Mémoire de DEA

présenté pour l'obtention du
DEA Systèmes de Télécommunications Numériques

Slim ESSID

Codeur audio paramétrique à bas débit basé sur un modèle "Sinusoïdes Amorties Exponentiellement + Transitoires + Bruit"

Encadré par Nicolas MOREAU

En collaboration avec Rémy BOYER

Octobre 2002
Département Traitement du Signal et des Images - ENST

Remerciements

Mes vifs remerciements vont tout d'abord à Nicolas MOREAU, Professeur à l'Ecole Nationale Supérieure des Télécommunications (ENST) pour avoir dirigé ce travail. Ses multiples compétences, son expérience en codage audio et son excellente capacité pédagogique m'ont permis de mener à bien ce projet.
Je remercie tout particulièrement Rémy BOYER, doctorant à l'ENST. Les innombrables discussions que nous avons eues, sa patience, son écoute, ses suggestions et son aide inestimable ont été des éléments essentielles dans l'aboutissement de ce travail qui n'aurait sûrement pas été ce qu'il est sans cette belle collaboration.
Je remercie également, la "population" du département de Traitement du Signal et des Images (TSI) pour l'accueil chaleureux qu'ils m'ont réservé et pour m'avoir permis de travailler dans d'aussi bonnes conditions.

Table des matières

Table des figures

Liste des tableaux

Liste des algorithmes

Notations

z^*	:	conjugué d'un nombre complexe z
$(.)^T$	:	opérateur transposé
$(.)^H$	:	opérateur conjugué transposé
$(.)^\dagger$	:	opérateur pseudo-inverse
rg(.)	:	rang d'une matrice
Im(.)	:	ensemble Image d'une matrice
Ker(.)	:	ensemble Noyau d'une matrice
vect{.}	:	espace vectoriel engendré
diag{.}	:	matrice diagonale
card{.}	:	cardinal d'un ensemble
$\|.\|$	:	norme l^2
$\Re(.)$	:	partie réelle
$\Im(.)$	:	partie imaginaire
SVD	:	décomposition en valeurs singulières
$O(.)$	:	de l'ordre de ...
$\mathbf{x}$	:	vecteur colonne
$\mathbf{X}$	:	matrice
$< .,. >$	:	produit scalaire
$\delta(n-p)$	:	masse de dirac au temps discret p
$\mathbf{A}(:,i)$	:	ième colonne de $\mathbf{A}$
$\mathbf{A}(i,:)$	:	ième ligne de $\mathbf{A}$
$\mathbf{A}(1:M,:)$	:	matrice obtenue en gardant les M premières lignes de $\mathbf{A}$

Introduction

Le domaine du codage audio (parole et musique ou son) a été très actif ces quinze dernières années. De nombreuses recommandations, dans le cadre de l'*UIT-T*[1] ou de l'*ETSI*[2], et normalisations, dans le cadre de l'*ISO/MPEG* [3], ont été effectuées dans le but de limiter la prolifération de systèmes incompatibles et de parvenir à un déploiement au niveau mondial des systèmes de compression. Une série de codeurs aux performances variables ont ainsi été mis en œuvre dans des contextes assez différents et pour des applications différentes. En effet, les études en codage de parole et de musique ont été effectuées séparément et conduites par des instances distinctes (l'UIT-T ou l'ETSI d'une part et l'ISO d'autre part) visant des objectifs différents (applications en téléphonie d'une part, diffusion d'autre part) et s'appuient sur des concepts tout aussi différents : les codeurs de parole font appel à la prédiction linéaire et se basent sur un modèle de production de la parole; les codeurs de musique utilisent une transformation temps-fréquence et exploitent les limitations du système auditif humain de manière à ce que le bruit de quantification des échantillons dans le domaine transformé soit inaudible. Cependant, la tendance actuelle est plutôt à la convergence des deux techniques, ce qui nous amène à énoncer la première problématique : est-il possible de réaliser un codeur traitant *indifféremment la parole et la musique* et donnant des résultats convenables aussi bien pour l'une que pour l'autre ?

La seconde problématique qui est posée concerne le champ d'applications potentielles demandant recours à des techniques de compression, dans un contexte industriel et économique qui, de prime abord, semble les faire tendre à devenir obsolètes. En effet, les progrès technologiques accomplis ces dernières années dans le domaine des télécommunications ont permis la généralisation de la numérisation des réseaux de transmission et de diffusion numériques, confortée par l'augmentation significative de la capacité des canaux de transmission et des moyens de stockage, au même titre que l'accroissement de la capacité de calcul des processeurs et micro-contrôleurs de traitement numérique du signal (DSP, *Digital Signal Processors*) et des PC (*Personal Computers*) : on a assisté d'une part à l'utilisation à grande échelle des fibres optiques, d'autre part à l'envahissement du secteur du stockage numérique par les CD (*Compact Discs*) et DVD (*Digital Versatile Discs*). Les fibres optiques ont une bande pas-

1. **U**nion **I**nternationale des **T**élécommunications - secteur de normalisation des **T**élécommunications.
2. **E**uropean **T**elecommunications **S**tandards **I**nstitue.
3. **I**nternational **S**tandardisation **O**rganisation,au sein de laquelle, le groupe WG11 de la sous-commission 29, connu sous l'acronyme MPEG (**M**oving **P**icture **E**xpert **G**roup), a pour mission le développement des normes multimédia basées sur l'audio et la vidéo numériques.

sante suffisamment importante[4] pour permettre la transmission sur le même support d'un nombre de signaux supérieur au nombre d'utilisateurs potentiels du canal. D'autre part, les CD et DVD peuvent contenir des enregistrements de plus d'une heure sans nécessiter aucune compression des données (sons et images). On se demande alors *quel rôle peut encore jouer la compression ?* Voici un début de réponse : certaines techniques de compression permettent de réduire les coûts des enregistrements pour une qualité identique à l'original. Par ailleurs, il existe des schémas de communications dans lesquels les canaux de transmissions ont une capacité limitée physiquement (tels que les liaisons par satellites ou les liaisons hertziennes) ou limitée par construction (tel que le réseau téléphonique où un utilisateur est limité à la bande de fréquence $300 - 3400Hz$). C'est ce dernier cas de figure qui nous intéresse le plus : le réseau téléphonique étant ce qu'il est, l'accès à Internet (et aux réseaux de données, de façon générale) se faisant par des millions d'utilisateurs via des liaisons par modem, il s'agit de répondre à une attente motivée de ces utilisateurs pour une application importante : le *streaming*. L'enjeux est de taille, puisque l'on doit pouvoir permettre la transmission de sons et d'images animées avec une qualité acceptable et en temps réel, sans pour autant monopoliser la bande passante allouée, sur des réseaux de communications classiques (RTC) qui sont par essence limités en bande passante ; et ce jusqu'à ce que l'UMTS et l'ADSL se soient généralisés à tous les utilisateurs, ce qui risque de nécessiter une période de temps assez longue. Cela implique que l'on doit disposer de codeurs performants fonctionnant à très bas débit, débit qui a été estimé par le groupe *MPEG-4* à 24 kbit/s pour la partie audio.

Le standard *MPEG-4* reprend justement ces concepts, à savoir les codeurs hybrides (parole et musique) et le streaming de qualité pour réseaux à bande limitée, puisqu'il se veut global, en ce sens qu'il concerne les 3 domaines suivants [1] :

- la télévision numérique;
- les applications graphiques interactives ;
- le multimédia intéractif (le World Wide Web, la distribution et l'accès au contenu).

Un codeur audio (parole et musique) à 4-16 kbit/s a même été normalisé : il s'agit en l'occurrence du codeur paramétrique bas débit *MPEG-4 HILN* (Harmonic and Individual Lines plus Noise) [2] qui, au vu de ses performances assez moyennes, demande à être amélioré .

Comment mesurer les performances d'un codeur ? L'analyse est faite suivant cinq critères principaux :

- le débit, qui reflète le degré de compression fourni par l'algorithme de codage;
- la qualité du signal restitué, déterminée par des tests subjectifs basés sur une moyenne des jugements exprimés par un certain nombre d'auditeurs avertis;
- la complexité algorithmique;
- le retard de transmission introduit par l'algorithme;
- la robustesse aux erreurs de transmission.

4. la bande passante est pratiquement infinie pour une fibre mono-mode ($> 10GHz/km$)...

L'analyse effectuée est avant tout liée à la bande passante du signal original ; on distingue 4 gammes essentielles de qualité qui sont rappelées dans le tableau 1.

Pour chaque gamme de qualité le tableau 2 donne la fréquence d'échantillonnage, le nombre de bits usuel utilisé pour la conversion analogique numérique, le débit nominal, les débits usuels et les taux de compression correspondants. On se place dans le cadre de ce travail dans la bande FM, c'est à dire que l'on utilise une fréquence d'échantillonnage des signaux audio de 32kHz. On verra que, compte tenu de la bande du signal concernée par la modélisation qui est réalisée, ce choix est complètement justifié.

Désignation	**Bande de fréquences**	**Caractéristiques**
Bande téléphonique	$300 - 3400Hz$	bonne intelligibilité mais naturel perturbé en parole, manque de bande passante pour la musique
Bande élargie	$50 - 7000Hz$	naturel respecté pour la parole, manque de bande passante pour la musique
Bande HiFi	$20 - 20000Hz$	mono-voie, stéréo ou encore 5.1 canaux spatialisés, excellente qualité aussi bien en parole qu'en musique

TAB. 1 – *Gammes de fréquences utilisées par les systèmes de compression [3]*

	Fe (kHz)	**B (bits)**	**Débit nominal (kbit/s)**	**Débit usuel (kbit/s)**	**Taux de compression**
Bande tél.	8	13	104	$64 - \ldots - 4 - \ldots$	$1.6 - \ldots - 26 - \ldots$
Bande élargie	16	14	224	$64 - \ldots - 16 - \ldots$	$3.5 - \ldots - 14 - \ldots$
Bande FM	32	16	512 mono-voie (1024 en stéréo)	$192 - \ldots - 64 - \ldots$	$2.6 - \ldots - 8 - \ldots$
Bande HiFi	44.1	16	705.6 (1411 en stéréo)	$192 - \ldots - 56 - \ldots$	$3.6 - \ldots - 12 - \ldots$

TAB. 2 – *Gammes de qualité et débits [3]*

De manière assez globale, on peut admettre que l'on dispose actuellement, aussi bien pour la parole en bande téléphonique que pour la musique en bande HiFi, de codeurs de qualité satisfaisante pour des taux de compression de l'ordre de 10. Les exemples les plus représentatifs sont sûrement, pour la parole, le codeur G.729 à 8 kbit/s [3], pour la musique, le codeur MPEG2-AAC [4] qui assure la "transparence" [5] à 64 kbit/s (dans le cas monophonique).

Dans l'optique de produire un codeur hybride aux performances comparables à celles des 2 codeurs cités plus haut, fonctionnant à des débits permettant le streaming [6], un appel à candidatures portant le titre "*Call for proposals for new tools for audio coding*" a été lancé en Janvier 2001 par le sous-groupe audio (ISO/IEC JTC1/SC29/WG11) afin d'aboutir à un codeur fonctionnant à un débit avoisinant les 24 kbit/s. Un codeur y répondant a même été proposé [5] et suit le processus de normalisation. Ce dernier présente encore des faiblesses au niveau de la restitution d'attaques de sons.

C'est dans ce contexte que ce travail a été réalisé dans le but de concevoir un codeur audio dans la lignée du HILN, qui réponde au cahier des charges de *MPEG-4* [6]. Le projet a été entrepris au sein du département de Traitement du Signal et des Images (TSI) de l'Ecole Nationale Supérieure des Télécommunications (ENST) dans le cadre de la thèse de Rémy Boyer associée au projet RNRT [7] "COHRAINTE" (COdage Hiérarchique et Robuste de sources Audiovisuelles et application INTErnet) impliquant, pour la partie audio l'ENST, par l'intermédiaire de Nicolas Moreau, et Dominique Massaloux pour France Télécom R&D.

La contribution de l'équipe du département TSI s'inscrit dans la philosophie des modèles sinusoïdaux initiés au MIT au début des années 80 pour coder de la parole en bande téléphonique [7], modèles qui ont été repris au département TSI [8] dans le contexte de modification/synthèse de la parole. Cette approche ne s'est pas limitée à la parole puisqu'elle a été largement exploitée dans les travaux de X. Serra [9] à Stanford dans un schéma d'analyse/synthèse de signaux musicaux.

L'essentiel du travail a porté sur une modification de ces modèles afin de mieux les adapter à la modélisation des signaux transitoires. Cette étude a été menée en étroite collaboration avec Rémy Boyer. Le présent rapport synthétise les différentes méthodes qui ont été mises en œuvre tout au long des derniers 18 mois passés ensemble. La conception d'une architecture de codage constitue l'axe principal de ma contribution personnelle. On propose un système de codage entièrement paramétrique à base de "Sinusoides Amorties Exponentiellement+Transitoires+Bruit". Le chapitre 1 sera consacré à un état de l'art des représentations paramétriques du signal audio. On présentera au chapitre 2 une vue générale du système de codage qui a été conçu en reportant la présentation des outils et justifications théoriques

5. on dit qu'on atteint la transparence lorsque l'auditeur n'est plus capable de distinguer la version codée de l'originale du signal audio

6. n'accaparant pas la bande passante allouée à l'utilisateur

7. **R**éseau **N**ational de **R**echerche en **T**élécommunications. http://www.telecom.gouv.fr/rnrt/suivi/cohrainte.htm

aux chapitres suivants. Le chapitre 3 traite donc du modèle EDS (Exponentially Damped Sinusoids), le chapitre 4 est dédié à la modélisation des transitoires, enfin la quantification et le codage des paramètres de modèle sont abordés au chapitre 5.

Chapitre 1

Représentations paramétriques du signal audio

Les systèmes de codage par transformée ont sans doute été les premiers à permettre de représenter efficacement le signal audio en bande HiFi à bas débit tout en assurant la transparence. Le point clé de ces systèmes est leur capacité à mettre en forme le bruit de quantification dans le plan temps/fréquence de manière à assurer que ce dernier ne sera pas perceptible par le système auditif humain en s'appuyant sur des principes de psychoacoustique [10, 11, 12]. Cependant, une dégradation de la qualité est accusée aux débits inférieurs à 64 kbit/s avec ce type de codecs perceptuels. En effet, ces derniers se mettent alors soit à réduire la largeur de bande audio, soit à introduire des artefacts fortement gênants résultant d'un défaut de bits suffisants pour coder toute la bande. Les défauts deviennent importants en dessous de 32 kbit/s, même en faisant appel aux techniques d'enrichissement spectral (SBR, Spectral Band Replication [13]). Par ailleurs, ce type de représentation du signal audio rend les modifications[1] assez complexes [14]. Il apparaît donc un besoin de mettre en œuvre des systèmes permettant d'atteindre des débits plus bas tout en garantissant une qualité de restitution satisfaisante et une malléabilité de traitement dans le domaine compressé. La solution qui va rapidement s'imposer est le recours aux modèles paramétriques, notamment, sinusoïdaux. Nous présentons dans ce chapitre un panorama de ces modèles.

1.1 Modèle Sinusoïdal

L'idée de représenter un signal audio à l'aide de modèles sinusoïdaux a été proposée au MIT dès le début des années 80 par Mc Aulay et Quatieri [7] pour coder de la parole à faible débit dans un schéma d'analyse/synthèse. Au cours de l'étape d'analyse, le signal est découpé en trames sur lesquels les amplitudes, les fréquences et les phases du modèle sont estimées. L'étape de synthèse consiste alors à interpoler les paramètres estimés afin d'assurer une évolution continue des formes d'ondes sinusoïdales aux abords des différentes trames. En notant $s(n,l)$ le segment du signal audio $s(n)$ analysé à la trame l de longueur N échantillons,

1. par exemple, le changement de vitesse de lecture laissant le pitch intact, ou vice-versa

avec $n \in 0,...,N-1$, le signal de modèle correspondant s'écrit

$$s_M(n,l) = \sum_{m=1}^{M} a_m(n,l) \cos\left[\phi_m(n,l)\right] = \sum_{m=1}^{M} a_m(n,l) \cos\left[2\pi f_m(n,l)n + \phi_m(0,l)\right] \tag{1.1}$$

où M est l'ordre de modélisation, $a_m(n,l)$ l'amplitude instantanée, $f_m(n,l)$ la fréquence instantanée et $\phi_m(n,l)$ la phase instantanée, obtenues par interpolation des paramètres $(\bar{a}_m(l))_l$, $(\bar{f}_m(l))_l$ et $(\bar{\phi}_m(l))_l$ tel qu'il sera décrit ultérieurement. Des modèles sinusoïdaux harmoniques ont également été mis en œuvre [16] exploitant les rapports harmoniques existant entre les composantes fréquentielles du signal afin de limiter le débit associé au codage des fréquences.

1.2 Modèle "Sinusoïdes+Bruit" (SB)

S'il est admis que les modèles sinusoïdaux sont performants dans la représentation de sons voisés, ils deviennent inefficaces en dehors de ce contexte, si l'on souhaite maintenir un niveau élevé de qualité de modélisation, qu'il s'agisse de sons non voisés pour le signal de parole ou de sons musicaux. Afin de pallier à ces limitations, Serra et Smith [15] ont développé un système pour l'analyse/synthèse du son basé sur une décomposition en composante déterministe plus une composante stochastique. Un signal déterministe est défini comme étant un signal dont l'évolution dans le temps est parfaitement prévisible par des mesures sur tout intervalle continu. Dans le cadre qui nous intéresse, la classe des signaux déterministes est réduite à celle de séries de M formes d'ondes données, typiquement des sinusoïdes, qui permettent de restituer les composantes tonales du signal, c'est à dire les pics présents dans le spectre. Un signal stochastique, ou bruit, est complètement décrit par sa densité spectrale de puissance. Lorsqu'un signal est considéré comme stochastique, il n'est nécessaire de préserver ni sa phase instantanée ni les détails de son spectre. Sous cette hypothèse, la composante stochastique du signal peut être vue comme le résultat du filtrage d'un bruit blanc par un filtre de réponse impulsionnelle variable dans le temps. On modélise donc le signal en deux étapes successives : la première consiste à déterminer des "composantes tonales", c'est à dire à déterminer les paramètres de la série de formes d'ondes représentant sa composante déterministe; la seconde est ensuite réalisée sur l'erreur $r(n)$ introduite par la première modélisation.

Avec les notations :

- M : ordre de la modélisation;
- $s_M(n,l)$: le signal obtenu à la suite de la première modélisation de $s(n,l)$ par M formes d'ondes Φ_M :

$$s_M(n,l) = \sum_{m=1}^{M} \Phi_m(n,l) \tag{1.2}$$

- $r(n,l)$: la partie stochastique du signal ou résiduel de première modélisation

$$r(n,l) = s(n,l) - s_M(n,l) \tag{1.3}$$

- $\hat{r}(n,l)$: modélisation de $r(n,l)$;

il vient :

$$\hat{s}(n,l) = s_M(n,l) + \hat{r}(n,l) \tag{1.4}$$

où $\hat{s}(n,l)$ est le signal synthétique correspondant à $s(n,l)$.

Cette décomposition en 2 étapes, initialement développée pour l'analyse/transformation/synthèse du signal de musique, a été très vite adoptée par la communauté audio [2, 8, 5] et notamment par Rodet et Depalle [17], s'avérant essentielle pour les modifications (pitch ou échelle temporelle) de haute qualité en bande HiFi. Le principe a été repris dans le cadre de la thèse de Ioannis Stylianou [18] au département TSI de l'ENST pour la modification de la parole et du locuteur. Plus récemment, cette approche a été adoptée dans le cadre du codeur bas débit MPEG-4 HILN [2] pour coder des signaux audio à très bas débit (de 4 à 16 kbits/s). Ce codeur utilise des modèles de source pour des composantes harmoniques, des sinusoïdes "individuelles" et pour un signal de bruit, d'où l'appellation "Harmonic and Individual Lines plus Noise".

1.3 Modèle "Sinusoïdes+Transitoires+Bruit" (STB)

Le modèle sinusoïdal est qualifié de stationnaire puisque les paramètres (amplitudes, phases et pulsations) sont supposés constants ou à variation lente au regard de la durée d'analyse. Ce modèle s'avère inefficace lors de la représentation de signaux transitoires, typiquement les attaques ou évanouissements de sons, qui constituent une composante essentielle des signaux audio, notamment des signaux de musique et plus particulièrement dans un contexte percussif (batterie, tambour, castagnettes, ...) où une mauvaise restitution des attaques est très préjudiciable à la qualité d'écoute. Le modèle "Sinusoïdes + Transitoires + Bruit" (STB) [19, 20] a donc été proposé afin d'améliorer la représentation de signaux à variations temporelles rapides. Différents systèmes de représentations des transitoires plus ou moins performants ont été élaborés. La mise en œuvre du modèle STB est essentiellement effectuée selon deux schémas distincts : un schéma en série ou un schéma en parallèle. Hamdy [19] effectue en premier lieu une modélisation sinusoïdale du signal, puis le signal résiduel est calculé et codé à l'aide d'un codeur par ondelettes. Dans [21] les transitoires sont détectés et modélisés spécifiquement à l'aide d'une approche paramétrique; le signal modélisé est ainsi soustrait de l'original, et le résiduel est analysé à l'aide d'un modèle "Sinusoïdes + Bruit" (SB). Levine [14] propose un schéma de représentation audio dans lequel les transitoires sont représentés à l'aide d'un codeur par transformée simplifié et les parties stables à l'aide du modèle SB. On "switch" ainsi d'un modèle paramétrique à une représentation non paramétrique sur l'ordre d'un détecteur de transitoires. Le système atteint des performances comparables à celles du codeur MPEG-AAC à 30 kbit/s tout en permettant des modifications dans le domaine compressé. Le débit global reste cependant élevé pour le cadre qui nous intéresse à cause du coût élevé de codage par transformée de la partie transitoire.

1.4 Modèles Sinusoïdaux Généralisés

Au cours des dernières années, un intérêt croissant a été porté aux différents modèles de représentation du signal. Plusieurs alternatives à la représentation classique de Fourier (modèle sinusoïdal) existent et différentes familles de formes d'ondes paramétriques, communément désignées par *dictionnaires*, ont été mises en œuvre. Le dictionnaire d'ondelettes en est sans doute l'exemple le plus connu. Un dictionnaire est vu comme une famille $\mathbf{G} = (\mathbf{g}_\gamma)_{\gamma\in\Gamma}$ de formes d'ondes, où γ est un index et Γ est l'ensemble de tous les index permis tel que $\Gamma \subseteq \mathbb{R}$. Il s'agit alors d'obtenir une décomposition du signal $s(n)$ de la forme

$$\mathbf{s} = \sum_{\gamma\in\Gamma} \alpha_\gamma \mathbf{g}_\gamma \tag{1.5}$$

ou de façon approximative

$$\mathbf{s} = \mathbf{s}_M + \mathbf{r}_M = \sum_{k=1}^{M} \alpha_{\gamma_k} \mathbf{g}_{\gamma_k} + \mathbf{r}_M \tag{1.6}$$

où M est l'ordre du modèle ainsi obtenu.

Les méthodes d'analyse/synthèse traditionnelles se basent sur l'utilisation de bases orthogonales telles que la base de Fourier, différentes bases de transformations en cosinus discrets ou les bases orthogonales d'ondelettes. Dans ces conditions le signal $\mathbf{s}_M$ est représenté par une combinaison linéaire de M formes d'ondes vues comme des vecteurs de $\mathbb{C}^N$ qui sont alors linéairement indépendants, d'où l'unicité de la représentation, *i.e*, des α_{γ_k} :

$$\alpha_{\gamma_k} = < \mathbf{g}_{\gamma_k}, \mathbf{s}_M > . \tag{1.7}$$

Cependant, la plupart des dictionnaires "modernes" sont *sur-complets (overcomplete)*, en ce sens qu'ils sont redondants, ils sont donc dits non-orthogonaux. De tels dictionnaires permettent de réaliser une décomposition de $\mathbf{s}$ avec une résolution supérieure. En revanche, l'unicité de la décomposition n'est plus assurée puisque la famille de vecteurs $(\mathbf{g}_{\gamma_k})_k$ n'est plus libre. Cela permet l'adaptation de la représentation, c'est à dire, la possibilité de choisir parmi toutes les représentations possibles celle qui répond à certains critères parmi les suivants :

- **compacité**: on définit la notion de "représentation compacte" associée au modèle $s_M(n)$ et à la norme 2 par la formulation suivante :

  soit $\varepsilon \in \mathbb{R}^+$ et $M \ll N$, le modèle est compacte si $\sum_{n\in T} |s(n) - s_M(n)|^2 \leq \epsilon \sum_{n\in T} |s(n)|^2$.

  On doit alors obtenir la représentation du signal la plus compacte, *i.e*, avec le moins de coefficients significatifs α_{γ_k}.
- **super-résolution**: on doit obtenir une résolution des paramètres des termes du développement de $s(n)$ largement supérieure à celle que l'on pourrait obtenir avec une approche traditionnelle.
- **complexité**: l'algorithme de décomposition devra avoir une complexité acceptable.

1.4.1 Dictionnaires et Modèles

Nous utilisons une terminologie introduite par Mallat et Zhang [22]. Un *dictionnaire* est une collection de formes d'ondes paramétriques $\mathbf{G} = (\mathbf{g}_\gamma)_{\gamma\in\Gamma}$. Les formes d'ondes $\mathbf{g}_\gamma$ sont des signaux à temps discret de durée N, vues comme des vecteurs de $\mathbb{C}^N$ et sont appelés *atomes*. Selon les dictionnaires, le paramètre γ peut indexer :

- la fréquence, auquel cas le dictionnaire est un dictionnaire de fréquence ou un dictionnaire de Fourier;
- les temps/échelle conjointement, auquel cas le dictionnaire est un dictionnaire temps/échelle;
- les temps/fréquence conjointement, auquel cas le dictionnaire est un dictionnaire temps/fréquence.

Ces dictionnaires peuvent être, respectivement, complets ou redondants, par suite ils contiennent, respectivement, exactement N atomes ou plus de N atomes. Il est également possible de considérer des dictionnaires *continus* contenant une infinité d'atomes. Un large éventail de dictionnaires a été proposé, nous nous intéresserons particulièrement à deux types de dictionnaires temps/fréquence, à savoir, le *dictionnaire de Gabor* et le *dictionnaire de Sinusoïdes Amorties Exponentiellement (EDS, Exponentially Damped Sinusoids).*

1.4.2 Les dictionnaires de fréquence

Un dictionnaire de fréquence ou de Fourier peut être défini comme une famille d'exponentielles complexes

$$g_\omega(n) = e^{i\omega n} \tag{1.8}$$

indexées par $\gamma = \omega$, avec $\omega \in [0,\pi[$.

1.4.3 Les dictionnaires temps/fréquence

Les atomes temps/fréquence permettent une meilleure représentation des signaux localisés à la fois en temps et en fréquence. L'occupation temporelle et spectrale d'un atome temps/fréquence (il s'agit en l'occurrence d'une EDS) est représentée sur la figure 1.1.

Une famille d'atomes temps/fréquence générale peut être générée par dilatations, translations et modulations d'une même fenêtre $g(n)$ supposée réelle. Pour tout paramètre d'échelle s, fréquence de modulation ω et translation u, on note $\gamma = (s,u,\omega) \in \mathbb{R}^3$ et on définit :

$$g_\gamma(n) = g(\frac{n-u}{s})e^{i\omega n}. \tag{1.9}$$

Si $g(n)$ est paire, $g_\gamma(n)$ est centrée autour de l'abscisse u. Son énergie est concentrée au voisinage du temps u, avec une dispersion temporelle Δu de l'ordre de s. Sa transformée de Fourier est localisée autour de la fréquence ω, avec une dispersion fréquentielle $\Delta\omega$ de l'ordre de $1/s$.

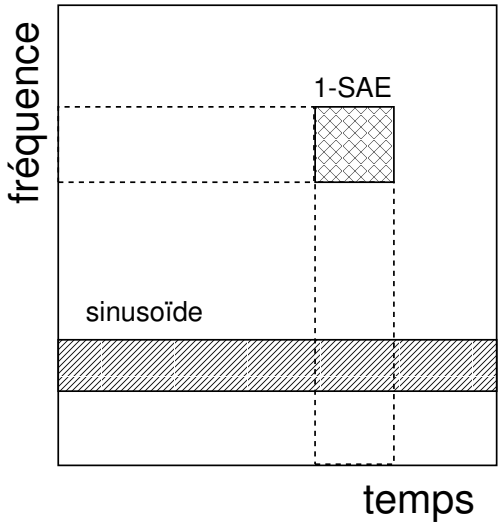

FIG. 1.1 – *Représentations temps/fréquence d'une EDS et d'une sinusoïde*

Atomes et Modèle de Gabor

En raison de ses propriétés optimales de localisation combinée en temps/fréquence, au sens du principe d'incertitude de Heisenberg, la fenêtre $g(n)$ est souvent une fenêtre gaussienne :

$$g(n) = g_G(n) = 2^{\frac{1}{4}} e^{-\pi n^2} \tag{1.10}$$

Les atomes temps/fréquence sont alors dits de *Gabor* (1946). Des exemples de ces atomes sont donnés à la figure 1.2 , avec $\gamma = (64{,}128{,}\frac{\pi}{8})$ en (a) et $\gamma = (32{,}128{,}\frac{\pi}{8})$ en (b). Notons qu'à cause de la symétrie paire de la fenêtre Gaussienne, les atomes de Gabor présentent un axe de symétrie à l'abscisse u.

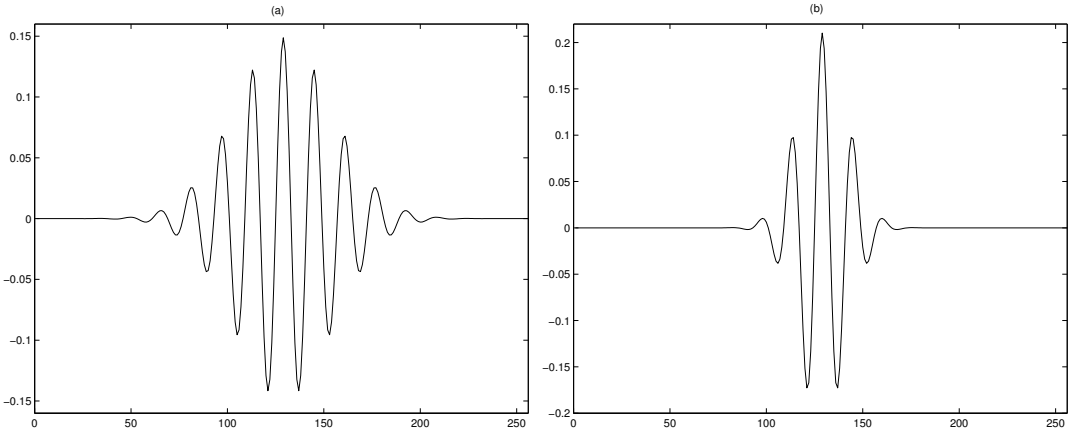

FIG. 1.2 – *Parties réelles d'atomes de Gabor complexes.*

Le modèle de Gabor s'obtient en écrivant

$$s_M(n) = \sum_{m=1}^{M} \alpha_{\gamma_m} \, g_G\left(\frac{n - u_m}{s_m}\right) e^{i\omega_m n}. \tag{1.11}$$

Atomes et Modèle "Sinusoïdes Amorties Exponentiellement" (EDS)

En posant $g(n) = e^n$, $d = \frac{1}{s}$ et $u = 0$, dans l'expression 1.9, on obtient les atomes EDS (Exponentially Damped Sinusoids) décrits par

$$g_\gamma(n) = e^{dn}\, e^{i\omega n}. \tag{1.12}$$

avec $\gamma = (d,\omega) \in \mathbb{R}^2$, d étant alors un coefficient d'amortissement. On obtient ainsi un modèle non-stationnaire assez simple qui de surcroît présente des propriétés structurelles intéressantes que l'on mettra à profit dans le choix de la stratégie de décomposition tel que décrit au chapitre 3. Des exemples de EDS sont donnés sur la figure 1.3 avec $d = -0.01$ et $d = 0.02$.

