection of glottal closure by jumps in the statistical properties of the speech signal*, Speech Communication **9** (1990), no. 5/6, 401–418. 28

[MG76] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, 1976. 40

[MKW94] C. Ma, Y. Kamp, and L. Willems, *A Frobenius Norm Approach to Glottal Closure Detection from the Speech Signal*, IEEE Trans. Speech and Audio Processing (1994). 27, 28, 31

[ML95] E. Moulines and J. Laroche, *Non-parametric techniques for pitch-scale and time-scale modification of speech*, Speech Communication (1995). 14

[MMY89] H. Murthy, K. Murthy, and B. Yegnanarayana, *Formant Extraction from Fourier Transform Phase*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1989, pp. 484–487. 33, 40

[Moo78] J. A. Moorer, *The Use of the Phase Vocoder in Computer Music Applications*, Journal of the Audio Engineering Society (1978). 15

[Mou90] E. Moulines, *Algorithmes de codage et de modification des paramètres prosodiques pour la synthèse de parole à partir de texte*, Ph.D. thesis, École Nationale Supérieur des Télécommunications (Paris), 1990. 11

[MQ86a] R. McAulay and T. Quatieri, *Phase Modeling and its Application to Sinusoidal Transform Coding*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1986. 22

[MQ86b] ———, *Speech Analysis/Synthesis based on a Sinusoidal Representation*, IEEE Trans. Acoust., Speech and Signal Processing **34** (1986), no. 4, 744–754. 15, 16, 83, 126, 143, 196, 71

[MV95] E. Moulines and W. Verhelst, *Speech coding and Synthesis*, ch. 15, Elsevier Science B.V., 1995. 9, 151

[NMEL95] J. Navarro-Mesa and I. Esquerra-Llucia, *A Time-Frequency Approach to Epoch Detection*, Proc. EUROSPEECH, 1995, pp. 405–407.

[OS75] A. Oppenheim and R. Schafer, *Digital Signal Processing*, Prentice Hall, 1975. 40, 21

[Oud98] M. Campedel Oudot, *Application du modèle sinusoides et bruit au codage, débruitage et à la modification des sons de parole*, Ph.D. thesis, École Nationale Supérieur des Télécommunications (Paris), 1998. 143, 147, 198

[Por80] M. Portnoff, *Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis*, IEEE Trans. Acoust., Speech and Signal Processing (1980). 15

[PR98] G. Peeters and X. Rodet, *Sinusoidal versus Non-Sinusoidal Signal Characterisation*, Proc. Digital Audio Effects (Barcelona (Spain)), 1998. 129

[PR99a] _____, *Non-Stationary Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum*, Proc. Int. Congr. Signal Proc. Applic. and Tech. (Orlando), 1999. iv, 37, 41, 127, 49, 50

[PR99b] _____, *SINOLA : A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum*, Proc. Int. Computer Music Conf. (Peking (China)), 1999. 37, 41, 42, 53, 86, 127, 207

[QM92] T. Quatieri and R. McAulay, *Shape Invariant Time-Scale and Pitch Modification of Speech*, IEEE Trans. Signal Processing **40** (1992), no. 3, 497–510. 198, 200, 73

[RD86] E. Robinson and T. Durrani, *Geophysical Signal Processing*, Prentice Hall, 1986. 21

[RD92] X. Rodet and P. Depalle, *A new additive synthesis method using inverse Fourier transform and spectral envelopes*, Proc. Int. Computer Music Conf., 1992. 195

[Rd96] G. Richard and C. d'Alessandro, *Analysis/Synthesis and Modification of the Speech Aperiodic Component*, Speech Communication (1996), no. 19, 221–244. 145

[Rod97] X. Rodet, *Musical Sound Signal Analysis/Synthesis : Sinusoidal+Residual and Elementary Waveform Models*, Proc. IEEE Symp. Time-Freq. and Time-Scale Anal., 1997. 119