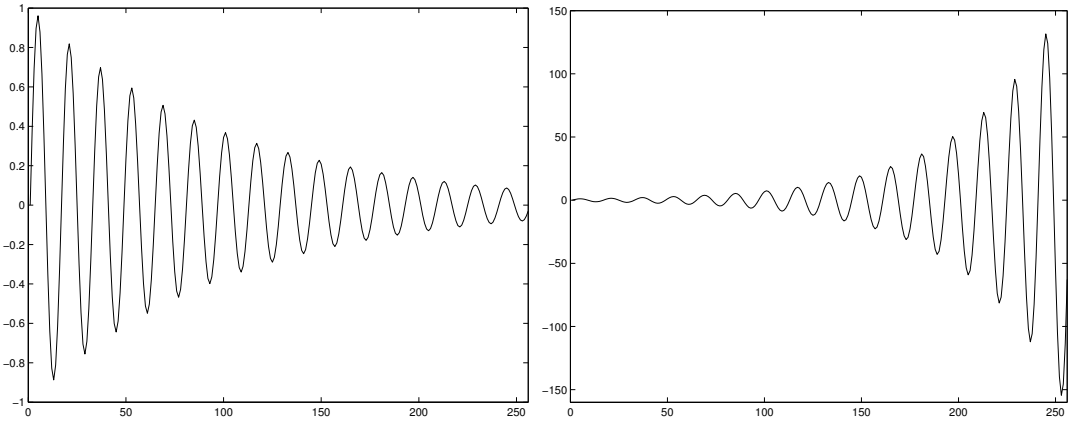

FIG. 1.3 – *Sinusoïdes Amorties Exponentiellement.*

L'expression du modèle est donnée par

$$s_M(n) = \sum_{m=1}^{M} \alpha_{\gamma_m}\, e^{d_m n}\, e^{i\omega_m n}. \tag{1.13}$$

Notons qu'il suffit de poser $d = 0$ pour retrouver le modèle de Fourier.

Autres dictionnaires

De très nombreuses variations existent dans la construction des dictionnaires [23, 24, 25]. Nous en donnons ici quelques exemples à titre indicatif. Goodwin [25] préconise l'utilisation de Sinusoïdes Amorties Retardées en introduisant un paramètre de translation u sur le mode des atomes de Gabor pour "casser" la symétrie de ces derniers (c.f. figure 1.4). Le but en est d'accélérer la convergence de l'algorithme de décomposition dans le cas des signaux à caractère non symétrique. D'autre part, afin de prendre en considération le comportement non stationnaire de la fréquence instantanée de certains signaux, les *atomes chirpés* ont été introduits [26]. Il s'agit d'une extension du dictionnaire de Gabor multi-échelle: ces atomes sont caractérisés, en plus des paramètres (s,u,ω), par un *paramètre de chirp* c grâce auquel leurs fréquences instantanées $\Omega(n) = \omega + c(n-u)$ varient linéairement avec le temps. Un

atome chirpé est ainsi décrit à l'aide d'un index (s,u,ω,c) :

$$g_{(s,u,\omega,c)}(n) = \frac{1}{\sqrt{s}}\, g(\frac{n-u}{s})\, e^{[i(\omega(n-u)+\frac{c}{2}(n-u)^2)]} \tag{1.14}$$

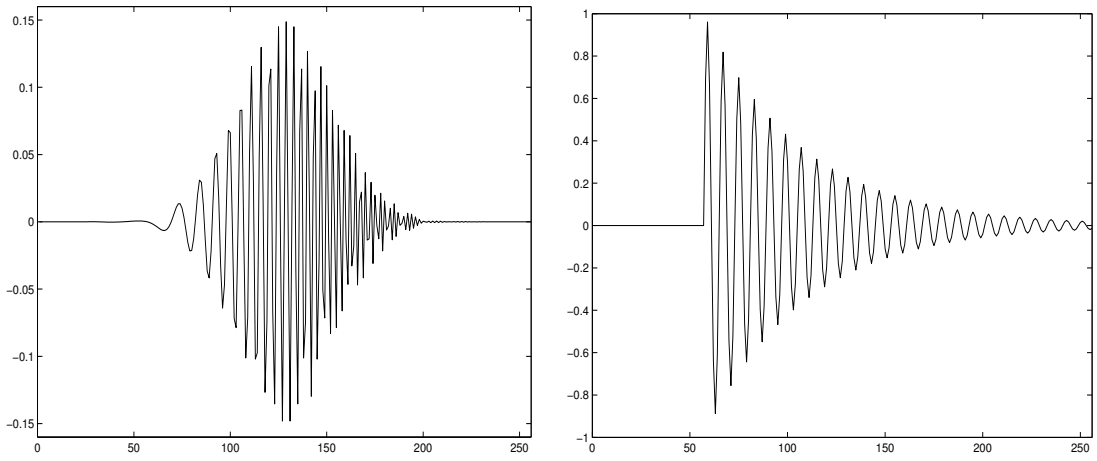

FIG. 1.4 – *Atome chirpé - Sinusoïde Amortie Retardée*

Signalons enfin qu'il est possible d'utiliser l'union de plusieurs dictionnaires $\mathbf{G}_i$ tels que : $\mathbf{G} = \bigcup_i \mathbf{G}_i$. De tels dictionnaires permettent de représenter efficacement des signaux de natures très diverses. On peut donner les exemples de dictionnaires composites suivants :

- en *paquets de cosinus* [24] où $\mathbf{G}_0$ est une base orthogonale de Fourier et $\mathbf{G}_1$ est une base formée d'atomes de Gabor [27];
- en *paquets d'ondelettes* [28] où $\mathbf{G}_0$ est une base orthogonale de Fourier et $\mathbf{G}_1$ est une base orthogonale d'ondelettes.

Au chapitre suivant, nous présentons une vue d'ensemble de l'architecture du système de codage qui a été élaboré. Le signal audio est segmenté et représenté par un modèle "EDS + transitoires + bruit". Les outils et discussions théoriques relatifs à la modélisation sont présentés aux chapitres suivants.

Chapitre 2

Vue d'ensemble de l'architecture du système de codage

Dans ce chapitre, on présente une vue générale du codeur en termes de blocs fonctionnels. On ne détaillera pas à ce stade les outils théoriques intervenant dans la décomposition du signal qui feront l'objet des chapitres suivants. Un schéma en blocs du codeur est présenté à la figure 2.1. Le système effectue une décomposition du signal audio en trois composantes : pseudo-stationnaire, transitoire et bruit. On désigne par *pseudo-stationnaire* la composante du signal qui est convenablement décrite par le modèle EDS, il s'agit en l'occurrence de la composante stationnaire étendue au cas des transitoires "doux", typiquement les plosives du signal de parole et les attaques lentes d'instruments.

Le signal audio à l'entrée du codeur est filtré par un filtre passe-haut pour éliminer les composantes aux fréquences inaudibles (en dessous de 5Hz) et la composante continue. Le filtre utilisé a pour fonction de transfert [5]

$$H(z) = \frac{-0.999643937167571 + 0.999643937167571 z^{-1}}{-0.999287874335142 z^{-1}}.$$

Le signal filtré est alors passé au bloc de détection des transitoires. On récupère en sortie une estimation des instants d'attaque de sons qui permet au bloc suivant de décider de la segmentation du signal. On commute d'un modèle de représentation EDS (*mode S*) à un modèle de représentation dédié aux transitoires (*mode T*) sous les ordres de ce même bloc. Les paramètres de modèle obtenus sont alors appariés le long de trajectoires afin de préparer la quantification et la synthèse du signal (c.f. chapitre 3). Une sélection de ces trajectoires est effectuée selon des considérations psychoacoustiques et l'on procède à la quantification des paramètres. On peut alors réaliser la synthèse de la composante déterministe du signal que l'on soustrait du signal original afin d'obtenir le résiduel de première modélisation assimilé à du bruit. Ce bruit est modélisé spécifiquement et les paramètres résultant sont quantifiés. Des codes de Huffman sont utilisés dans la représentation binaire du signal. La dernière étape consiste alors à formatter le bitstream.

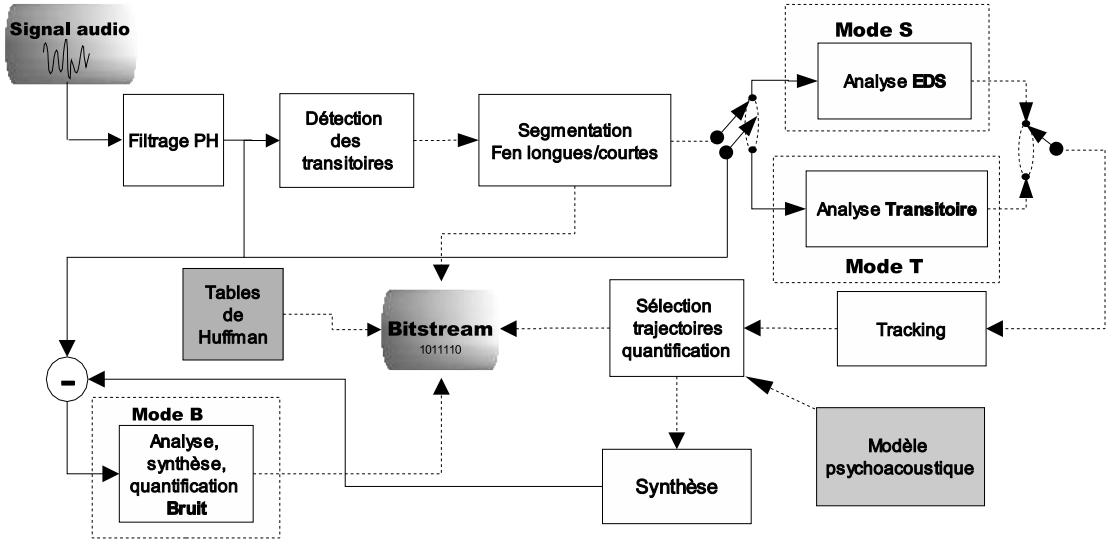

FIG. 2.1 – *Schéma en blocs de l'encodeur*

2.1 Détection, estimation des transitoires

La détection des transitoires est un élément clé de l'architecture de codage. Il s'agit en effet du bloc qui commande la segmentation du signal et dont dépend le choix du modèle de représentation. Le signal audio à l'entrée du codeur est en fait découpé en super-trames de taille égale à celle de 10 trames longues, *i.e*, 20480 échantillons, lesquels sont passées au bloc de détection des transitoires[1]. Outre le fait de détecter les transitoires, il importe d'estimer les instants d'attaques de façon précise. Différents systèmes ont été proposés dans des travaux précédents [20, 29, 14]. On distingue essentiellement deux types d'approches, celles se basant sur la variation de l'énergie du signal en amont et en aval de l'attaque et celles exploitant les variations de l'enveloppe du signal préalablement calculée. Des méthodes plus élaborés ont été conçues dans le contexte de la détection du rythme faisant appel à un pré-traitement du signal par un banc de filtres dans le but de relever le caractère transitoire présent dans certaines bandes et qui serait autrement noyé dans le signal "large bande". Signalons que toutes les méthodes recourent à l'utilisation d'un seuil au-dessus duquel le caractère transitoire est décidé, ce qui constitue souvent une limitation eu égard à la difficulté de fixer une valeur du seuil qui soit "universelle". Il est clair que les approches énergétiques deviennent peu efficaces dans le cas où l'énergie en amont de l'attaque varie peu par rapport à l'énergie en aval. Par ailleurs, il est généralement préférable d'effectuer une analyse par sous-bandes surtout dans le cas de pièces musicales polyphoniques pour produire un "effet de loupe" sur les attaques de sons qui apparaissent de façon plus prononcée dans certaines sous-bandes [29]. Un exemple en est donné à la figure 2.2. Il s'agit de 600 ms de signal de clavecin (2.2-(a)); remarquons que les attaques apparaissent plus clairement dans la sous-bande critique n°24 (2.2-(b)) où

1. Cette découpe a pour seul but de réduire la complexité du traitement

elles sont facilement détectées. La bande de fréquence concernée peut paraître très haute, mais un transitoire audio présente justement généralement de l'énergie sur la totalité de la bande audio.

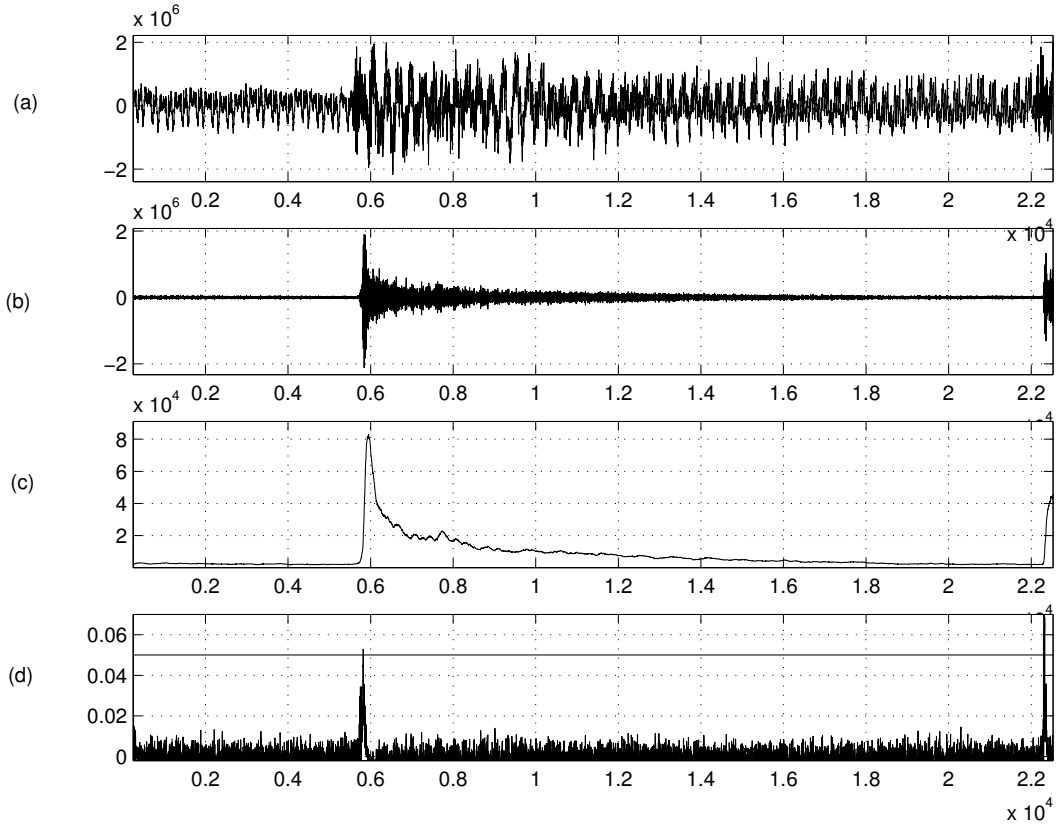

FIG. 2.2 – *Attaques de Clavecin (a) original, (b) sous-bande critique n° 24, (c) enveloppe, (d) FDR de l'enveloppe*

Nous adoptons donc une approche qui s'inspire de celle de Klapuri [29] et dont un schéma en blocs est présenté à la figure 2.3.
La première étape consiste à traiter le signal en entrée par le banc de filtres en bandes critiques [30] $\{h_k(n)\}$, $k = 1,...,24$ selon

$$s_k(n) = s(n) * h_k(n)$$

où les $(s_k(n))_k$ sont les signaux de sous-bandes. On effectue ainsi une découpe non uniforme de l'axe fréquentiel comme illustré sur la figure 2.4. Il est alors possible de procéder à une décimation des signaux de sous-bandes afin de réduire la complexité du traitement à suivre. Nous choisissons un facteur de décimation de 8 pour garder une bonne précision d'estimation des instants d'attaque.
On calcule alors les enveloppes $\nu_k(n)$ des signaux de sous-bandes à partir des signaux analytiques qui leur sont associés selon

$$\nu_k(n) = |y_k(n)| * f(n)$$

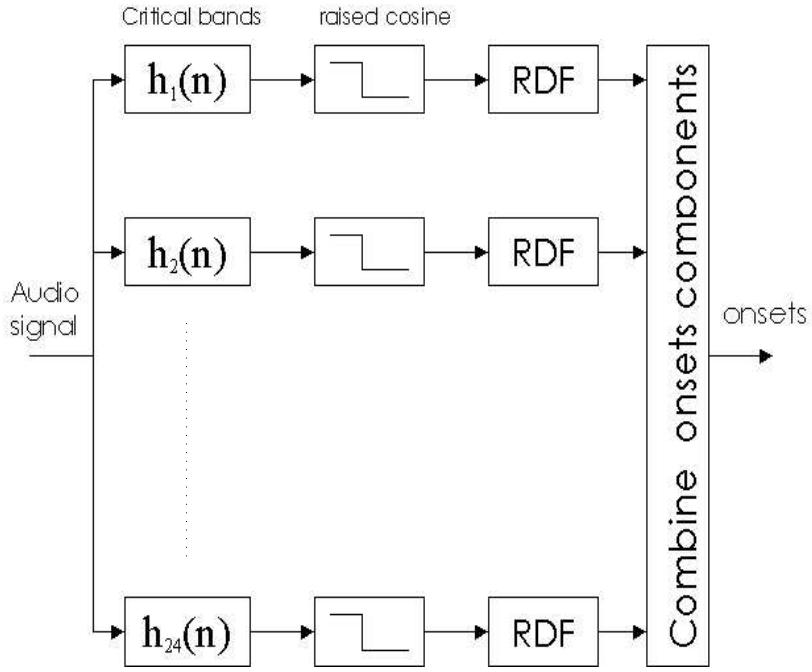

FIG. 2.3 – *Diagramme en blocs du système de détection des transitoires*

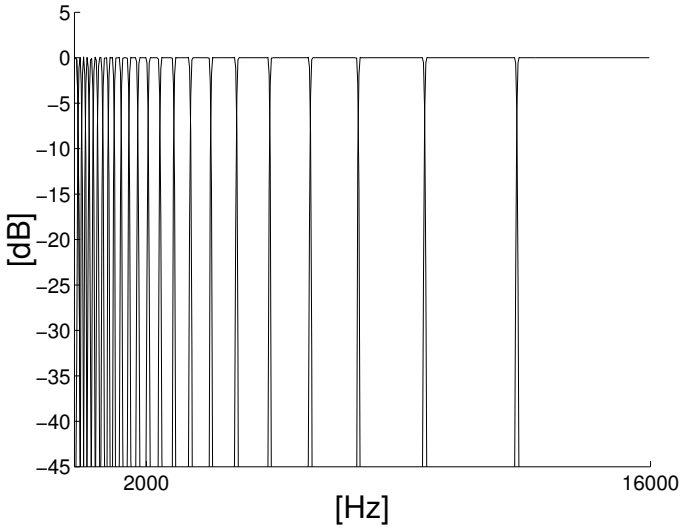

FIG. 2.4 – *Réponse fréquentielle du banc de filtres en bandes critiques*

où $f(n)$ est une demi-fenêtre de Hanning de taille 100 ms reproduisant l'intégration faite par l'oreille humaine [29] et $y_k(n)$ est le signal analytique de $s_k(n)$ tel que $y_k(n) = s_k(n) + i\Psi_k(n)$, avec $\Psi_k(n)$ la transformée de Hilbert du signal de sous-bande $s_k(n)$ (voir figure 2.2-(c)). Les attaques sont alors à rechercher parmi les maxima des dérivées des enveloppes. Afin d'adoucir les courbes obtenues, on considère en fait les dérivées des fonctions logarithmes prises sur les enveloppes, c'est à dire les Fonctions Différences Relatives (FDR) des $\nu_k(n)$ comme dans la figure 2.2-(d). L'étape suivante consiste à prendre une décision sur la validité des maxima

trouvés en tant qu'attaques de sons. Dans un premier temps, on ne garde que les candidats au-dessus d'un certain seuil. Il faut ensuite rassembler les données des différentes sous-bandes et l'on retient le maximum le plus puissant dans un voisinage de 50 ms. Les instants d'attaque sont alors donnés par les instants des maxima retenus. Le résultat de détection des transitoires sur un signal de glockenspiel est donné à la figure 2.5.

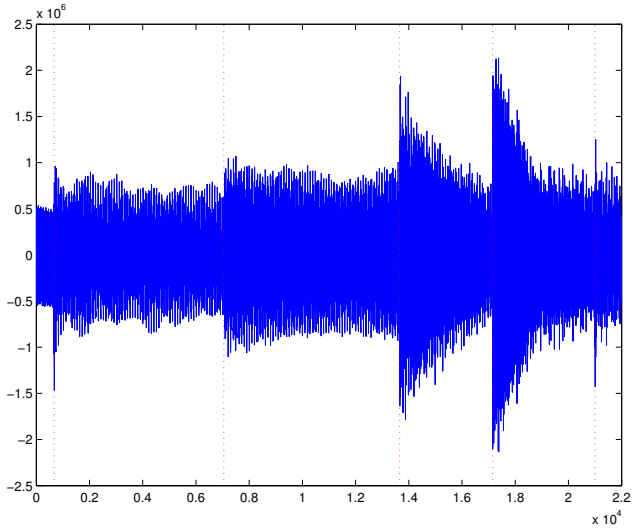

FIG. 2.5 – *Détection des attaques sur extrait de glockenspiel, les attaques sont indiquées par les lignes verticales en pointillés*

2.2 Segmentation du signal audio

Le signal est découpé en segments classés en pseudo-stationnaires ou fortement transitoires. Grâce à l'utilisation d'un modèle EDS, on peut utiliser des fenêtres d'analyse de taille importante sur les segments pseudo-stationnaires, taille supérieure à la durée de stationnarité généralement admise en codage audio (20 ms). La durée d'analyse est donc $N_l = 2048$ échantillons, soit 64 ms pour une fréquence d'échantillonnage f_e de 32 kHz [2], ce qui permet de maintenir une bonne résolution fréquentielle. Nous verrons dans le chapitre 4 que les performances de modélisation par EDS s'effondrent dans le contexte fortement transitoire. En conséquence, nous adoptons des fenêtres d'analyse courtes de taille $N_c = 256$ échantillons (8 ms) sur les segments fortement transitoires du signal sur le mode du codeur MPEG-AAC [12], privilégiant ainsi la résolution temporelle. On trouvera la justification de ce choix au chapitre 4. Les fenêtres d'analyse sont translatées tout le long des échantillons du signal avec un pas de $\frac{N_l}{2}$ pour les fenêtres longues et $\frac{N_c}{2}$ pour les fenêtres courtes. Les fenêtres d'analyse successives sont alors recouvrantes : chaque bloc d'échantillons long, respectivement court, d'indice l est superposé à 50% avec le bloc long, respectivement court, précédent d'indice $l-1$ et à 50% avec le bloc long, respectivement court, suivant d'indice $l+1$. Ce recouvrement

2. c'est cette fréquence qui a été privilégiée dans ce travail, nous verrons que le passage à une fréquence de 44.1 kHz est presque immédiat

est nécessaire afin de limiter les effets de bord. Sans cela, une variation brusque de l'erreur de modélisation entre deux blocs adjacents serait perceptible et donc fortement gênante à l'écoute. Le recouvrement permet de répartir le bruit de modélisation entre blocs adjacents de manière à réduire sa discontinuité au niveau des bords. Une alternative à cette technique de recouvrement a été proposée afin d'assurer la continuité du signal dans le contexte de la modélisation sinusoïdale [7]. Cette technique sera exposée au chapitre 3 et nous expliquerons pourquoi elle n'a pas été retenue pour le système de codage proposé.

On note $s(n,l)$ la partie du signal $s(n)$ dans la fenêtre d'analyse l, avec $n \in \{0,\ldots,N-1\}$, N étant le nombre d'échantillons dans une fenêtre. On a alors :

$$s(n,l) = R_N(n - Rl)s(n) \tag{2.1}$$

où R est le pas de translation et $R_N(n)$ est la fenêtre rectangulaire :

$$R_N(n) = \begin{cases} 1, & n = 0,1,\ \ldots,\ N-1, \\ 0, & \text{sinon.} \end{cases} \tag{2.2}$$

Si l'on désigne par $\hat{s}(n,l)$ le signal synthétique correspondant à $s(n,l)$, le signal reconstruit $\hat{s}(n)$ est obtenu par "recouvrement et addition" (*overlap-add*) selon :

$$\hat{s}(n) = \sum_l h(n - lR)\hat{s}(n,l) \tag{2.3}$$

où $h(n)$ est une fenêtre de pondération garantissant la condition de reconstruction [3]

$$\sum_l h(n - lR) = 1 \tag{2.4}$$

On utilise des fenêtres de Hanning telles que présentées à la figure 2.6. Ces fenêtres ont pour expression :

$$h(n) = \frac{1}{2} - \frac{1}{2}\cos\left(2\pi\frac{n}{N}\right). \tag{2.5}$$

Deux types de fenêtres de synthèse particuliers sont utilisés pour les transitions fenêtres longues/fenêtres courtes et *vice-versa*. La forme de ces fenêtres est donnée à la figure (2.7). Elles sont définies par :

$$h_{lc}(n) = \begin{cases} \frac{1}{2} - \frac{1}{2}\cos\left(2\pi\frac{n}{N_l}\right), & n \in [0,N_l/2-1]; \\ 1, & n \in [N_l/2,N_l - R_c - 1]; \\ \frac{1}{2} + \frac{1}{2}\cos\left(2\pi\frac{n+\frac{N_c}{2}}{N_c}\right), & n \in [N_l - R_c,N_l - 1] \end{cases}$$

pour la transition longues/courtes, et

$$h_{cl}(n) = \begin{cases} \frac{1}{2} - \frac{1}{2}\cos\left(2\pi\frac{n}{N_c}\right), & n \in [0,N_c/2-1]; \\ 1, & n \in [N_c/2,N_l - R_l - 1]; \\ \frac{1}{2} + \frac{1}{2}\cos\left(2\pi\frac{n+\frac{N_l}{2}}{N_l}\right), & n \in [N_l - R_l,N_l - 1] \end{cases}$$

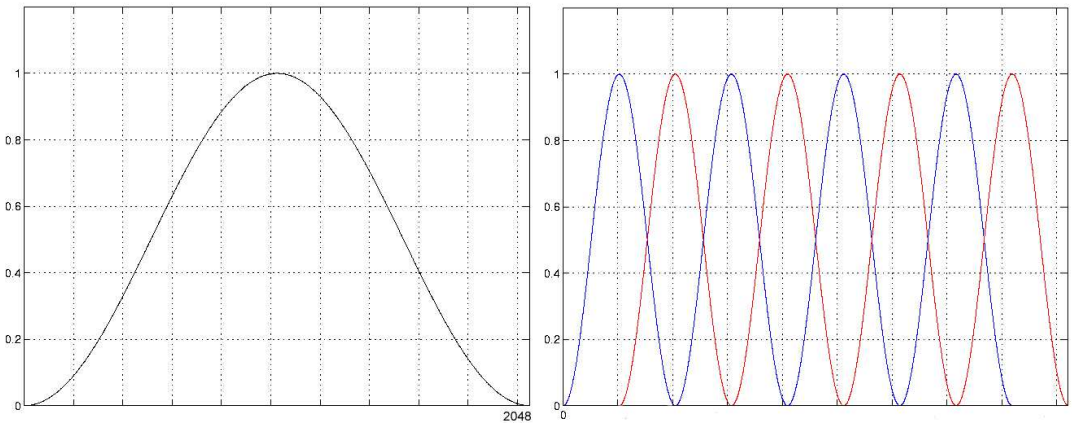

FIG. 2.6 – *Fenêtres de pondération de Hanning*

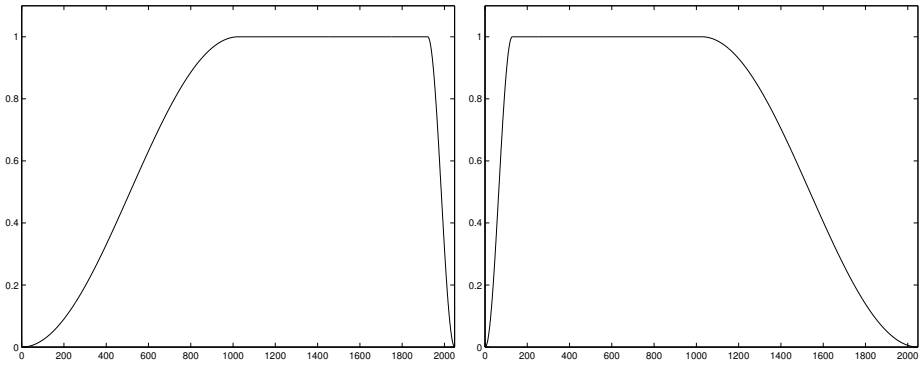

FIG. 2.7 – *Fenêtres de pondération de transitions longues/courtes et courtes/longues*

pour la transition courtes/longues.

Les parties fortement transitoires vont donc être analysées à l'aide des fenêtres courtes, le codeur fonctionnera en *mode T*, et les parties pseudo-stationnaires à l'aide des fenêtres longues et on dira que le codeur fonctionne en *mode S*. En mode T, une fenêtre longue est remplacée par 15 fenêtres courtes. Précisons que l'on peut dans certaines conditions utiliser 7 fenêtres courtes au lieu de 15 fenêtres afin d'adapter un modèle dédié aux transitoires forts précisément et uniquement sur les segments du signal concernés et de limiter l'augmentation du débit encourue (c.f. chapitre 5). En notant t_o l'estimation de l'instant de l'attaque, le mode 7 fenêtres courtes est adopté si $t_o \leq N_l - R_c$, sinon on utilise 15 fenêtres courtes.

2.3 Composante pseudo-stationnaire EDS et mode S

La partie non fortement transitoire du signal est modélisée par une somme de Sinusoïdes Amorties Exponentiellement. Une description détaillée du processus de modélisation sera donnée au chapitre 3. Signalons dès à présent, que l'extraction des paramètres se fait à

3. Cette condition est modifiée dans le cas où la fenêtre d'analyse n'est pas rectangulaire

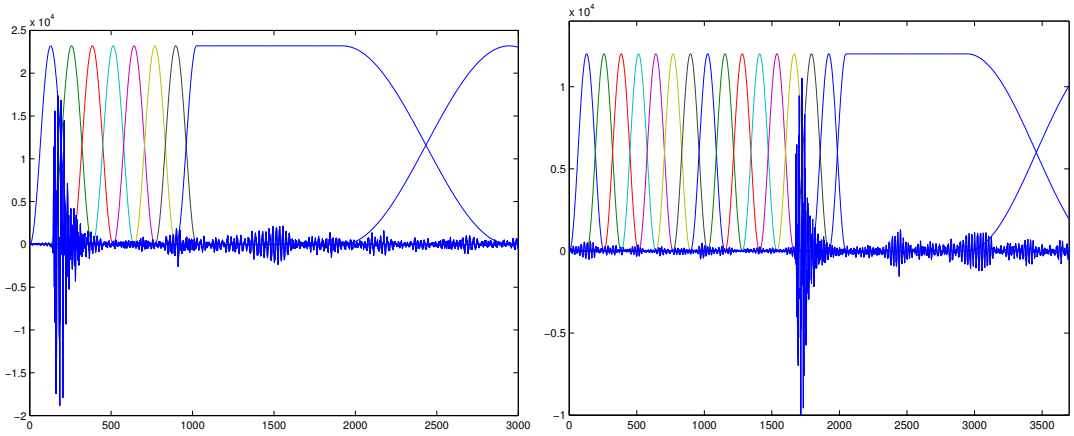

FIG. 2.8 – *Mode 7 fenêtres courtes - Mode 15 fenêtres courtes*

l'aide de deux classes de méthodes selon que l'on soit en mode S ou T. Une méthode itérative basée sur la Transformée de Fourier Discrète rapide (FFT) est utilisée sur les fenêtres longues permettant une très bonne résolution fréquentielle. 120 pics du spectre de puissance sont ainsi sélectionnés dans la bande 5Hz-10kHz. Nous verrons que nombre de ces pics sera abandonné lors de la phase de sélection des trajectoires sur la base de critères psychoacoustiques. Les coefficients d'amortissement sont déterminés grâce à une méthode de FFTs décalées [31] et les amplitudes et phases initiales sont résolues par une méthode des moindres carrés. Sur les fenêtres courtes, la FFT ne peut plus être utilisée eu égard à la mauvaise résolution fréquentielle (125 Hz). Nous faisons donc appel à une classe de méthodes dites haute-résolution permettant de garder de bonnes performances de modélisation sur ce type de fenêtres.

2.4 Composante transitoire et mode T

Les performances de modélisation par EDS devenant très faibles dans le cas de signaux fortement transitoire, il est nécessaire d'adapter un modèle prévu spécifiquement pour ce contexte. Un transitoire se retrouve toujours situé dans deux fenêtres d'analyse courtes successives compte tenu de la stratégie de segmentation adoptée (voir figure 2.8). Sur ces deux fenêtres, on effectue une modélisation basée sur le dictionnaire de Gabor. Les sons percussifs sont très bien reproduits à l'aide de ce modèle qui profite d'un paramètre de décalage permettant d'éviter les phénomènes de pré-écho. L'extraction des paramètres se fait grâce à l'algorithme Matching Pursuit qui sera décrit en détails au chapitre 4. Sur le reste des fenêtres courtes, on garde le modèle EDS en association avec une approche haute-résolution. On limite ainsi l'utilisation du modèle de Gabor aux seuls segments contenant effectivement les attaques. Nous verrons qu'une augmentation du débit est accusée dans ce cas, notamment à cause de la nécessité de coder les phases initiales.

2.5 Tracking

Lorsque les paramètres de modèle ont été estimés sur la trame l, ceux-ci sont appariés à l'intérieur de trajectoires inter-trames. Cette étape d'appariement ou tracking est réalisée dans une optique de quantification puisque les déviations des trajectoires de fréquence sont petites, sur des segments pseudo-stationnaires, ce qui suggère tout naturellement un schéma de quantification prédictif comme il sera vu au chapitre 5. De plus, cela permet de prendre des décisions quant à la pertinence de certains partiels sélectionnés : on sent, intuitivement, que des trajectoires de fréquence trop courtes ne sont pas très importantes et sont plutôt assimilables à du bruit; elles peuvent être abandonnées dans une perspective de compression. On verra que l'on pourra considérer les mesures de Rapports Signal à Masque (SMR) le long de ces trajectoires, pour sélectionner celles qui sont les plus perceptuellement significatives [14]. Mieux encore, le tracking permet, lors de l'étape de synthèse, d'abandonner les phases initiales des sinusoïdes, sur les parties pseudo-stationnaires. En effet, l'oreille est, dans ce cas, peu sensible aux valeurs de ces phases tant que leur continuité est assurée le long d'une trajectoire [7] (c.f. chapitre 3).

2.6 Synthèse

Les paramètres quantifiés permettent de reconstruire le signal audio de modèle. On est amené à réaliser deux types de synthèse. La première, au codeur, exploite les quatre informations quantifiées $[\bar{a}_m(l),\bar{d}_m(l),\bar{\omega}_m(l),\bar{\phi}_m(l)]$ de façon à reconstruire localement un signal en phase avec le signal original de telle sorte qu'une soustraction entre ces deux signaux ait un sens. On utilise, en l'occurrence, la technique OLA avec des fenêtres de pondération de Hanning comme décrit dans la section 2.2 et la synthèse d'une sinusoïde amortie à la trame l se fait selon :

$$s_m(n,l) = a_m(l)\exp(d_m(l)n)\cos(\omega_m(l)n + \phi_m(l)). \tag{2.6}$$

Le deuxième type de synthèse est effectué au décodeur en se contentant (sans dégradation auditive comme le montre des résultats des simulations) des trois informations $[\bar{a}_m(l),\bar{d}_m(l),\bar{\omega}_m(l)]$, sur les trames de signal pseudo-stationnaire, uniquement au niveau des continuations de trajectoire (c.f. chapitre 3). La synthèse est encore du type OLA mais une sinusoïde amortie est synthétisée, à la trame l, selon :

$$s_m(n,l) = a_m(l)\exp(d_m(l)n)\cos(\omega_m(l)n + \tilde{\phi}_m(l)). \tag{2.7}$$

en distinguant deux cas pour la valeur de $\tilde{\phi}_m(l)$:

1. *les paramètres* $[a_{m_t}(l),d_{m_t}(l),\omega_{m_t}(l),\phi_{m_t}(l)]$ *font partie d'une trajectoire t qui est en état de naissance à la trame l, c'est à dire qui est initiée à la trame l, ou d'une trame de transitoire*; dans ce cas

   $$\tilde{\phi}_{m_t}(l) = \phi_{m_t}(l)$$

   la phase $\phi_{m_t}(l)$ étant explicitement codée dans le bitstream;

2. *les paramètres* $[a_{m_t}(l),d_{m_t}(l),\omega_{m_t}(l),\phi_{m_t}(l)]$ *font partie d'une trajectoire qui est continuée depuis la trame* $l-1$, la phase $\phi_{m_t}(l)$ qui a été estimée lors de l'étape de modélisation n'a pas été codée dans le bitstream; on calcule alors $\tilde{\phi}_{m_t}(l)$ selon

$$\tilde{\phi}_{m_t}(l) = \tilde{\phi}_{p_t}(l-1) + \frac{3}{4}\omega_{p_t}(l-1)N - \omega_{m_t}(l)\frac{N}{4}.$$

Cette formule [5] permet d'assurer la continuité du signal d'une trame à l'autre. En fait, on n'assure pas une continuité parfaite de la phase instantané, on se contente de réaliser $\phi(N,l-1) \simeq \phi(0,l)$ et la synthèse OLA se charge du reste. Cette technique a été jugée très performante grâce aux simulations effectuées et a été préférée aux techniques de synthèse habituelles [7, 15].

Alors même que le signal synthétique est légèrement déphasé par rapport à l'original sur les segments pseudo-stationnaires, les tests d'écoute montrent que les dégradations sont faiblement perçues. Insistons sur le fait que cela ne concerne pas les segments transitoires sur lesquels les phases sont prises en compte et la version modélisée se superpose à l'originale.

2.7 Composante stochastique

Au terme de la synthèse de la composante déterministe du signal, un résiduel est calculé et modélisé comme étant du bruit. On peut alors faire l'hypothèse qu'il est bien décrit par l'allure générale de son spectre d'amplitude. Il n'est nullement nécessaire de coder sa phase instantanée ni les valeurs exactes des fréquences. On pourra donc se contenter de modéliser l'enveloppe de son spectre d'amplitude [15].

On suppose alors que le signal $r(n)$ est issu d'un processus auto régressif :

$$r(n) = \sum_{i=1}^{P} a_i r(n-i) + b(n) \tag{2.8}$$

avec $b(n)$ un bruit blanc centré de puissance σ^2. L'intérêt d'un tel modèle réside dans le fait que l'on va pouvoir représenter l'enveloppe spectrale $|R(k)|$ du résiduel en effectuant une simple opération de filtrage de bruit blanc. On ne code alors que les paramètres du filtre et une valeur de gain.

Les paramètres AR, *i.e*, les coefficients a_i, sont déterminés selon un critère moindres carrés par minimisation de la puissance moyenne de l'erreur de prédiction $\mathbf{e} = \mathbf{r} - \check{\mathbf{a}}$; $\check{\mathbf{a}}$ étant le signal prédit.

Le résiduel est ainsi analysé sur des fenêtres recouvrantes et synthétisé par une technique OLA. La puissance est normalisée à celle du signal. Notons que l'utilisation des fenêtres adaptatives de gain [5] permet d'améliorer sensiblement la qualité de modélisation. Les résultats de simulations montrent que la fusion du résiduel synthétique avec la composante déterministe de première modélisation donne lieu à un effet de bruit de souffle fortement gênant à l'écoute.

Plusieurs techniques de modélisation du résiduel ont été testées [32, 15, 14] et les mêmes artefacts ont été notés avec plus ou moins d'importance. Par conséquent, nous avons fait le choix de ne modéliser que la partie du bruit aux fréquences supérieures à 8kHz ce qui a permis d'éliminer l'impression de bruit de souffle et d'obtenir des résultats satisfaisants. Une technique de modélisation efficace de la composante stochastique sur toute la bande audio ne pourrait qu'améliorer la qualité. Nous pensons que l'utilisation de techniques d'enrichissement spectral serait très efficace [13].

Aux chapitres suivants nous nous intéressons aux outils et justifications théoriques qui ont mené à cette architecture de codage.

Chapitre 3

Modélisation du signal audio par somme de Sinusoïdes Amorties Exponentiellement

Les systèmes de modélisation de signaux audio par les méthodes *somme de sinusoïdes* s'avèrent efficaces pour modéliser des signaux stationnaires, présentant des variations lentes au regard de la durée d'analyse. Cependant, ces systèmes ne sont pas à même de représenter des signaux transitoires, typiquement les attaques ou évanouissements de sons, qui sont par nature localisés à la fois en temps et en fréquence. En effet, le modèle sinusoïdal souffre d'un défaut majeur car il réalise l'opération de modélisation à amplitude constante sur la durée d'une trame. Le suivi des variations d'enveloppe est alors effectué lors de la phase de synthèse au moyen d'interpolations linéaires des amplitudes estimés sur les trames successives [7], ce qui reste très approximatif. Il parait donc intéressant de produire une représentation du signal à partir d'une famille de formes d'ondes permettant l'étude de signaux à variation plus rapide. Dans cette optique, on fait le choix d'exploiter le modèle temps/fréquence "Sinusoïdes Amorties Exponentiellement" ou Exponentially Damped Sinusoids (EDS) [34] obtenu à partir du modèle sinusoïdal en introduisant un fenêtrage temporel par une exponentielle. Nous présentons, dans ce chapitre les propriétés structurelles de ce modèle et les stratégies d'expansion permettant l'extraction des paramètres.

3.1 Définition du modèle EDS

Le modèle M-EDS peut être vu comme une généralisation du modèle sinusoïdal d'ordre M dont on rappelle l'expression dans le cas complexe :

$$x_M(n) = \sum_{m=1}^{M} g_m(n) e^{i\omega_m n} \text{ avec } g_m(n) = a_m e^{i\phi_m} \tag{3.1}$$

où

- a_m est la m-ème amplitude réelle ;
- w_m est la m-ème pulsation ;

- ϕ_m est la m-ème phase initiale appartenant à l'ensemble $[-\pi,\pi[$;
- $g_m(n)$ est une fenêtre temporelle, choisie égale à une constante dans le cas du modèle sinusoïdal.

Le terme "$e^{i\omega_m n}$" est un noyau fréquentiel expliquant le caractère oscillant du signal $x_M(n)$. Le modèle M-EDS se déduit alors logiquement de l'expression (3.1) en choisissant un fenêtrage $g_m(n)$ non plus constant mais dépendant du temps et d'un paramètre réel d'atténuation d_m selon

$$g_m(n) = a_m e^{i\phi_m} e^{d_m n}. \tag{3.2}$$

Cette dernière expression nous conduit à donner l'expression du signal réel de modèle M-EDS, pour $n = 0, \ldots, N-1$

$$\begin{aligned} x_M(n) &= \sum_{m=1}^{M} a_m e^{d_m n} \cos(\omega_m n + \phi_m) \\ &= \tfrac{1}{2} \sum_{m=1}^{M} (\alpha_m z_m^n + \alpha_m^* z_m^{*n}) \end{aligned} \tag{3.3}$$

où $\alpha_m = a_m e^{i\phi_m}$ est la m-ème amplitude complexe et $z_m = e^{d_m + i\omega_m}$ est le m-ème pôle; $x_M(n)$ est un modèle M-pôles.

3.2 Analyse Temps/Fréquence du modèle EDS

3.2.1 Considérations structurelles

Le modèle Sinusoïdal d'ordre 1 possède un support temporel, Δ_t, constant et égal à la durée d'analyse, soit $\Delta_t = N$. On en déduit selon l'inégalité de Heisenberg, $\Delta_t \Delta_f \geq 1/(4\pi)$ liant supports temporel et fréquentiel, noté Δ_f, que le support fréquentiel d'un modèle sinusoïdal est minimal et vérifie $\Delta_f = 1/(4N\pi)$. On voit alors que ce modèle est tout particulièrement adapté pour la représentation de signaux localisés dans le domaine fréquentiel et sera à l'inverse très peu efficace sur la modélisation de signaux à support temporel réduit. Le modèle 1-EDS quant à lui présente un support temporel réduit Δ_t pour des valeurs d'atténuation non-nulles. Le modèle 1-EDS permet donc de modéliser de manière compacte tout événement du plan TF se produisant en *début* et *fin* du segment d'analyse. Sur la figure 3.1, on a représenté les "boites de Heisenberg" [22] pour le modèle sinusoidal d'ordre 1 et pour le modèle 1-EDS avec des valeurs d'atténuation négatives. On visualise sur les figures 3.2 et 3.3 le support temporel et fréquentiel (largeur du lobe) du modèle 1-EDS pour différentes valeurs d'atténuation.

3.2.2 Analyse Temps/Fréquence par banc de filtre

Dans cette partie, on s'attache à comparer les performances du modèle EDS à celles du modèle de Fourier (coefficients d'atténuation nuls) à *nombre de paramètres de modèle égal*. On rappelle l'expression du Rapport Signal à Bruit temporel défini pour la l-ème trame par

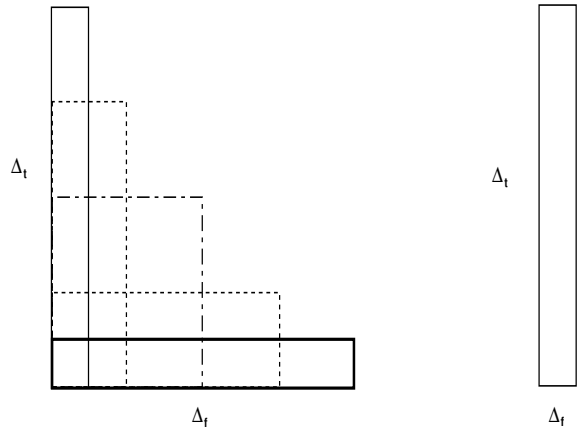

FIG. 3.1 – *Occupation de la ressource dans le plan Temps-Fréquence pour le modèle Sinusoïdal d'ordre 1 (à droite) et le modèle 1-EDS pour des atténuations négatives (à gauche).*

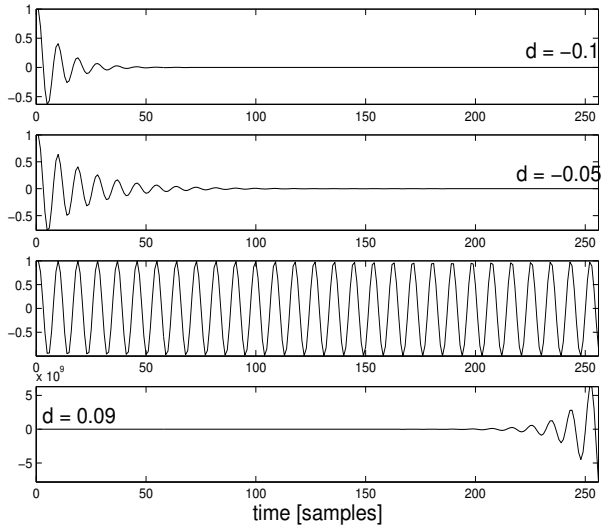

FIG. 3.2 – *Formes d'onde temporelle pour différentes valeurs d'atténuation*

$$\mathrm{RSB}_T^{(l)} = 10\log_{10}\frac{\sum_{n=0}^{N-1}|x(n,l)|^2}{\sum_{n=0}^{N-1}|x(n,l)-\hat{x}(n,l)|^2}\ [\mathrm{dB}]. \tag{3.4}$$

Ce critère doit être considéré avec beaucoup de prudence. En effet, quand bien même il est approprié dans le contexte de la modélisation temporelle de signaux, il perd toute sa valeur dans le cadre de la compression du signal audio stationnaires où le caractère fréquentiel prévaut sur les variations temporelles de la forme d'onde : c'est en fait ce qui justifie que l'on peut effectuer la synthèse de ce type de signaux ne prenant pas en compte les paramètres de phase initiale. Afin d'introduire une analyse "fréquentielle" du signal, on utilise un banc de 32 filtres polyphase pseudo-QMF effectuant une partition uniforme de l'axe des fréquences. Chaque bande est de largeur 500 Hz pour une fréquence d'échantillonnage de 32kHz. Ce dispositif n'est autre que le banc de filtre d'analyse de MPEG1 [11]. On peut voir, sur la

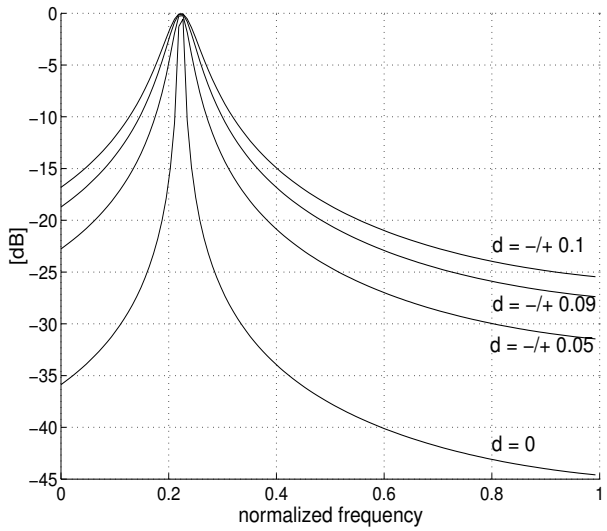

FIG. 3.3 – *Transformée de Fourier de 1-EDS pour différentes valeurs d'atténuation*

figure 3.4 le banc de filtre utilisé pour une fréquence d'échantillonnage de 32kHz.

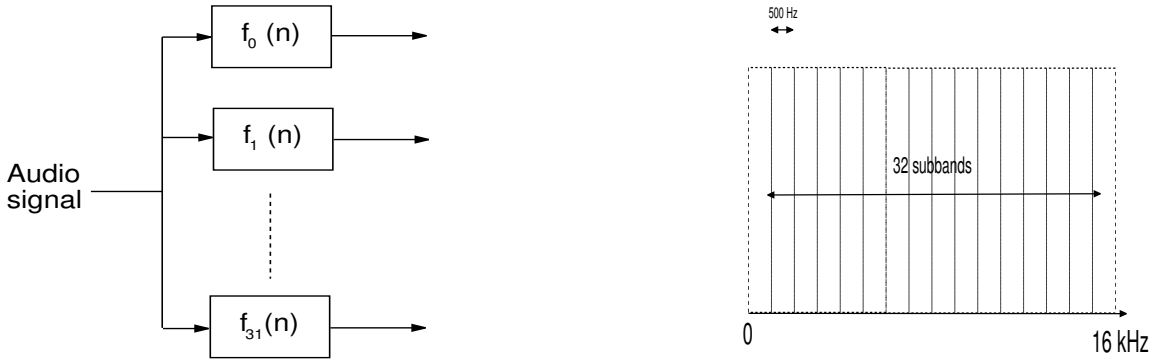

FIG. 3.4 – *Banc de 32 filtres - Partition idéale et régulière du spectre pour une fréquence d'échantillonnage de 32kHz*

Ensuite, dans chaque sous-bande, on applique le critère de RSB_T défini en (3.4) afin de caractériser la ressemblance temporelle des signaux modélisés. En ce sens, on exploite une représentation "Temps-Fréquence" du signal et l'utilisation du RSB_T dans la comparaison est pertinente car la synthèse est effectuée de façon identique par une technique OLA pour les deux modèles et en présence de phases initiales. Il est alors possible de re-formuler le critère par

$$\mathrm{RSB}_{TF}^{(l,b)} = 10\log_{10}\frac{\sum_{n=0}^{N-1}|x_b(n,l)|^2}{\sum_{n=0}^{N-1}|x_b(n,l)-\hat{x}_b(n,l)|^2}\ [\mathrm{dB}] \tag{3.5}$$

avec $b = 1,\ldots,32$. On note les signaux filtrés par

$$x_b(n,l) = f_b(n) * x(n,l),\ \ \hat{x}_b(n,l) = f_b(n) * \hat{x}(n,l). \tag{3.6}$$

$x_b(n,l)$ (resp. $\hat{x}_b(n,l)$) est le signal de sous-bande b issu du filtrage par $f_b(n)$ du signal original (resp. de modèle) considéré à la trame l. Notons que cette représentation est à mettre en relation avec le calcul des niveaux de puissance dans chacune des sous-bandes. En effet, un $\text{RSB}_{TF}^{(l,b)}$ dans une sous-bande b de puissance modérée ne sera pas déterminant pour l'analyse. A l'inverse, si une sous-bande possède une puissance élevée, l'analyse des résultats en terme de $\text{RSB}_{TF}^{(l,b)}$ sera très instructive quant au modèle considéré. Notons, enfin que le calcul de la puissance dans une sous-bande suit la formule suivante

$$P^{(l,b)} = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1}|x_b(n,l)|^2\right) \tag{3.7}$$

pour le signal original et

$$\hat{P}^{(l,b)} = 10\log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1}|\hat{x}_b(n,l)|^2\right) \tag{3.8}$$

pour les signaux de modèle. L'analyse des puissances est elle-même pertinente : un modèle restituant un niveau de puissance dans une sous-bande donnée plus proche de celui du signal original est plus performant qu'un autre ne modélisant pas l'énergie dans la sous-bande considérée. Signalons enfin que les résultats des simulations ont toujours été validés par des tests d'écoute.

On caractérise sur des exemples de signaux audio réels, le comportement du modèle paramétrique non-stationnaire EDS par rapport au modèle paramétrique stationnaire sinusoïdal classique, ici nommé modèle de Fourier. Une analyse par fenêtres rectangulaires recouvrantes à 50 % est effectuée sur le signal original. La synthèse est quant à elle réalisée au travers de l'utilisation de fenêtres de Hanning. Cette architecture d'analyse/synthèse, de type OLA (Overlap and Add), vérifie les conditions de reconstruction parfaite. On choisit une durée d'analyse de 512 échantillons soit 16 ms pour une fréquence d'échantillonnage de 32kHz. Les signaux tests sont :

- **Signal.1**, 517 ms (63 trames) de signal de parole : le mot 'matter' est prononcé par une voix féminine;
- **Signal.2**, 517 ms (63 trames) de signal de violon-trompette;
- **Signal.3**, 517 ms (63 trames) de signal de clochettes.

Ces signaux ont été choisis dans l'échantillon de signaux tests du "call for proposal" de MPEG4 [6]; ils sont représentatifs d'une classe plus large de phénomènes propres aux signaux audio. Ces signaux sont successivement représentés par le modèle de Fourier et le modèle EDS. Le choix des ordres de modélisation est le suivant

$$M_{EDS} = 35 \quad \text{et} \quad M_{Fourier} = 4/3\ M_{EDS} \approx 47 \tag{3.9}$$

Le rapport 4/3 s'explique par le fait que l'on souhaite avoir un nombre identique de paramètres de modèle (≈ 140) pour les deux modèles et pour tenir compte des paramètres

d'atténuation $\{d_m\}$ présents dans le modèle EDS et absents dans le modèle de Fourier.

Signal.1 : Parole

Sur la figure 3.5, on a reporté le signal original et les signaux modélisés.

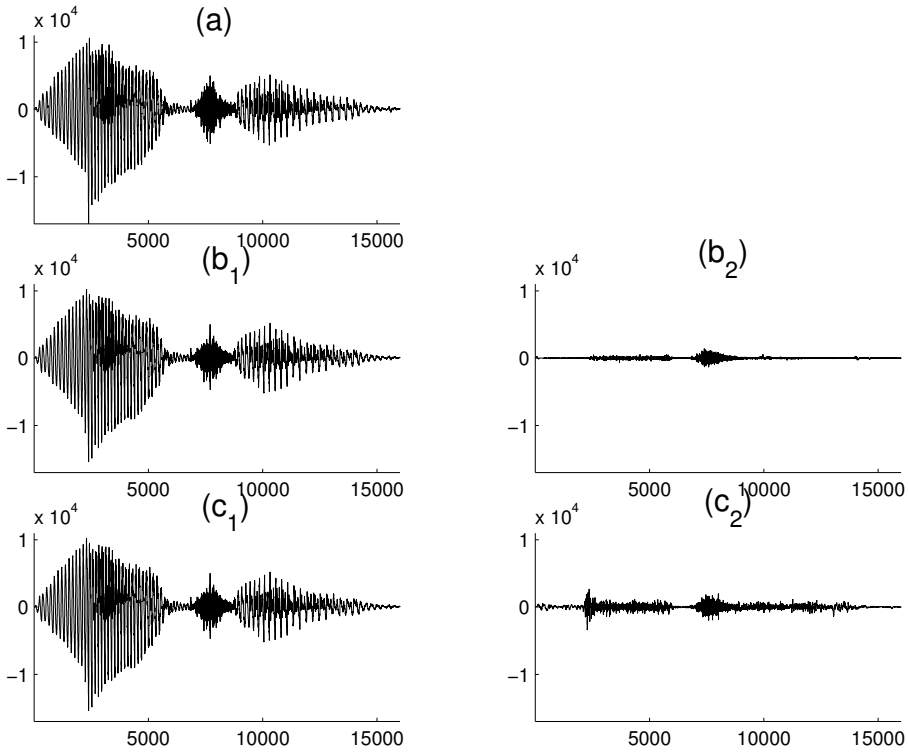

FIG. 3.5 – *512 ms de parole :* (a) *signal original,* (b_1) *signal EDS,* (b_2) *résiduel EDS,* (c_1) *signal de Fourier,* (c_2) *résiduel de Fourier*

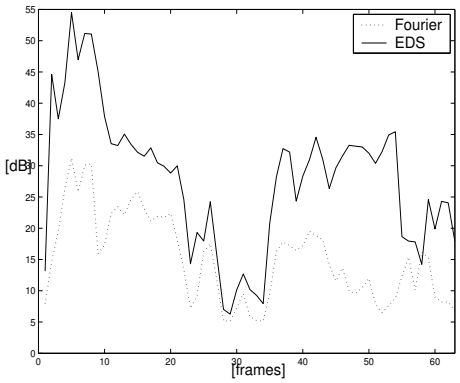

FIG. 3.6 – *512 ms de parole : Rapports Signaux à Bruit*

Dans la partie droite, figure les signaux résiduels, construits à partir de la soustraction du signal original et du signal de modèle. Sur la figure 3.6, est reportée la courbe de RSB_T

pour le signal de parole considéré et pour les deux modèles. Notons que le RSB_T du modèle EDS est toujours supérieure à celui du modèle de Fourier sur les zones voisées, non-voisées et sur les zones transitoires. Comme on l'a vu précédemment, le critère de RSB_T nous donne uniquement une information temporelle. Afin de caractériser fréquentiellement le modèle EDS et le modèle de Fourier on calcule une puissance (c.f. figure 3.7) et un RSB_{TF} (c.f. figure 3.8) par sous-bande. Le signal de parole est essentiellement basse fréquence, c'est à dire qu'il concentre une grande partie de sa puissance dans les sous-bandes d'index inférieur à 16 ($\approx$ 8kHz). Or sur cet intervalle fréquentiel, on peut voir que les RSB_{TF} par sous-bande du modèle EDS sont sensiblement (> 10 dB parfois) plus élevés que ceux relatifs au modèle de Fourier.

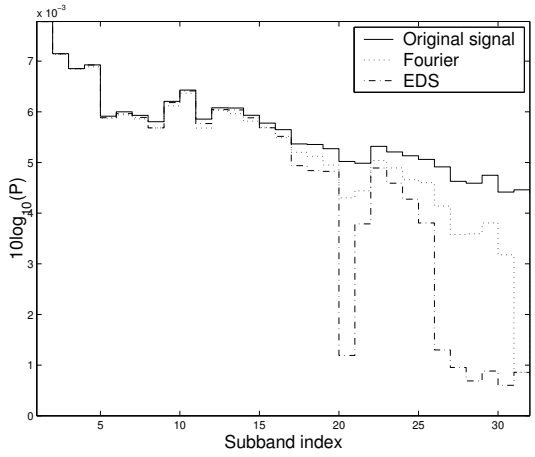

FIG. 3.7 – *512 ms de parole : puissance par sous-bande*

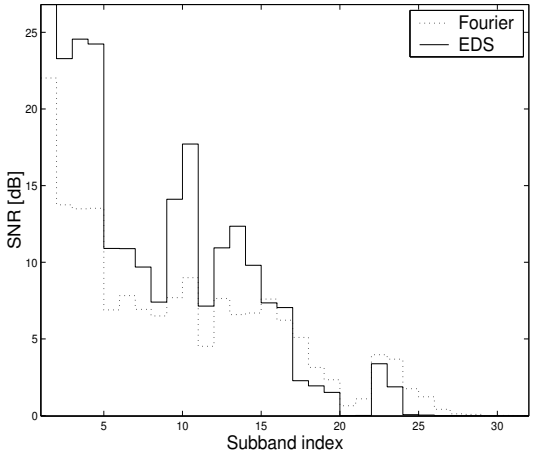

FIG. 3.8 – *512 ms de parole : RSB_{TF} par sous-bande*

Signal.2 : Violon-trompette

On réalise la même opération que pour l'exemple précédent mais cette fois-ci on considère un signal musical de violon + trompettes. Sur les figures 3.9 et 3.10 est représenté le signal original, ses versions modélisées et résiduelles et les mesures de RSB_T. Ici encore la mesure de RSB_T indique une modélisation temporelle par le modèle EDS plus efficace que celle obtenue par le modèle de Fourier. La puissance (c.f. figure 3.11) et le RSB_{TF} (c.f. figure 3.12) par sous-bande sont aussi instructifs car si on considère que les sous-bandes de plus haute puissance ($b < 10$, ≈ 5 kHz), le modèle EDS affiche des résultats supérieurs au modèle de Fourier. Notons que dans les sous-bandes indexée de 14 à 32, le modèle de Fourier semble se comporter mieux que le modèle EDS. Cependant, on considérera que ces sous-bandes peuvent être négligées du fait de leur faible puissance.