[Ros99] S. Rossignol, *Segmentation, Indexation et manipulation des signaux sonores*, Ph.D. thesis, Ircam (Université Paris VI), 1999. 211

[RW85] S. Roucos and A. Wilgus, *High Quality Time-Scale Modification for Speech*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1985. 9

[Sch98] D. Schwarz, *Spectral Envelopes in Sound Analysis and Synthesis*, Master's thesis, Universität Stuttgart, Fakultät Informatik, 1998. 198

[Smi98] J. Smith, *Bandlimited Interpolation - Introduction and Algorithm*, Tech. report, Center for Computer Research in Music and Acoustics (Stanford), http ://cm.stanford.edu/ jos/src/src.htm, 1998. 23

[SP00] N. Schnell and G. Peeters, *Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA)*, Proc. Int. Computer Music Conf. (Berlin (Germany)), 2000. 166, 5

[SS90] X. Serra and J. Smith, *Spectral Modeling Synthesis : a Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition*, Computer Music Journal **14** (1990), no. 4, 12–24. 15, 17, 84, 126, 205

[Str74] H. Strube, *Determination of the Instant of Glottal Closures from the Speech Wave*, J. Acoust. Soc. of America **56** (1974), no. 5, 1625–1629. 26, 27

[Sty96] Y. Stylianou, *Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole du Locuteur*, Ph.D. thesis, École Nationale Supérieur des Télécommunications (Paris), 1996. 17, 143, 180, 205, 37

[Sty98] _____, *Removing Phase Mismatches in Concatenative Speech Synthesis*, 3rd Speech Synthesis Workshop, Australia, 1998. 22, 47, 48

[SY95] R. Smits and B. Yegnanarayana, *Determination of Instants of Significant Excitation in Speech Using Group Delay Function*, IEEE Trans. Speech and Audio Processing (1995). 34, 41, 42, 185

[Td99] V. Tuan and C. d'Alessandro, *Robust Glottal Closure Detection using the Wavelet Transform*, Proc. EUROSPEECH, 1999. 22

[TG00] H. Thornburg and F. Gouyon, *A Flexible Analysis/Synthesis Method for Transients*, Proc. Int. Computer Music Conf., 2000, pp. 400–403. 53

[TP99] Z. Tychtl and J. Psutka, *Speech production based on the Mel-Frequency Cepstral Coefficients*, Proc. of Eurospeech, 1999. 14

[Ver98] W. Verhelst, *Overlap-Add Methods for Time-Scaling of Speech*, Speech Communication (1998). 9

[VLM97] T. Verma, S. Levine, and T. Meng, *Transient Modeling Synthesis : a flexible analysis/Synthesis tool for transient signals*, Proc. Int. Computer Music Conf., 1997. 53

[VR93] W. Verhelst and M. Roelands, *An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1993. 9

[WMG79] D. Wong, J. Markel, and A. Gray, *Least squares glottal inverse filtering from the acoustic speech waveform*, IEEE Trans. Acoust., Speech and Signal Processing **27** (1979), no. 4, 353–362. 26, 27

[WP94] Eric Walter and Luc Pronzato, *Identification de modèles paramétriques*, Masson, 1994. 147, 155

[YdD98] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, *An Iterative Algorithm for Decomposition of Speech Signals into Periodic and Aperiodic Components*, IEEE Trans. Speech and Audio Processing (1998). 42

[YMR91] B. Yegnanarayana, H. Murthy, and V. Ramachandran, *Processing of Noisy Speech using Modified Group Delay Functions*, Proc. Int. Conf. on Audio, Speech and Signal Proc., 1991, pp. 945–948. 33, 40

[YV98] B. Yegnanarayana and R. Veldhuis, *Extraction of Vocal-Tract System Characteristics from Speech Signal*, IEEE Trans. Speech and Audio Processing **6** (1998), no. 4, 313–327. 28, 33, 34

Index