Signal.3 : glockenspiel

Enfin on termine par un signal musical de clochettes présentant des zones de stationnarité et des attaques franches. Sur les figures 3.13 et 3.14 sont représentés le signal original, ses versions modélisées et résiduelles et les mesures de RSB_T. Ici encore la mesure de RSB_T

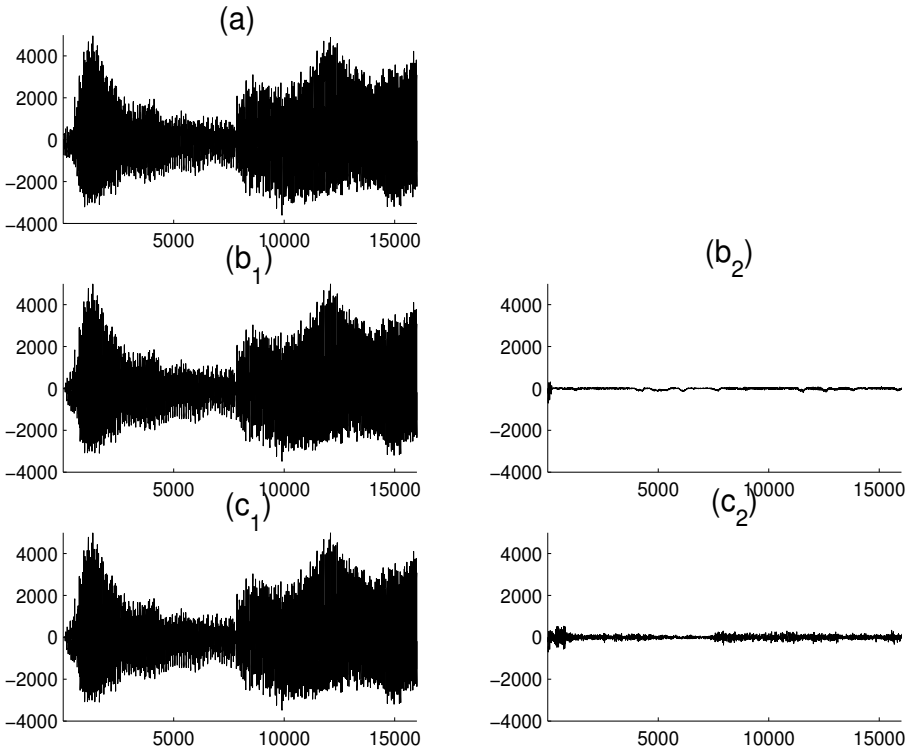

FIG. 3.9 – *512 ms de violon-trompette :* (a) *signal original,* (b_1) *signal EDS,* (b_2) *résiduel EDS,* (c_1) *signal de Fourier,* (c_2) *résiduel de Fourier*

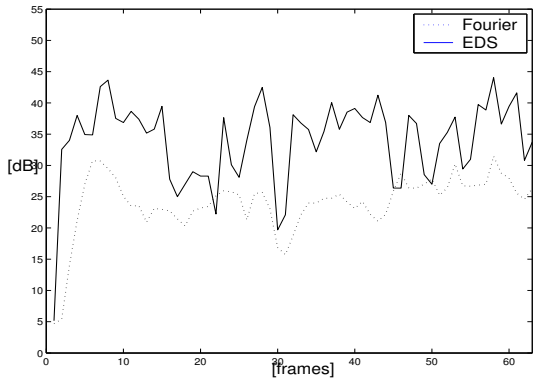

FIG. 3.10 – *512 ms de violon-trompette : Rapports Signaux à Bruit*

indique une modélisation temporelle par le modèle EDS plus efficace par rapport au modèle de Fourier, notamment sur les attaques. La puissance (c.f. figure 3.15) et le RSB_{TF} (c.f. figure 3.16) par sous-bande sont aussi instructifs car si on considère que les sous-bandes de plus haute puissance ($b < 25$, $\approx$ 13kHz), le modèle EDS affiche des résultats supérieurs au modèle de Fourier. Notons que cette analyse "Temps-Fréquence" a été effectuée sur des segments de 16 ms (512 échantillons). Les performances du modèle EDS auraient été encore supérieures à celle du modèle de Fourier si on avait choisi des fenêtres de 32 ms, voire 64 ms sur lesquelles le signal audio admet encore moins une représentation compacte par un modèle

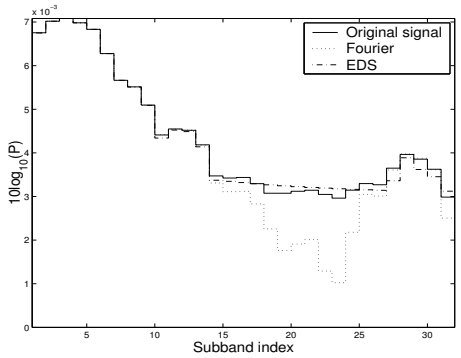

FIG. 3.11 – *512 ms de violon-trompette : puissance par sous-bande*

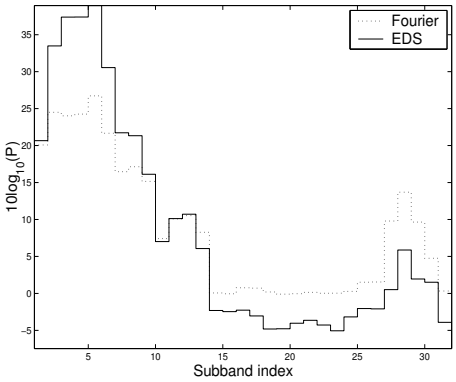

FIG. 3.12 – *512 ms de violon-trompette : RSB_{TF} par sous-bande*

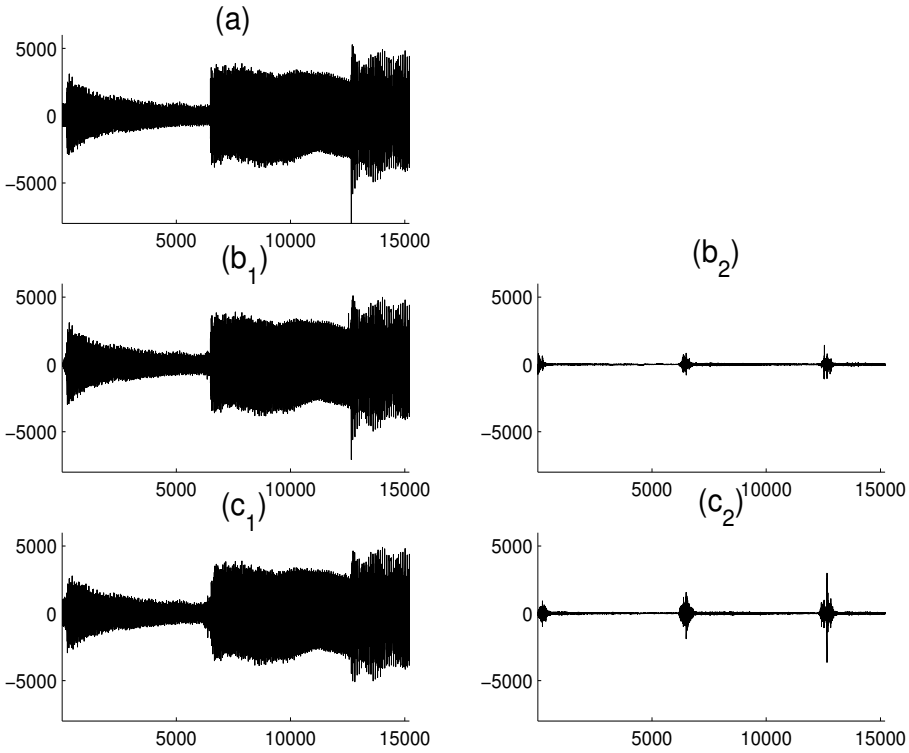

FIG. 3.13 – *512 ms de clochettes : (a) signal original, (b_1) signal EDS, (b_2) résiduel EDS, (c_1) signal de Fourier, (c_2) résiduel de Fourier*

stationnaire.

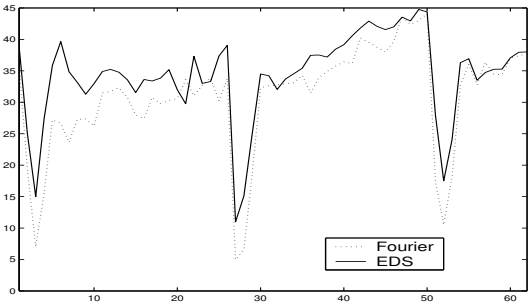

FIG. 3.14 – *512 ms de clochettes : Rapports Signaux à Bruit*

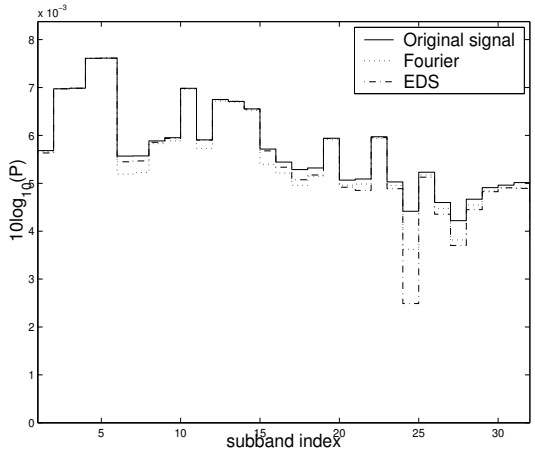

FIG. 3.15 – *512 ms de clochette : puissance par sous-bande*

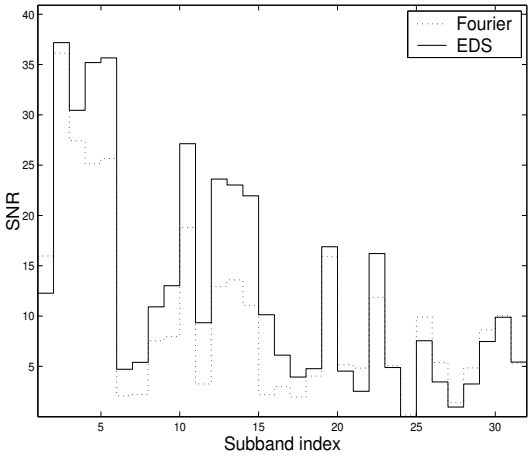

FIG. 3.16 – *512 ms de clochette : RSB_{TF} par sous-bande*

3.3 Estimation des paramètres de modèle

On souhaite modéliser le signal audio $x(n)$ par le modèle M-pôles défini précédemment en utilisant un critère moindres carrés

$$\varepsilon(\boldsymbol{\alpha},\mathbf{z}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{z}} \sum_{n=0}^{N-1} |x(n) - x_M(n)|^2 \tag{3.10}$$

où les paramètres de modèle sont le vecteur des amplitudes complexes $\boldsymbol{\alpha} = \frac{1}{2}(\alpha_1\ \alpha_1^* \dots \alpha_M\ \alpha_M^*)^T$, le vecteur des pôles $\boldsymbol{z} = (z_1\ z_1^* \dots z_M\ z_M^*)^T$ et l'ordre de modèle M. Notons que l'ordre de modèle dans une application de compression audio peut être fixé afin d'atteindre un objectif de débit, noté ρ (bit/s) selon

$$\hat{M} = \frac{N}{\mathcal{N}_{bit}\beta f_e}\ \rho \tag{3.11}$$

où $\mathcal{N}_{bit}$ est le nombre moyen de bits pour coder les paramètres de modèle $\{\alpha_m,z_m\}$, β est un réel compris entre 1 et 2 selon la technique de synthèse utilisée et f_e est la fréquence d'échantillonnage en Hertz.

Une étude sommaire de ce critère moindres carrés, nous apprend qu'il est linéaire en l'amplitude complexe $\boldsymbol{\alpha}$ et non linéaire en $\mathbf{z}$. On présentera donc la méthode de détermination de $\boldsymbol{\alpha}$, *i.e*, les amplitudes a_m et les phases ϕ_m, en supposant $\mathbf{z}$ connu puis on verra deux techniques d'estimation des pulsations ω_m et des coefficients d'amortissement d_m et on discutera de leurs performances relatives.

3.3.1 Détermination des amplitudes complexes $\{\alpha_m\}$

Soit le signal mesuré $\mathbf{x} = (s(0) \dots s(N-1))^T$ et son modèle $\mathbf{x}_M = (s_M(0) \dots s_M(N-1))^T$. On peut écrire

$$\mathbf{x}_M = \mathbf{Z}\boldsymbol{\alpha} \tag{3.12}$$

où $\mathbf{Z}$ est la matrice de Vandermonde de dimension $N \times 2M$ définie comme suit

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 & 1 & \dots & 1 & 1 \\ z_1 & z_1^* & z_2 & z_2^* & \dots & z_M & z_M^* \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ z_1^{N-1} & z_1^{*N-1} & z_2^{N-1} & z_2^{*N-1} & \dots & z_M^{N-1} & z_M^{*N-1} \end{pmatrix}$$

Il nous importe, donc, de calculer les paramètres du modèle en minimisant la puissance de l'erreur de modélisation $\mathbf{e} = \mathbf{x} - \mathbf{x_M}$, ce qui revient à résoudre

$$\arg\min_{\{\alpha_m\}} \|\mathbf{x} - \mathbf{Z}\boldsymbol{\alpha}\|^2 \tag{3.13}$$

On voit que la formulation précédente correspond à un formalisme de système linéaire en $\boldsymbol{\alpha}$. On va, par conséquent, pouvoir calculer le vecteur $\boldsymbol{\alpha}$ en faisant l'hypothèse que l'on a connaissance des M pulsations et facteurs d'amortissement. Il vient alors

$$\boldsymbol{\alpha} = \mathbf{Z}^\dagger \mathbf{x} \tag{3.14}$$

où $\mathbf{Z}^\dagger$ est la pseudo-inverse de $\mathbf{Z}$.

Il nous reste à déterminer les amplitudes réelles $\{a_m\}$ et les phases $\{\phi_m\}$ à l'aide des formules suivantes

$$\begin{cases} a_m = \sqrt{\alpha_m \alpha_m^*} \\ \phi_m = -\frac{j}{2} \ln\left(\frac{\alpha_m}{\alpha_m^*}\right). \end{cases} \tag{3.15}$$

3.3.2 Estimation des pôles $\{z_m\}$

Méthode basée sur la FFT

Classiquement, les pulsations sont estimées au travers d'une analyse spectrale en sélectionnant les M pics les plus énergétiques du périodogramme [7]. Le spectre est calculé en exploitant une fenêtre de pondération adéquate permettant de disposer d'une bonne résolution

fréquentielle tout en atténuant les lobes secondaires [35]. On fait en même temps appel au zero-padding pour augmenter la précision fréquentielle. En l'occurrence, on utilise, dans ce travail, une fenêtre de Hanning symétrique. Les pics sont choisis de manière à ce que leurs amplitudes soient localement maximales en prenant, éventuellement, en compte des critères psycho-acoustiques [15, 36]. Nous faisons le choix d'extraire les pulsations au travers d'une approche itérative qui présente de bonnes performances d'estimation, ce que l'on désignera par IA-EDS (Iterative Algorithm for EDS modeling). On pose

$$\begin{cases} x_0(n) = x(n) \\ x_{m+1}(n) = x_m(n) - a_m e^{d_m n} \cos(\omega_m n + \phi_m). \end{cases} \tag{3.16}$$

x_{m+1} est ainsi le $(m+1)$-ème signal résiduel obtenu suite à l'approximation du signal x_m par un 1-EDS. L'expression (3.16) peut être réécrite

$$x_{m+1}(n) = x(n) - \sum_{k=0}^{m} a_k e^{d_k n} \cos(\omega_k n + \phi_k). \tag{3.17}$$

On cherche donc à minimiser à chaque itération k la norme 2 du résiduel x_k, $k = 1,...,M+1$. On commence par estimer la pulsation selon

$$\omega_m = \arg \max_{\omega \in [0,\pi[} |X_m(\omega)| \tag{3.18}$$

où $X_m(\omega)$ est le spectre du signal $x_m(n)$ observé au travers de la fenêtre de pondération. Ensuite on détermine le m-ème coefficient d'atténuation par la méthode des FFTs décalées selon

$$d_m = \frac{1}{n_1} \ln \frac{\left| X_m^{(1)}(\omega_m) \right|}{\left| X_m^{(0)}(\omega_m) \right|} \tag{3.19}$$

où $X_m^{(0)}(\omega)$ et $X_m^{(1)}(\omega)$ sont les FFTs respectives des signaux suivants

$$\begin{array}{rcll} x_m^{(0)}(n) & = & x_m(n)w(n), & \text{pour} \quad n = 0,\ldots,N_1-1 \\ x_m^{(1)}(n) & = & x_m(n)w(n-n_1), & \text{pour} \quad n = n_1,\ldots,N-1 \end{array} \tag{3.20}$$

où $n_1 + N_1 = N$ et n_1 est un offset temporel choisi petit devant la taille d'analyse. $w(n)$ est une fenêtre de pondération ayant pour fonction d'isoler le pôle de pulsation positive de celui de pulsation négative [31]. On choisit ici une fenêtre de Blackman [37] qui permet une bonne résolution dynamique. Une variante de cette méthode peut être considérée en faisant varier le décalage n_s afin de déterminer plusieurs estimations du paramètre $d_m^{(s)}$ avec $s = 1,\ldots,S$. L'estimation du coefficient d'amortissement d_m est alors la valeur moyenne $1/S \sum_{s=1}^{S} d_m^{(s)}$. En posant, $n_s = s\bar{N}$ où $\bar{N} < N$, on a

$$d_m = \frac{1}{S\bar{N}} \sum_{s=1}^{S} \frac{1}{s} \ln \left| X_m^{(s)}(\omega_m) \right| - \frac{\ln \left| X_m^{(0)}(\omega_m) \right|}{\bar{N}} \tag{3.21}$$

où $X_m^{(s)}(\omega_m)$ est la FFT, en ω_m, du signal $x_m^{(s)}(n) = x_m(n)w(n-n_s)$ pour $n = n_s, \ldots, N-1$. Cette approche présente une robustesse accrue lorsque l'on se trouve être dans un environnement fortement non-stationnaire. Ayant la connaissance du m-ème pôle z_m, on peut former la matrice de Vandermonde de dimension $N \times 2$ suivante

$$\boldsymbol{V}_m = \begin{pmatrix} 1 & 1 \\ z_m & z_m^* \\ \vdots & \vdots \\ z_m^{N-1} & z_m^{*(N-1)} \end{pmatrix} \tag{3.22}$$

et résoudre le critère LS suivant

$$\arg\min_{\boldsymbol{\alpha}_m} \|\boldsymbol{x}_m - \boldsymbol{V}_m\boldsymbol{\alpha}_m\|_2^2 \iff \boldsymbol{\alpha}_m = \boldsymbol{V}_m^{\dagger}\boldsymbol{x}_m \tag{3.23}$$

où $\boldsymbol{x}_m = (x_m(0) \ldots x_m(N-1))^T$ et $\boldsymbol{\alpha}_m = \frac{1}{2}(\alpha_m \ \alpha_m^*)^T$. Notons que la pseudo-inverse de $\boldsymbol{V}_m$ existe si et seulement si la forme Hermitienne $\boldsymbol{V}_m^H\boldsymbol{V}_m$ est de rang plein. Or si on suppose que $z_m \in \mathbb{C}$, *i.e.*, $\omega_m \neq 0$, alors le rang de $\boldsymbol{V}_m$ est 2. Par conservation du rang, on en déduit que $\boldsymbol{V}_m^H\boldsymbol{V}_m$ est aussi de rang 2 or cette matrice est carrée de dimension 2×2, on en conclut que la pseudo-inverse de $\boldsymbol{V}_m$ existe sous la condition que le pôle soit complexe. Dans le cadre de la compression de signaux audio, le signal non oscillant porte peu d'information et sera systématiquement écarté.

Méthode Haute-Résolution : HR-EDS

Il s'agit d'utiliser une classe de méthodes dites "sous-espaces" se basant sur la factorisation (3.12) qui fait apparaître une structure de Vandermonde. Différents algorithmes ont été proposés dans la littérature; on peut citer les algorithmes ESPRIT [38], Matrix-Pencil [39] de Kung [40] ou encore MUSIC [41]. Une présentation synthétique de ces méthodes peut être trouvée dans [41]. Nous présentons dans ce qui suit l'algorithme de Kung qui a été retenu pour le système de codage.

Donnons d'abord la structure algébrique de l'espace des observations. Soit $\mathcal{O}$ cet espace tel que $dim\,(\mathcal{O}) = L$ et soit un signal audio de N échantillons, noté $\mathbf{x}$ tel que $\mathbf{x} \in \mathcal{O}$. $\mathcal{O}$ peut être décomposé en deux sous-espaces : $\mathcal{S}$, le sous-espace signal de dimension M et $\mathcal{B}$, le sous-espace bruit tels que $\mathcal{O} = \mathcal{S} \oplus \mathcal{B}$.

Le modèle M-pôles admet la représentation formulée ci-dessous

$$\boldsymbol{x_M(n)} = \boldsymbol{1_M}\mathbf{Z}^{\boldsymbol{n}}\boldsymbol{\alpha} \tag{3.24}$$

où on note

- $\mathbf{1}_M = (1 \ldots 1) \in \mathbb{R}^{1\times M}$
- $\mathbf{Z} = \mathrm{diag}\{z_1, \ldots, z_M\} \in \mathbb{C}^M$
- $\alpha = (\alpha_1 \ldots \alpha_M)^T \in \mathbb{C}^{M\times 1}$

On construit, alors, la matrice de Hankel suivante

$$\mathbf{H} = \mathcal{H}_L\left(\mathbf{x}_M\right) = \begin{pmatrix} x_M(0) & x_M(1) & \dots & x_M(L-1) \\ x_M(1) & x_M(2) & \dots & x_M(L) \\ \vdots & \vdots & & \vdots \\ x_M(N-L-1) & x_M(N-L) & \dots & x_M(N-1) \end{pmatrix}_{(N-L)\times L} \tag{3.25}$$

où $\mathcal{H}_L(.)$ est l'opérateur de Hankel et $\mathbf{x}_M$ est le vecteur de dimension N construit à partir des échantillons $x_M(n)$. En remplaçant la formulation du modèle par son expression matricielle, on est en mesure de donner la factorisation de la matrice $\mathbf{H}$

$$\begin{aligned} \mathbf{H} &= \begin{pmatrix} \mathbf{1}_M\alpha & \mathbf{1}_M\mathbf{Z}\alpha & \dots & \mathbf{1}_M\mathbf{Z}^{L-1}\alpha \\ \mathbf{1}_M\mathbf{Z}\alpha & \mathbf{1}_M\mathbf{Z}^2\alpha & \dots & \mathbf{1}_M\mathbf{Z}^{L}\alpha \\ \vdots & \vdots & & \vdots \\ \mathbf{1}_M\mathbf{Z}^{N-L-1}\alpha & \mathbf{1}_M\mathbf{Z}^{N-L}\alpha & \dots & \mathbf{1}_M\mathbf{Z}^{N-1}\alpha \end{pmatrix} \\ &= \begin{bmatrix} \mathbf{1}_M \\ \mathbf{1}_M\mathbf{Z} \\ \vdots \\ \mathbf{1}_M\mathbf{Z}^{N-L-1} \end{bmatrix} \begin{bmatrix} \alpha & \mathbf{Z}\alpha & \dots & \mathbf{Z}^{L-1}\alpha \end{bmatrix} \\ &= \mathbf{O}_{(N-L)\times M}\mathbf{C}_{M\times L} \end{aligned} \tag{3.26}$$

En définissant les deux opérateurs matriciels de shiftage de ligne : $(.)_\uparrow$, la première ligne est supprimée et $(.)_\downarrow$, la dernière ligne est supprimée tels que

$$\mathbf{O}_\uparrow = \begin{bmatrix} \mathbf{1}_M\mathbf{Z} \\ \vdots \\ \mathbf{1}_M\mathbf{Z}^{N-L-1} \end{bmatrix} \quad \text{et} \quad \mathbf{O}_\downarrow = \begin{bmatrix} \mathbf{1}_M \\ \mathbf{1}_M\mathbf{Z} \\ \vdots \\ \mathbf{1}_M\mathbf{Z}^{N-L-2} \end{bmatrix} \tag{3.27}$$

l'expression (3.26) nous permet de mettre en évidence la propriété d'invariance par shiftage de ligne (ou invariance rotationnelle) $\mathbf{O}_\downarrow\mathbf{Z} = \mathbf{O}_\uparrow$. Cela conduit alors à construire le produit $\mathbf{O}_\downarrow^\dagger\mathbf{O}_\uparrow$ pour en chercher une décomposition en valeurs propres. Soit une matrice $\mathbf{F}$ inversible et unitaire telle que $\mathbf{O}_\downarrow^\dagger\mathbf{O}_\uparrow = \mathbf{F}\mathbf{G}\mathbf{F}^H = \mathbf{Z}$. En se servant du fait que $\mathbf{F}$ est unitaire et de l'expression précédente, on peut dire que $\mathbf{G}$ et $\mathbf{Z}$ sont semblables et conservent donc le même spectre. Pour déterminer $\mathbf{Z}$ il suffit donc de diagonaliser $\mathbf{O}_\downarrow^\dagger\mathbf{O}_\uparrow$.

Choix de la matrice d'observation O. Il importe, maintenant de déterminer explicitement la matrice $\mathbf{O}$ puisque que la factorisation de la matrice $\mathbf{H}$ n'est pas directement calculable en pratique. Il existe plusieurs approches basées en général sur les factorisations QR ou SVD. On choisit ici cette dernière factorisation pour ses bonnes performances en présence de bruit. Rappelons que la SVD de la matrice $\mathbf{H}$ est $\sum_{l=1}^{L}\lambda_l\mathbf{u}_l\mathbf{v}_l^H$ où $\{\lambda_m\}$, $\{\mathbf{u}_l\}$

et $\{\mathbf{v}_l\}$ sont respectivement les valeurs singulières, les vecteurs singuliers à gauche et les vecteurs singuliers à droite. On choisit alors $\mathbf{O} = \mathbf{U}\mathbf{T}_M = [\mathbf{u}_1 \dots \mathbf{u}_M]$ où $\mathbf{T}_M$ est une matrice de sélection des M premières colonnes de $\mathbf{U}$. Ce choix se justifie par le fait que la SVD est dite "révélatrice de rang", c'est à dire que si $\mathbf{H}$ est de rang plein alors aucune des L valeurs singulières ne sera nulle. A l'inverse, si cette même matrice présente une déficience en rang, par exemple $rg(\mathbf{H}) = M$ alors $L - M$ valeurs singulières seront nulles et on associe, alors, les M vecteur singuliers gauches dominants $\{\mathbf{u}_l\}$ au sous-espace signal [42]. On en conclut que $Im\,(\mathbf{O}\mathbf{T}_M) = Im\,([\mathbf{u}_1 \dots \mathbf{u}_M]) = \mathcal{S}$ et donc que $(\mathbf{u}_1, \dots \mathbf{u}_M)$ est une base orthonormale.

Algorithme de séparation de sous-espaces. Pour des signaux réels, tronquer la SVD aux M premières valeurs n'est pas toujours évident car si on note Γ l'ensemble des L valeurs singulières $\{\lambda_l\}$ alors Γ est composée d'un continuum de valeurs décroissantes. On n'a donc pas l'assurance, si on écarte les $L-M$ plus petites valeurs singulières que la troncature d'ordre M soit optimale. Notons $e_\Gamma = \sum_{l=L-M+1}^{L} \lambda_l^2$ la norme deux du vecteur composé par les $L-M$ plus petites valeurs singulières, on cherche alors à rendre e_Γ minimale sous contrainte que la matrice $\mathbf{H}$ soit de rang M et conserve une structure de Hankel. Notons que cette dernière condition est importante car c'est elle qui permet la factorisation de l'expression (3.26). Ce problème d'optimisation peut être réécrit de manière itérative selon

$$\min_{\mathbf{H}_{k+1}} \|\mathbf{H}_{k+1} - \mathbf{H}_k\|_F \quad \text{s. c.} \quad \mathbf{H} \text{ de Hankel et } rg\,(\mathbf{H}_{k+1}) = M \tag{3.28}$$

et $\mathbf{H}_0 = \mathbf{H}$ pour l'initialisation. Notons que minimiser la norme de Frobenius de la différence de deux matrices de Hankel, respectivement de rang L et M avec $M \leq L$, revient très exactement à minimiser e_Γ selon le théorème d'Eckart et Young. L'algorithme CPM de la référence [43] peut alors être formulé en introduisant l'opérateur de troncature selon $\mathcal{T}_M(\mathbf{H}) = \sum_{m=1}^{M} \lambda_l \mathbf{u}_l \mathbf{v}_l^H$ et de moyennage sur les anti-diagonales $\mathcal{M}(\mathbf{H}) = \mathbf{h}$ où h_l est le résultat du moyennage sur la l-ème anti-diagonale de $\mathbf{H}$, soit

$$\mathbf{H}_k = [\mathcal{H}_L \circ \mathcal{M} \circ \mathcal{T}_M]\,(\mathbf{H}_{k-1}) \tag{3.29}$$

Cette récurrence peut être exprimée en fonction de son premier terme selon

$$\mathbf{H}_k = [\mathcal{H}_L \circ \mathcal{M} \circ \mathcal{T}_M]^k\,(\mathbf{H}) \tag{3.30}$$

en posant arbitrairement $[\mathcal{H}_L \circ \mathcal{M} \circ \mathcal{T}_M]^0\,(\mathbf{H}) = \mathcal{T}_M\,(\mathbf{H})$.

Conclusion et choix de la méthode

Deux méthodes de détermination des paramètres de modèle ont été exposées. La première est essentiellement itérative et en ce sens elle est sous-optimale. Cependant elle permet de déterminer les paramètres de modèle pour un coût par itération peu élevé en $O(N \log_2 N)$ pour la FFT plus $O(N)$ pour la pseudo-inverse. De plus, les méthodes itératives présentent une robustesse accrue aux erreurs d'estimation car celles-ci sont en partie corrigées lors des itérations ultérieures. La méthode sous-espace présentée réalise une estimation conjointe des

M coefficients d'atténuation et des M pulsations puis une résolution conjointe en M amplitudes complexes. Cette méthode bénéficie en outre de la propriété dite de Haute-Résolution, c'est à dire qu'elle permet de séparer deux pics fréquentiels à distance relative inférieure à la résolution fréquentielle $O(1/N)$ et cela même sur un nombre restreint d'échantillons (segment d'analyse court). D'autre part, une analyse sur des fenêtres courtes est tout à fait possible dans une optique d'amélioration de la résolution temporelle [44, 45] alors qu'avec une méthode basée sur la FFT, la limitation de résolution fréquentielle, nous contraint à une taille d'analyse minimale. Cette propriété intéressante sera exploitée lors de la représentation des transitoires par l'algorithme DYN-EDS (c.f. section 4.1.1). En contre partie, cette méthode se révèle coûteuse puisque la complexité algorithmique est dominée par le coût de la SVD en $O(N^3)$ ou en $O(NM^2)$ si on utilise un algorithme rapide de calcul itératif de cette dernière.

3.4 Procédure d'appariement/interpolation généralisée

La notion d'appariement des paramètres de modèle le long de trajectoires inter-trames, fut initialement introduite par McAulay et Quatieri [7] dans le schéma sinusoïdal permettant d'assurer la continuité du signal, essentielle d'un point de vue psychoacoustique, au travers d'interpolations d'amplitude, de phase et de fréquence. Dans cette section, on propose des améliorations à cette technique qui permettent de prendre en compte le facteur d'amortissement du modèle EDS ce qui amène une meilleure synthèse des signaux non-stationnaires.

3.4.1 Tracking ou appariement

Il s'agit d'apparier les paramètres des trames successives dont les fréquences relatives sont endessous d'un seuil maximum de déviation, traçant ainsi des trajectoires de paramètres telles qu'illustrées en figure 3.17. A la trame l, un quadruplet de paramètres $\{a_{m_t}(l),d_{m_t}(l),\omega_{m_t}(l),\phi_{m_t}(l)\}$ est apparié à $\{a_{p_t}(l+1),d_{p_t}(l+1),\omega_{p_t}(l+1),\phi_{p_t}(l+1)\}$ au sein de la trajectoire t, si

$$|\omega_{m_t}(l) - \omega_{p_t}(l+1)| < \Delta\omega_{max}$$

et

$$\{m_t,p_t\} = \arg\min_{\{m,p\}} |\omega_m(l) - \omega_p(l+1)|.$$

$\Delta\omega_{max}$ doit dépendre de la résolution fréquentielle de l'estimateur spectral utilisé ainsi que de la résolution fréquentielle du système auditif humain. On utilise généralement $\Delta\omega_{max} = \frac{\omega_m(l)}{10}$ [7, 14]. De plus, d'importantes déviations des trajectoires d'amplitudes doivent être évitées. Si à l'inter-trame $l/l+1$, $|a_{m_t}(l)\exp(d_{m_t}(l)N) - a_{p_t}(l+1)| > \Delta A_{max}$, la trajectoire t sera à ce niveau décomposée en deux trajectoires, une première "s'éteignant" et une deuxième "naissant". La naissance de trajectoires a lieu chaque fois qu'un quadruplet de paramètres EDS $\{a_m(l),d_m(l),\omega_m(l),\phi_m(l)\}$ se retrouve non incorporé aux trajectoires existantes. Elle consiste à initier une nouvelle trajectoire à la trame $l-1$ en ajoutant une sinusoïde fenêtrée par une demie fenêtre de Hanning croissante de même taille que la durée d'analyse et telle que $\omega_i(l-1) = \omega_m(l)$ et $\phi_i(l-1) = \phi_m(l) - \omega_m(l)N$. Cela permet la continuité du signal

à la trame l. De la même façon, lorsque l'appariement n'est pas possible, on considère que la trajectoire doit être "éteinte ou mise en veille" à l'entrée de la trame l et sa fréquence courante garde sa même valeur avec une amplitude nulle. Une telle trajectoire sera "tuée" si son temps de veille dépasse une durée fixée [15]. Dans la suite, on examine la technique de synthèse associée. Rappelons qu'il s'agit d'une alternative à la synthèse OLA et que l'on se place dans le cas d'une analyse à fenêtres non recouvrantes.

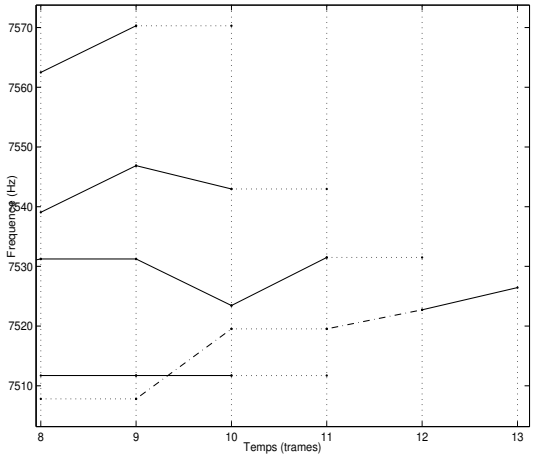

FIG. 3.17 – *Trajectoires de fréquences, (—) trajectoire continuée, (· · ·) trajectoire mise en veille, (-.-) trajectoire reprise après mise en veille*

3.4.2 Synthèse de la fréquence et de la phase

Il a été montré qu'il est possible, dans une perspective de compression, de ne pas prendre en compte les paramètres de phases initiales dans la synthèse de signaux stationnaires sans engendrer de dégradation de la qualité perceptuelle tant que la continuité de la phase instantanée est assurée [7, 15, 14] [1]. Le signal synthétique ne s'accorde pas alors nécessairement à l'original en terme de phase en ce sens qu'il peut être déphasé par rapport à ce dernier. Néanmoins, il est important de reconstruire une version "en phase" avec l'originale durant le processus de modélisation afin que la soustraction entre celles-ci ait un sens. A cette fin, on fait appel à l'interpolation polynômiale cubique [7]. Il n'en demeure pas moins qu'il est indispensable d'intégrer les paramètres de phases dans le cas d'attaques franches car les variations temporelles de la forme d'onde doivent alors être reproduites avec précision. On verra que l'interpolation cubique est mal adaptée à ce contexte puisqu'elle altère les caractéristiques temporelles du signal modélisé tout en créant du pré-écho. Commençons par décrire les techniques d'interpolation. A la trame l, le but est de reconstruire le signal

$$x_M(n,l) = \sum_{m=1}^{M} a_m(n,l) \cos\left(\phi(n,l)\right) \tag{3.31}$$

1. Ce type de synthèse est dit *phaseless synthesis*

où $\phi(n,l) = \omega_m(n,l)n + \phi_m(0,l)$ est la phase instantanée et $a_m(n,l) = e^{d_m(n,l)}$ est l'amplitude instantanée, pour $n = 0 \cdots N-1$. Pour alléger les notations on supprime dans la suite l'indice m en supposant un modèle d'ordre 1.

Reconstruction "phaseless"

Dans le cas de la reconstruction "phaseless", le signal synthétique est déphasé par rapport à l'original. L'information de phase n'est ni codée ni transmise mais la continuité de la phase instantanée doit être assurée. A la trame l, la pulsation instantanée $\omega(n,l)$ est obtenue par interpolation linéaire [7] des pulsations appariées $\bar{\omega}(l)$ et $\bar{\omega}(l-1)$ selon :

$$\omega(n,l) = \bar{\omega}(l-1) + \frac{\bar{\omega}(l) - \bar{\omega}(l-1)}{N} n \ , \ \ n = 0, \ldots, N-1 \tag{3.32}$$

La phase instantanée est alors la primitive de la fréquence instantanée vérifiant la condition aux limites $\phi(0,l) = \phi(N,l-1)$.

Reconstruction polynômiale cubique

L'interpolation des fréquences et des phases est assez délicate car ces interpolations doivent être faites conjointement. En effet, connaissant $[\bar{\omega}(l),\bar{\phi}(l)]$ et $[\bar{\omega}(l+1),\bar{\phi}(l+1)]$, on doit avoir

$$\begin{aligned}
\phi(0,l) &= \bar{\phi}(l) \\
\omega(0,l) &= \bar{\omega}(l) \\
\phi(0,l+1) &= \phi(N,l) = \bar{\phi}(l+1) \\
\omega(0,l+1) &= \omega(N,l) = \bar{\omega}(l+1).
\end{aligned}$$

Ces quatre conditions suggèrent le développement

$$\phi(n,l) = c_0(l) + c_1(l)n + c_2(l)n^2 + c_3(l)n^3$$

et par dérivation

$$f(n,l) = c_1(l) + 2c_2(l)n + 3c_3(l)n^2.$$

En remarquant que $\phi(N,l)$ est déterminé modulo 2π, on obtient les quatre équations

$$\begin{aligned}
c_0 &= \phi(0,l) = \bar{\phi}(l) \\
c_1 &= 2\pi f(0,l) = 2\pi \bar{f}(l) \\
c_0 + c_1 N + c_2 N^2 + c_3 N^3 &= \phi(N,l) + 2\pi Q = \bar{\phi}(l+1) + 2\pi Q \\
c_1 + 2c_2 N + 3c_3 N^2 &= 2\pi f(N,l) = 2\pi \bar{f}(l+1).
\end{aligned}$$

Les deux dernières équations s'écrivent

$$\begin{aligned}
c_2 N^2 + c_3 N^3 &= \bar{\phi}(l+1) - \bar{\phi}(l) - 2\pi \bar{f}(l) N + 2\pi Q = \alpha + 2\pi Q \\
2c_2 N + 3c_3 N^2 &= 2\pi[\bar{f}(l+1) - \bar{f}(l)] = \beta
\end{aligned}$$

avec $\alpha = \bar{\phi}(l+1) - \bar{\phi}(l) - \bar{\omega}(l+1)N$ et $\beta = \bar{\omega}(l+1) - \bar{\omega}(l)$ ce qui donne

$$\begin{aligned} c_2 &= \frac{3\alpha - \beta N + 6\pi Q}{N^2} \\ c_3 &= \frac{\beta N - 2\alpha - 4\pi Q}{N^3} \end{aligned}$$

Ces deux coefficients restent fonctions de Q. McAulay et Quatieri [7] proposent le critère

$$Q_{opt} = \arg\min_{Q} \int_0^{t_N} [\phi''(t,Q)]^2 dt.$$

où $\phi''(t,Q) = \frac{d^2\phi(t)}{dQ^2}$. La condition de minimalité s'écrit

$$\int_0^{t_N} (2c_2 + 6c_3 t)[\frac{12\pi}{(t_N)^2} - \frac{24\pi}{(t_N)^3}]dt = 0$$

soit

$$c_2(t_N)^2 + (3c_3 t_N - 2c_2)\frac{(t_N)^2}{2} - 6c_3\frac{(t_N)^3}{3} = 0.$$

On obtient finalement la condition qui consiste à choisir l'entier le plus proche de

$$Q_{opt} = \frac{\beta N - 2\alpha}{4\pi}.$$

3.4.3 Synthèse des amplitudes et des coefficients d'amortissement

On introduit une technique d'interpolation conjointe des paramètres d'amplitude et d'amortissement des trames successives conduisant à des représentations de meilleure qualité des signaux pseudo-stationnaires.

Le signal reconstruit doit vérifier les conditions aux limites suivantes :

$$\begin{aligned} a(N,l) &= a(0,l+1) \\ a(N+1,l) &= a(1,l+1) \end{aligned}$$

ce que l'on peut ré-écrire

$$\bar{a}(l)e^{d(l)N} = \bar{a}(l+1) \tag{3.33}$$

$$\bar{a}(l)e^{\bar{d}(l)(N+1)} = \bar{a}(l+1)e^{\bar{d}(l+1)}. \tag{3.34}$$

En se basant sur (3.33) et (3.34), on va déduire $d(n,l)$ sous la forme

$$d(n,l) = \delta_0(l) + \delta_1(l)n + \delta_2(l)n^2 + \delta_3(l)n^3. \tag{3.35}$$

Naturellement, $a(0,l) = \bar{a}(l)$, *i.e*, $d(0,l) = \ln(\bar{a}(l))$ d'où

$$\delta_0(l) = \ln(\bar{a}(l)). \tag{3.36}$$

En dérivant $\bar{d}(n,l)$, à l'instant $n = 0$, δ_1 peut être écrit

$$\delta_1(l) = \bar{d}(l). \tag{3.37}$$

La condition (3.33) implique

$$\delta_0(l) + \delta_1(l)N + \delta_2(l)N^2 + \delta_3(l)N^3 = \ln(\bar{a}(l+1)). \tag{3.38}$$

De (3.34) il vient

$$\delta_0(l) + \delta_1(l)(N+1) + \delta_2(l)(N+1)^2 + \delta_3(l)(N+1)^3 = \ln(\bar{a}(l+1)) + \bar{d}(l+1). \tag{3.39}$$

On peut alors déduire δ_2 et δ_3 :

$$\delta_2 = \frac{\bar{d}(l+1) - \bar{d}(l)}{2N+1} - [(N+1) + \frac{N^2}{2N+1}]\delta_3 \tag{3.40}$$

$$\delta_3 = \frac{1}{N+1}\left(\frac{\bar{d}(l)}{N} + \frac{\bar{d}(l+1)}{N+1}\right) + \frac{2N+1}{N^2(N+1)^2}\ln\left(\frac{\bar{a}(l)}{\bar{a}(l+1)}\right) \tag{3.41}$$

La figure 3.18 présente le résultat de la synthèse d'un partiel de trompette sur 10 trames de 1024 échantillons obtenue par interpolation cubique de la phase associée à une interpolation conjointe des amplitudes et des amortissements issus d'un modèle EDS en (a), et par une interpolation linéaire classique des amplitudes issus d'un modèle de Fourier [7] en (b). Notons que des variations plus précises de l'enveloppe sont permises avec notre technique. Les perfor-

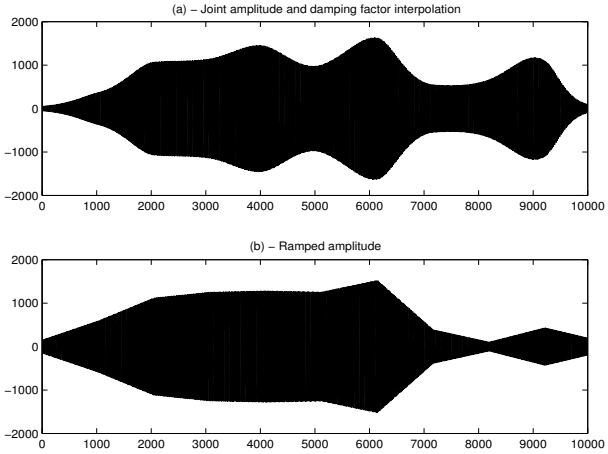

FIG. 3.18 – *Reconstruction de l'amplitude instantanée*

mances des deux approches sont comparées sur un segment de type transitoire doux de parole sur la figure 3.19 avec $M_{EDS} = 7$ et $M_{Fourier} = 10$. Notons que l'enveloppe du signal est globalement mieux reproduite avec l'interpolation conjointe comparativement à l'approche classique.

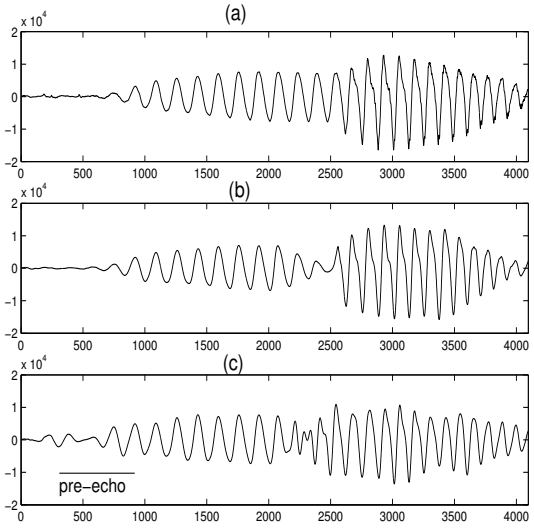

FIG. 3.19 – *Transitoire doux de parole - (a) Original, (b) EDS+interpolation conjointe des amplitudes et des amortissements, (c) Fourier+interpolation linéaire des amplitudes*

3.4.4 Synthèse du Signal

Une fois la reconstruction des paramètres effectuée, le signal peut être synthétisé selon (3.31). Notons cependant que l'interpolation cubique des phases altère la forme d'onde du signal synthétisé. Cela est observé alors même qu'une parfaite correspondance entre version originale et version modélisée est obtenue à l'intérieur des trames isolées. La raison est simple : la modulation des phases instantanées des différents partiels crée des phénomènes "d'interférence" entre les différents partiels ce qui déforme la forme d'onde résultante [46]. Cela est illustré sur la figure 3.20 présentant les résultats de synthèse de 200 ms de signal de parole par interpolation cubique des phases en (a), et à l'issue d'une analyse/synthèse OLA en (b). S'il est vrai que ce type de déformations est peu perceptible à l'écoute sur des signaux stationnaires en bande téléphonique, le phénomène devient fortement gênant pour la classe plus large de signaux pseudo-stationnaires et transitoires, et la qualité Hi-Fi qui est visée. Les tests d'écoute que nous avons effectués font apparaître un effet "tuyau" caractéristique de ce type de synthèse [46]. C'est la raison pour laquelle nous avons adopté une technique de synthèse OLA telle que décrite au chapitre 2 pour le système de codage.
Au chapitre suivant nous traitons le problème de la modélisation des signaux transitoires. Les limites du modèle EDS sont présentées et des solutions proposées pour aboutir une représentation de qualité de ce type de signaux.

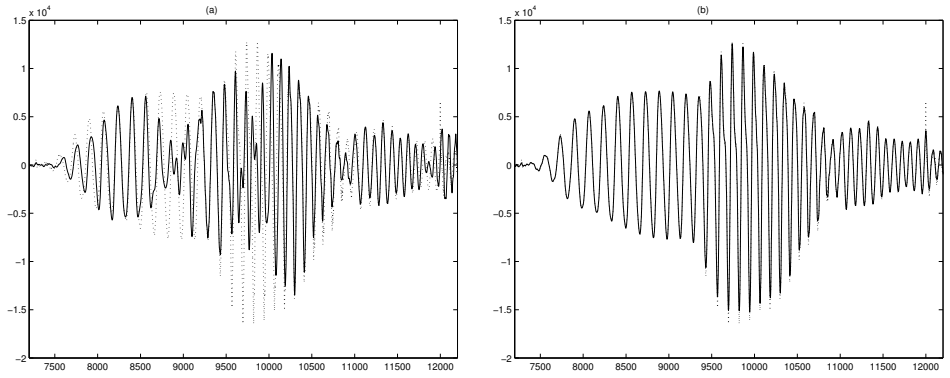

FIG. 3.20 – *Signal de parole modélisé par 20-EDS - original (· · ·) synthétique (—) , (a) synthèse par interpolation cubique de la phase, (b) synthèse OLA*

Chapitre 4

Modélisation des transitoires

Le modèle EDS présente de très bonnes performances sur la quasi-totalité des signaux. Cependant, dans un contexte fortement transitoire (typiquement percussif) une perte de qualité de modélisation est observée lorsque le début de l'attaque se situe loin du début du segment d'analyse. En effet, le modèle de EDS nécessaire pour atteindre une certaine qualité de modélisation devient non compacte [47, 48, 34]. Afin de mettre cela en évidence, on utilise un signal de castagnettes qui est souvent employé pour illustrer des phénomènes audio à caractère fortement transitoire [47, 49, 10]. La décroissance du Rapport Signal à Bruit (RSB_T) en fonction de la distance τ entre le début du transitoire et le début du segment de taille 512 échantillons (16 ms) est présentée à la figure 4.2 sur ce signal de castagnettes avec un ordre de modélisation de 35. Sur la figure 4.3, on relève les défauts du modèle EDS (les artefacts de modélisation) qui se matérialisent par deux aspects : la présence de pré-écho et la mauvaise restitution de la dynamique du signal. L'oreille étant très sensible à ce type de défaut, il est indispensable d'y remédier.

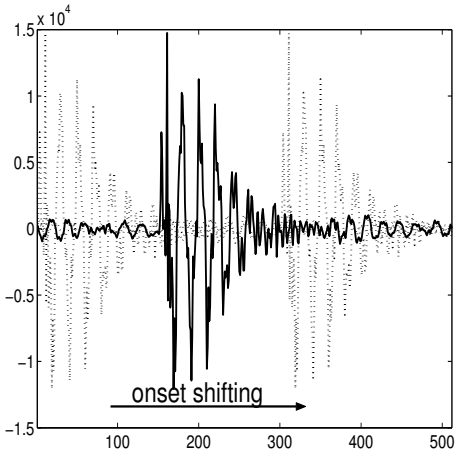

FIG. 4.1 – *Déplacement de l'attaque avec* $\tau = \{0{,}150{,}300\}$

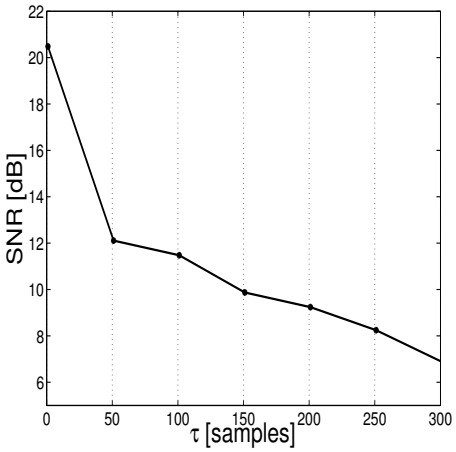

FIG. 4.2 – *RSB_T pour* $\tau = \{0, \ldots, 300\}$

Le phénomène de pré-écho peut être décrit par l'apparition d'une quantité d'énergie significative en amont de l'attaque, c'est à dire dans la région de faible énergie du signal, accompagnée

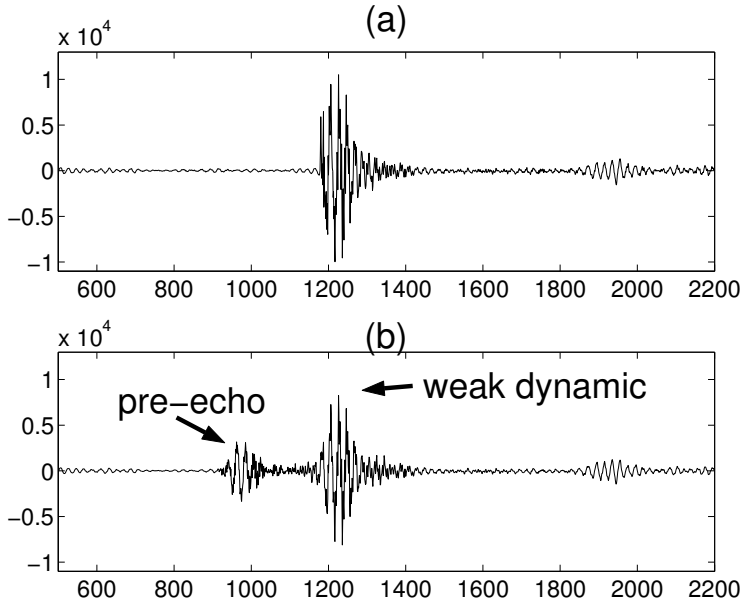

FIG. 4.3 – *Phénomène de pré-écho et mauvaise restitution de l'attaque pour un segment de castagnettes; (a) signal original ; (b) signal modélisé avec* $M = 35$

d'une mauvaise restitution de la dynamique du transitoire si bien que le caractère percussif est perdu à l'écoute. Nous avons développé des techniques permettant d'améliorer fortement le comportement du modèle EDS dans les conditions de transitions fortes que l'on présente dans la première partie de ce chapitre. Une deuxième partie sera consacrée à la mise en oeuvre du dictionnaire de Gabor dans la représentation des segments contenant l'attaque en mode de fonctionnement T du codeur.

4.1 Adaptation du modèle EDS au contexte fortement transitoire

Partant des limitations du modèle EDS dans la modélisation des signaux fortement transitoires, on propose deux méthodes de réduction de pré-écho. La première, nommée DYN-EDS [50], exploite un fenêtrage dynamique du signal afin de réduire la distance entre le début du transitoire et celui du segment d'analyse. Cette approche, inspirée des codeurs par transformé MPEG-AAC, est accompagnée d'une allocation dynamique de l'ordre de modélisation. La deuxième approche, exploite la représentation du signal dans un domaine transformé se prêtant mieux aux qualités intrinsèques du modèle EDS. Cette méthode, que l'on a baptisée FTA-EDS [48], permet d'annuler complètement le pré-echo.

4.1.1 Algorithme DYN-EDS

Un schéma en blocs du système de modélisation DYN-EDS est présenté à la figure 4.4.

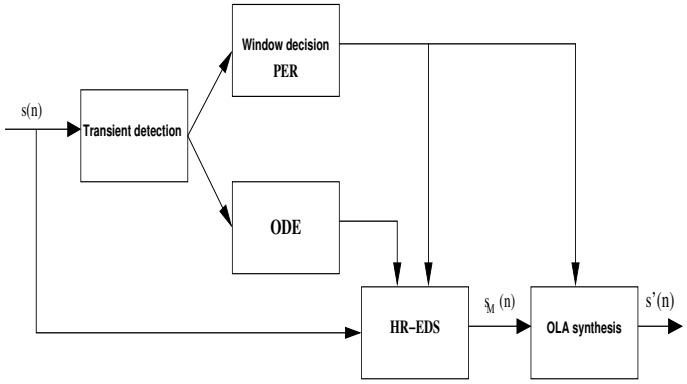

FIG. 4.4 – *Schéma en blocs du système DYN-EDS, PER : Pre-Echo Reduction, ODE : Onset Dynamic Enhancement*

Réduction du Pré-écho (PER)

On désigne la modélisation sur des fenêtres courtes, respectivement, longues par EDS(S), respectivement, EDS(L). Les fenêtres longues sont utilisées dans la modélisation des parties pseudo-stationnaires du signal. Notons qu'un faible recouvrement peut être utilisé entre les fenêtres longues sans pour autant dégrader les performances de modélisation puisque les signaux considérés sont parfaitement modélisés par EDS et que les discontinuités aux bords des fenêtres longues successives sont assez faibles. Les fenêtres courtes sont utilisées pour la modélisation des segments transitoires. Cependant, un recouvrement de 50% est maintenu entre ces dernières, ce qui permet d'assurer que l'attaque soit positionnée au début d'une fenêtre courte. Ainsi, le recouvrement entre fenêtres longues peut être identique à celui des fenêtres courtes ce qui évite l'utilisation de fenêtres de transitions et permet des implémentations simples et moins d'overhead. Lorsqu'un transitoire est détecté, on doit décider de passer à des fenêtres courtes ou non. En effet, si l'attaque est estimée à une distance inférieure à la longueur du recouvrement (c.f. figure 4.5-(α)), il n'est pas nécessaire de faire appel à des fenêtres courtes et EDS(L) reste approprié. Néanmoins, si l'attaque est détectée plus loin, la fenêtre longue est subdivisée en fenêtres courtes recouvrantes sur lesquelles le signal est modélisé (c.f. figure 4.5-(β)).

Restitution de la dynamique de l'attaque (ODE, Onset Dynamic Enhancement)

Un second degré de raffinement peut être atteint par le biais d'une allocation dynamique de l'ordre de modélisation autour de l'attaque. En effet, un transitoire type peut être décomposé en trois segments en terme de répartition énergétique : un premier segment de faible énergie, suivi d'un segment de forte dynamique qui représente l'attaque effective, et enfin un évanouissement de l'énergie. Il paraît alors naturel de faire varier l'ordre de modélisation en fonction du segment considéré. La stratégie d'allocation consiste à dédier la moitié des paramètres au segment d'attaque (2 fenêtres successives) et le reste aux autres segments.

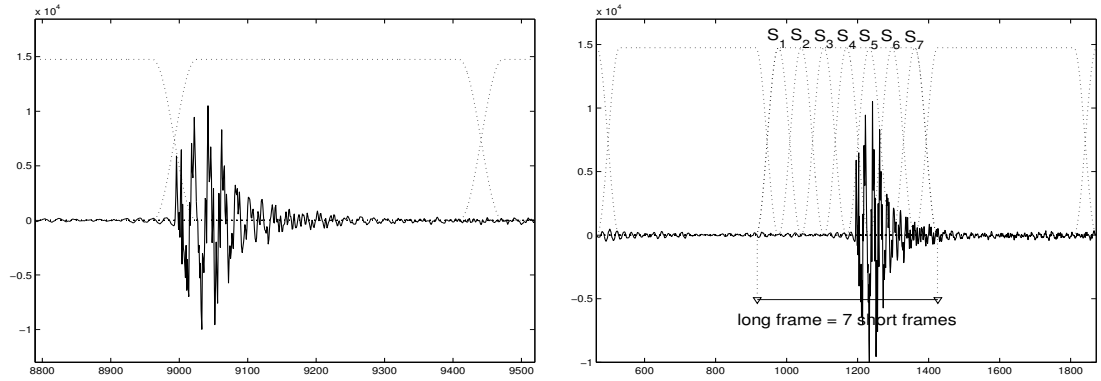

FIG. 4.5 – *(α) Fenêtres longues, (β) Fenêtrage dynamique du signal original*

Justification de l'approche HR

La méthode HR permet de produire une bonne estimation des paramètres sur un faible nombre d'échantillons (c.f. 3.3.2). L'utilisation de fenêtres courtes dans l'algorithme DYN-EDS est alors possible sans aucune perte de performance d'estimation tout en gardant une bonne résolution temporelle. Dans une analyse classique par FFT, la taille de la fenêtre d'analyse altère la résolution fréquentielle et une analyse par fenêtres courtes n'est pas suffisamment précise. Par exemple, sur des trames de 128 échantillons, la résolution n'est que de $f_e/N = 250Hz$. En plus, si les paramètres d'amortissement sont déduits par une méthode de FFTs décalées (c.f. IA-EDS) les erreurs d'estimation sur les paramètres de fréquence se propagent vers les estimations des $\{d_m\}$.

Simulation sur un transitoire audio type

Dans ce qui suit, la modélisation EDS par fenêtrage dynamique (EDS(L/S)) est comparée à la modélisation à taille de fenêtre fixe (EDS(L)). La longueur des fenêtres longues, respectivement, courtes, est 512 échantillons, respectivement 128 échantillons, *i.e.*, 16 et 2 ms pour $f_e = 32kHz$. Cela implique qu'une fenêtre longue est remplacée par 7 fenêtres courtes recouvrantes. Signalons que l'utilisation de fenêtres plus courtes conduirait à une représentation non compacte. Les ordres de modèle sont fixés à $M_L = 50$ pour EDS(L) et $M_S = 8 \simeq M_L/7$ pour EDS(S), assurant ainsi un nombre de paramètres constant sur des fenêtres de tailles différentes. Lorsqu'un transitoire est détecté et la décision d'utiliser des fenêtres courtes a été prise, il est possible de le localiser dans deux fenêtres d'analyse successives compte tenu de leur construction. Un ordre de modèle plus important est alors est utilisé sur ces deux fenêtres. Pour l'exemple présenté à la figure 4.6-(d), les ordres ont été choisis selon $M_{S_4} + M_{S_5} = [M_L/2]$. Le signal original, 17 ms d'un extrait de castagnettes, est présenté en figure 4.5 avec les deux possibilités de fenêtrage. Sur la figure 4.6-(b), le résultat de modélisation par EDS(L) sur des fenêtres à taille fixe est présenté. Le signal modélisé par fenêtrage dynamique est donné sur la figure 4.6-(c). Enfin, le résultat de modélisation par EDS(L/S) avec ODE (Onset Dynamic Enhancement) est présenté sur la figure 4.6-(d). On

peut facilement observer la forte présence de pré-écho sur la figure 4.6-(b), pré-écho qui est éliminé grâce à l'approche PER comme il peut être vu sur la figure 4.6-(c). En outre, on peut observé sur la figure 4.6-(d) qu'une meilleure restitution de la dynamique est obtenue par le biais de ODE. L'approche a ainsi été validée sur différents signaux transitoires grâce à des tests d'écoute.

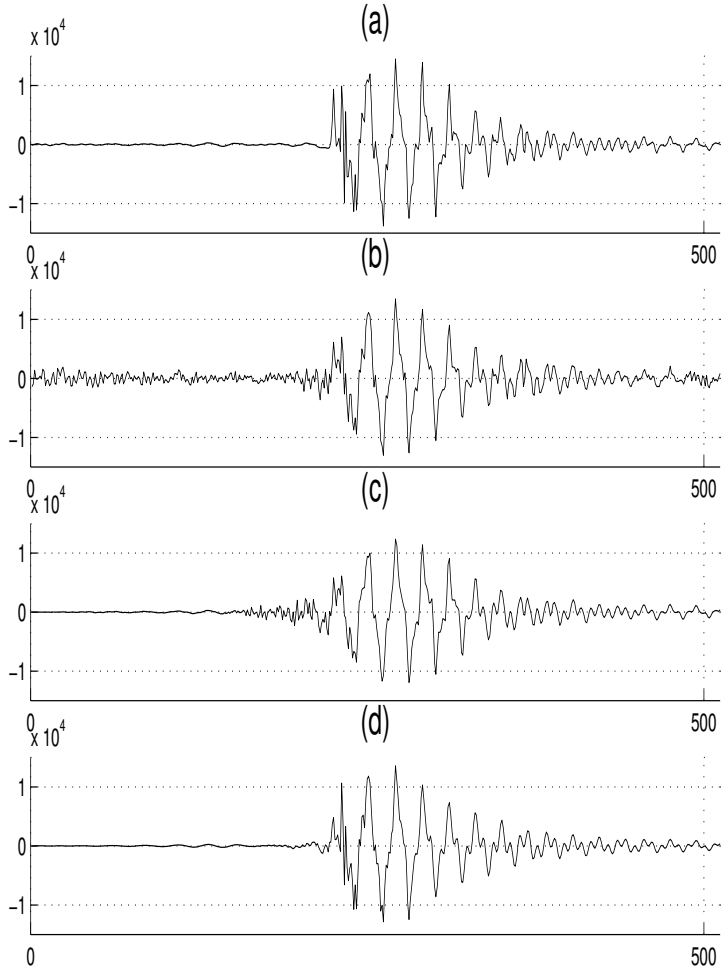

FIG. 4.6 – *(a) Signal original (b) EDS(L) (c) EDS(L/S) (d) EDS(L/S) avec ODE*

4.1.2 Algorithme FTA-EDS

Dualité entre transitoires et sinusoïdes

Les transformées fréquentielles telles que celle de Fourier en Cosinus Discrète (DCT-I-II-III-IV) ou encore en Sinus obéissent au principe de dualité temps/fréquence [20]. Une forme d'onde de type dirac dans le domaine temporel se transforme en forme d'onde oscillante dans le domaine fréquentiel et *vice versa*. Afin d'illustrer de principe, on considère la transformée de Fourier $X(\lambda)$ d'un transitoire fort synthétique, une Sinusoïde Amortie Retardée (DDS,

Damped & Delayed Sinusoid) [49] définie par

$$x(n) = ae^{i\phi}e^{(n-t)(i\omega+d)}\psi(n-t)$$

(c.f. figure 4.7-b), selon

$$X(\lambda) = ae^{i\phi}S(\lambda)e^{-i\lambda t},\ \lambda \in [0,\pi]$$

où

$$S(\lambda) = \frac{1-e^{(N-t)(i(\omega-\lambda)+d)}}{1-e^{i(\omega-\lambda)+d}}.$$

On note a,ϕ,ω,d,t, respectivement, l'amplitude, la phase, la pulsation, le facteur d'amortissement et le retard, et $\psi(n)$ la fonction de Heaviside, définie par $\psi(n) = 1$ pour $n \geq 0$ et 0 sinon. L'expression de $X(\lambda)$ indique qu'un retard t se traduit par un terme oscillant "$e^{-i\lambda t}$" dans le domaine fréquentiel. Ce résultat classique est illustré dans la figure 4.7.

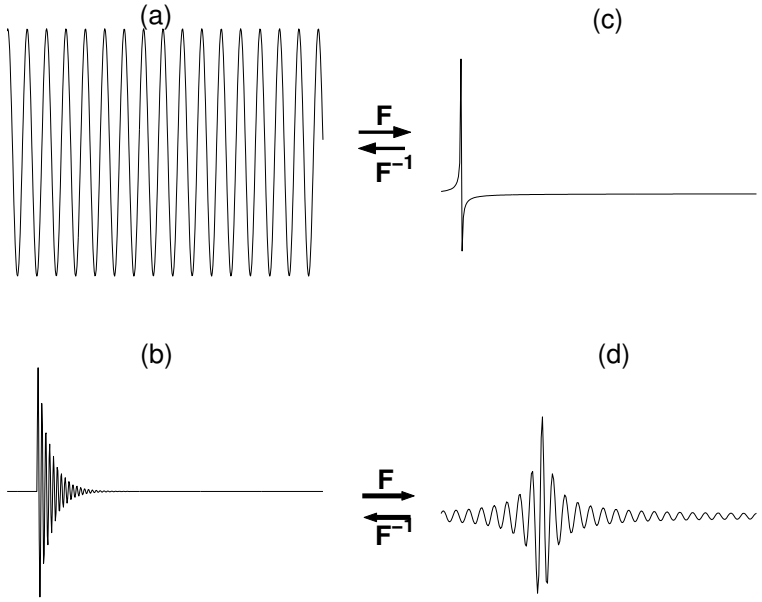

FIG. 4.7 – *Domaine temporel, (a) Sinusoïde (DDS avec $t = d = 0$) (b) DDS avec $t = 50$;(c) Domaine fréquentiel (d)* $\Re e\{X(\lambda)\}/2$

Ce principe en association avec le modèle EDS peut être mis à profit. En effet, un transitoire fort présente une singularité qui est similaire à celle du signal en figure 4.7-(b)-(d). Dans ce cas, une modélisation en temps par EDS du signal fournit des résultats peu satisfaisants. Par contre, comme le signal transformé en fréquence correspondant est essentiellement oscillant, il est avantageux de procéder à une modélisation en fréquence où le modèle EDS est plus performant. Le fait est que le signal de la figure 4.7-(d) est beaucoup mieux représenté par EDS comparativement au signal en 4.7-(b). Notons que cette approche est d'autant plus consistante que le terme $|dN|$ dans l'expression de $X(\lambda)$ est grand. On a alors $S(\lambda) \approx 1$. La situation idéale est réalisée quand la longueur d'analyse N est grande ou bien quand le signal s'évanouit de façon abrupte, *i.e.*, $|d|$ est grand. En d'autres termes, la longueur en temps du transitoire devrait être inférieure à la longueur d'analyse ce qui peut être réalisé par un choix

judicieux de N. Une estimation grossière du retard t est alors $\hat{t} = N\hat{\omega}/\pi$, où $\hat{\omega}$ est la valeur estimée de la pulsation ω du signal transformé.

Expression de la DCT-IV

Le signal transformé $\boldsymbol{X} \in \mathbb{R}^{N\times 1}$ est défini par

$$\boldsymbol{X} = \mathcal{F}_N(\boldsymbol{x}) = \boldsymbol{F}\boldsymbol{W}\boldsymbol{x} \tag{4.1}$$

où $\boldsymbol{W} = \text{diag}\{h(0),\ldots,h(N-1)\}$ et $\forall n,\ h(n) \neq 0$ est une fenêtre de pondération telle que $\boldsymbol{W}^{-1} = \text{diag}\{1/h(0),\ldots,1/h(N-1)\}$, *i.e*, $\boldsymbol{W}^{-1}\boldsymbol{W} = \boldsymbol{I}_N$. $\mathcal{F}_N(.) : \mathbb{R}^{N\times 1} \rightarrow \mathbb{R}^{N\times 1}$ est la DCT-IV fenêtrée. Notons que la matrice $N \times N$ $\boldsymbol{F} = (\vartheta_{nk})_{n,k}$, telle que $\vartheta_{nk} = \sqrt{2/N}\cos\left((n+0.5)(k+0.5)\pi/N\right)$, est réelle, symétrique et unitaire ($\boldsymbol{F}^{-1} = \boldsymbol{F} = \boldsymbol{F}^T$). Il vient alors $\boldsymbol{x} = \mathcal{F}_N^{-1} \circ \mathcal{F}_N(\boldsymbol{x})$. Le choix de la DCT-IV plutôt qu'une autre transformée fréquentielle se justifie par le fait qu'elle est bien adaptée aux signaux réels et ne demande aucune inversion.

Signal modélisé dans le domaine transformé

FTA-EDS désigne "Frequency Transform Algorithm with an EDS model", on en donne ici une description. En premier lieu, on calcule la transformée fréquentielle du signal transitoire $\boldsymbol{x}$ après fenêtrage. Ensuite, le signal est modélisé par EDS. En notant $\hat{\boldsymbol{X}}$ la version estimée du signal transformé $\boldsymbol{X}$, la transformation inverse et le fenêtrage sont enfin appliqués pour déduire le signal estimé en temps $\hat{\boldsymbol{x}}$ comme illustré sur la figure 4.8. La méthode basée sur HR-EDS (resp. IA-EDS) est résumée dans le tableau 4.1 (resp. tableau 4.2).

(1)	$\mathcal{H}_L(\boldsymbol{X})$	:=	$\mathcal{H}_L \circ \mathcal{F}_N(\boldsymbol{x})$
(2)	$\mathcal{H}_L(\boldsymbol{X})$	$\overset{\text{HR-EDS}}{\longrightarrow}$	$\hat{\boldsymbol{X}}$
(3)	$\hat{\boldsymbol{x}}$	:=	$\mathcal{F}_N^{-1}(\hat{\boldsymbol{X}})$

TAB. 4.1 – *FTA-EDS avec HR-EDS*

(1)	$\boldsymbol{X}$	:=	$\mathcal{F}_N(\boldsymbol{x})$
(2)	$\boldsymbol{X}$	$\overset{\text{IA-EDS}}{\longrightarrow}$	$\hat{\boldsymbol{X}}$
(3)	$\hat{\boldsymbol{x}}$	:=	$\mathcal{F}_N^{-1}(\hat{\boldsymbol{X}})$

TAB. 4.2 – *FTA-EDS avec IA-EDS*

Simulation sur extrait de castagnettes

On reprend le signal de castagnettes présenté précédemment. On utilise un ordre de modélisation $M = 40$ et une fenêtre de pondération $h(n)$ de Hamming.
Considérons les figures 4.9-(b)-(c)-(d). On relève l'absence totale de pré-écho et la bonne restitution de la dynamique avec FTA-EDS (c.f. figure 4.9-(d)) par rapport à l'algorithme de

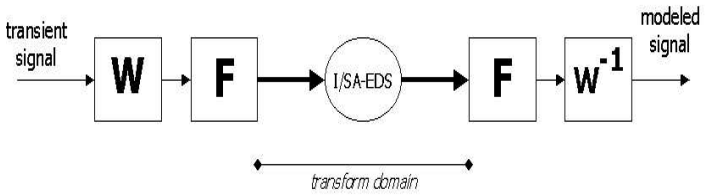

FIG. 4.8 – *Diagramme en blocs de FTA-EDS*

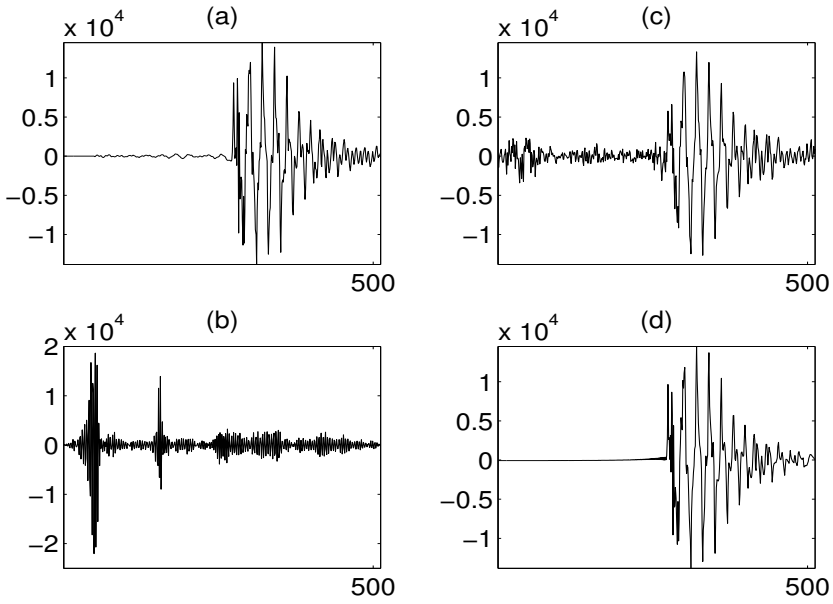

FIG. 4.9 – *Attaque de castagnettes, (a) signal original, (b) signal transformé (c) IA-EDS ($M = 40$), (d) FTA-EDS ($M = 40$)*

base (c.f. figure 4.9-(c)). La figure 4.10 permet de rendre compte du comportement des deux algorithmes. L'ordre de modélisation est progressivement augmenté, $M = \{2; 10; 20; 35\}$ en allant de haut en bas.

On note que l'algorithme de base modélise du signal sur toute la durée de la fenêtre d'analyse ce qui génère du pré-écho dès les premiers stades de modélisation (c.f. figure 4.10-(e)). Au contraire, FTA-EDS représente uniquement les parties consistantes du signal (c.f. figures 4.10-(a),(b),(c) et (d)).

Limites de l'algorithme proposé

Sur la figures 4.11, on montre la modélisation d'un segment oscillant de glockenspiel. Ce type de signal mêle des attaques franches à des parties très oscillantes. On remarque l'effondrement des performances de FTA-EDS sur ce segment oscillant du signal en temps se traduisant dans le domaine fréquentiel par des transitoires forts, lesquels sont mal représentés par EDS.

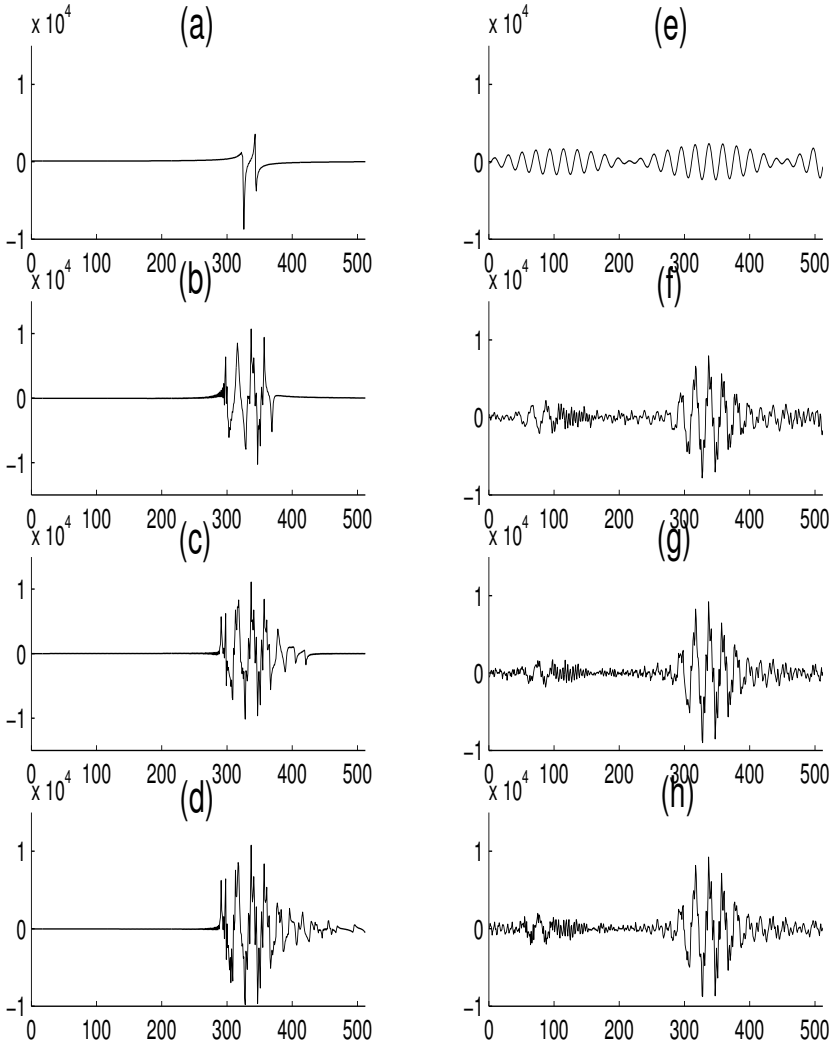

FIG. 4.10 – *Attaque de castagnettes, modélisation progressive ($M = 2; 10; 20; 35$), partie droite (a),(b),(c),(d) : FTA-EDS, partie gauche (e),(f),(g),(h) : IA-EDS*

4.1.3 Comparaison des méthodes et conclusion

Pour une meilleure modélisation des signaux fortement transitoires basée sur le modèle EDS, on a exposé deux approches. Sur les figures 4.12 et 4.13 sont représentés les modélisations temporelles et les spectrogrammes de 1.6 secondes de signal de castagnettes. Les figures 4.13-b et 4.12-b présentent la modélisation du signal original par 50-EDS. On peut remarquer sur la figures 4.12-b la mauvaise restitution de la dynamique du signal au niveau des attaques par rapport au signal original (c.f. figure 4.12-a). Sur la figure 4.13-b, on remarque que le support temporel des attaques est plus "large" que celui du signal original (c.f. figure 4.13-a). Ce phénomène caractérise le signal de pré-écho vu au travers d'une représentation Temps-Fréquence. Les figures 4.12-c et 4.13-c montrent la modélisation par la technique DYN-EDS du signal test, pour un ordre $M = 50$, on peut constater par rapport aux figures 4.12-b et 4.13-b, que les attaques sont plus nettes (pré-écho réduit) sur le spectrogramme et que le

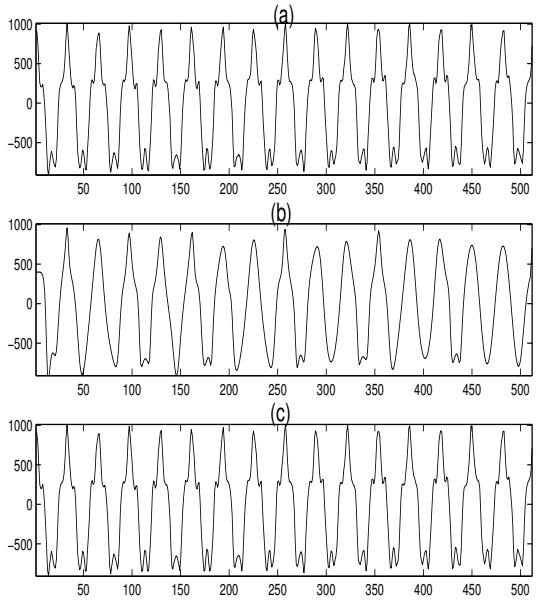

FIG. 4.11 – *Signal de glockenspiel, (a) signal original, (b) FTA-EDS ($M = 35$), (c) IA-EDS ($M = 35$)*

signal modélisé présente des dynamiques au niveau des attaques de castagnettes plus fidèles au signal original. Cette amélioration est fortement perceptible lorsque l'on se livre à l'écoute de cette séquence. Pour terminer, on a effectué les mêmes visualisations (forme temporelle et spectrogramme) pour l'approche FTA-EDS sur les figures 4.12-d et 4.13-d. On fixe l'ordre à $M = 50$. On peut remarquer que les attaques sont encore plus nettes sur la représentation Temps-Fréquence et que la dynamique des attaques est restituée avec encore plus de fidélité qu'avec la méthode DYN-EDS. On constate à l'écoute un progrès par rapport à l'approche DYN-EDS.

Ces résultats d'observation sont à mettre en corrélation avec les figures 4.14 et 4.15. Sur ces figures, on a représenté les résultats de puissances et de RSB_{TF} par sous-bandes obtenus en suivant la méthodologie de la section 3.2. Remarquons pour commencer que contrairement aux signaux de la section 3.2, la distribution de puissance (c.f. figure 4.14) est assez homogène. Cette observation corrobore le fait que les signaux transitoires possèdent, en général, une puissance relativement importante dans les hautes fréquences de par l'existence de brusques variations temporelles. Il s'en suit que nous ne négligeons pas cette fois les mesures de RSB_{TF} dans les bandes les plus hautes. Sur la figure 4.15, on peut voir que le modèle EDS marque le pas sur l'ensemble de la bande 0 à 16kHz. Notons aussi, pour ce modèle, l'effondrement des performances au niveau des puissances de sous-bandes pour $b > 23$, $\approx$ 12kHz (c.f. figure 4.14). Cela implique que ce modèle ne réussit pas à capturer les variations temporelles les plus rapides du signal test. La méthode DYN-EDS permet, à la fois de rehausser le niveau des puissances de sous-bandes (c.f. figure 4.14) et d'améliorer les résultats de RSB_{TF} dans les sous-bandes hautes par rapport au modèle EDS. Notons que dans les sous-bandes basses,

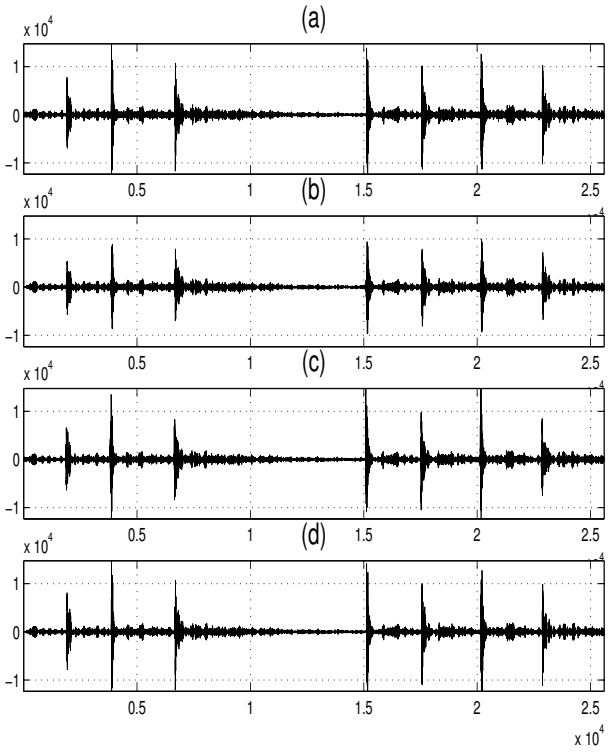

FIG. 4.12 – *Formes d'ondes temporelles, (a) Signal de castagnettes, (b) Signal modélisé par 50-EDS, (c) Signal modélisé par 50-DYN-EDS, (d) Signal modélisé par 50-FTA-EDS*

elle fait mieux que la méthode EDS et parfois que la méthode FTA-EDS. Enfin, la méthode FTA-EDS montre une courbe de puissance par sous-bande (c.f. figure 4.14) pratiquement confondue avec le celle du signal original et affiche les meilleurs résultats en terme de RSB_{TF} (c.f. bandes d'index 2 à 9 et 17 à 32 sur la figure 4.15).

On conclut de ce qui précède que la méthode FTA-EDS présente des performances supérieures aux méthodes DYN-EDS et EDS sur ce type de signaux. Cependant, ces performances s'effondrent dès que le transitoire est mêlé à du signal fortement oscillant ce qui est fortement pénalisant pour une application de codage s'adressant au signal audio dans toute sa diversité et sans hypothèses à priori sur le contenu. C'est la raison pour laquelle cet algorithme a été abandonné dans le cadre du système de codage. La deuxième approche (DYN-EDS) est simple dans son concept puisqu'elle étend au modèle paramétrique pseudo-stationnaire EDS les techniques de fenêtrage et d'allocation dynamique de débit des codeurs par transformée de la famille MPEG-AAC. Elle améliore nettement les résultats de modélisation EDS de base mais se trouve limitée par la difficulté d'automatiser la procédure de répartition de l'ordre du modèle pour les signaux les plus variés et les performances idéales sont atteintes pour des ordres de modélisation qui restent élevés pour le débit visé.

Dans la section suivante nous nous intéressons à la mise en oeuvre du modèle de Gabor pour la modélisation des transitoires. Rappelons que ce modèle présente un atout majeur puisqu'il

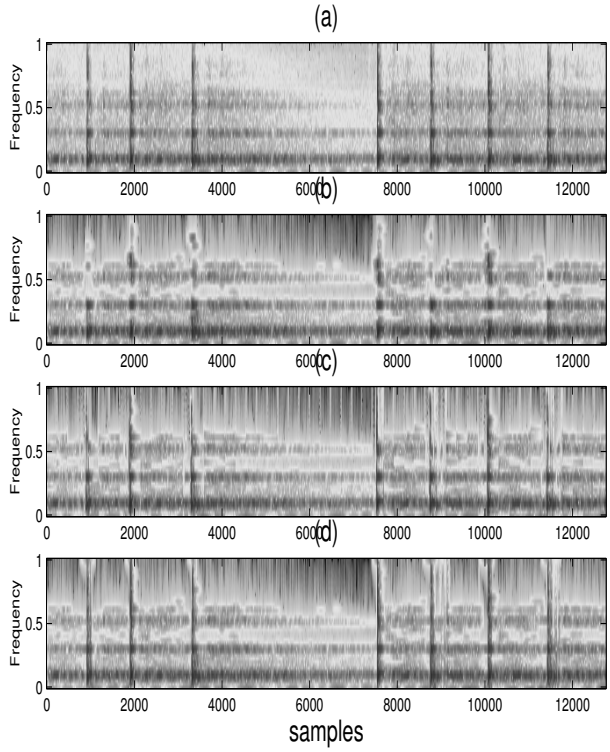

FIG. 4.13 – *Spectrogrammes, (a) Signal de castagnettes, (b) Signal modélisé par 50-EDS, (c) Signal modélisé par 50-DYN-EDS, (d) Signal modélisé par 50-FTA-EDS*

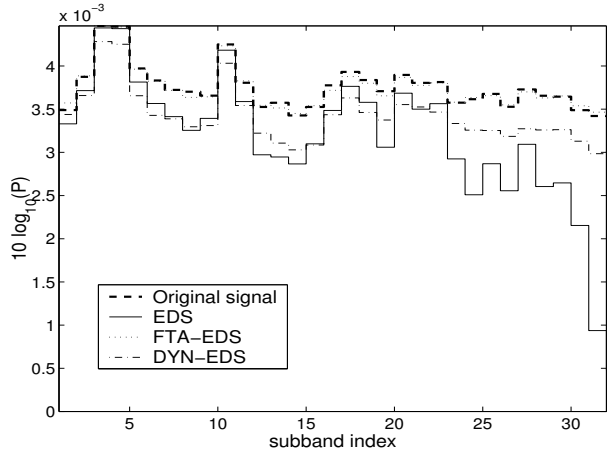

FIG. 4.14 – *Signal de castagnettes : puissances de sous-bandes* $P^{(r,b)}$

intègre un paramètre de décalage. Nous décrirons la stratégie d'expansion qui est utilisée et la façon dont on l'exploite dans le cadre du codeur.

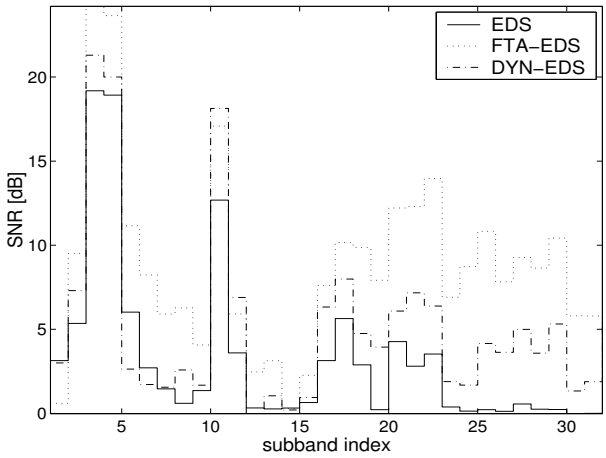

FIG. 4.15 – *Signal de castagnettes : $RSBs_{TF}^{(r,b)}$ de sous-bandes*

4.2 Utilisation du modèle de Gabor

On rappelle l'expression du modèle de Gabor donné au chapitre 1 :

$$x_M(n) = \sum_{m=1}^{M} \alpha_{\gamma_m} \, g_G \left(\frac{n - u_m}{s_m} \right) e^{i\omega_m n} \tag{4.2}$$

avec $\gamma_m = (s_m,u_m,\omega_m) \in \mathbb{R}^3$. Cette définition du modèle avec un paramètre de décalage rend impossible l'utilisation de méthodes Haute-Résolution. Il est alors nécessaire de procéder à une discrétisation de l'index γ pour ensuite exploiter les approches dites de décomposition atomique. Les paramètres sont échantillonnés selon [22]:

$$s = 2^j, 0 < j \leq \log_2 N \tag{4.3}$$

$$u = p2^{j-1}, 0 \leq p < \frac{N}{2^{j-1}} \tag{4.4}$$

$$\omega = \frac{k\pi}{2^j}, 0 \leq k < 2^{j+1} \tag{4.5}$$

Pour un signal de N échantillons, on considère donc $O(\log_2 N)$ échelles et $O(N)$ couples (u,ω) par échelle. Le dictionnaire de Gabor multi-échelle comprend donc $O(N \log_2 N)$ atomes. On note D le nombre total d'atomes du dictionnaire et on suppose ces atomes de norme unitaire.

Différentes méthodes de décomposition ont été proposées parmi lesquelles on peut citer le Best Basis Selection (BB) [28] et le Basis Pursuit (BP) [24] qui consistent à décomposer le signal sur une une base optimale au sens de la minimisation d'une fonction de coût, pour le BB et de la norme dans l^1 du vecteur formé par les coefficients de projection, pour le BP; cette base étant orthogonale dans le cas *BB*. L'approche que nous exploitons et qui a connu un large succès grâce au compromis qualité de décomposition/complexité algorithmique qu'elle réalise, est la poursuite adaptative ou Matching Pursuit.

4.2.1 Poursuite adaptative ou "Matching Pursuit" (MP)

Présentation générale

Le *Matching Pursuit* est un algorithme itératif introduit par Mallat et Zhang [22] qui réalise une décomposition du signal sur un ensemble d'atomes choisis dans un dictionnaire. Le principe a auparavant été utilisé en codage de parole, dans les codeurs CELP, pour la sélection des vecteurs d'excitation [51] , [52]. Le but est de calculer un développement linéaire de $\mathbf{s}$ de manière à l'approcher au mieux au sens de la minimisation de la norme de l'erreur résiduelle de modélisation. Connaissant les M vecteurs $\mathbf{g}_{\gamma_1},\ldots,\mathbf{g}_{\gamma_M}$ qui approximent au mieux le signal, il est facile de calculer les coefficients du développement. Ces coefficients ne sont autres que les coefficients de projection de $\mathbf{s}$ sur le sous-espace $vect\{\mathbf{g}_{\gamma_1},\ldots,\mathbf{g}_{\gamma_M}\}$. Malheureusement, les vecteurs optimaux sont à priori inconnus et doivent être déterminés simultanément avec les gains. L'algorithme optimal demanderait que l'on envisage toutes les combinaisons possibles de M vecteurs parmi D pour n'en retenir que celle qui fournit le critère minimal, ce qui impliquerait une complexité trop importante. On procède donc par approximations successives du signal $\mathbf{s}$, par projections orthogonales sur des éléments de $\mathbf{G}$.

Soit $\mathbf{g}_{\gamma_0} \in \mathbf{G}$, $\mathbf{s}$ peut être décomposé en :

$$\mathbf{s} =< \mathbf{s},\mathbf{g}_{\gamma_0} > \mathbf{g}_{\gamma_0} + \mathbf{r} \tag{4.6}$$

où $\mathbf{r}$ est le résidu après approximation de $\mathbf{s}$ dans la direction de $\mathbf{g}_{\gamma_0}$. Il est clair que $\mathbf{g}_{\gamma_0}$ est orthogonal à $\mathbf{r}$, d'où :

$$\|\mathbf{s}\|^2 = | < \mathbf{s},\mathbf{g}_{\gamma_0} > |^2 + \|\mathbf{r}\|^2 \tag{4.7}$$

Pour minimiser $\|\mathbf{r}\|$, il faut donc choisir $\mathbf{g}_{\gamma_0} \in \mathbf{G}$ tel que $| < \mathbf{s},\mathbf{g}_{\gamma_0} > |$ soit maximum.

Le MP est un algorithme itératif qui effectue une sous-décomposition du résidu $\mathbf{r}$ en le projetant sur un vecteur de $\mathbf{G}$ qui "match" le mieux $\mathbf{r}$ (comme cela a été fait pour $\mathbf{s}$). Cette procédure est ainsi répétée à chaque itération sur les résidus suivants. Partant des conditions initiales :

$$\begin{cases} \mathbf{s}^{(0)} = 0 \\ \mathbf{r}_0 = \mathbf{s} \end{cases} \tag{4.8}$$

on suppose que l'on a calculé le résidu d'ordre k $\mathbf{r}_k$, pour $k \geq 0$. On choisit un élément $\mathbf{g}_{\gamma_k} \in \mathbf{G}$ qui approxime au mieux le résidu $\mathbf{r}_k$ que l'on sous-décompose en :

$$\mathbf{r}_k =< \mathbf{r}_k,\mathbf{g}_{\gamma_k} > \mathbf{g}_{\gamma_k} + \mathbf{r}_{k+1} \tag{4.9}$$

ce qui définit le résidu d'ordre $k+1$. Puisque $\mathbf{r}_{k+1}$ est orthogonal à $\mathbf{g}_{\gamma_k}$

$$\|\mathbf{r}_k\|^2 = | < \mathbf{r}_k,\mathbf{g}_{\gamma_k} > |^2 + \|\mathbf{r}_{k+1}\|^2 \tag{4.10}$$

Poussons cette décomposition jusqu'à l'ordre m. $\mathbf{s}$ s'écrit alors sous la forme d'une somme

$$\mathbf{s} = \sum_{k=0}^{m-1} (\mathbf{r}_k - \mathbf{r}_{k+1}) + \mathbf{r}_m \tag{4.11}$$

et d'après (4.9)

$$\mathbf{s} = \sum_{k=0}^{m-1} < \mathbf{r}_k, \mathbf{g}_{\gamma_k} > \mathbf{g}_{\gamma_k} + \mathbf{r}_m \tag{4.12}$$

De même $\|\mathbf{s}\|^2$ est décomposé selon

$$\|\mathbf{s}\|^2 = \sum_{k=0}^{m-1} | < \mathbf{r}_k, \mathbf{g}_{\gamma_k} > |^2 + \|\mathbf{r}_m\|^2 \tag{4.13}$$

ce qui traduit la conservation de l'énergie de $\mathbf{s}$.

Le nombre final M d'itérations pour un signal donné $\mathbf{s}$ dépend de la précision ε désirée sur $\mathbf{s}$ et vérifie par suite :

$$\|\mathbf{r}_M\| = \|\mathbf{s} - \sum_{k=0}^{M-1} < \mathbf{r}_k, \mathbf{g}_{\gamma_k} > \mathbf{g}_{\gamma_k}\| \leq \varepsilon \|\mathbf{s}\| \tag{4.14}$$

ce qui compte tenu de (4.13) est équivalent à

$$\|\mathbf{s}\| - \sum_{k=0}^{M-1} | < \mathbf{r}_k, \mathbf{g}_{\gamma_k} > |^2 \leq \varepsilon^2 \|\mathbf{s}\| \tag{4.15}$$

Explicitons le choix de l'atome optimal à chaque itération. A la k-ème itération on sélectionnera :

$$\mathbf{g}_{\gamma_k} = \arg \min_{\mathbf{g}_\gamma \in \mathbf{G}} \|\mathbf{r}_{k+1}\|^2 \tag{4.16}$$

Afin d'alléger les notations on pose $\mathbf{g}_{\gamma_k} = \mathbf{g}_k$. Grâce au principe d'orthogonalité

$$< \mathbf{r}_{k+1}, \mathbf{g}_k > = < \mathbf{r}_k - \alpha_k \mathbf{g}_k, \mathbf{g}_k > = (\mathbf{r}_k - \alpha_k \mathbf{g}_k)^H \mathbf{g}_k = 0 \tag{4.17}$$

$$\Rightarrow \alpha_k = \frac{< \mathbf{g}_k, \mathbf{r}_k >}{< \mathbf{g}_k, \mathbf{g}_k >} = \frac{< \mathbf{g}_k, \mathbf{r}_k >}{\|\mathbf{g}_k\|^2} = < \mathbf{g}_k, \mathbf{r}_k > \tag{4.18}$$

où la dernière égalité est assurée par le fait que les atomes sont tous de norme unitaire. La norme de $\mathbf{r}_{k+1}$ peut alors être exprimée comme suit

$$\|\mathbf{r}_{k+1}\|^2 = \|\mathbf{r}_k\|^2 - \frac{| < \mathbf{g}_k, \mathbf{r}_k > |^2}{\|\mathbf{g}_k\|^2} = \|\mathbf{r}_k\|^2 - |\alpha_k|^2 \tag{4.19}$$

qui est minimisé en maximisant $|\alpha_k|^2 = | < \mathbf{g}_k, \mathbf{r}_k > |^2$. Cela équivaut simplement à choisir l'atome donnant la plus grande amplitude du coefficient de corrélation $|\alpha_k|$. Ainsi (4.16) peut être réécrit

$$\mathbf{g}_k = \arg \max_{\mathbf{g}_k \in \mathbf{G}} | < \mathbf{g}_k, \mathbf{r}_k > | \tag{4.20}$$

Notons que grâce à (4.19) la norme du résiduel décroît au fur et à mesure des itérations de l'algorithme, étant entendu que le dictionnaire est complet et tant qu'un modèle exact n'a pas été atteint.

Dans la décomposition du signal $\mathbf{s}$, le MP se poursuit jusqu'à ce que l'énergie du résiduel soit en-dessous d'un certain seuil ou jusqu'à ce que l'on atteigne une quelconque autre condition

d'arrêt jugée pertinente. Au bout de M itérations on obtient ainsi l'approximation compacte de $\mathbf{s}(n)$

$$s(n) \simeq s_M(n) = \sum_{k=1}^{M-1} \alpha_k g_k(n) \tag{4.21}$$

D'après (4.19) l'erreur quadratique moyenne d'un tel modèle tend vers zéro quand le nombre d'itérations devient important. Cette convergence implique que M itérations donnent un modèle d'ordre M acceptable, modèle qui n'est toutefois pas toujours optimal au sens d'un critère moindres carrés à cause de la contrainte itérative de l'algorithme.

Matching Pursuit en sous-espaces conjugués (MPsec) et décompositions de signaux réels

Quand le dictionnaire est constitué d'atomes complexes, les coefficients de corrélation sont complexes. Pour un signal réel, une décomposition sur des atomes complexes n'est pas envisageable. En effet, un tel signal n'est pas approximé dans le MP par des paires d'atomes conjugués, comme on pourrait l'espérer, parce qu'un atome et son conjugué ne sont pas orthogonaux. Il est donc nécessaire de considérer des développements des signaux réels en fonction d'atomes réels tels que

$$g_{(\gamma,\phi)}(n) = \frac{K_{(\gamma,\phi)}}{\sqrt{s}} g(\frac{n-u}{s}) \cos(\omega t + \phi) \tag{4.22}$$

où la constante $K_{(\gamma,\phi)}$ est ajustée de manière à garder $\|\mathbf{g}_{\gamma,\phi}\| = 1$.

La phase ϕ qui était cachée dans les nombres complexes apparaît maintenant explicitement comme étant un paramètre des atomes réels. Dans le cas complexe les atomes ne sont indexés que par trois paramètres (s,u,ω) et la phase d'un atome dans le développement est donnée par sa corrélation. Par contre, un dictionnaire réel requière la phase ϕ comme index supplémentaire. ϕ n'est pas fournie par le calcul de la corrélation, comme dans le cas complexe; elle doit être discrétisée et incorporée comme paramètre du dictionnaire dans la poursuite de façon analogue aux autres paramètres. Cela se traduit par un dictionnaire de dimension plus importante et donc par une recherche plus complexe dans le dictionnaire.

Ces nombreux inconvénients peuvent être dépassés en utilisant un dictionnaire complexe et en effectuant une poursuite en sous-espaces conjugués [25]. Nous commençons par présenter la poursuite en sous-espaces puis nous traitons le cas des sous-espaces conjugués qui nous intéresse le plus.

▷ *Poursuite en sous-espaces*

Il s'agit de trouver, à chaque itération du MP, un sous-espace de dimension assez petite et qui soit optimal en ce sens qu'il possède une structure susceptible de simplifier la poursuite. A la k-ème itération, on cherche donc à trouver une matrice $(\mathbf{G}_{\gamma_l})_{(N\times R)}$ dont les R colonnes

sont des vecteurs du dictionnaire qui minimisent le résiduel $\mathbf{r}_{k+1} = \mathbf{r}_k - \mathbf{G}_{\gamma_l}\alpha_l$, où α_l est maintenant un vecteur $R \times 1$ de poids.

La formulation en dimension R est alors similaire à la formulation mono-dimensionnelle. La contrainte d'orthogonalité $< \mathbf{r}_k - \mathbf{G}_{\gamma_l}\alpha_l, \mathbf{G}_{\gamma_l} >= 0$ implique une solution pour les poids

$$\boldsymbol{\alpha} = (\mathbf{G}_{\gamma_l}^H \mathbf{G}_{\gamma_l})^{-1} \mathbf{G}_{\gamma_l}^H \mathbf{r}_k \tag{4.23}$$

L'énergie du résiduel est donnée par

$$< \mathbf{r}_{k+1}, \mathbf{r}_{k+1} >=< \mathbf{r}_k, \mathbf{r}_k > - \mathbf{r}_k^H \mathbf{G}_{\gamma_l} (\mathbf{G}_{\gamma_l}^H \mathbf{G}_{\gamma_l})^{-1} \mathbf{G}_{\gamma_l}^H \mathbf{r}_k \tag{4.24}$$

qui est minimisée en choisissant $\mathbf{G}_{\gamma_l}$ de façon à maximiser le second terme. Cette approche est clairement coûteuse sauf si $\mathbf{G}_{\gamma_l}$ est formée par des vecteurs orthogonaux ou si elle possède une autre structure spéciale.

▷ *Cas des sous-espaces conjugués*

On considère le sous-espace de dimension 2 engendré par un atome et son conjugué. Les deux colonnes de $\mathbf{G}_{\gamma_l}$ sont ici simplement un atome $\mathbf{g}$ et son conjugué $\mathbf{g}^*$. On a alors :

$$\begin{aligned} r_{k+1}(n) &= r_k(n) - \alpha_k(1) g_k(n) - \alpha_k(2) g_k^*(n) \\ &= r_k(n) - \alpha_k(1) g_k(n) - \alpha_k^*(1) g_k^*(n) \\ &= r_k(n) - 2\Re(\alpha_k(1) g_k(n)) \end{aligned} \tag{4.25}$$

Avec cette démarche, on obtient des décompositions du signal de la forme

$$\mathbf{s} \simeq 2 \sum_{k=1}^{K} \Re(\alpha_k(1) \mathbf{g}_k) \tag{4.26}$$

Cette méthode fournit donc des décompositions réelles de signaux réels à l'aide d'un dictionnaire formé d'atomes complexes. Signalons que dans le cas de dictionnaires constitués à la fois d'atomes complexes et d'atomes purement réels, les atomes réels doivent être traités indépendamment des différents sous-espaces conjugués puisque la formulation ci-dessus n'est plus valide, étant donné qu'alors, $\mathbf{g}$ et $\mathbf{g}^*$ sont linéairement dépendants et qu'ils n'engendrent pas de sous-espaces de dimension 2.

Réduction de la complexité du MP, Matching Pursuit Rapide

L'algorithme MP lui-même est peu coûteux en termes d'opérations élémentaires mais la complexité globale est déterminée essentiellement par le nombre d'atomes présents dans le dictionnaire. Cette complexité devient très contraignante lorsque l'on utilise des tailles de fenêtres supérieures à 512 échantillons : un dictionnaire de Gabor discrétisé selon 4.3, 4.4 et 4.5 prévu pour des fenêtres de 1024 échantillons, contient 20480 atomes, ce qui augmente considérablement le nombre de projections à effectuer. La complexité algorithmique de M

Algorithme 1: MPsec
Entrée: $\mathbf{x}$, $\mathbf{G}$, M, $\mathbf{x}_M := 0$, $\mathbf{r}_0 := 0$
Sortie: $\mathbf{x}_M$

(1) **foreach** $k = 1 : M$

(2) $\quad \mathbf{g}_{\gamma_k} := \arg\max_{\gamma\in\Gamma} \left| \frac{<\mathbf{g}_\gamma,\mathbf{r}_k> - <\mathbf{g}_\gamma,\mathbf{g}_\gamma^*><\mathbf{g}_\gamma,\mathbf{r}_k>^*}{1-|<\mathbf{g}_\gamma,\mathbf{g}_\gamma^*>|^2} \right|$

(3) $\quad \alpha_k(1) := \frac{<\mathbf{g}_{\gamma_k},\mathbf{r}_k> - <\mathbf{g}_{\gamma_k},\mathbf{g}_{\gamma_k}^*><\mathbf{g}_{\gamma_k},\mathbf{r}_k>^*}{1-|<\mathbf{g}_{\gamma_k},\mathbf{g}_{\gamma_k}^*>|^2}$

(4) $\quad \mathbf{r}_k := \mathbf{r}_k - 2\Re(\alpha_k(1)\mathbf{g}_{\gamma_k})$

(5) $\quad \mathbf{x}_M := \mathbf{x}_M + 2\Re(\alpha_k(1)\mathbf{g}_{\gamma_k})$

itérations de MP est en fait $O(MN\log_2^2(N))$; elle est dominée par le coût du calcul des produits scalaires avec les atomes complexes. Des algorithmes rapides existent et permettent de ramener cette complexité à $O(MN)$. On trouvera dans [26] une description détaillée des techniques employées donnant lieu à l'algorithme Matching Pursuit Rapide qui est exploité au sein de notre codeur. Signalons juste que l'on met à profit la Transformée de Fourier Rapide pour calculer les produits scalaires; en outre, on considère, lors de la poursuite, des sous-dictionnaires de maxima locaux, réduisant ainsi l'espace de recherche à chaque itération. On peut également utiliser une formule de mise à jour déduite de (4.9)

$$< \mathbf{g},\mathbf{r}_{k+1} >=< \mathbf{g},\mathbf{r}_k > -\alpha_k < \mathbf{g},\mathbf{g}_k > \tag{4.27}$$

où le seul terme à calculer pour la mise à jour de la corrélation $< \mathbf{g},\mathbf{r}_{k+1} >$ est $< \mathbf{g},\mathbf{g}_k >$ que l'on peut pré-calculer et stocker pour toutes les valeurs de k.

Matching Pursuit Orthogonal OMP

Le Matching Pursuit Orthogonal [22] [1] permet d'assurer que le résiduel $\mathbf{r}_k$ est orthogonal à tous les $k-1$ vecteurs du dictionnaire déjà sélectionnés dans le modèle et doit ainsi assurer la convergence de l'algorithme [2] en au plus D itérations. L'algorithme initial est modifié comme suit : on orthonormalise à l'itération k le vecteur $\mathbf{g}_{\gamma_k}$ par rapport aux $k-1$ vecteurs $\mathbf{g}_{\gamma_1},\ldots,\mathbf{g}_{\gamma_{k-1}}$ déjà sélectionnés, par le procédé de Gram-Schmidt, soit

$$\mathbf{g}_{\gamma_k}^{\perp} = \frac{\mathbf{g}_{\gamma_k} - \mathbf{P}_{\mathcal{V}_{k-1}}\mathbf{g}_{\gamma_k}}{\left\|\mathbf{g}_{\gamma_k} - \mathbf{P}_{\mathcal{V}_{k-1}}\mathbf{g}_{\gamma_k}\right\|_2} \tag{4.28}$$

où $\mathbf{P}_{\mathcal{V}_{k-1}}$ est le projecteur orthogonal sur le sous-espace $\mathcal{V}_{k-1} = vect\{\mathbf{g}_{\gamma_1},\ldots,\mathbf{g}_{\gamma_{k-1}}\}$ et l'on re-projette $\mathbf{r}_k$ sur $\mathbf{g}_{\gamma_k}^{\perp}$.

4.2.2 Modélisation du signal audio par MP

Simulation sur un segment audio de trompette

Considérons la poursuite présentée à la figure 4.16. Il s'agit de 256 échantillons d'un signal de trompette échantillonné à 32kHz modélisé à l'aide d'un dictionnaire formé d'atomes de

1. là encore le principe a été utilisé en codage de parole [52] , [33]

2. En fait, le MP nécessite une infinité d'itérations pour reconstruire $\mathbf{s}$ parfaitement ; OMP permet de s'assurer que la poursuite cesse après un nombre fini d'étapes [26]

Algorithme 2: OMPsec
Entrée: $\mathbf{x}$, $\mathbf{G}$, M, $\mathbf{x}_M := 0$, $\mathbf{r}_0 := 0$
Sortie: $\mathbf{x}_M$

(1) **foreach** $k = 1 : M$

(2) $\quad \mathbf{g}_{\gamma_k} := \arg\max_{\gamma\in\Gamma} \left| \frac{<\mathbf{g}_\gamma,\mathbf{r}_k> - <\mathbf{g}_\gamma,\mathbf{g}_\gamma^*><\mathbf{g}_\gamma,\mathbf{r}_k>^*}{1-|<\mathbf{g}_\gamma,\mathbf{g}_\gamma^*>|^2} \right|$

(3) $\quad \mathbf{g}_{\gamma_k}^{\perp} := \frac{\mathbf{g}_{\gamma_k} - \mathbf{P}_{\mathcal{V}_{k-1}}\mathbf{g}_{\gamma_k}}{\left\|\mathbf{g}_{\gamma_k} - \mathbf{P}_{\mathcal{V}_{k-1}}\mathbf{g}_{\gamma_k}\right\|_2}$

(4) $\quad \alpha_k(1) := \frac{<\mathbf{g}_{\gamma_k}^{\perp},\mathbf{r}_k> - <\mathbf{g}_{\gamma_k}^{\perp},\mathbf{g}_{\gamma_k}^{\perp *}><\mathbf{g}_{\gamma_k}^{\perp},\mathbf{r}_k>^*}{1-|<\mathbf{g}_{\gamma_k}^{\perp},\mathbf{g}_{\gamma_k}^{*\perp}>|^2}$

(5) $\quad \mathbf{r}_k := \mathbf{r}_k - 2\Re(\alpha_k(1)\mathbf{g}_{\gamma_k})$

(6) $\quad \mathbf{x}_M := \mathbf{x}_M + 2\Re(\alpha_k(1)\mathbf{g}_{\gamma_k})$

Gabor complexes et d'exponentielles complexes par une approche sous-espaces conjugués. Les approximations successives des résiduels à différentes étapes de la poursuite sont illustrées. A gauche sont donnés, le résiduel à l'itération k en gris, son approximation en noir et le résiduel qui s'en suit à l'itération $k+1$ en dessous . A droite, le signal original est présenté avec le signal synthétique en cours de construction. L'approximation devient plus précise au fil des itérations. Aux premiers stades de la poursuite, l'algorithme parvient à une estimation de la forme et de l'énergie globale du signal par des atomes d'échelle importante. Il en résulte une forte atténuation de l'énergie du résiduel. Aux stades suivants, des atomes de petite échelle (de support plus réduit) sont sélectionnés afin de raffiner l'estimation et de "matcher" les détails du signal. Le MP vu dans le domaine spectral est illustré à la figure 4.17 : à droite le spectre de puissance en dB du signal original avec celui du signal synthétique en cours de construction, à gauche celui du résiduel à l'itération k et celui correspondant à l'atome sélectionné. Le MP commence par atténuer les pics les plus énergétiques et finit par ceux d'énergie plus faible. La figure 4.18 montre la décroissance de la puissance du résiduel en fonction de l'avancement de l'algorithme.

Discussion de la méthode

L'algorithme MP est un algorithme qui effectue la décomposition d'un problème d'optimisation de dimension M en M problèmes d'optimisation de dimension 1. Cette approche est, donc, sous-optimale par essence mais présente l'intérêt de ne faire aucune hypothèse sur le signal étudié et permet d'utiliser une grande variété de formes d'onde pour composer son dictionnaire. De plus, comme toutes les méthodes travaillant en temps, elle ne soufre pas du problème de résolution fréquentielle d'une méthode d'analyse spectrale. Cependant, le pas de discrétisation des paramètres dans le dictionnaire conditionne ses performances. On peut souligner la capacité de cet algorithme à corriger aux itérations avancées les artefacts de modélisation créés à une itération courante. Cette propriété procure au MP une bonne robustesse lors de la modélisation de signaux de natures très diverses. Notons enfin que l'utilisation de l'OMP permet d'accélérer la vitesse de convergence de la norme du résiduel [22] mais ne se justifie que pour des dictionnaires sur-complets [3] puisque le fait de contenir des atomes

3. On définit un tel dictionnaire par la construction d'une famille de vecteurs génératrice mais non-libre.

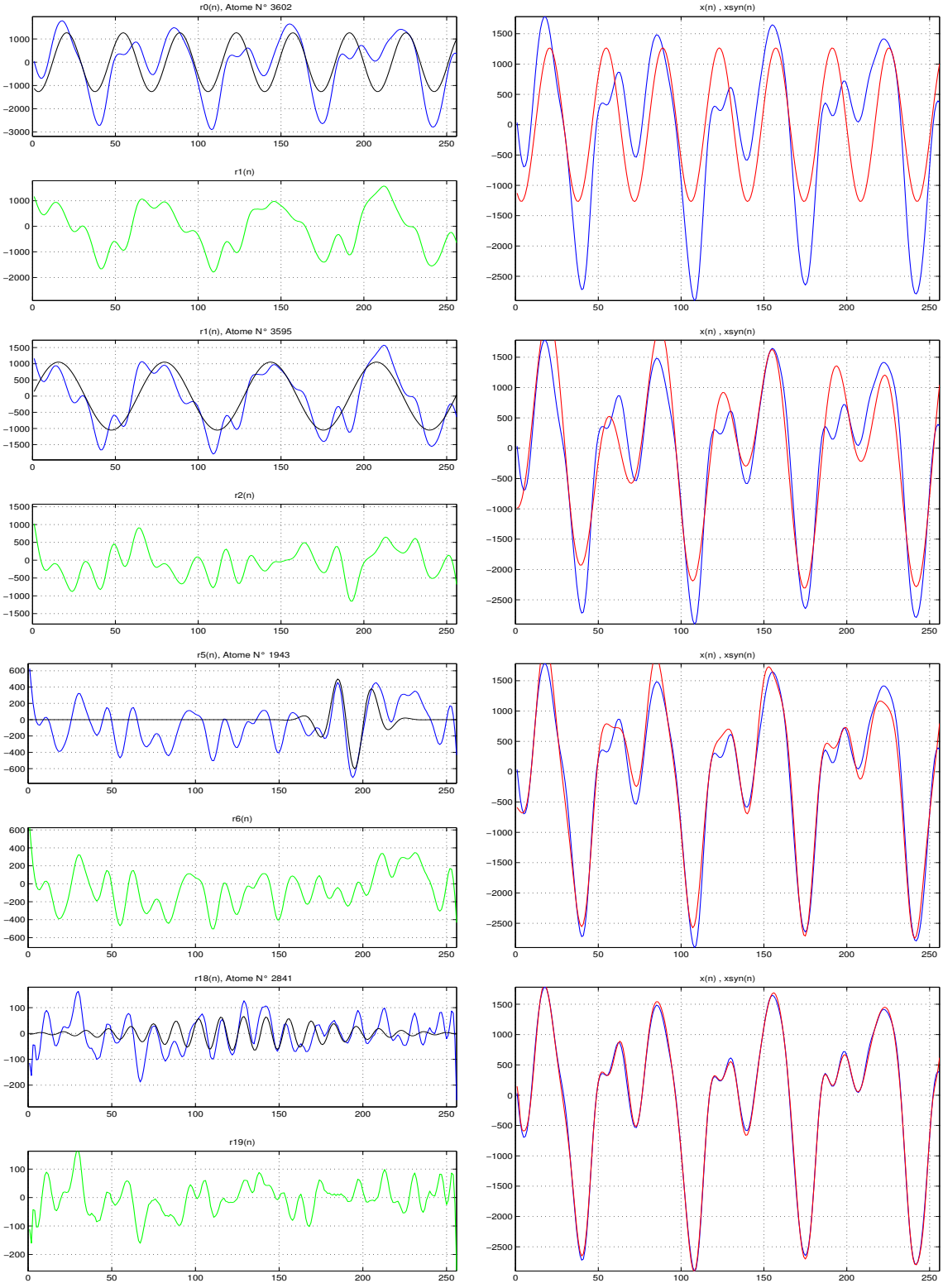

FIG. 4.16 – *Modélisation d'un signal de trompette par atomes de Gabor et exponentielles complexes.*

s'exprimant comme combinaison linéaire d'autres atomes du même dictionnaire implique une perte d'orthogonalité au sein de cette famille et nécessite donc, lors de l'extraction de la base

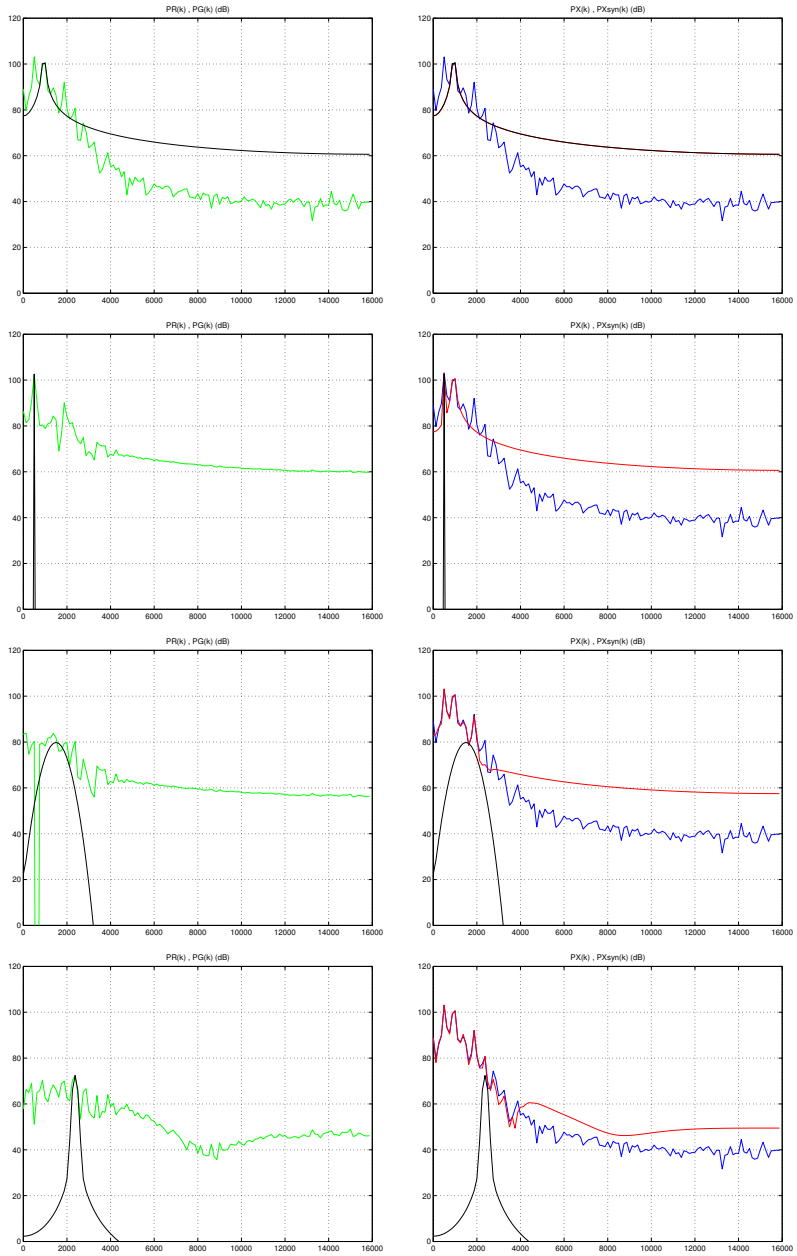

FIG. 4.17 – *MP vu dans le domaine spectral*

du sous-espace signal, une ré-orthogonalisation de celle-ci.

On expose dans [53, 48] une étude comparative des performances du MP comparativement à une approche HR. On retiendra qu'une méthode HR présente intrinsèquement de meilleures performances de modélisation. Cependant, s'agissant de modéliser des signaux fortement transitoires, le modèle de Gabor impose sa supériorité sur les modèles se prêtant aux approches HR (c'est-à-dire les modèles Sinusoïdal et EDS), d'où la nécessité de recourir au MP. En effet, on montre qu'à *débit égal*, MP-Gabor[4] permet une modélisation du signal audio fortement

4. MP-Gabor désigne l'utilisation d'un modèle de Gabor en association avec le MP

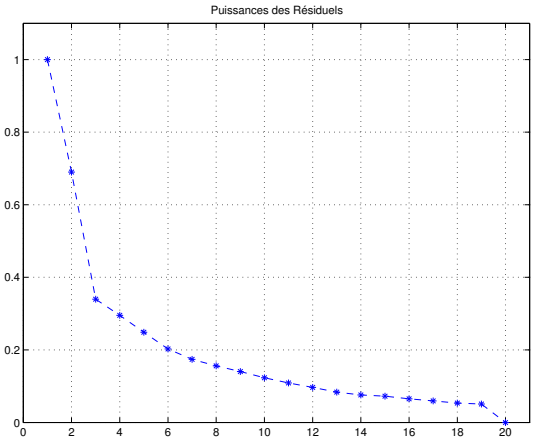

FIG. 4.18 – *Puissances des résiduels successifs normalisés par la puissance du signal original*

transitoire de qualité supérieure à celle obtenue par EDS et DYN-EDS.

4.2.3 MP-Gabor Vs DYN-EDS

On présente sur les figure 4.19 et 4.20 les résultats de modélisation d'un extrait de castagnettes par MP-Gabor en (b) et par DYN-EDS en (c) à nombre de paramètres équivalent. On voit que les attaques sont parfaitement reproduites avec MP-Gabor qui présente des performances supérieures à DYN-EDS péchant par une moins bonne restitution des dynamiques des attaques. Afin de s'en convaincre considérons les puissances par sous-bandes de l'original

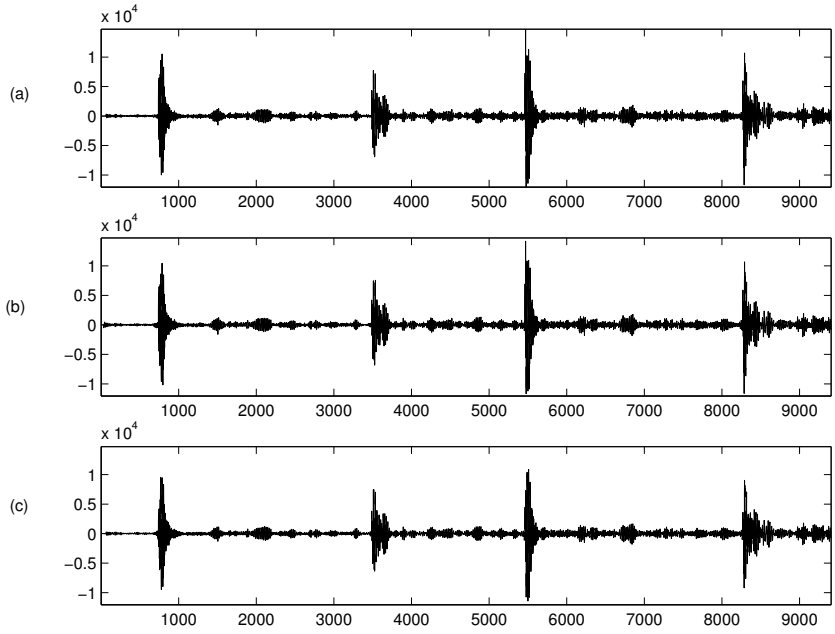

FIG. 4.19 – *Signal de castagnettes, (a) original, (b) modélisé par MP-Gabor, (c) modélisé par DYN-EDS*

avec les versions modélisées ainsi que les RSB_{TF} obtenus par MP-Gabor comparativement à

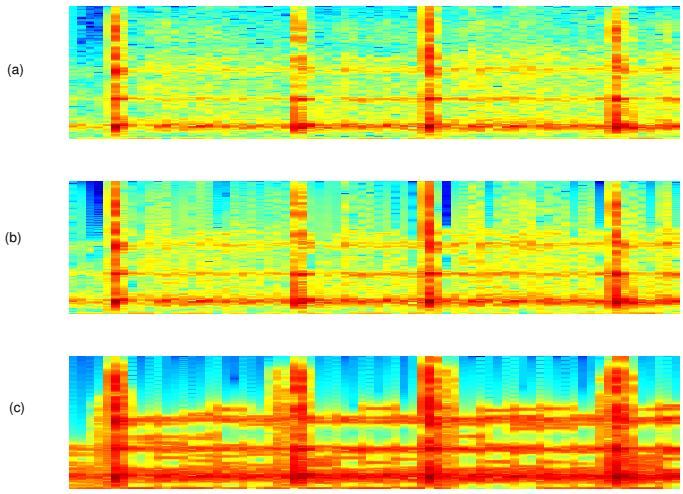

FIG. 4.20 – *Spectrogrammes du signal de castagnettes, (a) original, (b) modélisé par MP-Gabor, (c) modélisé par DYN-EDS*

DYN-EDS, présentés dans la figure 4.21. La supériorité de MP-Gabor apparaît clairement, ce qui est confirmé par les tests d'écoute. Même si la complexité de départ du MP est im-

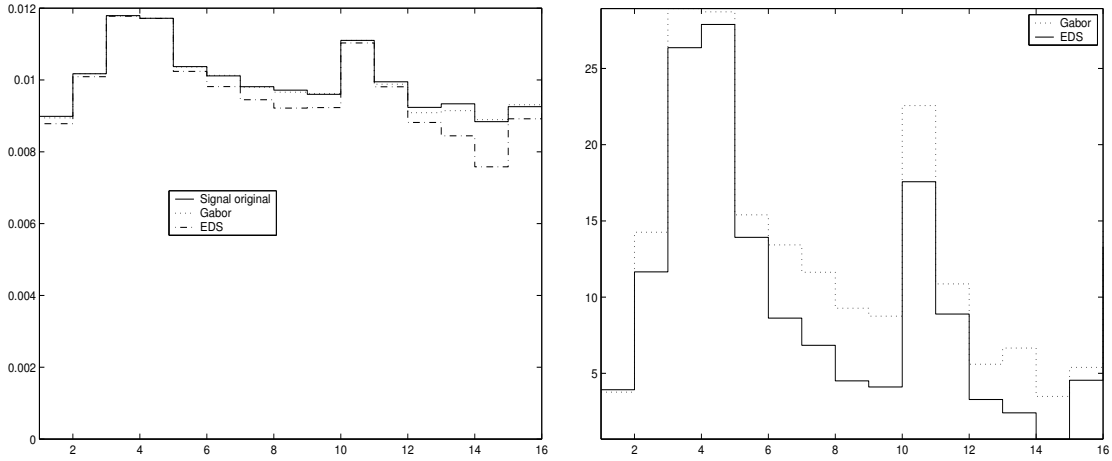

FIG. 4.21 – *Signal de castagnettes, Puissances par sous-bandes - RSBs par sous-bandes.*

portante, l'utilisation du MPR sur des fenêtres de taille 256 échantillons rend cet algorithme très avantageux et justifie complètement son utilisation dans le système de codage telle que décrite au chapitre 2, permettant d'atteindre de hautes performances de modélisation sur les signaux fortement transitoires.

Au chapitre suivant, nous nous intéressons aux techniques de quantification employées et au codage binaire des paramètres. Ces opérations sont précédées par une étape de sélection des trajectoires de paramètres selon des critères perceptuels.

Chapitre 5

Quantification

Les différents paramètres de modèle (fréquences, amplitudes, amortissements et phases pour le modèle EDS, et fréquences, amplitudes, phases, retards et facteurs d'échelle pour le modèle de Gabor) doivent être quantifiés avant d'être codés dans le bitstream. Dans un premier temps, on sélectionne les paramètres pertinents en se basant sur des critères psychoacoustiques tel que décrit dans la section 5.2. Signalons qu'à ce stade du travail, la sélection ne concerne pas les paramètres issus des segments transitoires. Ensuite, on traduit les amplitudes et les fréquences en échelles perceptuellement significatives, respectivement une échelle dB-SPL et une échelle ERB [10]. Une quantification scalaire uniforme est alors réalisée et les nouveaux paramètres sont codés relativement à certaines valeurs de référence. On commence par une brève présentation de la notion de masquage en psychoacoustique que l'on exploite dans la sélection des trajectoires.

5.1 Seuils de masquage

On sait d'après des études psychoacoustiques qu'un son perçu, de puissance et de fréquence données, possède un pouvoir masquant vis-à-vis des sons de puissance inférieure se trouvant dans la même plage de fréquence [10]. On dit alors que ce son *masque* les autres. La courbe de masquage d'un son est définie par les valeurs en dB qu'un son test doit prendre pour que ce dernier soit juste audible aux alentours de la fréquence du son masquant. On distingue deux types de masquage : celui dû à un son tonal et celui dû à un bruit à bande étroite. Le seuil de masquage est déduit des courbes de masquage calculées à partir de toutes les composantes spectrales du signal, un exemple est donné dans la figure 5.1 . On définit alors le Rapports Signal à Masque (SMR, Signal to Mask Ratio) comme le logarithme de la distance séparant l'amplitude d'un son au seuil de masquage en dB. Les expériences de psychoacoustique montrent que le SMR de sinusoïdes pures est beaucoup plus grand que celui de bruit à bande étroite [55]. Dans le standard MPEG-AAC [12] une région fréquentielle ne contenant que des sinusoïdes pures présente un SMR de 18dB alors qu'une région de bruit à bande étroite présente un SMR de 6dB. C'est justement le modèle psychoacoustique MPEG-AAC que l'on utilise dans notre codeur. Nous allons pouvoir exploiter les valeurs de SMR pour juger de la pertinence d'un partiel EDS et plus généralement d'une trajectoire.

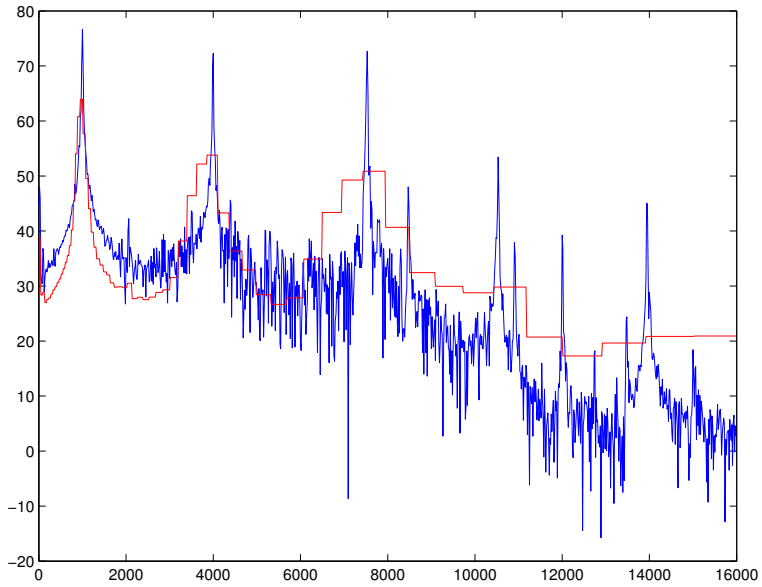

FIG. 5.1 – *Seuil de masquage MPEG-AAC sur spectre de glockenspiel*

5.2 Sélection des trajectoires

Il nous importe lors de cette étape de sélectionner les paramètres EDS, autrement dit de décider lesquels seront gardés et codés et lesquels seront abandonnés. En effet, il est inutile de chercher à représenter par EDS un contenu qui n'est pas perçu par l'oreille ou pouvant être assimilé à un processus stochastique, de même qu'il est inutile de coder d'éventuelles erreurs d'estimation (typiquement des paramètres provenant de lobes secondaires de pics spectraux) pour ne garder que les composantes tonales stables. Différentes propositions ont été faites dans ce sens. Comme dit précédemment, le "réflexe naturel" consiste à éliminer les trajectoires trop courtes (1 ou 2 trames) qui s'apparentent plutôt à un modèle de bruit. Par ailleurs, certains schémas sinusoïdaux réalisent la sélection des paramètres en se basant sur les SMRs considérés sur les trames les unes indépendamment des autres, sans exploiter les variations le long d'une trajectoire [56, 2, 46]. La stratégie de décision consiste alors à éliminer les composantes sinusoïdales en-dessous du seuil de masquage. Nous adoptons la technique proposée par Levine [14] qui exploite un suivi des valeurs de SMR sur une trajectoire. On procède en 2 étapes :

1. on élimine les composantes EDS (une composante est donnée par un quadruplet de paramètres $[a_m,d_m,\omega_m,\phi_m]$) présentant des valeurs de SMR inférieurs à SMR_{min};
2. on sélectionne alors les trajectoires présentant un SMR moyen en temps et une durée assez significatifs.

La première étape permet d'abandonner les partiels qui sont masqués dans le signal; n'étant pas perçus par l'oreille, il est inutile de les coder. La deuxième étape est plus délicate puisqu'il s'agit de prendre en compte la longueur de la trajectoire et la valeur de SMR moyen dans le temps pour aboutir à une décision sur la validité de tous les paramètres la constituant. Il s'agit d'éliminer ou de préserver toute une trajectoire du même coup. On considère qu'une

trajectoire ne sera pas perçue par l'oreille si son SMR moyen dans le temps est inférieur à un certain seuil et qu'elle représente du bruit si elle est trop courte. Par conséquent, la décision est prise en fonction de la position du point (longueur de trajectoire, SMR moyen) par rapport à une droite délimitant le plan en 2 zones de validité comme illustré sur la figure 5.2. Lorsqu'une trajectoire est telle qu'elle se positionne dans le demi-plan grisé, elle est gardée et codée sinon elle est éliminée. Signalons que la pente et l'ordonnée à l'origine de la

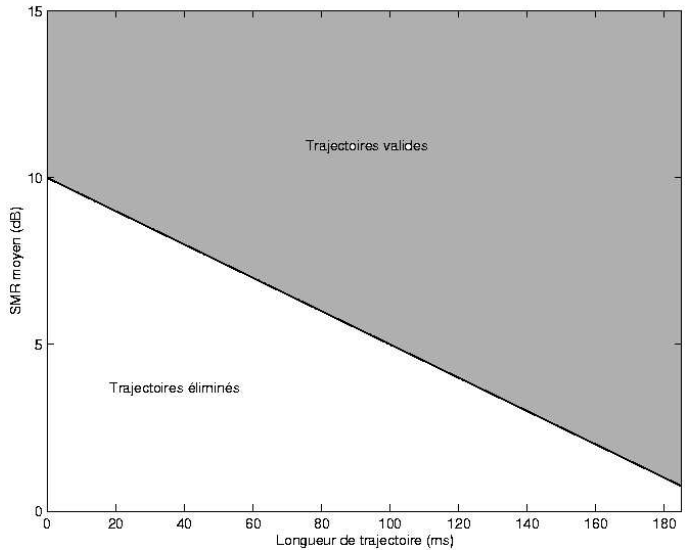

FIG. 5.2 – *Critère de sélection des trajectoires*

droite dépendent du modèle psychoacoustique utilisé. Ils ont été choisis suite à des simulations sur différents signaux. D'avantage de tests sur une variété plus large d'échantillons sonores permettrait sûrement de parvenir à un réglage plus optimal de ces paramètres. 50% des paramètres estimés, en moyenne, peuvent ainsi être abandonnés à ce stade.

5.3 Quantification des paramètres de modèle

Le schéma de quantification qui a été retenu est pratiquement identique à celui proposé dans [5] pour ce qui est des paramètres de fréquence, d'amplitude et de phase. Nous en présentons ici une description en signalant que cette étape reste à finaliser et qu'un travail d'optimisation permettrait probablement d'améliorer les performances de codage.

5.3.1 Quantification des paramètres du modèle EDS

Amplitudes

Les amplitudes dans l'intervalle $[1{,}2^{15}-1]$ sont quantifiés uniformément sur une échelle logarithmique selon

$$\bar{a}_m = \left\lfloor \frac{\log(a_m)}{2\log(a_b)} + 0.5 \right\rfloor$$

avec $a_b = 1.0218$. L'erreur maximale sur l'amplitude est alors de $\pm\frac{1.5}{8}$dB et $\bar{a}_m \in [0{,}241]$. Une précision de $\pm$1.5dB est généralement suffisante [46], on pourra donc utiliser un niveau de granularité plus grossier en considérant jusqu'à une valeur sur 8 dans l'intervalle [0,241].

Fréquences

Les fréquences exprimées en Hz dans l'intervalle $[0, \frac{f_e}{2}]$ sont traduites en échelle ERB avec L_f niveaux de quantification par ERB selon

$$\bar{f}_m = \lfloor L_f \mathrm{erb}(f_m) + 0.5 \rfloor$$

Le choix de $L_f = 91.2$ aboutit à des valeurs de fréquences quantifiées dans l'intervalle [0,3612] pour la bande 0Hz-16kHz à une fréquence d'échantillonnage de 32 kHz, ce qui correspond à une précision de $\pm$0.011 ERB. Là encore une précision de $\pm$0.088 ERB est généralement suffisante et l'on peut considérer jusqu'à une valeur sur 8 dans l'intervalle [0,3612].

Phases initiales

Les phases exprimées en radians dans l'intervalle $[-\pi,\pi[$ sont quantifiées sur une échelle linéaire avec une erreur maximale $\phi_e = \frac{\pi}{32}$ rd selon

$$\bar{\phi}_m = \left\lfloor \frac{\phi_m}{2\phi_e} + 0.5 \right\rfloor$$

et les valeurs de $\bar{\phi}_m$ égales à +16 sont transformées en -16 ($\pi = -\pi$). L'intervalle résultant est alors $[-16{,}15]$ ce qui nécessite 5 bits.

Coefficients d'Amortissements

On considère les coefficients d'amortissement dans l'intervalle $[-0.001{,}0.001]$. Ceux-ci sont quantifiés uniformément sur une échelle linéaire selon

$$\bar{d}_m = \lfloor 10^6 d_m + 0.5 \rfloor$$

ce qui entraîne une erreur maximale de 0.001 sur les coefficients d'amortissement et l'intervalle résultant est $[-1000{,}1000]$. En pratique, une erreur beaucoup plus importante peut être tolérée et l'on peut prendre jusqu'à une valeur sur 16 dans l'intervalle $[-1000{,}1000]$.

5.3.2 Quantification des paramètres du modèle de Gabor

Les paramètres issus du modèle de Gabor sont en partie déjà quantifiés eu égard à la structure du dictionnaire donnée à la section 4.2. Il s'agit d'une structure en arbre que l'on va mettre à profit pour la mise en œuvre de codage entropique (c.f. section 5.4). Les seuls paramètres à quantifier sont donc les coefficients de corrélations complexes α_m que l'on considère comme des amplitudes a_m et des phases ϕ_m ($\alpha_m = a_m \exp i\phi_m$). Ces amplitudes et ces phases sont quantifiées de façon assez grossière sur le mode des paramètres EDS.

5.4 Codage entropique des paramètres

5.4.1 Paramètres EDS

Naissances des trajectoires

Fréquences La fréquence la plus basse $\bar{f}_{min}(l)$ dans une trame l est directement codée à l'aide des codes de Huffman de la table H1(f). Les autres fréquences de la même trame sont codés relativement à $\bar{f}_{min}(l)$. Les valeurs $\Delta\bar{f}_m = \bar{f}_m - \bar{f}_{min}(l)$ sont ainsi codées à l'aide de la table de Huffman H2(f).

Amplitudes On procède pareillement pour les amplitudes : les $\bar{a}_{min}(l)$ sont codés à l'aide de la table H1(a) et les $\Delta\bar{a}_m = \bar{a}_m - \bar{a}_{min}(l)$ à l'aide de H2(a).

Phases initiales et Coefficients d'amortissement Ils sont codés directement en utilisant les tables de Huffman H1(ϕ) et H1(d).

Continuations des trajectoires

Les 3 paramètres amplitude, coefficient d'amortissement et fréquence peuvent être codés relativement à leurs prédécesseurs le long d'une trajectoire. Prenons l'exemple de la fréquence. A la trame l, la fréquence $\bar{f}_{m_t}(l)$ appartenant à la trajectoire t est codé relativement à la fréquence $\bar{f}_{p_t}(l-1)$ à laquelle elle est appariée. On code donc $\Delta\bar{f}_t = \bar{f}_{m_t}(l) - \bar{f}_{p_t}(l-1)$ à l'aide de la table de Huffman H3(f). Les tables H3(a) et H3(d) sont utilisées pour les amplitudes et les amortissements. Les phases ne sont pas codées en continuation. Signalons qu'à des fins de rafraîchissement, on peut considérer des trames de référence sur lesquelles tous les paramètres sont directement codés et ce même au niveau des continuations des trajectoires.

Table de Huffman	Paramètres
H1(f)	Fréquences de référence pour naissances
H2(f)	Fréquences relatives pour naissances
H3(f)	Fréquences relatives en continuations
H1(a)	Amplitudes de référence pour naissances
H2(a)	Amplitudes relatives pour naissances
H3(a)	Amplitudes relatives en continuations
H1(d)	Amortissements de référence pour naissances
H3(d)	Amortissements relatifs en continuations
H1(ϕ)	Phases initiales

TAB. 5.1 – *Tables de Huffman pour le modèle EDS*

5.4.2 Paramètres de Gabor

Les paramètres de Gabor (s,u,ω) sont codés au travers des paramètres (j,p,k) donnés par (4.3), (4.4) et (4.5). Les paramètres d'amplitude et de phase sont codés de la même façon que

pour la naissance de trajectoires du modèle EDS.

Facteurs d'échelle Le facteur d'échelle correspondant à la première itération de MP dans la trame l est codé directement par la table de Huffman $\mathrm{H}_1(\mathrm{s})$ et les autres valeurs sont codées relativement à celui-ci à l'aide de la table $\mathrm{H}_2(\mathrm{s})$. Compte tenu de la construction du dictionnaire, l'ensemble des valeurs prises par u et ω dépend de la valeur prise par s (c.f. (4.3), (4.4) et (4.5)). Des tables de Huffman différentes, fonction de s seront donc utilisées pour coder ces retards et pulsations.

Retards et pulsations Les retards, respectivement pulsations, correspondant à une même valeur de s sont codés les uns relativement à leurs prédécesseurs dans l'ordre d'apparition dans la poursuite. Soient $(p_{j_i})_{(1\leq i\leq I_s)}$, respectivement, $(k_{j_l})_{(1\leq l\leq L_s)}$, la suite de paramètres de retard, respectivement de pulsation, associés à une même valeur de facteur d'échelle s; on code directement p_{j_1} par $\mathrm{H}_s^0(\mathrm{p})$, respectivement k_{j_1} par $\mathrm{H}_s^0(k)$. Ensuite, on considère les $\Delta p_{j_i} = p_{j_i} - p_{j_1}$, $2 \leq i \leq I_s$, respectivement les $\Delta k_{j_l} = k_{j_l} - k_{j_1}$, $2 \leq l \leq L_s$, que l'on code à l'aide des tables de Huffman $\mathrm{H}_s(\mathrm{p})$, respectivement $\mathrm{H}_s(k)$.

Table de Huffman	Paramètres
H4(a)	Amplitudes de référence
H5(a)	Amplitudes relatives
H2(ϕ)	Phases initiales
$\mathrm{H}_s^0(\mathrm{p})$	Retards de référence
$\mathrm{H}_s^0(k)$	Pulsations de référence
$\mathrm{H}_s(\mathrm{p})$	Retards relatifs
$\mathrm{H}_s(k)$	Pulsations relatives

TAB. 5.2 – *Tables de Huffman pour le modèle de Gabor*

5.5 Conclusion

L'opération de quantification et de codage est encore en cours de traitement au moment de l'écriture de ce rapport. La stratégie de quantification des paramètres issus du modèle de bruit qui sera adoptée est celle proposée dans [5]. Les premières estimations effectuées permettent d'envisager un débit se situant autour de 24 kbit/s pour une fréquence d'échantillonnage de 32 kHz. Nous pensons que des efforts seront nécessaires afin de procéder à un réglage plus précis des paramètres intervenant dans la sélection des trajectoires et la quantification, notamment les niveaux de granularité à adopter. Enfin, il serait très utile d'adapter un modèle psychoacoustique plus approprié [14].

Conclusion

Dans le cadre de ce travail réalisé en étroite collaboration avec Rémy BOYER, doctorant au département TSI de l'ENST, nous avons proposé une architecture de codage entièrement paramétrique basé sur un modèle "Sinusoïdes Amorties Exponentiellement + Transitoires + Bruit" pour la bande 0-16kHz (fréquence d'échantillonnage de 32kHz) et visant un débit avoisinant les 24kbit/s. Signalons que le passage à une fréquence de 44.1kHz est aisément réalisable puisque la modélisation de la composante déterministe ne concerne que la bande de fréquence 5Hz-10kHz, le reste de la bande étant modélisé par du bruit.

Un ensemble de méthodes nouvelles a été développé dans ce contexte, notamment rattachées à la modélisation des signaux audio transitoires, donnant lieu à des publications qui sont présentées en annexe de ce document.

L'architecture de codage retenue suite à de nombreuses simulations s'articule autour d'un bloc de segmentation commandé par un puissant détecteur de transitoires. Ce bloc permet le passage de fenêtres d'analyse longues (2048 échantillons) destinées aux segments pseudo-stationnaires du signal à des fenêtres courtes (256 échantillons) destinées aux segments de signal transitoire. Nous avons défini un signal pseudo-stationnaire comme un signal correctement modélisé par un modèle EDS, c'est à dire aboutissant à un modèle EDS compact, et la définition de modèle compacte a également été donnée.

Les segments pseudo-stationnaires sont ainsi représentés par le modèle EDS au travers d'une méthode par FFTs décalées en mode de fonctionnement du codeur que l'on désigne par mode S. En mode T, les deux segments successifs contenant l'attaque sont modélisés par des atomes de Gabor en faisant appel à l'algorithme Matching Pursuit, et le reste des segments courts est décomposé en somme de EDS grâce à une approche Haute-résolution, en l'occurrence, l'algorithme de Kung.

La composante déterministe du signal est synthétisé suite à une opération de quantification en se basant sur une approche d'appariement des composantes du signal le long de trajectoires. Cet appariement permet en outre une sélection des paramètres perceptuellement pertinents en ayant recours à des mesures de SMR moyens dans le temps le long des différentes trajectoires. D'importantes améliorations peuvent être apportées à ce niveau en utilisant un modèle psychoacoustique mieux adapté au contexte et en réalisant une optimisation des seuils in-

tervenant dans la décision en rapport à la validité des paramètres à coder. De meilleures performances de codage pourraient sans doute être atteintes en effectuant des réglages de la granularité des niveaux de quantification décidés.

La composante stochastique est calculée par soustraction du signal déterministe synthétisé du signal original. Cette composante est vue comme du bruit et codée par une technique LPC au dessus de 5kHz. Ce choix de ne représenter que la partie haute fréquence du signal résiduel a été fait suite à de nombreuses tentatives infructueuses de représenter la totalité de la bande. En effet, cela s'est traduit par un effet de bruit de souffle très désagréable sur le signal synthétique. Des efforts restent donc à fournir afin d'améliorer la représentation de la composante bruit.

Les premiers résultats demeurent encourageants puisque nous avons pu constater, grâce à des tests informels, que des auditeurs "naïfs" n'étaient pas capables de percevoir les dégradations introduites par nos algorithmes pour une bonne partie des signaux tests.

Bibliographie

[1] International Organization for Standardization, "Overview of the MPEG-4 standard", *ISO/IEC JTC1/SC29/WG11 N4030*, March 2001.

[2] **H. Purnhagen, N. Meine**, "HILN - The MPEG-4 parametric audio coding tools", *Laboratorium für Informationstechnologie*, University of Hanover, February 2000.

[3] **A. Le Guyader, P. Philippe, J.B Rault**, "Synthèse des normes de codage de la parole et du son (UIT-T, ETSI et ISO/MPEG)", *Annales des télécommunications*, tome 55, No 9-10, pp 421-556, sep-oct 2000.

[4] International Organization for Standardization, *ISO/IEC 14496-3 (Information technology - very low bitrate audio-visual coding)*, 1998.

[5] **W. Oomen**,"MPEG-4 Audio Extension 1, Working Draft", *ISO/IEC JTC1/SC29/WG11 MPEG2001/N4379*, December 2001.

[6] International Organization for Standardization, "Call for proposals for new tools for audio coding", *ISO/IEC JTC1/SC29/WG11 MPEG2001/N3793*, January 2001.

[7] **R.J. McAulay, T.F. Quatieri**, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. on ASSP*, Vol. 34, No 4, August 1986.

[8] **J. Laroche, Y. Stylianou, E. Moulines**, "HNM : a simple, efficient harmonic+noise model for speech", *IEEE Workshop on Applications of Signal Processing to Audio ans Acoustics*, 1993.

[9] **X. Serra**, "A system for sound analysis/synthesis based on a deterministic plus stochastic decomposition", PhD thesis, CCRMA Department of Music, Stanford 1989.

[10] **T. Painter**, "Perceptual Coding of Digital Audio", *Proceedings of the IEEE*, vol. 88, No 4, April 2000.

[11] **K. Banderburg and G. Stoll**, "ISO-MPEG-1 Audio : a generic standard for coding of high-quality digital audio", *JASA*, Vol. 42, October 1994

[12] Norme Internationale ISO/CEI, *IS 13818-7 (MPEG-2 Advanced Audio Coding, AAC)*, Avril 1997.

[13] Mp3Pro/Spectral Band Replication (SBR), *http://www.mp3-tech.org/*.

[14] **S. Levine**, *Audio Representations for Data Compression and Compressed Domain Processing*, PhD thesis, University of Stanford, 1998.

[15] **X. Serra, J. Smith III**, "Spectral Modeling Synthesis : A Sound System Based on a Deterministic plus Stochastic Decomposition", *Computer Music Journal*, Vol. 14, No 4, Winter 1990.

[16] **R.J. McAulay and T.F. Quatieri**, "Speech Coding and Synthesis", *Elsevier Science*, chapter 4, 1995.

[17] **X. Rodet**, "Musical sound Analysis/Synthesis : Sinudoidal + Residual and Elementary Waveform Models", *TFTE*, 1997.

[18] **Ioannis Stylianou**, "Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole et du Locuteur", thèse de doctorat, Télécom Paris, 1996.

[19] **K. Hamdy, M. Ali and H. Tewfik**, "Low bit rate high quality audio coding with combined harmonic and wavelet representations". *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta. 1996.

[20] **T. S. Verma**, "A perceptually based audio signal model with applications to scalable audio compression", PhD thesis, Department of Music, Stanford 1999.

[21] **T.S. Verma, T.H.Y. Meng**,"A 64Kbps to 85Kbps scalable audio coder",*Proc. of the IEEE ICASSP*, 2000.

[22] **S.G. Mallat, Z. Zhang**, "Matching Pursuit With Time-Frequency Dictionaries", *IEEE Trans. on SP*, Vol. 41, No 12, December 1993.

[23] **S. Mallat**, "A Wavelet Tour of Signal Processing", *Academic Press, second edition*, 1999.

[24] **S.S. Chen, D.L. Donoho, M.A. Saunders**, "Atomic Decomposition by Basis Pursuit", Technical Report, Stanford, 1995.

[25] **M. M. Goodwin, M. Vetterli**, "Matching pursuit and atomic signal models based on recursive filter banks", *IEEE transactions on SP*, vol. 47, No 7, July 1999.

[26] **R. Gribonval**, "Approximations non-linéaires pour l'analyse de signaux sonores", thèse de doctorat, IRISA-INRIA, 1999.

[27] **S. Qian, D. Chen**, "Signal representation using adaptive normalized gaussian functions", *Signal Processing*, 36 (1994) 1-11.

[28] **R.R. Coifman, M.V. Wickerhauser**, "Entropy-based algorithms for best basis selection",*IEEE Trans. on I.T.*, Vol 38 Issue: 2 Part: 2 , March 1992.

[29] **A. Klapuri**, "Sound onset detection by applying psychoacoustic knowledge", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, 1999.

[30] **M. Goodwin**, "Nonuniform Filterbank Design for Audio Signal Modeling", *IEEE*, CNMAT-Berkeley, 1997.

[31] **P. O'Shea**, "The use of sliding spectral windows for parameter estimation in power system disturbance monitoring", *IEEE Trans. on PES*, 2000.

[32] **M. Goodwin**, "Residual Modeling in Music Analysis-Synthesis", *Proceedings of the IEEE ICASSP*, 1996.

[33] **N. Moreau**, "Techniques de Compression des signaux", *Masson, Collection technique et scientifique des télécommunications*, 1995.

[34] **R. Boyer, S. Essid and N. Moreau**, "Non-stationary signal parametric modeling techniques with an application to low bitrate audio coding" *Proc. of IEEE Int. Conf. Signal Processing*, August 2002.

[35] **M. Kunt**, "Traitement numérique des signaux", *Presses Polytechniques et Universitaires Romandes, traité d'électricité*, 1989.

[36] **T. Painter, A. Spanias**, "Perceptual Component Selection in Sinusoidal Coding of Audio", PhD thesis project, Arizona State University, 2001.

[37] **B. Porat**, *A course in Digital Signal Proecessing*, John Wiley & Sons, New York, 1997.

[38] **R. Roy and T. Kailath**, "ESPRIT-Estimation of Signal Parameters Via Rotational Invariance Techniques", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 7, July 1989.

[39] **Y. Hua, T.K. Sarkar**, "Matrix Pencil for Estimating Parameters of Exponentially Damped/Undamped Sinusoids in Noise", *IEEE Trans. on ASSP*, Vol. 38, No 5, May 1990.

[40] **S.Y. Kung, K.S. Arun, D.V. Bhaskar Rao**, "State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem", *J. OPT. SOC. AM.*, Vol. 73, No 12, December 1983.

[41] **A.J. Van Der Veen, ED F. Deprettere, A. Lee Swindlehurst**, "Subspace-Based Signal Analysis Using Singular Value decomposition, *Proc. of the IEEE*, Vol. 81, No. 9, September 1993.

[42] **G.H. Golub and C.F. Van Loan**, *Matrix Computation*, North Oxford Academic, Oxford, second edition, 1983.

[43] **J.A. Cadzow**, "Signal Enhancement - A Composite Property Mapping Algorithm", *IEEE Trans. on ASSP*, Vol. 36, No 1, January 1988.

[44] **J. Laroche**, "The use of the matrix pencil method for the spectrum analysis of musical signal", *J. Acoust. Soc. Am.*, 94(4), October 1994.

[45] **S. M. kay**, *Modern Spectral Estimation*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[46] **A.C. den Brinker, E.G.P. Schuijers and A.W.J. Oomen**, "Parametric Coding for High-Quality Audio", *Proc. of the 112th Convention, Audio Engineering Society*, May 2002.

[47] **J. Nieuwenhuijse, R. Heusdens and E.F. Deprettere**, "Robust Exponential Modeling of Audio Signal", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, May 1998.

[48] **R. Boyer and S. Essid**, "Transient modeling with a Frequency-Transform Subspace Algorithm and "Transient + Sinusoidal" scheme" *Proc. of IEEE Int. Conf. on Digital Signal Processing*, July 2002.

[49] **R. Boyer and K. Abed-Meraim**, "Audio transients modeling by Damped & Delayed Sinusoids (DDS)", *Proc. of IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, May 2002.

[50] **R. Boyer, S. Essid and N. Moreau**, "Dynamic temporal segmentation in parametric non-stationary modeling for percussive musical signals", *IEEE Int. Conf. on Multimedia and Expo*, August 2002.

[51] **N. Moreau, P. Dymarski**, "Sucessive Orthogonalizations in the Multistage CELP C", *IEEE*, 1992.

[52] **N. Moreau, P. Dymarski**, "Selection of Excitation Vectors for the CELP Coders", *IEEE Trans. on SP* , Vol. 2, No 1, PART I, January 1994.

[53] **R. Boyer, S. Essid, N. Moreau**, "Exploration de techniques modernes de modélisation adaptées à du codage audio bas-débit", *Journées d'Etudes et d'Echanges : Compression et Représentation des Signaux Audiovisuels*, Novembre 2001.

[54] **T. Painter, A. Spanias**, "Perceptual Component Selection in Sinusoidal Coding of Audio", PhD thesis project, Arizona State University, 2001.

[55] **E. Zwicker, E. Feltkeller**, "Psycho-acoustique, l'oreille récepteur d'information", *Masson, Collection technique et scientifique des télécommunications*, 1981.

[56] **M. Ali**, "Adaptive Signal Representation with Application in Audio Coding", Ph.D. thesis, University of Minnesota, 1996.